

Ewout W. Steyerberg

# Clinical Prediction Models

A Practical Approach to  
Development, Validation, and  
Updating



Springer

# Statistics for Biology and Health

## *Series Editors:*

M. Gail  
K. Krickeberg  
J. Samet  
A. Tsiatis  
W.Wong

# Statistics for Biology and Health

---

- Bacchieri/Cioppa: Fundamentals of Clinical Research  
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations  
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints  
Duchateau/Janssen: The Frailty Model  
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS  
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.  
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor  
Hougaard: Analysis of Multivariate Survival Data  
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.  
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.  
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text, 2nd ed.  
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.  
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.  
Lazar: The Statistical Analysis of fMRI Data  
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations  
Martinussen/Scheike: Dynamic Regression Models for Survival Data  
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators  
Nielsen: Statistical Methods in Molecular Evolution  
O'Quigley: Proportional Hazards Regression  
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software  
Proschan/LanWittes: Statistical Monitoring of Clinical Trials: A Unified Approach  
Siegmund/Yakir: The Statistics of Gene Mapping  
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations  
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics  
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modelling in the Manville Asbestos Case  
Steyerberg: Clinical Prediction Models: A Practical Approach to Model Development, Validation, and Updating  
Sun: The Statistical Analysis of Interval-censored Failure Time Data  
Therneau/Grambsch: Modelling Survival Data: Extending the Cox Model  
Ting: Dose Finding in Drug Development  
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models  
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL  
Zhang/Singer: Recursive Partitioning in the Health Sciences  
Zuur/ Ieno/ Smith: Analysing Ecological Data

Ewout W. Steyerberg

# Clinical Prediction Models

A Practical Approach to Development,  
Validation, and Updating



Springer

E.W. Steyerberg  
Department of Public Health  
Erasmus MC  
3000 CA Rotterdam  
The Netherlands

*Series Editors*

M. Gail  
National Cancer Institute  
Bethesda, MD 20892  
USA

A. Tsiatis  
Department of Statistics  
North Carolina State University  
Raleigh, NC 27695  
USA

K. Krickeberg  
Le Chatelet  
F-63270 Manglieu  
France

W. Wong  
Department of Statistics  
Stanford University  
Stanford, CA 94305-4065  
USA

J. Sarnet  
Department of Epidemiology  
School of Public Health  
Johns Hopkins University  
615 Wolfe Street  
Baltimore, MD 21205-2103  
USA

ISBN: 978-0-387-77243-1      e-ISBN: 978-0-387-77244-8  
DOI: 10.1007/978-0-387-77244-8

Library of Congress Control Number: 2008929620

© 2009 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

[springer.com](http://springer.com)

*For Aleida, Matthijs, Laurens and Suzanne*

# Preface

Prediction models are important in various fields, including medicine, physics, meteorology, and finance. Prediction models will become more relevant in the medical field with the increase in knowledge on potential predictors of outcome, e.g. from genetics. Also, the number of applications will increase, e.g. with targeted early detection of disease, and individualized approaches to diagnostic testing and treatment. The current era of evidence-based medicine asks for an individualized approach to medical decision-making. Evidence-based medicine has a central place for meta-analysis to summarize results from randomized controlled trials; similarly prediction models may summarize the effects of predictors to provide individualized predictions of a diagnostic or prognostic outcome.

## Why Read This Book?

My motivation for working on this book stems primarily from the fact that the development and applications of prediction models are often suboptimal in medical publications. With this book I hope to contribute to better understanding of relevant issues and give practical advice on better modelling strategies than are nowadays widely used.

Issues include:

- (a) Better predictive modelling is sometimes easily possible; e.g. a large data set with high quality data is available, but all continuous predictors are dichotomized, which is known to have several disadvantages.
- (b) Small samples are used:
  - Studies are underpowered, with unreliable answers to difficult questions such as “Which are the most important predictors in this prediction problem?”
  - The problem of small sample size is aggravated by doing a complete case analysis which discards information from nearly complete records. Statistical imputation methods are nowadays available to exploit all available information.

- Predictors are omitted that should reasonably have been included based on subject matter knowledge. Modelers rely too much on the limited data that they have available in their data set, instead of wisely combining information from several sources, such as medical literature and experts in the field.
- Stepwise selection methods are abundant, which are especially risky in small data sets.
- Modelling approaches are used that require higher numbers. Data-hungry techniques, such as tree modelling and neural network modelling, should not be used in small data sets.
- No attempts are made towards validation, or validation is done inefficiently. For example, a split-sample approach is followed, leading to a smaller sample for model development and a smaller sample for model validation. Better methods are nowadays available and should be used far more often.

(c) Claims are exaggerated:

- Often we see statements such as ‘the predictors were identified’; in many instances such findings may not be reproducible and may largely represent noise.
- Models are not internally valid, with overoptimistic expectations of model performance in new patients.
- One modern method with a fancy name is claimed as being superior to a more traditional regression approach, while no convincing evidence exists, and a suboptimal model strategy was followed for the regression model.
- Researchers are insufficiently aware of overfitting, implying that their apparent findings are merely coincidental (“the curse of dimensionality”).

(d) Poor generalizability:

- If models are not internally valid, we cannot expect them to generalize.
- Models are developed for each local situation, discarding earlier findings on effects of predictors and earlier models; a framework for continuous improvement and updating of prediction models is required.

In this book; I try to suggest many small improvements in modelling strategies. Combined, these improvements hopefully lead to better prediction models.

## Intended Audience

Readers should have a basic knowledge of biostatistics, especially regression analysis, but no strong background in mathematics is required. The number of formulas is deliberately kept small. Usually a bottom-up approach is followed in teaching regression analysis techniques, starting with model assumptions, estimation methods, and basic interpretation. This book is more top-down: given that we want to predict an outcome, how can we best utilize regression techniques?

Three levels of readers are envisioned:

- (a) The core intended audience is formed by epidemiologists and applied biostatisticians who want to develop or apply a prediction model. Both students and professionals should find practical guidance in this book, especially by the proposed seven steps to develop a valid model (Part II).
- (b) A second group is formed by clinicians, policy makers, and health care professionals who want to judge a study that presents a prediction model. This book should aid them in a critical appraisal, providing explanations of terms and concepts that are common in publications on prediction models. They should try to read chapters of particular interest, or read the main text of the chapters. They can skip the examples and more technical sections (indicated with\*).
- (c) A third group includes more theoretical researchers, such as (bio)statisticians and computer scientists, who want to improve the methods that we use in prediction models. They may find inspiration for further theoretical work and simulation studies in this book. Many of the methods in prediction modelling are not fully developed yet, and common sense underlies some of the proposed approaches in this book.

## Other sources

Many excellent text books exist on regression analysis techniques, but these usually do not have a focus on modelling strategies for prediction. The main exception is Frank Harrell's book "Regression Modelling Strategies".<sup>174</sup> He brings advanced biostatistical concepts to practical application, supported by the Design and Hmisc libraries for S+ software (nowadays: packages for R). Harrell's book may however be too advanced for clinical and epidemiological researchers. This also holds for the Hastie, Tibshirani, and Friedman's quite thorough text book "The Elements of Statistical Learning".<sup>181</sup> These books are very useful for a more in-depth discussion of statistical techniques and strategies. Harrell's book provided the main inspiration for the presented work here. Another good companion book is the Vittinghoff et al. book on "Regression Methods in Biostatistics".<sup>472</sup>

Various sources at the internet can be used that explain terms used in this book. Frank Harrell has a glossary at his web site: [<http://biostat.mc.vanderbilt.edu/twiki/pub/Main/ClinStat/glossary.pdf>]. Other useful sources include [[http://www.aiaccess.net/e\\_gm.htm](http://www.aiaccess.net/e_gm.htm)] and Wikipedia.

## Structure

It has been found that people learn by example, by checklists, and by own discovery. Therefore I provide many examples throughout the text, including the essential computer code and output. I also suggest a checklist for prediction modelling (Part II).

Own discovery is possible with exercises per chapter, with data sets provided at the book's web site: <http://www.clinicalpredictionmodels.org>.

Many statistical techniques and approaches are readily possible with any modern software package. Personally, I work with SPSS for simple, straightforward analyses, but this package is insufficient for more advanced analyses which are essential in prediction modelling. The SAS computer package is more advanced, but may not be so practical for some. A package such as Stata is very suitable. It may be similar in capabilities to S-plus software, which was my preferred program for advanced prediction modelling since a stay at Duke University in 1996. The R software is very similar in nature to S-plus, and has several additional advantages: the software is for free, and innovations in biostatistical methods become readily available for R. Therefore, R is the natural choice as the software accompanying this book. R software is available at <http://www.cran.r-project.org>, with help files and a tutorial.

Some R commands are provided in this book; full programs can be downloaded from a web site (<http://www.clinicalpredictionmodels.org>). This web site also provides a number of data sets that can be downloaded for application of the described techniques. I provide data files in SPSS format that can readily be imported in R and other packages. Further, comments on the text can be submitted electronically.

# Acknowledgements

Many have made small to large contributions to this book. I'm very grateful to all. Frank Harrell has been a source of inspiration for my research in the area of clinical prediction models, together with Hans van Houwelingen, who has developed many of the theoretical innovations that are presented in this book. At my department (Public Health, Erasmus MC), Dik Habbema has encouraged me to further develop my interests in prognosis and prediction research, together with my close coworker René Eijkemans. Hester Lingsma was very supportive in the last phase of finishing this book. Several NIHES MSc students, PhD students, colleagues, and external reviewers made specific comments on various chapters. These include Lex Burdorf, Sarwa Darwish Murad, Sonja Deurloo, René Eijkemans, Frank Harrell, Yolanda van der Graaf, Cecile Janssens, Mike Kattan, Michael Koller, Carl Moons, Jeroen van der Net, Saskia Pluijm, Yvonne Vergouwe, Andrew Vickers, and many others.

I thank John Kimmel from Springer for his encouragement and practical support since we first met in Boston in 2005. I specifically would like to thank investigators who allowed their data sets to be made available for didactic purposes, including Kerry Lee (Duke University) and the GUSTO-I investigators. Finally, I thank my family, in particular my beloved wife Aleida, for their warm and ongoing support, and for allowing me to devote time, often at nights and weekends, to work on this book.

Rotterdam

Ewout Steyerberg

# Contents

<b>Preface .....</b>	vii
<b>Acknowledgements .....</b>	xi
<b>1 Introduction .....</b>	1
1.1 Prognosis and Prediction in Medicine .....	1
1.1.1 Prediction Models and Decision-Making .....	1
1.2 Statistical Modelling for Prediction .....	2
1.2.1 Model Uncertainty .....	3
1.2.2 Sample Size .....	4
1.3 Structure of the Book .....	5
1.3.1 Part I: Prediction Models in Medicine .....	5
1.3.2 Part II: Developing Valid Prediction Models .....	6
1.3.3 Part III: Generalizability of Prediction Models .....	6
1.3.4 Part IV: Applications .....	7
1.3.5 Questions and Exercises .....	7
<b>Part I Prediction Models in Medicine</b>	
<b>2 Applications of Prediction Models .....</b>	9
2.1 Applications: Medical Practice and Research .....	11
2.2 Prediction Models for Public Health .....	12
2.2.1 Targeting of Preventive Interventions .....	12
2.2.2 Example: Incidence of Breast Cancer .....	12
2.3 Prediction Models for Clinical Practice .....	13
2.3.1 Decision Support on Test Ordering .....	13
2.3.2 Example: Predicting Renal Artery Stenosis .....	14
2.3.3 Starting Treatment: the Treatment Threshold .....	15
2.3.4 Example: Probability of Deep Venous Thrombosis .....	16
2.3.5 Intensity of Treatment .....	16
2.3.6 Example: Defining a Poor Prognosis Subgroup in Cancer .....	18

2.3.7	Cost-Effectiveness of Treatment . . . . .	18
2.3.8	Delaying Treatment . . . . .	19
2.3.9	Example: Spontaneous Pregnancy Chances . . . . .	19
2.3.10	Surgical Decision-Making . . . . .	21
2.3.11	Example: Replacement of Risky Heart Valves . . . . .	21
2.4	Prediction Models for Medical Research . . . . .	23
2.4.1	Inclusion and Stratification in an RCT . . . . .	23
2.4.2	Example: Selection for TBI Trials . . . . .	24
2.4.3	Covariate Adjustment in an RCT . . . . .	25
2.4.4	Gain in Power by Covariate Adjustment . . . . .	26
*2.4.5	Example: Analysis of the GUSTO-III Trial . . . . .	27
2.4.6	Prediction Models and Observational Studies . . . . .	27
2.4.7	Propensity Scores . . . . .	28
2.4.8	Example: Statin Treatment Effects . . . . .	28
2.4.9	Provider Profiling . . . . .	29
2.4.10	Example: Ranking Cardiac Outcome . . . . .	29
2.5	Concluding Remarks . . . . .	30
<b>3</b>	<b>Study Design for Prediction Models . . . . .</b>	<b>33</b>
3.1	Study Design . . . . .	33
3.2	Cohort Studies for Prognosis . . . . .	33
3.2.1	Retrospective Designs . . . . .	35
3.2.2	Example: Predicting Early Mortality in Oesophageal Cancer . . . . .	35
3.2.3	Prospective Designs . . . . .	35
3.2.4	Example: Predicting Long-Term Mortality in Oesophageal Cancer . . . . .	36
3.2.5	Registry Data . . . . .	36
3.2.6	Example: Surgical Mortality in Oesophageal Cancer . . . . .	37
3.2.7	Nested Case–Control Studies . . . . .	37
3.2.8	Example: Perioperative Mortality in Major Vascular Surgery . . . . .	38
3.3	Studies for Diagnosis . . . . .	38
3.3.1	Cross-Sectional Study Design and Multivariable Modelling . . . . .	38
3.3.2	Example: Diagnosing Renal Artery Stenosis . . . . .	38
3.3.3	Case–Control Studies . . . . .	39
3.3.4	Example: Diagnosing Acute Appendicitis . . . . .	39
3.4	Predictors and Outcome . . . . .	39
3.4.1	Strength of Predictors . . . . .	39
3.4.2	Categories of Predictors . . . . .	40
3.4.3	Costs of Predictors . . . . .	40
3.4.4	Determinants of Prognosis . . . . .	41
3.4.5	Prognosis in Oncology . . . . .	41

3.5	Reliability of Predictors . . . . .	42
3.5.1	Observer Variability . . . . .	42
3.5.2	Example: Histology in Barrett's Oesophagus . . . . .	42
3.5.3	Biological Variability . . . . .	43
3.5.4	Regression Dilution Bias . . . . .	43
3.5.5	Example: Simulation Study on Reliability of a Binary Predictor . . . . .	43
3.5.6	Choice of Predictors . . . . .	44
3.6	Outcome . . . . .	44
3.6.1	Types of Outcome . . . . .	44
3.6.2	Survival Endpoints . . . . .	45
3.6.3	Example: Relative Survival in Cancer Registries . . . . .	45
3.6.4	Composite End Points . . . . .	46
3.6.5	Example: Mortality and Composite End Points in Cardiology . . . . .	46
3.6.6	Choice of Prognostic Outcome . . . . .	46
3.6.7	Diagnostic End Points . . . . .	47
3.6.8	Example: PET Scans in Oesophageal Cancer . . . . .	47
3.7	Phases of Biomarker Development . . . . .	47
3.8	Statistical Power . . . . .	48
3.8.1	Statistical Power to Identify Predictor Effects . . . . .	49
3.8.2	Examples of Statistical Power Calculations . . . . .	49
3.8.3	Statistical Power for Reliable Predictions . . . . .	50
3.9	Concluding Remarks . . . . .	51
<b>4</b>	<b>Statistical Models for Prediction . . . . .</b>	<b>53</b>
4.1	Continuous Outcomes . . . . .	53
4.1.1	Examples of Linear Regression . . . . .	54
4.1.2	Economic Outcomes . . . . .	54
4.1.3	Example: Prediction of Costs . . . . .	54
4.1.4	Transforming the Outcome . . . . .	54
4.1.5	Performance: Explained Variation . . . . .	55
4.1.6	More Flexible Approaches . . . . .	55
4.2	Binary Outcomes . . . . .	57
4.2.1	R <sup>2</sup> in Logistic Regression Analysis . . . . .	58
4.2.2	Calculation of R <sup>2</sup> on the Log Likelihood Scale . . . . .	58
4.2.3	Models Related to Logistic Regression . . . . .	60
4.2.4	Bayes Rule . . . . .	61
4.2.5	Example: Calculations with Likelihood Ratios . . . . .	62
4.2.6	Prediction with Naïve Bayes . . . . .	63
4.2.7	Examples of Naïve Bayes . . . . .	65
4.2.8	Calibration and Naïve Bayes . . . . .	65
4.2.9	Logistic Regression and Bayes . . . . .	65
4.2.10	More Flexible Approaches to Binary Outcomes . . . . .	65

4.2.11	Classification and Regression Trees . . . . .	67
4.2.12	Example: Mortality in Acute MI Patients . . . . .	67
4.2.13	Advantages and Disadvantages of Tree Models . . . . .	67
4.2.14	Trees as Special Cases of Logistic Regression Modelling . . . . .	69
4.2.14	Other Methods for Binary Outcomes . . . . .	70
4.2.15	Summary on Binary Outcomes . . . . .	71
4.3	Categorical Outcomes . . . . .	71
4.3.1	Polytomous Logistic Regression . . . . .	72
4.3.2	Example: Histology of Residual Masses . . . . .	72
4.3.3	Alternative Models . . . . .	73
4.3.4	Comparison of Modelling Approaches . . . . .	74
4.4	Ordinal Outcomes . . . . .	74
4.4.1	Proportional Odds Logistic Regression . . . . .	75
4.4.2	Alternative: Continuation Ratio Model . . . . .	77
4.5	Survival Outcomes . . . . .	77
4.5.1	Cox Proportional Hazards Regression . . . . .	77
4.5.2	Predicting with Cox . . . . .	78
4.5.3	Proportionality Assumption . . . . .	78
4.5.4	Kaplan–Meier Analysis . . . . .	79
4.5.5	Example: NFI After Treatment of Leprosy . . . . .	79
4.5.6	Parametric Survival . . . . .	80
4.5.7	Example: Replacement of Risky Heart Valves . . . . .	80
4.5.8	Summary on Survival Outcomes . . . . .	81
4.6	Concluding Remarks . . . . .	81
<b>5</b>	<b>Overfitting and Optimism in Prediction Models . . . . .</b>	<b>83</b>
5.1	Overfitting and Optimism . . . . .	83
5.1.1	Example: Surgical Mortality in Oesophagectomy . . . . .	84
5.1.2	Variability within One Centre . . . . .	84
5.1.3	Variability between Centres: Noise vs. True Heterogeneity . . . . .	85
5.1.4	Predicting Mortality by Centre: Shrinkage . . . . .	87
5.2	Overfitting in Regression Models . . . . .	87
5.2.1	Model Uncertainty: Testimation . . . . .	87
5.2.2	Other Biases . . . . .	89
5.2.3	Overfitting by Parameter Uncertainty . . . . .	90
5.2.4	Optimism in Model Performance . . . . .	90
5.2.5	Optimism-Corrected Performance . . . . .	92
5.3	Bootstrap Resampling . . . . .	92
5.3.1	Applications of the Bootstrap . . . . .	93
5.3.2	Bootstrapping for Regression Coefficients . . . . .	93
5.3.3	Bootstrapping for Optimism Correction . . . . .	94
5.3.4	Calculation of Optimism-Corrected Performance . . . . .	95

5.3.5	Example: Stepwise Selection in 429 Patients . . . . .	96
5.4	Cost of Data Analysis . . . . .	97
5.4.1	Example: Cost of Data Analysis in a Tree Model . . . . .	98
5.4.2	Practical Implications . . . . .	98
5.5	Concluding Remarks. . . . .	99
<b>6</b>	<b>Choosing Between Alternative Statistical Models. . . . .</b>	<b>101</b>
6.1	Prediction with Statistical Models . . . . .	101
6.1.1	Testing of Model Assumptions and Prediction. . . . .	102
6.1.2	Choosing a Type of Model . . . . .	102
6.2	Modelling Age–Outcome Relationships. . . . .	103
6.2.1	Age and Mortality After Acute MI . . . . .	103
6.2.2	Age and Operative Mortality . . . . .	103
6.2.3	Age–Outcome Relationships in Other Diseases . . . . .	106
6.3	Head-to-Head Comparisons . . . . .	107
6.3.1	StatLog Results . . . . .	107
6.3.2	GUSTO-I Modelling Comparisons. . . . .	108
6.3.3	GUSTO-I Results . . . . .	109
6.4	Concluding Remarks. . . . .	110

## Part II Developing Valid Prediction Models

<b>7</b>	<b>Dealing with Missing Values . . . . .</b>	<b>113</b>
7.1	Missing Values in Predictors. . . . .	115
7.1.1	Inefficiency of Complete Case Analysis. . . . .	116
7.1.2	Interpretation of Analyses with Missing Data . . . . .	117
7.1.3	Missing Data Mechanisms . . . . .	117
7.1.4	Summary Points . . . . .	118
7.2	Regression Coefficients Under MCAR, MAR, and MNAR. . . . .	118
7.2.1	R Code . . . . .	120
7.3	Missing Values in Regression Analysis . . . . .	121
7.3.1	Imputation Principle . . . . .	121
7.3.2	Simple and More Advanced Single Imputation Methods . . . . .	122
7.3.3	Multiple Imputation . . . . .	123
7.4	Defining the Imputation Model. . . . .	124
7.4.1	Transformations of Variables . . . . .	125
7.4.2	Imputation Models for SI . . . . .	125
7.4.3	Summary Points . . . . .	126
7.5	Simulations of Imputation Under MCAR, MAR, and MNAR. . . . .	126
7.5.1	Multiple Predictors . . . . .	127
7.6	Imputation of Missing Outcomes . . . . .	128
7.7	Guidance to Missing Values in Prediction Research . . . . .	129

7.7.1	Patterns of Missingness . . . . .	129
7.7.2	Simple Approaches . . . . .	130
7.7.3	Maximum Fraction of Missing Values Before Omitting a Predictor . . . . .	131
7.7.4	Single or Multiple Imputation for Predictor Effects? . . . . .	131
7.7.5	Single or Multiple Imputation for Predictions? . . . . .	132
7.7.6	Reporting of Missing Values in Prediction Research . . . . .	133
7.8	Concluding Remarks . . . . .	134
7.8.1	Summary Statements . . . . .	135
7.8.2	Currently Available Software and Challenges . . . . .	136
<b>8</b>	<b>Case Study on Dealing with Missing Values . . . . .</b>	<b>139</b>
8.1	Introduction . . . . .	139
8.1.1	Aim . . . . .	139
8.1.2	Patient Selection . . . . .	140
8.1.3	Selection of Potential Predictors . . . . .	140
8.1.4	Coding and Time Dependency of Predictors . . . . .	141
8.2	Missing Values in the IMPACT Study . . . . .	142
8.2.1	Missing Values in Outcome . . . . .	142
8.2.2	Quantification of Missingness of Predictors . . . . .	143
8.2.3	Patterns of Missingness . . . . .	144
8.3	Imputation of Missing Predictor Values . . . . .	147
8.3.1	Correlations Between Predictors . . . . .	147
8.3.2	Imputation Model . . . . .	147
8.3.3	Distributions of Imputed Values . . . . .	149
8.4	Estimating Adjusted Effects . . . . .	149
8.4.1	Adjusted Analysis for Complete Predictors: Age and Motor Score . . . . .	151
8.4.2	Adjusted Analysis for Incomplete Predictors: Pupils . . . . .	154
8.5	Multivariable Analyses . . . . .	155
8.6	Concluding Remarks . . . . .	155
<b>9</b>	<b>Coding of Categorical and Continuous Predictors . . . . .</b>	<b>159</b>
9.1	Categorical Predictors . . . . .	159
9.1.1	Examples of Categorical Coding . . . . .	160
9.2	Continuous Predictors . . . . .	161
9.2.1	Examples of Continuous Predictors . . . . .	161
9.2.2	Categorization of Continuous Predictors . . . . .	162
9.3	Non-Linear Functions for Continuous Predictors . . . . .	163
9.3.1	Polynomials . . . . .	164
9.3.2	Fractional Polynomials . . . . .	164
9.3.3	Splines . . . . .	165
9.3.4	Example: Functional Forms with RCS or FP . . . . .	166
9.3.5	Extrapolation and Robustness . . . . .	166

9.4	Outliers and Truncation . . . . .	167
9.4.1	Example: Glucose Values and Outcome of TBI . . . . .	168
9.5	Interpretation of Effects of Continuous Predictors . . . . .	170
9.5.1	Example: Predictor Effects in TBI . . . . .	171
9.6	Concluding Remarks . . . . .	172
9.6.1	Software . . . . .	172
<b>10</b>	<b>Restrictions on Candidate Predictors . . . . .</b>	<b>175</b>
10.1	Selection Before Studying the Predictor–Outcome Relationship . . . . .	175
10.1.1	Selection Based on Subject Knowledge . . . . .	175
10.1.2	Example: Too Many Candidate Predictors . . . . .	176
10.1.3	Meta-Analysis for Candidate Predictors . . . . .	176
10.1.4	Example: Predictors in Testicular Cancer . . . . .	176
10.1.5	Selection Based on Distributions . . . . .	177
10.2	Combining Similar Variables . . . . .	177
10.2.1	Example: Coding of Comorbidity . . . . .	178
10.2.2	Assessing the Equal Weights Assumption . . . . .	178
10.2.3	Logical Weighting . . . . .	179
10.2.4	Statistical Combination . . . . .	180
10.3	Averaging Effects . . . . .	180
10.3.1	Example: Chlamydia Trachomatis Infection Risks . . . . .	180
10.3.2	Example: Acute Surgery Risk Relevant for Elective Patients? . . . . .	180
10.4	Case study: Family History for Prediction of a Genetic Mutation . . . . .	181
10.4.1	Clinical Background and Patient Data . . . . .	181
10.4.2	Similarity of Effects . . . . .	182
10.4.3	CRC and Adenoma in a Proband . . . . .	184
10.4.4	Age of CRC in Family History . . . . .	185
10.4.5	Full Prediction Model for Mutations . . . . .	186
10.5	Concluding Remarks . . . . .	187
<b>11</b>	<b>Selection of Main Effects . . . . .</b>	<b>191</b>
11.1	Predictor Selection . . . . .	191
11.1.1	Reduction Before Modelling . . . . .	191
11.1.2	Reduction While Modelling . . . . .	192
11.1.3	Collinearity . . . . .	192
11.1.4	Parsimony . . . . .	193
11.1.5	Should Non-Significant Variables Be Removed? . . . . .	193
11.1.6	Summary Points . . . . .	194
11.2	Stepwise Selection . . . . .	194
11.2.1	Stepwise Selection Variants . . . . .	194
11.2.2	Stopping Rules in Stepwise Selection . . . . .	195

11.3	Advantages of Stepwise Methods . . . . .	196
11.4	Disadvantages of Stepwise Methods . . . . .	197
11.4.1	Instability of selection . . . . .	197
11.4.2	Biased Estimation of Coefficients . . . . .	199
11.4.3	Bias of Stepwise Selection and Events Per Variable . . . . .	199
11.4.4	Misspecification of Variability . . . . .	201
11.4.5	Exaggeration of P-Values . . . . .	204
11.4.6	Predictions of Worse Quality Than from a Full Model . . . . .	204
11.5	Influence of Noise Variables . . . . .	205
11.6	Univariate Analyses and Model Specification . . . . .	206
11.6.1	Pros and Cons of Univariate Pre-Selection . . . . .	207
11.6.2	Testing of Predictors within Domains . . . . .	207
11.7	Modern Selection Methods . . . . .	207
11.7.1	Bootstrapping for Selection . . . . .	208
11.7.2	Bagging and Boosting . . . . .	208
11.7.3	Bayesian Model Averaging (BMA) . . . . .	208
11.7.4	Practical Advantages of BMA . . . . .	209
11.7.5	Shrinkage of Regression Coefficients to Zero . . . . .	210
11.8	Concluding Remarks . . . . .	210
<b>12</b>	<b>Assumptions in Regression Models: Additivity and Linearity . . . . .</b>	<b>213</b>
12.1	Additivity and Interaction Terms . . . . .	213
12.1.1	Potential Interaction Terms to Consider . . . . .	214
12.1.2	Interactions with Treatment . . . . .	214
12.1.3	Other Potential Interactions . . . . .	215
12.1.4	Example: Time and Survival After Valve Replacement . . . . .	216
12.2	Selection, Estimation and Performance with Interaction Terms . . . . .	216
12.2.1	Example: Age Interactions in GUSTO-I . . . . .	217
12.2.2	Estimation of Interaction Terms . . . . .	217
12.2.3	Better Prediction with Interaction Terms? . . . . .	219
12.2.4	Summary Points . . . . .	220
12.3	Non-linearity in Multivariable Analysis . . . . .	220
12.3.1	Multivariable Restricted Cubic Splines (RCS) . . . . .	220
12.3.2	Multivariable Fractional Polynomials (FP) . . . . .	221
12.3.3	Multivariable Splines in GAM . . . . .	222
12.4	Example: Non-Linearity in Testicular Cancer Case Study . . . . .	222
12.4.1	Details of Multivariable FP and GAM Analyses . . . . .	224
12.4.2	GAM in Univariate and Multivariable Analysis . . . . .	224
12.4.3	Predictive Performance . . . . .	226
12.4.4	R code for Non-Linear Modelling . . . . .	227
12.5	Concluding Remarks . . . . .	227
12.5.1	Recommendations . . . . .	228

<b>13 Modern Estimation Methods . . . . .</b>	231
13.1 Predictions from Regression and Other Models . . . . .	231
13.2 Shrinkage . . . . .	232
13.2.1 Uniform Shrinkage . . . . .	233
13.2.2 Uniform Shrinkage in GUSTO-1 . . . . .	233
13.3 Penalized Estimation. . . . .	234
13.3.1 Penalized Maximum Likelihood Estimation. . . . .	234
13.3.2 Penalized ML in Sample4 . . . . .	235
13.3.3 Shrinkage, Penalization, and Model Selection . . . . .	238
13.4 Lasso . . . . .	238
13.4.1 Estimation of Lasso Model . . . . .	238
13.4.2 Lasso in GUSTO-I . . . . .	239
13.4.3 Predictions after Shrinkage . . . . .	239
13.4.4 Model Performance after Shrinkage . . . . .	240
13.5 Concluding Remarks. . . . .	240
<b>14 Estimation with External Information . . . . .</b>	243
14.1 Combining Literature and Individual Patient Data . . . . .	243
14.1.1 Adaptation Method 1 . . . . .	244
14.1.2 Adaptation Method 2 . . . . .	244
14.1.3 Estimation . . . . .	245
14.1.4 Simulation Results . . . . .	245
14.1.5 Performance of Adapted Model . . . . .	247
14.1.6 Improving Calibration . . . . .	247
14.2 Example: Mortality of Aneurysm Surgery . . . . .	248
14.2.1 Meta-Analysis . . . . .	248
14.2.2 Individual Patient Data Analysis . . . . .	249
14.2.3 Adaptation Results . . . . .	250
14.3 Alternative Approaches . . . . .	251
14.3.1 Overall Calibration . . . . .	251
14.3.2 Bayesian Methods: Using Data Priors to Regression Modelling . . . . .	251
14.3.3 Example: Predicting Neonatal Death . . . . .	252
14.3.4 Example: Mortality of Aneurysm Surgery . . . . .	252
14.4 Concluding Remarks . . . . .	253
<b>15 Evaluation of Performance . . . . .</b>	255
15.1 Overall Performance Measures . . . . .	255
15.1.1 Explained Variation: $R^2$ . . . . .	255
15.1.2 Brier Score . . . . .	257
15.1.3 Example: Performance of Testicular Cancer Prediction Model . . . . .	257
15.1.4 Overall Performance Measures in Survival . . . . .	258

15.1.5	Decomposition in Discrimination and Calibration . . . . .	259
15.1.6	Summary Points . . . . .	259
15.2	Discriminative Ability . . . . .	260
15.2.1	Sensitivity and Specificity of Prediction Models . . . . .	260
15.2.2	Example: Sensitivity and Specificity of Testicular Cancer Prediction Model . . . . .	260
15.2.3	ROC Curve . . . . .	260
15.2.4	R <sup>2</sup> vs. c . . . . .	262
15.2.5	Box Plots and Discrimination Slope . . . . .	264
15.2.6	Lorenz Curve . . . . .	264
15.2.7	Discrimination in Survival Data . . . . .	267
15.2.8	Example: Discrimination of Testicular Cancer Prediction Model . . . . .	268
15.2.9	Verification Bias and Discriminative Ability . . . . .	269
15.2.10	R Code . . . . .	269
15.3	Calibration . . . . .	270
15.3.1	Calibration Plot . . . . .	270
15.3.2	Calibration in Survival . . . . .	271
15.3.3	Calibration-in-the-Large . . . . .	271
15.3.4	Calibration Slope . . . . .	272
15.3.5	Estimation of Calibration-in-the-Large and Calibration Slope . . . . .	272
15.3.6	Other Calibration Measures . . . . .	273
15.3.7	Calibration Tests . . . . .	274
15.3.8	Goodness-of-Fit Tests . . . . .	274
15.3.9	Calibration of Survival Predictions . . . . .	276
15.3.10	Example: Calibration in Testicular Cancer Prediction Model . . . . .	276
15.3.11	Calibration and Discrimination . . . . .	278
15.3.12	R Code . . . . .	278
15.4	Concluding Remarks . . . . .	278
15.4.1	Bibliographic Notes . . . . .	279
16	Clinical Usefulness . . . . .	281
16.1	Clinical Usefulness . . . . .	281
16.1.1	Intuitive Approach to the Cutoff . . . . .	282
16.1.2	Decision-Analytic Approach to the Cutoff . . . . .	282
16.1.3	Error Rate and Accuracy . . . . .	283
16.1.4	Accuracy Measures for Clinical Usefulness . . . . .	284
16.1.5	Decision Curves . . . . .	284
16.1.6	Examples of NB in Decision Curves . . . . .	285
16.1.7	Example: Clinical Usefulness of Prediction Model for Testicular Cancer . . . . .	286

16.1.8	Decision Curves for Testicular Cancer Example . . . . .	287
16.1.9	Verification Bias and Clinical Usefulness. . . . .	288
16.1.10	R Code . . . . .	289
16.2	Discrimination, Calibration, and Clinical Usefulness. . . . .	289
16.2.1	Aim of the Prediction Model and Performance Measures . . . . .	290
16.2.2	Summary Points . . . . .	291
16.3	From Prediction Models to Decision Rules . . . . .	291
16.3.1	Performance of Decision Rules . . . . .	292
16.3.2	Treatment Benefit in Prognostic Subgroups . . . . .	294
16.3.3	Evaluation of Classification Systems . . . . .	294
16.4	Concluding Remarks. . . . .	295
16.4.1	Bibliographic notes . . . . .	296
<b>17</b>	<b>Validation of Prediction Models. . . . .</b>	<b>299</b>
17.1	Internal vs. External Validation, and Validity . . . . .	299
17.2	Internal Validation Techniques . . . . .	300
17.2.1	Apparent Validation . . . . .	300
17.2.2	Split-Sample Validation . . . . .	301
17.2.3	Cross-Validation . . . . .	302
17.2.4	Bootstrap Validation . . . . .	303
17.3	External Validation Studies. . . . .	304
17.3.1	Temporal Validation . . . . .	305
17.3.2	Example: Development and Validation of a Model for Lynch Syndrome . . . . .	306
17.3.3	Geographic Validation . . . . .	307
17.3.4	Fully Independent Validation . . . . .	308
17.3.5	Reasons for Poor Validation . . . . .	309
17.4	Concluding Remarks. . . . .	310
<b>18</b>	<b>Presentation Formats . . . . .</b>	<b>313</b>
18.1	Prediction vs. Decision Rules . . . . .	313
18.2	Clinical Prediction Models . . . . .	315
18.2.1	Regression Formula . . . . .	315
18.2.2	Confidence Intervals for Predictions. . . . .	316
18.2.3	Nomograms. . . . .	317
18.2.4	Score Chart . . . . .	319
18.2.5	Tables with Predictions . . . . .	320
18.2.6	Specific Formats . . . . .	321
18.3	Case Study: Clinical Prediction Model for Testicular Cancer Model . . . . .	321
18.3.1	Regression Formula from Logistic Model . . . . .	321
18.3.2	Nomogram . . . . .	324

18.3.3	Score Chart . . . . .	324
18.3.4	Coding with Categorization . . . . .	327
18.3.5	Summary Points . . . . .	327
18.4	Clinical Decision Rules. . . . .	328
18.4.1	Regression Tree. . . . .	328
18.4.2	Score Chart Rule. . . . .	328
18.4.3	Survival Groups . . . . .	329
18.4.4	Meta-Model. . . . .	329
18.5	Concluding Remarks. . . . .	330

### Part III Generalizability of Prediction Models

19	Patterns of External Validity . . . . .	333
19.1	Determinants of External Validity . . . . .	335
19.1.1	Case-Mix. . . . .	335
19.1.2	Differences in Case-Mix. . . . .	336
19.1.3	Differences in Regression Coefficients. . . . .	336
19.2	Impact on Calibration, Discrimination, and Clinical Usefulness. . . . .	337
19.2.1	Simulation Set-Up. . . . .	338
19.2.2	Performance Measures . . . . .	339
19.3	Distribution of Predictors . . . . .	340
19.3.1	More- or Less-Severe Case-Mix According to X . . . . .	340
19.3.2	Example: Interpretation of Testicular Cancer Validation . . . . .	341
19.3.3	More or Less Heterogeneous Case-Mix According to X . . . . .	341
19.3.4	More- or Less-Severe Case-Mix According to Z . . . . .	342
19.3.5	More or Less Heterogeneous Case-Mix According to Z . . . . .	344
19.4	Distribution of Observed Outcomes $Y$ . . . . .	344
19.5	Coefficients $\beta$ . . . . .	345
19.5.1	Coefficient of Linear Predictor $< 1$ . . . . .	345
19.5.2	Coefficients Different. . . . .	346
19.5.3	R Code . . . . .	346
19.5.4	Influence of Different Coefficients . . . . .	347
19.5.5	Other Scenarios of Invalidity . . . . .	348
19.5.6	Summary of Patterns of Invalidity . . . . .	348
19.6	Reference Values for Performance . . . . .	349
19.6.1	Calculation of Reference Values. . . . .	349
19.6.2	R Code . . . . .	350
19.6.3	Performance with Refitting. . . . .	350
19.6.4	Examples: Testicular Cancer and TBI . . . . .	351

19.7	Estimation of Performance .....	352
19.7.1	Uncertainty in Validation of Performance .....	352
19.7.2	Estimating Standard Errors in Validation Studies.....	354
19.7.3	Summary Points .....	354
19.8	Design of External Validation Studies .....	355
19.8.1	Power of External Validation Studies .....	355
19.8.2	Required Sample Sizes for Validation Studies .....	356
19.8.3	Summary Points .....	357
19.9	Concluding Remarks.....	358
<b>20</b>	<b>Updating for a New Setting .....</b>	<b>361</b>
20.1	Updating the Intercept.....	361
20.1.1	Simple Updating Methods .....	362
20.1.2	Bayesian Updating .....	362
20.2	Approaches to More-Extensive Updating.....	363
20.2.1	A comparison of Eight Updating Methods.....	364
20.3	Case Study: Validation and Updating in GUSTO-I .....	366
20.3.1	Validity of TIMI-II Model for GUSTO-I .....	366
20.3.2	Updating the TIMI-II Model for GUSTO-I .....	368
20.3.3	Performance of Updated Models .....	369
20.3.4	R Code for Updating Methods .....	370
20.4	Shrinkage and Updating .....	371
20.4.1	Example: Shrinkage towards Re-calibrated Values in GUSTO-I.....	371
20.4.2	R code for Shrinkage and Penalization in Updating....	372
20.5	Sample Size and Updating Strategy .....	373
20.5.1	Simulations of Sample Size, Shrinkage, and Updating Strategy .....	374
20.6	Validation and Updating of Tree Models .....	376
20.6.1	Example: Tree Modelling in Testicular Cancer .....	377
20.7	Validation and Updating of Survival Models .....	378
20.7.1	Case Study: Validation of a Simple Indexfor Non-Hodgkin's Lymphoma .....	379
20.7.2	Updating the Prognostic Index .....	380
20.7.3	Re-calibration for Groups by Time Points .....	380
20.7.4	Re-calibration with a Cox Regression Model.....	381
20.7.5	Parametric Re-calibration .....	382
20.7.6	Summary Points .....	384
20.8	Continuous Updating .....	384
20.8.1	A Continuous Updating Strategy .....	385
20.8.2	Example: Continuous Updating in GUSTO-I.....	386
20.9	Concluding Remarks.....	388

<b>21 Updating for Multiple Settings . . . . .</b>	391
21.1 Differences Between Settings . . . . .	391
21.1.1 Testing for Calibration-in-the Large . . . . .	391
21.1.2 Illustration of Heterogeneity in GUSTO-I . . . . .	392
21.1.3 Updating for Better Calibration-in-the Large . . . . .	393
21.1.4 Empirical Bayes Estimates . . . . .	394
21.1.5 Illustration of Updating in GUSTO-I . . . . .	394
21.1.6 Testing and Updating of Predictor Effects . . . . .	396
21.1.7 Heterogeneity of Predictor Effects in GUSTO-I . . . . .	396
21.1.8 R Code for Random Effect Analyses . . . . .	397
21.2 Provider Profiling . . . . .	398
21.2.1 Indicators for Differences Between Centres . . . . .	398
21.2.2 Ranking of Centres . . . . .	399
21.2.3 Example: Provider Profiling in Stroke . . . . .	401
21.2.4 Testing of Differences Between Centres . . . . .	401
21.2.5 Estimation of Differences Between Centres . . . . .	402
21.2.6 Uncertainty in Differences . . . . .	403
21.2.7 Ranking of Centres . . . . .	404
21.2.8 Essential R Code for Provider Profiling . . . . .	405
21.2.9 Guidelines for Provider Profiling . . . . .	406
21.3 Concluding Remarks . . . . .	406
21.3.1 Bibliographic Notes . . . . .	407

## Part IV Applications

<b>22 Prediction of a Binary Outcome: 30-Day Mortality After Acute Myocardial Infarction . . . . .</b>	411
22.1 GUSTO-I Study . . . . .	411
22.1.1 Acute Myocardial Infarction . . . . .	411
22.1.2 Treatment Results from GUSTO-I . . . . .	412
22.1.3 Prognostic Modelling in GUSTO-I . . . . .	412
22.2 General Considerations of Model Development . . . . .	415
22.2.1 Research Question and Intended Application . . . . .	415
22.2.2 Outcome and Predictors . . . . .	416
22.2.3 Study Design and Analysis . . . . .	416
22.3 Seven Modelling Steps in GUSTO-I . . . . .	417
22.3.1 Data Inspection . . . . .	417
22.3.2 Coding of Predictors . . . . .	418
22.3.3 Model Specification . . . . .	418
22.3.4 Model Estimation . . . . .	418
22.3.5 Model Performance . . . . .	419
22.3.6 Model Validation . . . . .	419
22.3.7 Presentation . . . . .	420

22.4	Validity .....	421
22.4.1	Internal Validity: Overfitting.....	421
22.4.2	External Validity: Generalizability .....	421
22.4.3	Summary Points .....	421
22.5	Translation into Clinical Practice .....	422
22.5.1	Score Chart for Choosing Thrombolytic Therapy .....	422
22.5.2	Predictions for Choosing Thrombolytic Therapy .....	423
22.5.3	Covariate Adjustment in GUSTO-I .....	424
22.6	Concluding Remarks.....	425
<b>23</b>	<b>Case Study on Survival Analysis: Prediction of Secondary Cardiovascular Events .....</b>	<b>427</b>
23.1	Prognosis in the SMART Study .....	427
23.1.1	Patients in SMART .....	428
23.2	General Considerations in SMART .....	429
23.2.1	Research Question and Intended Application.....	429
23.2.2	Outcome and Predictors .....	429
23.2.3	Study Design and Analysis.....	432
23.3	Data Inspection Steps in the SMART Cohort.....	432
23.4	Coding of Predictors .....	435
23.4.1	Extreme Values .....	435
23.4.2	Transforming Continuous Predictors .....	436
23.4.2	Combining Predictors with Similar Effects .....	437
23.5	Model Specification .....	438
23.5.1	Selection .....	440
23.6	Model Estimation, Performance, Validation, and Presentation ..	440
23.6.1	Model Estimation .....	440
23.6.2	Model Performance.....	442
23.6.3	Model Validation: Stability .....	442
23.6.4	Model Validation: Optimism.....	444
23.6.5	Model Presentation .....	444
23.7	Concluding Remarks.....	444
<b>24</b>	<b>Lessons from Case Studies .....</b>	<b>447</b>
24.1	Sample Size.....	447
24.1.1	Example: Sample Size and Number of Predictors .....	447
24.1.2	Number of Predictors .....	448
24.1.3	Potential Solutions .....	449
24.2	Validation .....	450
24.2.1	Examples of Internal and External Validation .....	450
24.3	Subject Matter Knowledge .....	451
24.4	Data Sets .....	452
24.4.1	GUSTO-I Prediction Models .....	453
24.4.2	Modern Learning Methods in GUSTO-I .....	453

24.4.3	Modelling Strategies in Small Data Sets from GUSTO-I .....	453
24.4.4	SMART Case Study .....	453
24.4.5	Testicular Cancer Case Study .....	455
24.4.6	Abdominal Aortic Aneurysm Case Study.....	455
24.4.7	Traumatic Brain Injury Data Set.....	459
24.5	Concluding Remarks.....	459
<b>References</b>	.....	463
<b>Index</b>	.....	487

# Chapter 1

## Introduction

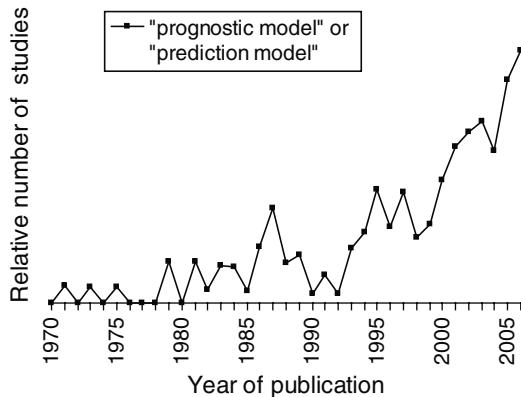
### 1.1 Prognosis and Prediction in Medicine

Prognosis is central to medicine. All diagnostic and therapeutic actions aim to improve prognosis:

- *Screening*: If we screen for early signs of disease, we may, for example, find cancers early in their course of disease, and treat them better than when they were detected later. But whether screening is useful depends on the improvement in prognosis that is achieved compared to a no screening strategy. Some cancers may not have caused any impact on life expectancy, while side-effects of treatment may be substantial.
- *Diagnosis*: If we do a diagnostic test, we may detect an underlying disease. But some diseases are not treatable, or the natural course might be very similar to what is achieved with treatment.
- *Therapy*: New treatments become available nearly every day, but their impact on prognosis is often rather limited, despite high hopes at early stages. “Magic bullets” are rare. Treatment effects are often small relative to the effects of determinants of the natural history of a disease, such as the patient’s age. The individual benefits need to exceed any side effects, harms and economic costs.

#### 1.1.1 *Prediction Models and Decision-Making*

Physicians and health policy makers need to make predictions on the prognosis of a disease, or the likelihood of an underlying disease, in their decision-making on screening and treatment of disease in high-risk groups, diagnostic work-up (e.g. ordering another, possibly risky or expensive test), and choice of therapy. Traditionally, the probabilities of diagnostic and prognostic outcomes were implicitly assessed for such decision-making. Medicine was much more subjective than in the current era of “evidence-based medicine,” which can be defined as “the conscientious, explicit and judicious use of current best evidence in making decisions



**Fig. 1.1** Studies in PubMed with the terms “prognostic model” or “prediction model” in the title, published between 1970 and 2005, as a fraction of the total number of studies in PubMed (a total of 676,000 in 2005)

about the care of individual patients.”<sup>317, 362</sup> Evidence-based medicine applies the scientific method to medical practice.<sup>160</sup>

Another development is that we are moving towards “shared decision-making,” where physicians and patients both actively participate in deciding on choices for diagnostic tests and therapeutic interventions.<sup>84</sup> For shared decision-making, adequate communication about risks and benefits is a pre-requisite.

Clinical prediction models may provide the evidence-based input for shared decision-making, by providing estimates of the individual probabilities of risks and benefits.<sup>249</sup> Clinical prediction models are also referred to as clinical prediction rules, prognostic models, or nomograms.<sup>351</sup> Clinical prediction models combine a number of characteristics (e.g., related to the patient, the disease, or treatment) to predict a diagnostic or prognostic outcome. Typically, a limited number of predictors is considered (say between 2 and 20). Publications with clinical prediction models have increased steeply over recent years (Fig. 1.1).

## 1.2 Statistical Modelling for Prediction

Prediction is primarily an estimation problem. For example: What is the risk of dying of this patient within 30 days after an acute myocardial infarction? Or, what is the expected 2-year survival rate for this patient with oesophageal cancer? Prediction is also about testing of hypotheses. For example, is age a predictor of 30-day mortality after an acute myocardial infarction? How important is nutritional status for survival of a patient with oesophageal cancer? Or more general: what are the most important predictors in a certain disease? Are some predictors correlated with each other, such that their apparent predictive effects are explained by other

predictor variables? The latter question comes close to aetiological research, where biases such as confounding are among the major concerns of epidemiologists.

Statistical models may serve to address both estimation and hypothesis testing questions. In the medical literature, much emphasis has traditionally been given to the identification of predictors. Over 60,000 papers have been published with the terms “predictor” or “prognostic factor” (PubMed, January 2007). It is nowadays widely recognized that the prognostic value of a predictor has to be shown in addition to already known, easily measurable predictors.<sup>386</sup> For example, the prognostic value of a new genetic marker would need to be assessed for additional value over classical, well-established predictors.<sup>225</sup> Such thorough evaluations are however still less common, and require statistical modelling.

Statistical models summarize patterns of the data available for analysis. In doing so, it is inevitable that assumptions have to be made. Some of these assumptions can be tested, for example, whether predictor effects work in an additive way, and whether continuous predictors have reasonably linear effects. Testing of underlying assumptions is especially important if specific claims are made on the effect of a predictor (Chaps. 4, 6, and 12).

Statistical models for prediction can be discerned in main classes: regression, classification, and neural networks.<sup>181</sup> The characteristics of alternative models are discussed in Chaps. 4 and 6. The main focus in this book is on regression models, which are the most widely used in the medical field. We consider situations where the number of candidate predictor variables is limited, say below 25. This is in contrast to research in areas such as genomics (genetic effects), proteomics (protein effects), or metabolomics (metabolite effects). In these areas more complex data are generated, with larger numbers of candidate predictors (often  $> 10,000$ ). Moreover, we assume that subject knowledge is available, from previous empirical studies and from experts on the topic (e.g. medical doctors treating patients with the condition under study).

### 1.2.1 *Model Uncertainty*

Statistical modelling to make predictions encounters various challenges, including dealing with model uncertainty and limited sample size. Model uncertainty arises from the fact that we usually do not fully pre-specify a model before we fit it to a data set.<sup>69, 101</sup> An iterative process is often followed with model checking and model modification. On the other hand, standard statistical methods assume that a model was pre-specified. In that case, parameter estimates such as regression coefficients, their corresponding standard errors, 95% confidence intervals, and  $p$ -values are largely unbiased. When the structure of a model was at least partly based on findings in the data, bias may occur, and we underestimate the uncertainty of conclusions drawn from the model.

Fortunately, some statistical tools have become available which help to study model uncertainty. Especially, a statistical re-sampling procedure named “bootstrapping” is helpful for many aspects of model development and validation.<sup>108</sup> The bootstrap hence is an important tool in prediction research (Chaps. 5 and 17).

### 1.2.2 Sample Size

A sufficient sample size is important to address any scientific question with empirical data. First, we have to realize that the effective sample size may often be much smaller than indicated by the total number of subjects in a study.<sup>174</sup> For example, when we study complications of a procedure that occur with an incidence of 0.1%, a study with 10,000 patients will contain only 10 events. The number 10 determines the effective sample size in such a study. In small samples, model uncertainty may be large, and we may not be able to derive reliable predictions from a model.

Second, a large sample size facilitates many aspects of prediction research. For example, large-scale international collaborations are increasingly set up to allow for the identification of gene-disease associations.<sup>211</sup> For multivariable prognostic modelling, a large sample size allows for selection of predictors with simple automatic procedures such as stepwise methods with  $p<0.05$  and reliable testing of model assumptions. An example is the prediction of 30-day mortality after an acute myocardial infarction, where Lee et al. derived a prediction model with 40,830 patients of whom 2,850 died.<sup>255</sup> This example will be used throughout this book, with a thorough description in Chap. 22. In practice, we often have relatively small samples available. For example, a review of 31 prognostic models in traumatic brain injury showed that 22 were based on samples with less than 500 patients.<sup>307</sup> The main challenges are hence with the development of a good prediction model with a relatively small study sample.

Third, with small sample size we have to be prepared to make stronger modeling assumptions. For example, Altman illustrates the use of a parametric test (ANOVA) to compare 3 groups with 8, 9, and 5 patients in his seminal text “Practical statistics for medical research”.<sup>8</sup> With larger samples, we would more readily switch to a non-parametric test such as a Kruskal–Wallis test. With small sample size, we may have to assume linearity of a continuous predictor (Chap. 9) and no interaction between predictors (Chap. 13). We will subsequently have limited power to test deviations from these model assumptions. It hence becomes more important what our starting point of the analysis is. From a Bayesian viewpoint, we could say that our prior information becomes more important, since the information contributed by our study is limited.

Fourth, we have to match our ambitions in research questions with the effective sample size that is available. When the sample size is very small, we should only ask relatively simple questions, while more complex questions can be addressed with larger sample sizes. A question such as: “What are the most important predictors in this prediction problem” is actually more complex than a question such as “What are the predictions of the outcome given this set of predictors” (Chap. 11). Table 1.1 lists questions on predictors (known or determined from the data?), functional form (known or determined from the data?), and regression coefficients (known or determined from the data?) and the consequence for the required sample size in a study.

**Table 1.1** Stages of development of regression models and implications for modelling approach and required sample size (see [http://e-collection.ethbib.ethz.ch/ecol-pool/incoll/incoll\\_102.pdf](http://e-collection.ethbib.ethz.ch/ecol-pool/incoll/incoll_102.pdf))

Predictors known?	Functional form known?	Coefficients known?	Approach	Required sample size
-	-	-	Development from scratch	Very large
+	-	-	Specification of transformations and/or interactions	Large
+	+	-	Estimated regression coefficients	Modest
+	+	+	Validation and updating	Modest

## 1.3 Structure of the Book

This book consists of four parts. Part I provides background on developing and applying prediction models in medicine. Part II is central for model development, while Part III focuses on applicability in external settings and advanced issues related to model modification and model extension (“updating”). Part IV is practical in nature with a detailed description of predictive modelling in two case studies, some lessons learned for model development, and a description of medical problems with publicly available data sets.

### 1.3.1 Part I: Prediction Models in Medicine

This book starts with an overview of various applications of prediction models in clinical practice and in medical research (Chap. 2). Next, we note that the quality of a statistical model depends to a large extent on the design and quality of the data used in the analysis. A sophisticated analysis cannot salvage a poorly designed study, or poor data collection procedures. Several considerations are presented around the design of cohort studies for prognostic models, and cross-sectional studies for diagnostic models (Chap. 3). Various statistical techniques, each having their strengths and limitations can be considered for a prediction model. An overview of more and less flexible models for different types of outcomes is presented in Chap. 4. Unfortunately, prediction models commonly suffer from a methodological problem, which is known as “overfitting.” This means that idiosyncrasies in the data are fitted rather than generalizable patterns.<sup>174</sup> A model may hence not be applicable to new patients, even when the setting of application is very similar to the development setting. Statistical optimism is discussed with possible solutions in Chap. 5. Chapter 6 discusses considerations in choosing between alternative models, and presents some empirical comparisons on the quality of predictions derived with alternative modelling techniques.

### 1.3.2 Part II: Developing Valid Prediction Models

The core of this book is a proposal for seven steps to consider in developing valid prediction models with regression analysis. We present a checklist for model development, which is intended to give a structure to model building and validation.

In Chaps. 7–18 we discuss seven modelling steps.

1. A preliminary step is to carefully consider the prediction problem: what are the research questions, what is already known about predictors? Next, we consider the data under study: how are the predictors defined, what is the outcome of interest? An important issue is that missing values will occur in at least some of the predictors under study. We discuss and propose approaches to deal with missing values in Chaps. 7 and 8.
2. When we start on building a prediction model, the first issue is the coding of predictors for a model; several choices need to be considered on categorical variables and continuous variables (Chaps. 9 and 10).
3. We then move to the most thorny issue in prediction modelling: how to specify a model (Chaps. 11 and 12). What predictors should we include, what are the pros and cons of stepwise selection methods, and how should we deal with assumptions in models such as additivity and linearity?
4. Once a model is specified, we need to estimate model parameters. For regression models, we estimate coefficients for each predictor or combination of predictors in the model. We consider classical and more modern estimation methods for regression models (Chaps. 13 and 14). Several techniques are discussed which aim to limit the overfitting of a model to the available data.
5. For a specified and estimated model, we need to determine the quality. Several performance measures are commonly used, as discussed in Chap. 15. Most relevant to clinical practice is whether the model is useful; this can be quantified with some more novel performance measures (Chap. 16).
6. Since overfitting is a central problem in prediction modelling, we need to consider the validity of our model for new patients. In Chap. 17, we concentrate on statistical techniques to evaluate the internal validity of a model, i.e., for the underlying population that the sample originated from. Internal validation addresses statistical problems in the specification and estimation of a model (“reproducibility”).<sup>222</sup>
7. A final step to consider is the presentation of a prediction model. Regression formulas can be used, but many alternatives are possible for easier applicability of a model (Chap. 18).

### 1.3.3 Part III: Generalizability of Prediction Models

Generalizability (or external validity) of a model relates to the applicability of a model to a different setting.<sup>222</sup> External validity of a model cannot be expected if there is no internal validity. Steps 1–7 in Part II support the development of

internally valid prediction models. The performance may be lower when a model is applied in a new setting because of genuine differences between the new setting and the development setting. Examples of a different setting include a hospital different from the development hospital, a more recent time period, and a different selection of patients. We systematically consider patterns of invalidity that may arise when externally validating a model (Chap. 19).

To improve predictions for a new setting, we need to consider whether we can make modifications and extensions to the model. Various parsimonious techniques are available to achieve such updating (Chap. 20). When several settings are considered, we may use more advanced updating methods, including Empirical Bayes methods. Moreover, we may specifically be interested in ranking of providers of care (“provider profiling,” Chap. 21).

### ***1.3.4 Part IV: Applications***

A central case study in this book is formed by the GUSTO-I trial. Patients in this trial suffered from an acute myocardial infarction. We study 30-day mortality in relation to various predictors.<sup>255</sup> Overfitting is not a concern in the full data set ( $n=40,830$  patients, 2,850 died within 30 days), but modelling is more challenging in small parts of this data set, which are made publicly available for applying the concepts and techniques presented in this book. We discuss the logistic regression model developed from the GUSTO-I patients in Chap. 22.

A further case study concerns a survival problem. We aim to predict secondary cardiovascular events among a hospital-based cohort. The seven steps to develop a prediction model are systematically considered (Chap. 23).

Finally, we try to give some practical advice on the main issues in prediction modelling, and describe the medical problems used throughout the text and available data sets (Chap. 24).

### ***1.3.5 Questions and Exercises***

Each chapter ends with a few questions to test insight in the material presented. Furthermore, practical exercises are available from the book’s web site (<http://www.clinicalpredictionmodels.org>), involving work with data sets in R software (<http://www.cran.r-project.org>).

# **Part I**

## **Prediction Models in Medicine**

# Chapter 2

## Applications of Prediction Models

**Background** In this chapter, we consider several areas of application of prediction models in public health, clinical practice, and medical research. We use several small case studies for illustration.

### 2.1 Applications: Medical Practice and Research

Broadly speaking, prediction models are valuable for medical practice and for research purposes (Table 2.1). In public health, prediction models may help to target preventive interventions to subjects at relatively high risk of having or developing a disease. In clinical practice, prediction models may inform patients and their treating physicians on the probability of a diagnosis or a prognostic outcome. Prognostic estimates may for example be useful for planning of remaining life-time in terminal disease; or give hope for recovery if a good prognosis is expected after an acute event such as a stroke. Classification of a patient according to his/her risk may also be useful for communication among physicians. A key condition for this type of application of a prediction model is that predictions are reliable. This means that when a 10% risk is predicted, on average 10% of patients with these characteristics should have the outcome (“calibration”, Chap. 4 and 15).

In the diagnostic work-up, predictions can be useful to estimate the probability that a disease is present. When the probability is relatively high, treatment is indicated; if the probability is low, no treatment is indicated and further diagnostic testing may be considered necessary. In therapeutic decision-making, treatment should only be given to those who benefit from the treatment. Prognostic predictions may support the weighing of harms vs. individual benefits. If risks of a poor outcome are relatively low, the maximum benefit will also be relatively low. Any harm, such as a side effect of treatment, may then readily outweigh any benefits. The claim of prediction models is that better decisions can be made with a model than without.

In research, prediction models may assist in the design and analysis of randomized trials. Models are also useful to control for confounding variables in observational research, either in traditional regression analysis or with modern approaches such

**Table 2.1** Some areas of application of clinical prediction models

Application area	Example in this chapter
<i>Public health</i>	
Targeting of preventive interventions	
Incidence of disease	Models for (hereditary) breast cancer
<i>Clinical practice</i>	
Diagnostic work-up	
Test ordering	Probability of renal artery stenosis
Starting treatment	Probability of deep venous thrombosis
Therapeutic decision-making	
Surgical decision making	Replacement of risky heart valves
Intensity of treatment	More intensive chemotherapy in cancer patients
Delaying treatment	Spontaneous pregnancy chances
<i>Research</i>	
Inclusion in an RCT	Traumatic brain injury
Covariate adjustment in an RCT	Primary analysis of GUSTO-III
Confounder adjustment with a propensity score	Statin effects on mortality
Case-mix adjustment	Provider profiling

as “propensity scores”. Several areas of application are discussed in the next sections.

## 2.2 Prediction Models for Public Health

### 2.2.1 Targeting of Preventive Interventions

Various models have been developed to predict the future occurrence of disease in asymptomatic subjects in the population. Well-known examples include the Framingham risk functions for cardiovascular disease.<sup>487</sup> The Framingham risk functions underpin several of the current policies for preventive interventions. For example, statin therapy is only considered for those with relatively high risk of cardiovascular disease. Similarly, prediction models have been developed for breast cancer, where more intensive screening or chemoprophylaxis can be considered for those at elevated risk.<sup>130,131</sup>

### \*2.2.2 Example: Incidence of Breast Cancer

In 1989, Gail et al. presented a by now famous risk prediction model for developing breast cancer.<sup>131</sup> The model was based on case-control data from the Breast Cancer Detection Demonstration Project (BCDDP). The BCDDP recruited 280,000 women from 1973 to 1980 who were monitored for 5 years. From this cohort, 2,852 white women developed breast cancer and 3,146 controls were selected, all with complete risk factor information. The model includes age at menarche, age at first live birth,

number of previous biopsies, and number of first-degree relatives with breast cancer. Individualized breast cancer probabilities were calculated from information on relative risks and the baseline hazard rate in the general population. The calculations accounted for competing risks (the risk of dying from other causes).

The predictions were validated later on other data sets from various populations, with generally favorable conclusions.<sup>83,94</sup> Practical application of the original model involved cumbersome calculations and interpolations. Hence, more easily applicable graphs were created to estimate the absolute risk of breast cancer for individual patients for intervals of 10, 20, and 30 years.<sup>33</sup> The absolute risk estimates have been used to design intervention studies, to counsel patients regarding their risks of disease, and to inform clinical decisions, such as whether or not to take tamoxifen to prevent breast cancer.<sup>132</sup>

Other models for breast cancer risk include the Claus model, which is useful to assess risk for familial breast cancer.<sup>74</sup> This is breast cancer that runs in families but is not associated with a known hereditary breast cancer susceptibility gene. Unlike the Gail model, the Claus model requires the exact ages at breast cancer diagnosis of first or second-degree relatives as an input.

Some breast cancers are caused by a mutation in a breast cancer susceptibility gene (BRCA), referred to as hereditary breast cancer. A suspicious family history for hereditary breast cancer includes many cases of breast and ovarian cancers, or family members with breast cancers under age 50. Simple tables have been published to determine the risk of a BRCA mutation, based on specific features of personal and family history.<sup>127</sup> Another model considers the family history in more detail (BRCA PRO<sup>323</sup>). It explicitly uses the genetic relationship in families, and is therefore labeled a Mendelian model. Calculations are based on Bayes' theorem. BRCA PRO was shown to perform at least as good as experienced genetic counselors.<sup>116</sup>

Friedenson provides an interesting overview of risk models in breast cancer and their clinical implications (Table 2.2).<sup>128</sup> Various measures are possible to reduce breast cancer risk, including behavior (e.g. exercise, weight control, alcohol intake) and medical interventions (e.g. tamoxifen use).

## 2.3 Prediction Models for Clinical Practice

### 2.3.1 *Decision Support on Test Ordering*

Prediction models may be useful to estimate the probability of an underlying disease, such that we can decide on further testing. When a diagnosis is very unlikely, no further testing is indicated, while more tests may be indicated when the diagnosis is not yet sufficiently certain for decision-making on therapy. Further testing usually involves one or more imperfect tests (sensitivity below 100%, specificity below 100%). Ideally, a gold standard test is available (sensitivity=100%, specificity=100%). A gold standard test is the diagnostic test that is regarded as definitive

**Table 2.2** Risk factors in four prediction models for breast cancer: two for breast cancer incidence, two for presence of mutation in BRCA1 or BRCA2 genes<sup>128</sup>

Risk factor	Gailmodel	Clausmodel	Myriad tables	BRCA PRO model
Woman's personal information				
Age	+	+	+	+
Race/ethnicity	+			
Ashkenazi Jewish			+	+
Breast biopsy	+			
Atypical hyperplasia	+			
Hormonal factors				
Age at menarche	+			
Age at first live birth	+			
Age at menopause	+			
Family history				
1st degree relatives with breast cancer	+	+	Age <50/≥50	Age for all affected
2nd degree relatives with breast cancer		+	Age <50/≥50	Age for all affected
1st or 2nd degree with ovarian cancer			+	Age for all affected
Bilateral breast cancer				+
Male breast cancer				+
Outcome predicted	Incident breast cancer			BRCA 1/2 mutation

in determining whether a subject has the disease. The gold standard test may not be suitable to apply in all subjects suspected of the disease because it is burdensome (e.g. invasive), or costly.

### \*2.3.2 Example: Predicting Renal Artery Stenosis

Renal artery stenosis is a rare cause of hypertension. The gold standard for diagnosing renal artery stenosis, renal angiography, is invasive and costly. Krijnen et al. aimed to develop a prediction rule for renal artery stenosis from clinical characteristics. The rule might be used to select patients for renal angiography.<sup>243</sup> Logistic regression analysis was performed with data from 477 hypertensive patients who underwent renal angiography. A simplified prediction rule was derived from the regression model for use in clinical practice. Age, sex, atherosclerotic vascular disease, recent onset of hypertension, smoking history, body mass index, presence of an abdominal bruit, serum creatinin concentration, and serum cholesterol level were selected as predictors. The diagnostic accuracy of the regression model was similar to that of renal scintigraphy, which had a sensitivity of 72% and a specificity of 90%. The conclusion was that this clinical prediction rule can help to select

patients for renal angiography in an efficient manner by reducing the number of angiographic procedures without the risk for missing many renal artery stenoses. The modelling steps summarized here will be described in more detail in Part II.

An interactive Excel program is available to calculate diagnostic predictions for individual patients. Figure 2.1 shows the example of a 45-year-old male with recent onset of hypertension. He smokes, has no signs of atherosclerotic vascular disease, a BMI<25, no abdominal bruit is heard, serum creatinin is 112 µmol/L, and serum cholesterol is not elevated. According to a score chart (see Chap. 18), the sum score was 11, corresponding to a probability of stenosis of 25%. According to exact logistic regression calculations, the probability was 28% [95% confidence interval 17–43%].

### 2.3.3 Starting Treatment: the Treatment Threshold

Decision analysis is a method to formally weigh pros and cons of decisions. For starting treatment after diagnostic work-up, a key concept is the treatment threshold. This threshold is defined as the probability where the expected benefit of treatment is equal to the expected benefit of avoiding treatment. If the probability of the diagnosis is lower than the threshold, no treatment is the preferred decision, and if the probability of the diagnosis is above the threshold, treatment is the preferred decision.<sup>325</sup> The threshold is determined by the relative weight of false-negative vs. false-positive decisions. If a false-positive decision is much less important than a false-negative decision, the threshold is low. For example, a 1:100 ratio leads to a 1% threshold. On the other hand, if false-positive decisions confer serious risks, the threshold should be higher. Further details on the threshold concept are beyond the scope of this book, but the issue returns when we discuss the performance of prediction models with decision curves<sup>469</sup> (Chap. 16).

	A	B	C	D	E	F	G	H
<b>Prediction rule for renal artery stenosis</b>								
<b>Predictor</b>								
3 Predictor								
4 Smoking		former or current =1						
5 Current age		years						
6 Gender		male = 1						
7 Atherosclerotic vascular disease*		yes = 1						
8 Onset of hypertension within 2 years		yes = 1						
9 Body mass index >= 25 kg/m <sup>2</sup>		yes = 1						
10 Presence of abdominal bruit		yes = 1						
11 Serum creatinine concentration		µmol/L						
12 Serum cholesterol level > 6.5 mmol/L**		yes = 1						
17 Sumscore								
18 Predicted probability of renal artery stenosis			Formula	Score chart				
19 Confidence interval			28%	25%				
20			17%	-	43%			See figure for graphical illustration
21 * femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery								
22 ** or cholesterol lowering therapy								

Fig. 2.1 Prediction rule for renal artery stenosis as implemented in an Excel spreadsheet

Note that a single treatment threshold applies only when all diagnostic work-up is completed, including all available tests for the disease. If more tests can still be done, a more complex decision analysis needs to be performed to determine the optimal choices on tests and treatments. We then have two thresholds: a low threshold between no treatment and further testing; and a higher threshold between further testing and treatment. This concept is illustrated with the diagnosis of deep venous thrombosis using ultrasound.

### \*2.3.4 Example: Probability of Deep Venous Thrombosis

A systematic review of 54 studies indicated that individual clinical features are of limited value in diagnosing deep venous thrombosis (DVT). Characteristics such as previous DVT, malignant disease, recent immobilization, and recent surgery only modestly increased the probability of DVT.<sup>144</sup> A clinical prediction rule developed by Wells et al. combines nine signs, symptoms and risk factors to categorize patients as having low, moderate or high probability of DVT.<sup>482</sup> This rule stratifies a patient's probability of DVT much better than individual findings.<sup>144</sup>

Patients who are found to be at low pretest probability ("score  $\leq 1$ ") can have DVT safely excluded (1) on the basis of a single negative ultrasound result, or (2) a negative plasma D-dimer test. Patients who are at increased pretest probability ("score  $> 1$ ") require both a negative ultrasound result, and a negative D-dimer test to exclude DVT.<sup>481</sup> A possible diagnostic algorithm is shown in Fig. 2.2.<sup>369</sup>

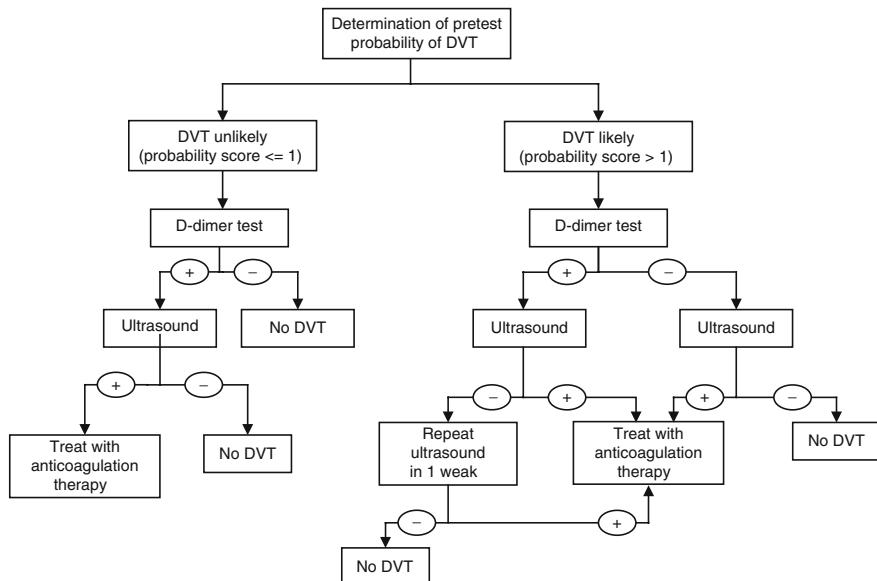
### 2.3.5 Intensity of Treatment

Prognostic estimates are also important to guide decision-making once a diagnosis is made. Decisions include, for example, more or less intensive treatment approaches. The framework for decision-making based on prognosis is very similar to that based on diagnostic probabilities as discussed before.

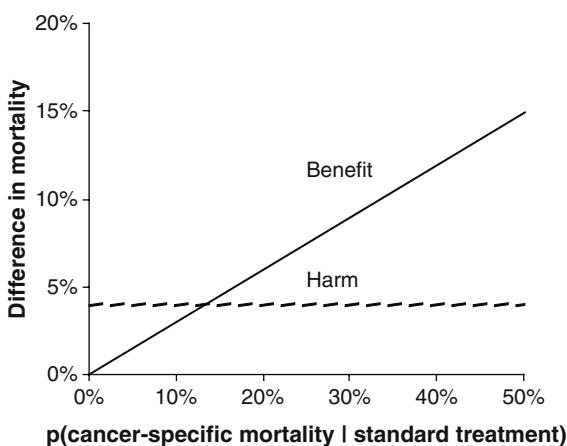
A treatment should only be given to a patient if a substantial gain is expected, which exceeds any risks and side effects (Fig. 2.3). Glasziou and Irwig illustrate this approach with a case study in anticoagulants and risk of atrial fibrillation.<sup>138</sup> Anticoagulants are very effective in reducing the risk of stroke in patients with non-rheumatic atrial fibrillation. However, using these drugs increases the risk of serious bleedings. Hence, the risk of stroke has to outweigh the bleeding risk before treatment is considered.

The specific calculation of the net benefit of a treatment requires various steps:<sup>138</sup>

- (1) Estimate benefit and harm: randomized controlled trials (RCTs) may often provide the most reliable source for relative risk estimates for both benefits and harms of treatment.



**Fig. 2.2** A possible diagnostic algorithm for patients suspected of DVT with D-dimer testing and ultrasound imaging<sup>369</sup>



**Fig. 2.3** Graphical illustration of weighing benefit and harm of treatment. Benefit of treatment (reduction in absolute risk) increases with cancer-specific mortality (relative risk set to 0.7). Harm of treatment (excess absolute risk, e.g. due to toxicity of treatment) is assumed to be constant at 4%. Net benefit occurs only when the cancer-specific mortality given standard treatment is above the threshold of 11%<sup>451</sup>

- (2) Check assumptions of relative benefit and absolute harm: subgroup effects of treatment may exist, both for benefit and harm, which invalidate the simple decision-analytic model in Fig. 2.3.
- (3) Weigh up benefit and harm: If the assumptions of relative risk reduction and constant harm are fulfilled the predicted benefit needs to be weighed up against the potential harm. This results in a graph as Fig. 2.3, with actual numbers on the Y-axis.
- (4) Predict patient's risk: To identify patients who should expect benefit to be greater than harm, we need to predict each patient's risk. Prognostic models are important for this step.

### \*2.3.6 Example: Defining a Poor Prognosis Subgroup in Cancer

As an example we consider high-dose chemotherapy (HD-CT) as first line treatment to improve survival of patients with non-seminomatous testicular cancer.<sup>451</sup> Several non-randomized trials reported a higher survival for patients treated with HD-CT as first line treatment (including etoposide, ifosfamide, cisplatin) with autologous stem cell support, compared to standard-dose (SD) chemotherapy (including bleomycin, etoposide, cisplatin). However, HD-CT is related to a higher toxicity, both during treatment (e.g. granulocytopenia, anaemia, nausea/vomiting, diarrhoea), shortly after treatment (e.g. pulmonary toxicity), and long after treatment (e.g. leukemia, cardiovascular disease). HD-CT should therefore only be given to patients with a relatively poor prognosis.

We can specify the threshold for such a poor prognosis group by weighing expected benefit against harms. Benefit of HD-CT treatment is the reduction in absolute risk of cancer mortality. Benefit increases linearly with risk of cancer mortality, if we assume that patients with the highest risk have most to gain. Harm is the increase in absolute risk of treatment mortality (e.g. related to toxicity) due to treatment. The level of harm is the same for all patients, assuming that the toxicity of treatment is independent of prognosis. Patients are candidates for more aggressive treatment when their risk of cancer mortality is above the threshold, i.e. when benefit is higher than harm (Fig. 2.3).

### 2.3.7 Cost-Effectiveness of Treatment

Cost-effectiveness of treatment also directly depends on prognosis. Treatments may not be cost-effective if the gain is small (for patients at low risk), and the costs high (e.g. for all patients the same drug costs are made). For example, statin therapy should only be given to those at increased cardiovascular risk.<sup>157</sup> And more aggressive thrombolysis should only be used in those patients with an acute myocardial infarction (AMI) who are at increased risk of 30-day mortality.<sup>63</sup> Many other examples can be

found, where the relative benefit of treatment is assumed to be constant across various risk groups, and the absolute benefit hence increases with higher risk.

Another approach is to search for differential treatment effects among subgroups of patients. The assumption of a fixed relative benefit is then relaxed. Some patients respond well to a certain treatment and others do not. Patient characteristics such as age, or the specific type of disease, may interact with treatment response. Effects of drugs are affected by the drug metabolism, which is, e.g. mediated by cytochrome P450 enzymes and drug transporters (P-glycoprotein).<sup>103</sup> Research in the field of pharmacogenomics aims to further understand the relation between an individual patient's genetic make-up (genotype) and the response to drug treatment, such that response can better be predicted.<sup>45</sup> Cost-effectiveness will vary depending on the likelihood of response to treatment.

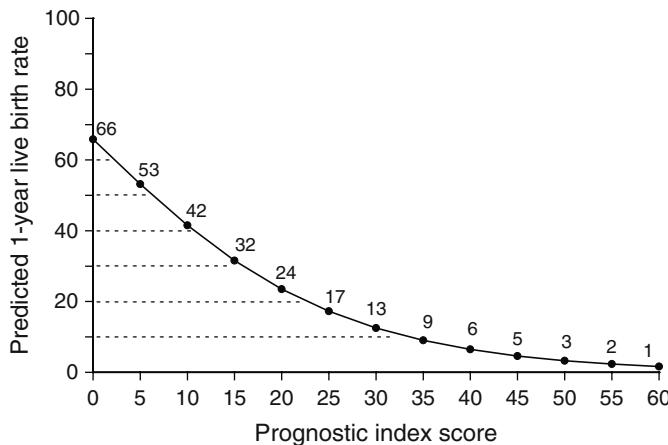
### ***2.3.8 Delaying Treatment***

In medical practice, prediction models may provide information to patients and their relatives, such that they have realistic expectations of the course of disease. A conservative approach can sometimes be taken, which means that the natural history of the disease is followed. For example, many men may opt for a watchful waiting strategy if a probably unimportant ("indolent") prostate cancer is detected.<sup>227,424</sup> Or women may be reassured on their pregnancy chances if they have relatively favourable characteristics.

### ***\*2.3.9 Example: Spontaneous Pregnancy Chances***

Several models have been published for the prediction of spontaneous pregnancy among subfertile couples.<sup>76,111,393</sup> A "synthesis model" was developed for predicting spontaneous conception leading to live birth within 1 year after start of follow-up based on data from three previous studies.<sup>205</sup> This synthesis models hence had a broader empirical basis than the original models. The predictors included readily available characteristics such as the duration of subfertility, women's age, primary or secondary infertility, percentage of motile sperm, and whether the couple was referred by a general practitioner or by a gynaecologist (referral status). The chance of spontaneous pregnancy within 1 year can easily be calculated. First a prognostic index score is calculated. The score corresponds to a probability, which can be read from a graph (Fig. 2.4).

For example, a couple with a 35-year-old woman (7 points), 2-year duration of infertility (3 points), but with one child already (secondary infertility, 0 points), normal sperm motility (0 points), and directly coming to the gynecologist (secondary care couple, 0 points), has a total score of 10 points. This corresponds to a chance of becoming pregnant of 42%.



**Fig. 2.4** Score chart to estimate the chance of spontaneous pregnancy within 1 year after intake resulting in live birth. *Upper part:* calculating the score; *lower part:* predicting 1-year pregnancy rate.<sup>205</sup> Procedure: circle the subfertility score for each of the variables, transfer to rightmost column and add to get the prognostic index score. Insert the score in the figure below to read off the chance of spontaneous pregnancy within 1 year resulting in live birth

	Subfertility Score					
Woman's age (years)	21–25	26–31	32–35	36–37	38–39	40–41
Score	<b>0</b>	<b>3</b>	<b>7</b>	<b>10</b>	<b>13</b>	<b>15</b>
Duration of subfertility (yrs)	1	2	3–4	5–6	7–8	.....
Score	<b>0</b>	<b>3</b>	<b>7</b>	<b>12</b>	<b>18</b>	.....
Type of subfertility	Secondary			Primary		
Score	<b>0</b>			<b>8</b>		
Motility (%)	≥60	40–59	20–39	0–19		
Score	<b>0</b>	<b>2</b>	<b>4</b>	<b>6</b>		.....
Referral status	Secondary care			Tertiary care		
Score	<b>0</b>			<b>4</b>		.....
				Prognostic index score (Sum)		.....

Most couples who have tried for more than 1 year to become pregnant demand immediate treatment.<sup>205</sup> In their judgment, further waiting is senseless because they consider themselves as infertile. Moreover, the psychological pressure caused by feelings of uncertainty and frustration may increase a desire for immediate action. In addition, most couples overestimate the success of assisted reproduction, such as

in vitro fertilization, and underestimate the related risks. The estimations of spontaneous pregnancy leading to live birth can be a tool in advising these couples in the following manner. If the chances are low, e.g. below 20%, there is no point in further waiting, and advising the couple to quickly undergo treatment is realistic. In contrast, if the chances are favourable, e.g. above 40%, the couple should be strongly encouraged to wait for another year, because there is a substantial chance of success.

### ***2.3.10 Surgical Decision-Making***

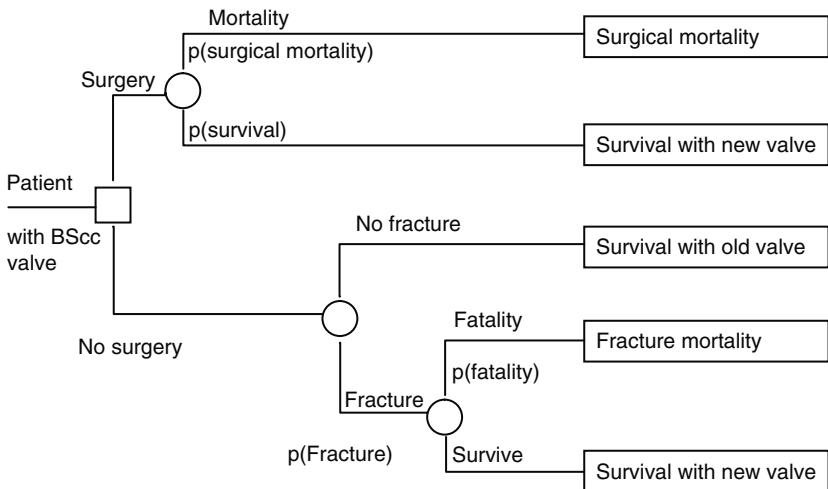
In surgery, it is typical that short-term risks are taken to reduce long-term risks. Short-term risks include both morbidity and mortality. The surgery aims to reduced long-term risks that would occur in the natural history. Acute situations include surgery for trauma, and for acute conditions such as a ruptured aneurysm (a widened artery). Elective surgery is done for many conditions, and even for such planned and well-prepared surgery, the short-term risk and burden are never zero. In oncology, increased surgical risks typically lead to the choice for less risky treatments, e.g. chemotherapy or radiation, or palliative treatments. For example, in many cancers, older patients and those with comorbidity do less often undergo surgery.<sup>6,169,207</sup>

Many prognostic models have been developed to estimate short-term risks of surgery, e.g. 30-day mortality. These models vary in complexity and accuracy. Also, long-term risks have been modeled explicitly for various diseases, although it is often hard to find a suitable group of patients for the natural course of a disease without surgical intervention. As an example, we consider a surgical decision problem on replacement of risky heart valves (Fig. 2.5). Prognostic models were used to estimate surgical mortality, individualized risk of the specific valve, and individual survival.<sup>37,415,449</sup>

### ***\*2.3.11 Example: Replacement of Risky Heart Valves***

Björk–Shiley convexo–concave (BScc) mechanical heart valves were withdrawn from the market in 1986 after reports of mechanical failure (outlet strut fracture). Worldwide, approximately 86,000 BScc valves had been implanted by then. Fracture of the outlet strut occurs suddenly and is often lethal.<sup>448</sup> Therefore, prophylactic replacement by another, safer valve, may be considered to avert the risk of fracture. Decision analysis is a useful technique to weigh the long-term loss of life expectancy due to fracture against the short-term surgical mortality risk (Fig. 2.5). The long-term loss of life expectancy due to fracture depends on three aspects:

1. The annual risk of fracture, given that a patient is alive
2. The fatality of a fracture
3. The annual risk of death (survival).



**Fig. 2.5** Schematic representation of surgical decision-making on short-term vs. long-term risk in replacement of a risky BScc heart valve. *Square* indicates a decision, *circle* a chance node. Predictions ('p') are needed for four probabilities: surgical mortality, long-term survival, fracture, and fatality of fracture

This long-term loss of life expectancy has to be weighed against the risk of surgical mortality. If the patient survives surgery, the fracture risk is assumed to be reduced to zero. Predictive regression models were developed for each aspect, based on the follow-up experience from 2,263 patients with BScc valves implanted between 1979 and 1985 in The Netherlands.<sup>223,415</sup> We considered 50 fractures that had occurred during follow-up and 883 patients who died (excluding fractures).

The risk of fracture is the key consideration in this decision problem. But the low number of fractures makes predictive modelling challenging, and various variants of models have been proposed. A relatively detailed model included four traditional predictors (age, position (aortic/mitral), type ( $70^\circ$  opening angle valves had higher risks than  $60^\circ$  valves), size (larger valves had higher risks)), and two production characteristics.<sup>415</sup> The fatality of a fracture depended on the age of the patient, and the position (higher fatality in aortic position). Survival was related to age, gender, position of the valve, and also to time since implantation. This meant that patients of a given age (e.g. 50 years), had higher risks when the implantation of the valve was longer ago (e.g. implantation at age 35 vs 45 years). Finally, surgical risk was modelled in relation to age and position of the valve. This was a relatively rough approach, since many more predictors are relevant, and a later prediction model was much more detailed.<sup>454</sup>

The results of this decision analysis depended strongly on age: replacement was only indicated for younger patients, who have lower surgical risks, and a higher long-term impact of fracture because of longer survival (Table 2.3). Also, the posi-

**Table 2.3** Patient characteristics used in the decision analysis of replacement of risky heart valves<sup>415</sup>

Characteristic	Surgical risk	Survival	Fracture	Fatality fracture
Patient related				
Age (years)	+	+	+	+
Sex (male/female)		+		
Time since implantation (years)		+		
Valve related				
Position (aortic/mitral)	+	+	+	+
Opening angle (60°/70°),			+	
Size (<29 mm or >=29 mm)			+	
Production characteristics			+	
Type of prediction model	Logistic regression	Poisson regression	Poisson regression	Logistic regression

tion of the valve affects all four aspects (surgical risk, survival, fracture, fatality). Before, results were presented as age-thresholds for eight subgroups of valves: by position (aortic/mitral), by type (70°/60°), and by size (large/small).<sup>449</sup> The more recent analysis was so detailed that individualized calculations were necessary, which were performed for all patients who were alive in The Netherlands in 1998. The recommendations from this decision analysis were rather well followed in clinical practice.<sup>455</sup>

## 2.4 Prediction Models for Medical Research

In medical research, prediction models may serve several purposes. In experimental studies, such as a randomized controlled trial (RCT), predictive baseline characteristics may assist in the inclusion and stratification of patients, and improve the statistical analysis. In observational studies, adequate controlling for confounding factors is essential.

### 2.4.1 *Inclusion and Stratification in an RCT*

In a RCT, prognostic estimates may be used for selection of subjects for the study. Traditionally, a set of inclusion and exclusion criteria is applied to define the subjects for the RCT. Some criteria aim to create a more homogeneous group according to expected outcome. Traditionally, all inclusion criteria have to be fulfilled, and none of the exclusion criteria. Alternatively, some prognostic criteria can be combined in a prediction model, with selection based on individualized predictions. This leads to a more refined selection.

Stratification is often advised in RCTs for the main prognostic factors.<sup>18,338,496</sup> In this way, balance is obtained between arms of a trial with respect to baseline prognosis. This may facilitate simple, direct comparisons of treatment results, especially for smaller RCTs, where some imbalance may readily occur. Prediction models may refine stratification of patients, especially when many prognostic factors are known.

### \*2.4.2 Example: Selection for TBI Trials

As an example, we consider the selection of patients for RCTs in traumatic brain injury (TBI). Patients above 65 years of age and those with non-reacting pupils are often excluded because of a high likelihood of a poor outcome. Indeed we find a higher than 50% mortality at 6-month follow-up in patients fulfilling either criterion (Table 2.4). Hence, we can simply select only those less than 65 years with at least one reacting pupil (Table 2.5, part A). However, we can use a prognostic model for more efficient selection that inclusion based on separate criteria. A simple logistic regression model with “age” and “pupils” can be used to calculate the probability of mortality in a more detailed way. If we aim to exclude those with a predicted risk over 50%, this leads to an age limit of 30 years for those without any pupil reaction, and an age limit of 76 years for those with any pupil reaction (Table 2.5, part B). So, patients under 30 years of age can always be included, and patients between 65 and 75 years can be included if they have at least one reacting pupil (Table 2.5).

**Table 2.4** Analysis of outcome in 7,143 patients with severe moderate traumatic brain injury according to reactive pupils and age dichotomized at age 65 years<sup>276</sup>

	>= 1 Reactive pupil		Non-reactive pupils	
	<65	>=65 years	<65	>=65 years
6-month mortality	926/5101 (18%)	159/284 (56%)	849/1644 (52%)	97/114 (85%)

**Table 2.5** Selection of patients with two criteria (age and reactive pupils) in a traditional way (A) and according to a prognostic model (probability of 6-month mortality < 50%, B)

Pupillary reactivity	No reactivity >=1 pupil	A: Traditional selection		B: Prognostic selection			
		<65	>=65 years	<30	30–75	>=76 years	
		Exclude	Exclude	Include	Exclude	Exclude	Include

### 2.4.3 Covariate Adjustment in an RCT

Even more important is the role of prognostic baseline characteristics in the analysis of an RCT. The strength of randomization is that comparability is created between treated groups both with respect to observed *and unobserved* baseline characteristics (Fig. 2.6). No systematic confounding can hence occur in RCTs. But some observed baseline characteristics may be strongly predictive of outcome. Adjustment for such covariates has several advantages:<sup>133,182,188,190,339,348</sup>

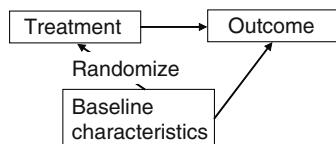
1. To reduce any distortion in the estimate of treatment effect that occurred by random imbalance between groups
2. To improve the precision of the estimated treatment effect
3. To increase the statistical power for detection of a treatment effect

Remarkably, covariate adjustment works differently for linear regression models and generalized linear models (e.g. logistic, Cox regression, Table 2.6).

1. For randomized clinical trials the randomization guarantees that the bias is zero a priori, both for observed and unobserved baseline characteristics. However, random imbalances may occur, generating questions such as: What would have been the treatment effect had the two groups been perfectly balanced? We may think of this distortion as a bias a posteriori, since it affects interpretation similarly as in observational epidemiological studies.

Regression analysis is an obvious technique to correct for such random imbalances. When no imbalances have occurred for predictors considered in a regression model, the adjusted and unadjusted estimates of the treatment effect would be expected to be the same. This is indeed the case in linear regression analysis. Remarkably, in generalized linear models such as logistic regression, the adjusted and unadjusted estimates of a treatment effect are not the same, even when predictors are

**Fig. 2.6** Schematic representation of the role of baseline characteristics in an RCT. By randomization, there is no systematic link between baseline characteristics and treatment. Baseline characteristics are still important, since they are prognostic for the outcome



**Table 2.6** Comparison of adjustment for predictors in linear and generalized linear models (e.g. logistic regression) in estimation and testing of treatment effects, when predictors are completely balanced

Method	Effect estimate	Standard error	Power
Linear model	Identical	Decreases	Increases
Generalized linear model	Further from zero	Increases	Increases

completely balanced<sup>133</sup> (see Questions 2.3 and 22.2). Adjusted effects are expected to be further from zero (neutral value, OR further from 1). This phenomenon is referred to as a “stratification effect”, and does not occur with linear regression.<sup>403</sup>

2. With linear regression, adjustment for important predictors leads to an improvement in precision of the estimated treatment effect, since part of the variance in the outcome is explained by the predictors. Contrary, in generalized linear models such as logistic regression, the standard error of the treatment effect always increases with adjustment.<sup>348</sup>
3. In linear regression, adjusted analyses provide more power to the analysis of treatment effect, since the standard error of the treatment effect is smaller. For a generalized linear model such as logistic regression, the effect of adjustment on power is not so straightforward. It has however been proven that the expected value of the treatment effect estimate increases more than the standard error. Hence, the power for detection of a treatment effect is larger in an adjusted logistic regression analysis compared to an unadjusted analysis.<sup>348</sup>

#### **2.4.4 Gain in Power by Covariate Adjustment**

The gain in power by covariate adjustment depends on the correlation between the baseline covariates (predictors) and the outcome. For continuous outcomes, this correlation can be indicated by Pearson’s correlation coefficient ( $r$ ). Pocock et al. showed that in the continuous outcome situation, the sample size can be reduced with  $1 - r^2$ , to achieve the same statistical power with a covariate adjusted analysis as an unadjusted analysis.<sup>339</sup> A very strong predictor may have  $r=0.7$  ( $r^2 50\%$ ), e.g. a baseline covariate of a repeated measure such as blood pressure, or a questionnaire score. The required number of patients is then roughly halved. The saving is less than 10% for  $r=0.3$  ( $r^2 9\%$ ).<sup>339</sup>

Similar results have been obtained in empirical evaluations with dichotomous outcomes, where Nagelkerke’s  $R^2$ <sup>309</sup> was used to express the correlation between predictor and outcome.<sup>188,190,403</sup> The reduction in sample size was slightly less than

**Table 2.7** Illustration of reduction in sample size with adjustment for baseline covariates with dichotomous outcomes

Application area	Correlation baseline–outcome	Reduction in sample size
Acute MI: 30-day mortality <sup>403</sup>		
Age adjustment	$R^2 13\%$	12%
17 predictor adjustment	$R^2 25\%$	19%
Traumatic brain injury: 6-month mortality <sup>189</sup>		
3 predictor model	$R^2 30\%$	25%
7 predictor model	$R^2 40\%$	30%

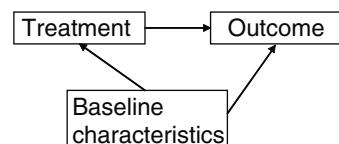
$1 - R^2$  in simulations for mortality among acute MI patients<sup>403</sup> and among TBI patients<sup>189</sup> (Table 2.7).

### \*2.4.5 Example: Analysis of the GUSTO-III Trial

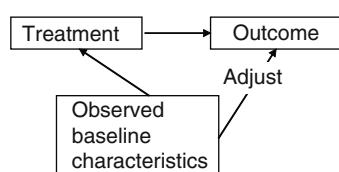
The GUSTO-III trial considered patients with an acute myocardial infarction.<sup>4</sup> The outcome was 30-day mortality. The protocol pre-specified a prognostic model for the primary analysis of the treatment effect. This model combined age, systolic blood pressure, Killip class, heart rate, infarct location, and age-by-Killip-class interaction. These predictors were previously found to comprise 90% of the predictive information of a more complex model for 30-day mortality in the GUSTO-I trial.<sup>255</sup> A review of RCTs published in the major medical journals after the year 2000 shows that covariate adjustment is used in approximately 50% of the cases.<sup>339</sup>

### 2.4.6 Prediction Models and Observational Studies

Confounding is the major concern in epidemiological analyses of observational studies. When treatments are compared, groups are often quite different because of a lack of randomization. Subjects with specific characteristics are more likely to have received a certain treatment than other subjects (“indication bias”, Fig. 2.7). If these characteristics also affect the outcome, a direct comparison of treatments is biased, and may merely reflect the lack of initial comparability (“confounding”). Instead of treatment, many other factors can be investigated for their causal effects. Often, randomization is not possible, and observational studies are the only possible design. Dealing with confounding is an essential step in such analyses.



**Fig. 2.7** Schematic representation of confounding in an observational study. Baseline characteristics act as confounders since they are related to the treatment and to the outcome



**Fig. 2.8** Schematic representation of adjustment for baseline characteristics in an observational study. By adjustment, we aim to correct for the systematic link between observed baseline characteristics and outcome, hence answering the question what the treatment effect would be if observed baseline characteristics were similar between treatment groups

Regression analysis is a commonly used method to control for imbalances between treatment groups, e.g. with logistic or Cox regression.<sup>235</sup> Many baseline characteristics can be simultaneously adjusted for (Fig. 2.8). Similarly, regression analysis can be used to control for confounders in aetiological research.

### 2.4.7 Propensity Scores

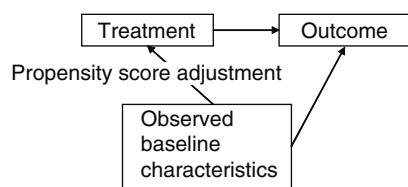
A problem arises when the outcome is relatively rare. Constructing a regression model with many predictors is then problematic. This may lead to biased and inefficient estimates of the difference between groups in the adjusted analysis.<sup>66</sup> An alternative in the setting of rare outcomes is to use a propensity score.<sup>55</sup> The propensity score defines the probability that a subject receives a particular treatment (“Tx”) given a set of confounders:  $p(\text{Tx} | \text{confounders})$ . For calculation of the propensity score, the confounders are usually used in a logistic regression model to predict the treatment, without including the outcome.<sup>60,359</sup> The propensity score is subsequently used in a second stage as a summary confounder (Fig. 2.9). Approaches in this second stage are matching on propensity score, stratification of propensity score (usually by quantile), and inclusion of the propensity score with treatment in a regression model for the outcome.<sup>89</sup>

Empirical comparisons provided no indication of superiority of propensity score methods over conventional regression analysis for confounder adjustment.<sup>381,429</sup> Simulation studies however suggest a benefit of propensity scores in the situation of few outcomes relatively to the number of confounding variables.<sup>66</sup>

### \*2.4.8 Example: Statin Treatment Effects

Seeger et al. investigated the effect of statins on the occurrence of acute myocardial infarction (AMI).<sup>378</sup> They studied members of a Community Health Plan with a recorded  $\text{LDL} > 130 \text{ mg dl}^{-1}$  at any time between 1994 and 1998. Members who initiated therapy with a statin were matched using propensity scores to members who did not initiate statin therapy. The propensity score predicted the probability of sta-

**Fig. 2.9** Schematic representation of propensity score adjustment for baseline characteristics in an observational study. The propensity score estimates the probability of receiving treatment. By subsequent adjustment for the propensity score, we mimic an RCT, since we removed the systematic link between baseline characteristics and treatment. We can however only include observed baseline characteristics, and have no control over unobserved characteristics



**Table 2.8** The effect of statins on the occurrence of acute myocardial infarction<sup>378</sup>

	Confounders	N with AMI	HR [95% CI]
Unadjusted	–	325 vs. 124	2.1 [1.5–3.0]
Propensity score adjusted	52 main effects, 6 quadratic terms	77 vs. 114	0.69 [0.52–0.93]

tin initiation. Scores were estimated using a logistic regression model that included 52 variables and 6 quadratic terms (Table 2.8). Statin initiators were matched to a noninitiator within a 0.01 caliper of propensity. Initiators for whom no suitable noninitiator could be found were excluded, leaving 2,901 matched initiators out of 4,144 initiators (70%). The 4,144 statin initiators had a higher prevalence of established coronary heart disease risk factors than did unmatched noninitiators. The follow-up of these unmatched cohorts identified 325 AMIs in the statin initiator group and 124 in the noninitiator group (hazard ratio 2.1, 95% confidence interval 1.5–3.0). The propensity score-matched cohorts ( $2 \times n=2,901$ ) were very similar with respect to 51 of the 52 baseline characteristics. There were 77 cases of AMI in statin initiators compared with 114 in matched non-initiators (hazard ratio 0.69, 95% confidence interval 0.52–0.93). The authors hence conclude that statin use in the members of this Community Health Plan was beneficial on the occurrence of AMI, but warn that predictors that are not part of the model may remain unbalanced between propensity score matched cohorts, leading to residual confounding.

#### 2.4.9 Provider Profiling

Another area of application of prediction models is in the comparison of outcomes from different hospitals (or other providers of care, “provider profiling”).<sup>47</sup> The quality of health care providers is being compared by their outcomes, which are considered as performance indicators. Simple comparisons between providers may obviously be biased by differences in case-mix; for example, academic centers may see more severe patients, which accounts for poorer outcome on average. Prediction models are useful for case-mix adjustment in such comparisons.

#### \*2.4.10 Example: Ranking Cardiac Outcome

New York State was among the first to publicly release rankings of outcome of coronary artery bypass surgery by surgeon and hospital. Such cardiac surgery report cards have been criticized because of their methodology.<sup>136</sup> Adequate risk adjustment is nowadays better possible with sophisticated prediction models. An example is a model published by Krumholz et al., who present a prediction model for 30-day mortality rates among patients with AMI.<sup>245</sup> The model used information from

administrative claims and aimed to support profiling of hospital performance. They analyzed 140,120 cases discharged from 4,664 hospitals in 1998. They compared the model from claims data with a model using medical record data and found high agreement. They also found adequate stability over time (data from years 1995 to 2001). The final model included 27 variables and had an area under the receiver operating characteristic curve of 0.71. The authors conclude that this administrative claims-based model is as adequate for profiling hospitals as a medical record model. Chapter 21 provides a more in-depth discussion of this research area.

## 2.5 Concluding Remarks

We have discussed several areas of potential application of prediction models, including public health (targeting of preventive interventions), clinical practice (diagnostic work-up, therapeutic decision making), and research (design and analysis of RCTs, confounder adjustment in observational studies). More types of application can probably be thought of. Obtaining predictions from a model has to be separated from obtaining insights in the disease mechanisms and patho-physiological processes. Such insights are related to the estimated effects of predictors in a model. Often, prediction models serve the latter purpose too, but the primary aim considered in this book is outcome prediction.

## Questions

### 2.1 Examples of applications of prediction models

- (a) What is a recent application of a prediction model that you encountered? Search PubMed [<http://www.ncbi.nlm.nih.gov/sites/entrez>] if nothing comes to mind.
- (b) How could you use a prediction model in your own research or in your clinical practice?

### 2.2 Cost-effectiveness

How could prediction models contribute to targeting of treatment and to increasing cost-effectiveness of medical care?

### 2.3 Covariate adjustment in an RCT

What are the purposes of covariate adjustment in an RCT? Explain and distinguish between logistic and linear regression.

### 2.4 Propensity score

- (a) What is the definition of a propensity score?
- (b) Explain the difference between adjustment for confounders through regression analysis and through a propensity score.
- (c) When is propensity score specifically appropriate? See papers by Braiman and by Cepeda.<sup>55,66</sup>

# Chapter 3

## Study Design for Prediction Models

**Background** In this chapter, we consider several issues in the design of studies for prediction research. These include the selection of subjects or patients for a cohort study, strengths and limitations of case series from a single center, from registries, or prospective trials. We further discuss issues in choosing predictors and outcome variables for prediction models. An important question is often how large a study needs to be for sufficient statistical power. Power considerations are given for studying effects of specific predictors, and for developing a prediction model that can provide reliable predictions. We use several case studies for illustration.

### 3.1 Study Design

Prognostic studies are inherently longitudinal in nature. They are most often performed in cohorts of patients, who are followed over time for an outcome to occur. The cohort is defined by the presence of one or more particular characteristics, e.g. having a certain disease, living in a certain place, having a certain age, or simply being born alive. For example, we may follow a cohort of patients with an acute myocardial infarction for long-term mortality according to ECG characteristics.<sup>335</sup>

Diagnostic studies are most often designed as a cross-sectional study, where predictive patient characteristics are related to an underlying diagnosis. The study group is defined by the presence of a particular symptom or sign that makes the subject suspected of having a particular (target) disease. Typically, the subjects undergo the index test and subsequently a reference test to establish the “true” presence or absence of the target disease, over a short time span. For example, we may aim to diagnose those with an acute myocardial infarction among patients presenting at an emergency department.<sup>142</sup>

### 3.2 Cohort Studies for Prognosis

Several types of cohort studies can be used for prognostic modelling. The most common type may be a single-center retrospective cohort study (Table 3.1). In this case, patients are identified from hospital records between certain dates, for example,

**Table 3.1** Study designs for prognostic studies

Design	Characteristics	Strengths	Limitations
Retrospective	Often single-centre studies	Simple, low costs Well-defined selection of patients Prospective recording of predictors Prospective assessment of outcome according to protocol	Selection of patients Definitions and completeness of predictors Outcome assessment not by protocol Poor generalizability because of stringent in- and exclusion criteria
Prospective	Often multicentre RCT	Simple, low costs Prospective recording of predictors Prospective assessment of outcome according to protocol	Outcome assessment not by protocol
Registry	Complete coverage of an area/participants covered by insurance	Simple, low costs Prospective recording of predictors	Selection of controls critical Definitions and completeness of predictors Outcome assessment not by protocol
Case-control	Efficient when outcome relatively rare	Simple, low costs	

those diagnosed between January 1, 1997, and December 31, 2003. These patients were followed over time for the outcome, but the investigator looks back in time (hence we may use the label “retrospective study”<sup>463</sup>).

### ***3.2.1 Retrospective Designs***

Strengths of a retrospective study design include its simplicity and feasibility. It is a design with relatively low costs, since patient records can often easily be searched, especially with modern hospital information systems or electronic patient records. A limitation is the correct identification of patients, which has to be done in retrospect. If some information is missing, or was incorrectly recorded, this may lead to a selection bias. Similarly, the recording of predictors has to have been reliable to be useful for prediction modelling. Finally, the outcome has to be reliable. This may be relatively straightforward for outcomes such as survival, where some deaths will be known from hospital records. But additional confirmation of vital status may often be required from nationwide statistical bureaus for a complete assessment of survival status. Other outcomes, e.g. related to functional status, may not be available at the time points that we wish to analyse. Finally, single centre studies may be limited by their sample size, which is a key problem in prediction research. Multicentre, collaborative studies can address this sample size issue. Moreover, the representativeness of the prediction model will then be better.

### ***\*3.2.2 Example: Predicting Early Mortality in Oesophageal Cancer***

As an example, we consider outcome prediction in oesophageal cancer. A retrospective chart review was performed of 120 patients treated in a single institution between January 1, 1997, and December 31, 2003.<sup>252</sup> The patients had palliative treatment, which means therapy that relieves symptoms, but does not alter the course of the disease. A stent was placed in the oesophagus because of malignancy-related dysphagia (difficulty in swallowing). The authors studied 30-day mortality, which occurred in an unspecified number of patients (probably around 10%,  $n=12$ ).<sup>252</sup>

### ***3.2.3 Prospective Designs***

In a prospective study, we can better check specific inclusion and exclusion criteria. The investigator is said to age with the study population (hence the label “prospective study”). We can use clear and consistent definitions of predictors, and assess

patient outcomes at pre-defined time points. Prospective cohort studies are therefore preferable to analyses in retrospective series.

Prospective cohort studies are sometimes solely set up for prediction modeling, but a more common design is that prediction research is done in data from randomized clinical trials (RCTs), or from prospective before–after trials. The strengths are in the well-defined selection of patients, the prospective recording of predictors, usually with quality checks, and the prospective assessment of outcome. Sample size is usually reasonably large. A limitation of data from (randomized) trials may be in the selection of patients. Often stringent inclusion and exclusion criteria are used, which may limit the generalizability of a model developed on such data. On the other hand, RCTs are often performed in multiple centres, sometimes from multiple countries or continents. Benefits of the multi-centre design include that consensus has to be reached on definition issues for predictors and outcome, and that generalizability of findings will be increased. This is in contrast to single centre studies, which only reflect predictive relationships from one specific setting.

A topic of debate is whether we should only use patients from an RCT who are randomized to a conventional treatment or placebo (the “control group”). If we combine randomized groups we assume that no specific subgroup effects are relevant for the prognostic model. This may generally be reasonable. Moreover, the prognostic effect of a treatment is usually small compared to prognostic effects of other predictors.

### \*3.2.4 Example: Predicting Long-Term Mortality in Oesophageal Cancer

In another study of outcome in oesophageal cancer, data from an RCT (“SIREC”,  $n=209^{197}$ ) were combined with other prospectively collected data ( $n=396$ ).<sup>414</sup> Long-term mortality was studied after palliative treatment with a stent or radiation (“brachytherapy”).

### 3.2.5 Registry Data

Prognostic studies are often performed with registry data, for example cancer registries, or insurance databases. Data collection is prospective, but not primarily for prediction research. The level of detail may be a limitation for prognostic analyses. For example, the well-known US-based cancer registry (Surveillance, Epidemiology and End Results, SEER) contains information on cancer incidence, mortality, patient demographics, and tumour stage. It has been linked to the Medicare data base for information on comorbidity<sup>233</sup> and treatment (surgery,<sup>80</sup>

chemotherapy,<sup>478</sup> radiotherapy<sup>471</sup>). Socio-economic status (SES) is usually based on median income as available at an aggregated level.<sup>24</sup> SEER-Medicare does not contain detailed information on performance status, which is an important factor for medical decision-making and for survival of cancer patients. Also, staging may have some measurement bias.<sup>118</sup>

Another problem may occur when reimbursement depends on the severity that is scored for a patient. This may pose an upward bias on the recording of comorbidities in claims databases for example.

The outcomes for prognostic analyses usually suffer from the same limitations as retrospective studies, since usually no pre-defined assessments are made. Outcomes are therefore often limited to survival, although other adverse events can sometimes also be derived.<sup>105,394</sup> Strengths of prognostic studies with registry data include large sample sizes, and representativeness of patients (especially with population-based cancer registries). Such large databases may especially be useful for studying predictive relationships of a limited number of predictors with survival.

### **\*3.2.6 Example: Surgical Mortality in Oesophageal Cancer**

The SEER-Medicare database was used to analyze 30-day mortality in 1,327 patients undergoing surgery for oesophageal cancer between 1991 and 1996. Predictive patient characteristics included age, comorbidity (cardiac, pulmonary, renal, hepatic, and diabetes), preoperative therapy, and a relatively low hospital volume, which were combined in a simple prognostic score. Validation was done in another registry, and in a hospital series.<sup>423</sup>

### **3.2.7 Nested Case–Control Studies**

A prospectively designed, nested case–control study is sometimes an efficient option for prediction research. A case–control design is especially attractive when the outcome is relatively rare, such as incident breast cancer.<sup>131</sup> For example, if 30-day mortality is 1%, it is efficient to determine detailed predictors in all patients who died, but for example 4% of the controls (1:4 case–control ratio). A random sample of controls is used as comparison for the cases. If the outcome is well defined, such as survival, selection bias cannot be a problem. Assessment of details of predictors is in retrospect, which is a limitation. If a prediction model is developed, the average outcome incidence has to be adjusted for final calculation of probabilities, while the regression coefficients can be based on the case–control study.<sup>131</sup>

### \*3.2.8 Example: Perioperative Mortality in Major Vascular Surgery

An interesting example is the analysis of perioperative mortality in patients undergoing major vascular surgery.<sup>340</sup> Predictors were determined in retrospect from a detailed chart review in all cases (patients who died), and in selected controls (patients who did survive surgery). Controls had surgery just before and just after the case. Hence a 1:2 ratio was achieved for cases against controls.

## 3.3 Studies for Diagnosis

### 3.3.1 Cross-Sectional Study Design and Multivariable Modelling

Ideally, a diagnostic study considers a well-defined cohort of patients suspected of a certain diagnosis, e.g. an acute myocardial infarction.<sup>238</sup> Such a diagnostic study then resembles a prognostic cohort study. The cohort is here defined by the suspicion of having (rather than actually having) a disease. The outcome is the underlying diagnosis. The study may therefore be labelled cross-sectional, since the predictor–outcome relationships are studied at a single point in time. Several characteristics may be predictive of the underlying diagnosis. For a model, we should start with considering simple characteristics such as demographics, and symptoms and signs obtained from patient history. Next, we may consider simple diagnostic tests, and finally invasive and/or costly tests.<sup>295</sup> The diagnosis (presence or absence of the target disease) should be established by a reference test or standard. This test used to be called “gold” standard, but no method is 24 carat gold. The result of the reference test is preferably interpreted without knowledge of the predictor and diagnostic test values. Such blinding prevents information bias (or incorporation, or “diagnostic review” bias).<sup>296</sup>

A common problem in diagnostic evaluations is the incomplete registration of all predictive characteristics. Not all patients may have undergone the entire diagnostic work-up, especially if they are considered as at low risk of the target disease. Similarly, outcome assessment may be incomplete, if a test is used as a gold standard which is selectively performed.<sup>343</sup> These problems are especially prominent in diagnostic analyses on data from routine practice.<sup>313</sup> Prospective studies are hence preferable, since these may use a pre-specified protocol for systematic diagnostic work-up and reference standard testing.

### \*3.3.2 Example: Diagnosing Renal Artery Stenosis

A cardiology database was retrospectively reviewed for patients who underwent coincident screening abdominal aorta angiography to detect occult renal artery stenosis. In a development set, stenosis was observed in 128 of 635 patients. This 20%

prevalence may be an overestimate if patients underwent angiography because of suspicion of stenosis.<sup>347</sup>

### **3.3.3 Case–Control Studies**

Diagnostic studies sometimes select patients on the presence or absence of the target disease as determined by the reference test. Hence patients without a reference standard are not selected. In fact, a case–control study is performed, where cases are those with the target disease, and controls those without. This design has a number of limitations, especially related to the representativeness of the selected patients for all patients who are suspected of the diagnosis of interest. Selection bias is the most important limitation. Indeed, empirical evidence is now available on the bias that arises in diagnostic studies, especially by including non-consecutive patients in a case–control design, non-representative patients (severe cases compared to healthy controls), and when data are collected retrospectively.<sup>259,361</sup>

### **\*3.3.4 Example: Diagnosing Acute Appendicitis**

C-reactive protein (CRP) has been used for the diagnosis of acute appendicitis. Surgery and pathology results constituted the reference test for patients with a high CRP. Patients with a low CRP were not operated on and clinical follow-up determined whether they were classified as having acute appendicitis. As low-grade infections with low CRPs can resolve spontaneously, this verification strategy fails to identify all false-negative test results. In this way, the diagnostic performance of CRP will be overestimated.<sup>259</sup>

## **3.4 Predictors and Outcome**

### **3.4.1 Strength of Predictors**

For a well-performing prediction model, strong predictors have to be present. Strength is a function of the association of the predictor with the outcome, and the distribution of the predictor. For example, a dichotomous predictor with an odds ratio of 2.0 is more relevant for a prediction model than a dichotomous predictor with an odds ratio of 2.5, when the first predictor is distributed in a 50:50 ratio (50% prevalence of the predictor), and the second 1:99 (1% prevalence of the predictor). Similarly, continuous predictors have to cover a wide range to make them relevant for prediction.

When some characteristics are considered as key predictors, these have to be registered carefully, with clear definitions and preferably no missing values. This is

usually best possible in a prospective study, with a protocol and pre-specified data collection forms.

### ***3.4.2 Categories of Predictors***

Several categories of predictors have been suggested for prediction models.<sup>174</sup> These include

- Demographics (e.g. age, sex, race, socio-economic status)
- Type and severity of disease (e.g. principal diagnosis, presenting characteristics)
- History characteristics (e.g. previous disease episodes, risk factors)
- Comorbidity (concomitant diseases)
- Physical functional status (e.g. Karnofsky score, WHO performance score)
- Subjective health status and quality of life (psychological, cognitive, psychosocial functioning)

The relevance of these categories will depend on the specifics of the application. Publications tend to group predictors under general headings, see for example, the predictors in the GUSTO-I model (Chap. 22).<sup>255</sup> Of note, definitions of predictors may vary from study to study.<sup>492</sup> Socioeconomic status (SES) can be defined in many ways, considering a patient's working status, income, and/or education. Also, SES indicators are sometimes not determined at the individual level, but for example at census tract level ("ecological SES", e.g. in analyses of SEER-Medicare data<sup>24,404</sup>). Race/ethnicity can be defined in various ways, and sometimes be self-reported rather than determined by certain pre-defined rules. Comorbidity definitions and scoring systems are still under development.<sup>91,126,201</sup> Variation in definitions is a serious threat to the generalizability of prediction models.<sup>16</sup>

Another differentiation is to separate the patient's condition from his/her constitution. Condition may be reflected in type and severity of disease, history characteristics, comorbidity, physical and subjective health status. Constitution may especially be related to demographics such as age and gender. For example, the same type of trauma (reflected in patient condition) affects patients of different ages differently (constitution).

In the future, genetic characteristics will be used more widely in a prediction context. Inborn variants of the human genome, such as polymorphisms and mutations, may be considered as indicators of the patient's constitution. Other genetic characteristics, for example the genomic profile of a malignant tumour, may better be thought of as indicators of subtypes of tumours, reflecting condition.

### ***3.4.3 Costs of Predictors***

Predictors may require different costs, in monetary terms, but also in burden for a patient. In a prediction context, it is evident that information that is easy to obtain

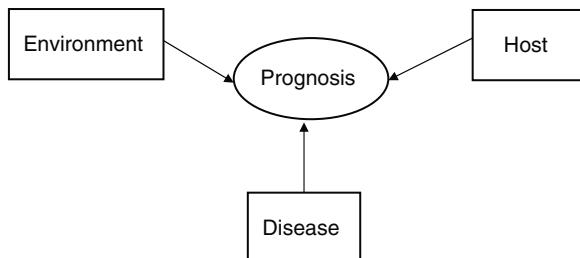
should be considered before information that is more difficult to obtain. Hence, we should first consider characteristics such as demographics and patient history, followed by simple diagnostic tests, and finally invasive and/or costly tests. Expensive genetic tests should hence be considered for their incremental value over classical predictors rather than alone.<sup>225</sup> Such an incremental evaluation is well possible with predictive regression models, where a model is first considered without the test, and subsequently a model with the test added.<sup>399</sup>

### ***3.4.4 Determinants of Prognosis***

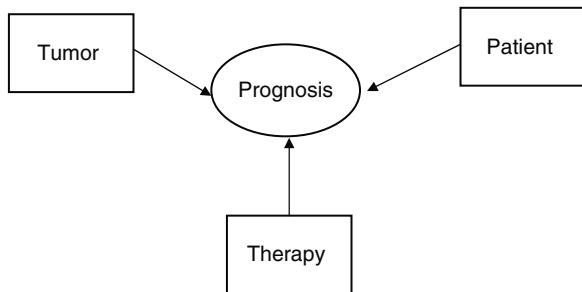
Prognosis can also be viewed in a triangle of interacting causes (Fig 3.1). Predictors may be separated as related to environment (e.g. socio-economic conditions, health care access and quality, climate), the host (e.g. demographic, behavioral, psychosocial, premorbid biologic factors), and disease (e.g. imaging, pathophysiologic, genomic, proteomic, metabolomic factors).<sup>184</sup>

### ***3.4.5 Prognosis in Oncology***

For prognosis in oncology, it has been proposed to separate factors related to the patient, the tumour and to treatment (Fig.3.2).<sup>186</sup> Examples of patients characteristics include demographics (age, sex, race/ethnicity, SES), comorbidity, functional status. Tumour characteristics include the extent of disease (e.g. reflected in TNM stage), pathology, and sometimes values of tumour markers in the blood. Treatment may commonly include (combinations of) surgery, chemotherapy, and radiotherapy.



**Fig. 3.1** Prognosis may be thought of as determined by predictors related to environment, host and disease<sup>184</sup>



**Fig. 3.2** Prognosis of a patient with cancer may be thought of as determined by predictors related to the tumour, the patient, and therapy<sup>186</sup>

## 3.5 Reliability of Predictors

### 3.5.1 *Observer Variability*

We generally prefer predictors that are well defined and reliably measured by any observer. In practice, observer variability is a problem for many measurements.<sup>185,246</sup> Disciplines include, for example pathologists, who may unreliably score tissue specimens for histology, cell counts, colouring of cells, and radiologists, who, for example, score X-rays, CT scans, MRI scans, and ultrasound measurements. This variability can appropriately be measured with kappa statistics.<sup>248</sup> The interobserver and intraobserver variability can be substantial, which will be reflected in low kappa values.

### \*3.5.2 *Example: Histology in Barrett's Oesophagus*

Barrett's oesophagus is a pre-malignant condition. Surgery is sometimes performed in high-grade dysplasia, whereas other physicians defer treatment until adenocarcinoma is diagnosed. The agreement between readings of histology in Barrett's oesophagus for high-grade dysplasia or adenocarcinoma was only fair, with kappa values around 0.4.<sup>314</sup> The agreement between no dysplasia and low-grade dysplasia had been reported as even lower.<sup>389</sup> Because of observer variability, sometimes a central review process is organized, where an expert reviews all readings. This should be done independently and blinded for previous scores. Subsequently a rule has to be determined for the final score, for example that only the expert score is used, or that an additional reader is required in case of disagreement. Also, consensus procedures can be set up with experts only, for example with scoring by two experts, and involvement of a third if these disagree.<sup>230</sup> Some use the unreliability of classical pathology as an argument for using modern biomarkers.<sup>247</sup>

### 3.5.3 Biological Variability

Apart from observer variability, some measurements are prone to biological variability. A well-known example is blood pressure, where a single measurement is quite unreliable.<sup>318</sup> Usually at least two measurements are made, and preferably more, with some spread in time. Again, definitions have to be clear (e.g. position of patient at the measurement, time of day).

### 3.5.4 Regression Dilution Bias

The effect of unreliable scoring by observers, or biological variability, generally is a dilution of associations of predictors with the outcome. This has been labelled “regression dilution bias”, and methods have been proposed to correct for this bias.<sup>257</sup> A solution is to repeat unreliable measurements, either by the same observer (e.g. use the mean of three blood pressure measurements), or different observers (e.g. double reading of mammograms by radiologists). Practical constraints may limit such procedures.

### \*3.5.5 Example: Simulation Study on Reliability of a Binary Predictor

Suppose we have a binary predictor that we measure with noise. Suppose two observers make independent judgments of the predictor. Their judgments agree with the true predictor status with sensitivity of 80% (observer scores 1 if true = 1) and specificity of 80% (observer scores 0 if true = 0, Table 3.2). If both observers score the predictor independently and without correlation, the observers agree with each other with a kappa of only 0.36 (Table 3.3).

The true predictor status predicts outcome well, with an odds ratio of 4. The observed predictor status has a diluted predictive effect, with odds ratio 2.25. Similarly, the discriminative ability is diluted ( $c$  statistic decreases from 0.67 to 0.60, Table 3.4).

**Table 3.2** Sensitivity and specificity for observers in determining the true predictor status (sensitivity = specificity = 80%)

		True predictor status	
		0 N(col%)	1 N(col %)
Observer	0	750 (80%)	187 (20%)
	1	187 (20%)	750 (80%)

**Table 3.3** Agreement between observer 1 and observer 2 ( $\kappa = 0.36$ )

		Observer 2	
		0	1
Observer 1	0	637	300
	1	300	637

**Table 3.4** Association with outcome for the true predictor status and observed predictor status (by observer 1 or 2, Table 3.3)

		Outcome		Odds ratio	$c$ statistic
		0 $N$ (row%)	1 $N$ (row%)		
True predictor status	0	625 (67%)	312 (33%)	4.0	0.67
	1	312 (33%)	625 (67%)		
Observer	0	562 (60%)	375 (40%)	2.25	0.60
	1	375 (40%)	562 (60%)		

### 3.5.6 Choice of Predictors

In aetiological research we may often aim for the best assessment of an exposure variable. We will be concerned about various information biases that may occur. In the context of a prediction model we can be much more pragmatic. If we aim to develop a model that is applicable in daily practice, we should use definitions and scorings that are in line with daily practice. For example, if medical decisions on surgery are made considering local pathology reports, without expert review, the local pathology report should be considered for a prediction model applicable to the local setting. As illustrated, such less reliable assessments will affect the performance of a predictive model, since predictive relationships are disturbed. If misclassification is at random, a dilution of the relationship occurs (Table 3.4). On the other hand, if measurements are more reliable in clinical practice than in a research setting, e.g. repeated assessments of blood pressure, we might argue that a correction has to be made in the prediction model. In practice, prediction models tend to include predictors that are quite readily available, not too costly to obtain, and can be measured with reasonable precision.

## 3.6 Outcome

### 3.6.1 Types of Outcome

The outcome of a prediction model should be relevant, either from an applied medical perspective or from a research perspective. From a medical perspective, “hard” end points are generally preferred. Especially mortality is often used as an end point in prognostic research. Mortality risks are relevant for many acute and chronic

conditions, and for many treatments, such as surgery. In some diseases, mortality may not be a relevant outcome. Other outcomes include non-fatal events (e.g. disease recurrence), patient centred outcomes (e.g. scores on quality of life questionnaires), or wider indicators of burden of disease (e.g. absence from work, Table 3.5, based on Hemingway<sup>184</sup>).

### **3.6.2 *Survival End points***

When cause-specific mortality is considered, a reliable assessment of the cause of death is required. If cause of death is not known, relative survival can be calculated.<sup>166,167</sup> This is especially popular in cancer research. Mortality in the patients with a certain cancer is compared with the background mortality from the general population. The difference can be thought of as mortality due to the cancer.

The pros and cons of relative survival estimates are open to debate. Some have proposed to also study conditional survival for patients already surviving for some years after diagnosis. These measures may sometimes be more meaningful for clinical management and prognosis than 5-year relative survival from time of diagnosis.<sup>139,214</sup> Others have proposed that median survival times are better indicators of survival than 5-year relative survival rates, especially when survival times are short.<sup>319</sup>

### **\*3.6.3 *Example: Relative Survival in Cancer Registries***

Five-year relative survival was studied for patients enrolled in the SEER registry in the period 1990–1999.<sup>139</sup> The 5-year relative survival rate for persons diagnosed with cancer was 63%, with substantial variation by cancer site and stage at diagnosis.

**Table 3.5 Examples of prognostic outcomes<sup>184</sup>**

Prognostic outcome	Example	Characteristics
Fatal events	All-cause, or cause-specific	Hard end point, relevant in many diseases, but sometimes too infrequent for reliable statistical modeling
Non-fatal events	Recurrence of tumor, cardiovascular events (e.g. myocardial infarction, revascularization)	Somewhat softer end point, reflecting decision-making by physicians, increases power for analysis
Patient centered	Symptoms, functional status, health-related quality of life, utilities	Subjective end point, focused on the patients themselves; often used as secondary end point
Wider burden	Absence from work because of sickness	Especially of interest from an economical point of view

Five-year relative survival increased with time since diagnosis. For example, for patients diagnosed with cancers of the prostate, female breast, corpus uteri, and urinary bladder, the relative survival rate at 8 years after diagnosis was over 75%.

Similar analyses were performed with registry data from the Eindhoven region, where it was found that patients with colorectal, melanoma skin, or stage I breast cancer could be considered cured after 5–15 years, whereas for other tumours survival remained poorer than the general population.<sup>214</sup>

### **3.6.4 Composite End Points**

Sometimes composite end points are defined, which combine mortality with non-fatal events. Composite end points are especially popular in cardiovascular research (see also Chap. 23). For example, the Framingham models have been used to predict incident cardiovascular disease in the general population. A popular Framingham model (the Wilson model) defines cardiovascular events as fatal or non-fatal myocardial infarction, sudden death, or angina pectoris (stable or unstable).<sup>487</sup> Composite end points have the advantage of increasing the effective sample size and hence the power for statistical analyses.

### **\*3.6.5 Example: Mortality and Composite End Points in Cardiology**

A prediction model was developed in 949 patients with decompensated heart failure. The outcome was 60-day mortality or the composite end point of death or rehospitalization at 60 days. The discriminatory power of the model was substantial for mortality ( $c$  statistic 0.77) but less for the composite end point ( $c$  statistic 0.69).<sup>121</sup> These findings are in line with prediction of acute coronary syndromes, where predictive performance was better for mortality than for a composite end point of mortality or myocardial (re)infarction.<sup>43</sup> The case study in Chap. 23 also considers a composite end point.

### **3.6.6 Choice of Prognostic Outcome**

The choice of a prognostic outcome should be guided by the prediction problem, but the outcome should be measured as reliable as possible. Prediction models may be developed with pragmatic definitions of predictors, since this may resemble the future use of a model. But the outcome should be determined with similar rigour as in an aetiological study or randomized clinical trial. In the future, decisions are to be based on the predictions from the model. Predictions hence need to be based on robust statistical associations with an accurately determined outcome.

If there is a choice between binary and continuous outcomes, the latter are preferred from a statistical perspective, since they provide more power in the analysis. Also, ordered outcomes provide more power than binary outcomes. In practice, binary outcomes are however very popular, making logistic regression and Cox regression the most common techniques for prediction models in medicine.

### ***3.6.7 Diagnostic End Points***

The outcome in diagnostic research naturally is the underlying disease, which needs to be defined according to a reference standard.<sup>48,49,238,296</sup> The reference standard can sometimes be anatomical, e.g. findings at surgery. Other definitions may include blood or spinal fluid cultures (e.g. in infectious diseases), results of high-quality diagnostic tests such as angiography (e.g. in coronary diseases), and histological findings (e.g. in oncology). Methods are still under development on how to deal with the absence of an acceptable reference standard. In such situations the results of the diagnostic test can, for example, be related to relevant other clinical characteristics and future clinical events.<sup>360</sup>

The relevance of the underlying diagnosis may be high when treatment and prognosis depends directly on the diagnosis. This is for example the case with testing for genetic defects such as trisomy 21 (Down syndrome). However, often a diagnosis covers a spectrum of more and less severe disease, and longer-term outcome assessment would be desirable. This is especially relevant in the evaluation of newer imaging technology, which may detect disease that remained previously unnoticed.<sup>34,266</sup>

### ***\*3.6.8 Example: PET Scans in Oesophageal Cancer***

In oesophageal cancer, positron-emission tomography (PET) scans provide additional information on extent of disease compared to CT scanning alone.<sup>316,495</sup> However, the clinical relevance of the additionally detected metastases can only be determined in a comparative study, preferably a randomized controlled trial. Diagnosing more metastases is not sufficient to make PET/CT clinically useful.<sup>462</sup>

## **3.7 Phases of Biomarker Development**

Pepe has proposed a phased approach to developing predictive biomarkers, in particular for early detection of cancer<sup>332</sup> (Table 3.6). These phases are also relevant to the development of future prediction models, which may add novel biomarkers to traditional clinical characteristics. The development process begins with small studies focused on classification performance and ends with large studies of impact on

**Table 3.6** Phases of development of a biomarker for cancer screening<sup>332</sup>

Phase	Objective	Study design
1. Preclinical exploratory	Promising directions identified	Case-control (convenient samples)
2. Clinical assay and validation	Determine if a clinical assay detects established disease	Case-control (population based)
3. Retrospective longitudinal	Determine if the biomarker detects disease before it becomes clinical. Define a “screen positive” rule	Nested case-control in a population cohort
4. Prospective screening	Extent and characteristics of disease detected by the test; false referral rate	Cross-sectional population cohort
5. Cancer control	Impact of screening on reducing the burden of disease on the population	Randomized trial

populations. The aim is to select promising markers early while recognizing that early studies do not answer the ultimate questions that need to be addressed.

As an example, Pepe considers the development of a biomarker for cancer screening. Phase 1 is exploratory and may consider gene expression arrays or protein mass spectrometry that yields high-dimensional data for biomarker discovery. Reproducibility between laboratories is an aspect to consider before moving on to phase 2, where a promising biomarker is compared between population-based cases with cancer and population-based controls without cancer. Phase 3 is a more thorough evaluation in a case-control study to determine if the marker can detect subclinical disease. In phase 4, the marker may be applied prospectively as a screening test in a population. Finally, the overall impact of screening is addressed in phase 5 by measuring effects on clinically relevant outcomes such as mortality.

The study design implications are also shown in Table 3.6. In the exploratory phase 1 it may be acceptable to use “convenient samples”, which will likely lead to spectrum bias in the assessment of the biomarker. In phase 2, population-based samples are desired for a simple case-control design. In phase 3, we require samples taken from cancer patients before their disease became clinically apparent. A nested case-control study design can be efficient for data from a cohort study. For phase 4, a prospective cohort study is required to determine the characteristics and treatability of early detected disease. Finally, an RCT is desired for unbiased assessment of the impact of screening.

### 3.8 Statistical Power

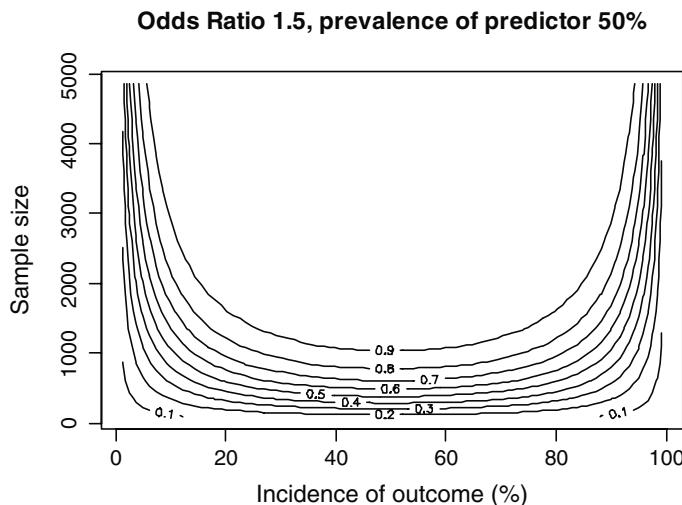
An important issue is how large a study needs to be for sufficient statistical power to address the primary research question. Power considerations are given for studying effects of a specific predictor, and for developing a prediction model that can provide reliable predictions.

### 3.8.1 Statistical Power to Identify Predictor Effects

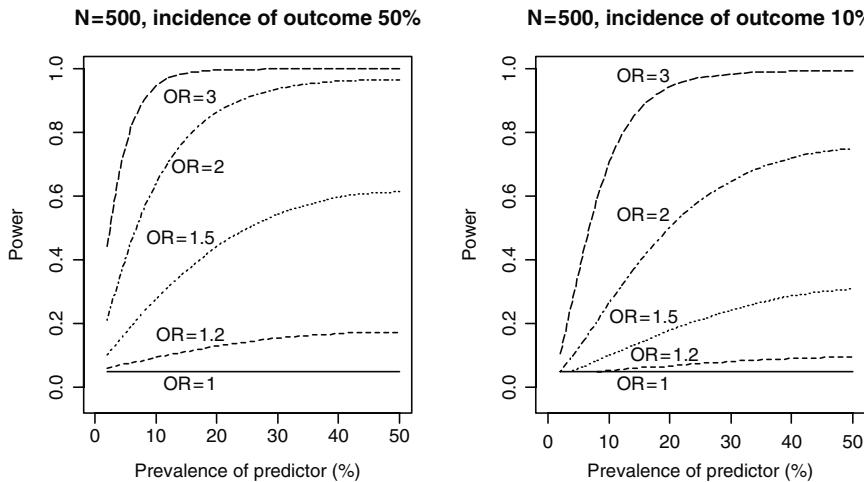
We may primarily be interested in the effect of a specific predictor on a diagnostic or prognostic outcome. We may then aim to test the effect of this predictor for statistical significance. This leads to similar sample size considerations as for testing of treatment effects, e.g. in the context of an RCT. Sample size calculations are straightforward for such univariate testing. The required sample size is determined by choices for the acceptable Type I and Type II error. The Type I error is usually set at 5% for statistical significance. The Type II error determines the power, and may, e.g. be set at 20% for 80% power. Other considerations are the variability of the effect estimate. For binary predictors of a binary outcome, the prevalence of the predictor and the incidence of the outcome are important. Finally, the magnitude of the effect determines the required sample size, with larger sample size required to detect smaller effects.

### \*3.8.2 Examples of Statistical Power Calculations

Sample size calculations can be performed for most types of regression models with standard software. For illustration, we consider the statistical power for a binary predictor of a binary outcome (Fig. 3.3). We find that the required sample size increases steeply with a very low or very high incidence of the outcome. With



**Fig. 3.3** Power corresponding to sample sizes for incidence of the outcome ranging from 0 to 100%. A binary predictor was considered with 50% prevalence with odds ratio 1.5



**Fig. 3.4** Power in relation to prevalence of a binary predictor, for odds ratios from 1 to 3 in samples with 500 subjects. Incidences of the outcome were 50% (left panel) and 10% (right panel)

an odds ratio of 1.5, 80% power requires approximately 2,000 subjects at a 10% incidence, 1,000 subjects at 20% incidence, and 800 subjects at 50% incidence.

Next, we illustrate that statistical power is related to the prevalence of a binary predictor (Fig. 3.4). We consider odds ratios from 1 to 3, as may often be encountered in medical prediction research. In a sample size of 500 subjects, 250 with and 250 without the outcome, 80% power is reached with prevalences of 16% and 5.5% for odds ratios of 2 and 3, respectively. Odds ratios of 1.2 and 1.5 require sample sizes of 3,800 and 800 at 50% prevalence, respectively. With 10% incidence of the outcome, power is substantially lower (Fig. 3.4, right panel). An odds ratio of 3 now requires 18% instead of 5.5% prevalence of the predictor for 80% power. Without an effect (OR=1), statistical significance is by definition expected in 5%.

### 3.8.3 Statistical Power for Reliable Predictions

Instead of focusing on predictors, we can consider the reliability of predictions that are provided by a prediction model. Some rules of thumb have been proposed, supported by simulation studies. The sample size requirements are commonly formulated as events per variable (EVP). The minimum EVP for obtaining good predictions may be 10.<sup>175,326,327</sup> Clinical prediction models that are constructed with EVP less than 10 are overfitted, and may perform poorer than a simpler model which considers fewer predictors, such that the EVP is at least 10 (see further illustration in Chap. 24). EVP values for reliable selection of predictors from a larger set of candidate predictors may be as large as 50 (events per candidate predictor, see Chap. 11). For

pre-specified models, shrinkage may not be required with EPV of at least 20 (Chap. 13).<sup>410</sup> Validation studies may need to include at least 100 events (details in Chap. 19).<sup>465</sup>

Other EPV values may apply for specific circumstances. Regression analyses can technically well be performed with lower EVP values. Adjusted analyses of an exposure variable may be performed with EPV less than 10 when we only aim to correct for confounding.<sup>473</sup>

### 3.9 Concluding Remarks

Prognostic studies are ideally designed as prospective cohort studies, where the selection of patients and definition of predictors is pre-specified. Data from randomized clinical trials may often be useful, although representativeness of the included patients for the target population should be considered as a limitation. Data may also be used from retrospective designs, registries, and case–control studies, each with their strengths and limitations. Diagnostic studies are usually cross-sectional in design, and should prospectively select all patients who are suspected of a disease of interest. In practice, designs are still frequent where patients are selected by a reference test which is not performed in all patients.

Predictors should be defined pragmatically, and cover the relevant domains for prediction of outcome in a disease. The outcome of a prediction model should be measured with high accuracy. Hard end points such as mortality are often preferred but statistical power considerations may motivate the use of composite and other end points.

## Questions

### 3.1 Cohort studies

One could argue that both diagnostic and prognostic studies are examples of cohort studies.

- (a) What is the difference between diagnostic and prognostic outcomes in such cohorts?
- (b) What is the implication for the statistical analysis?

### 3.2 Prospective vs. retrospective designs (Sect. 3.2)

Prospective study designs are generally noted as preferable to retrospective designs. What are the pros and cons of prospective vs. retrospective designs?

### 3.3 Accuracy of predictors and outcome (Sect. 3.5 and 3.6)

- (a) Why do we need to be more careful with reliable assessment of outcome than reliable assessment of predictors?
- (b) What is the effect of imprecise measurement of a predictor?

### 3.4 Composite end points (Sect. 3.6.4)

Composite end points are often motivated by the wish to increase statistical power for analysis. What is the price that we pay for this increase in term of assumptions on predictive relationships? See a recent JCE paper for a detailed discussion.<sup>140</sup>

### 3.5 Statistical power (Figs. 3.3 and 3.4)

- (a) What is the required total sample size for 50% power at 10%, 30%, or 50% incidence of the outcome in Fig. 3.3?
- (b) What is the similarity between Fig. 3.3 and 3.4 with respect to the ranges of the incidence of the outcome or prevalence and associated statistical power?

# Chapter 4

## Statistical Models for Prediction

**Background** In this chapter, we consider statistical models for different types of outcomes: binary, unordered categorical, ordered categorical, continuous, and survival data. We discuss common statistical models in medical research such as the linear, logistic, and Cox regression model, and also simpler approaches and more flexible extensions, including regression trees and neural networks. Details of the methods are found in many excellent texts. We focus on the most relevant aspects of these models in a prediction context. All models are illustrated with case studies. In Chap. 6, we will discuss aspects of choosing between alternative statistical models.

### 4.1 Continuous Outcomes

Continuous outcomes have traditionally received most attention in texts on regression modelling, with the ordinary least square model (“linear regression”) as the reference statistical model.<sup>64,137,232,281,472</sup> Continuous outcome are quite common in medical, epidemiological, and economical studies, but not so often considered for clinical prediction models.

The linear regression model can be written as

$$y = \alpha + \beta_i \times x_i + \text{error},$$

where  $\alpha$  refers to the intercept,  $\beta_i$  to the set of regression coefficients that relate one or more predictors  $x_i$  to the outcome  $y$ . The error is calculated as observed  $y$  – predicted  $y$  ( $\hat{y}$ ). This difference is also known as the residual for the prediction of  $y$ . We assume that the residuals have a normal distribution, and do not depend on  $x_i$  (“homoscedasticity”).

The outcome  $y$  is hence related to a *linear combination* of the  $x_i$  variables with the estimated regression coefficients  $\beta_i$ . This is an important property, which is also seen in *generalized* linear models, such as the logistic regression model.

### \*4.1.1 Examples of Linear Regression

An example of a medical outcome is blood pressure. We may want to predict the blood pressure after treatment with an anti-hypertensive or other intervention.<sup>241,460</sup> Also, quality of life scales may be relevant to evaluate.<sup>242</sup> Such scales are strictly speaking only ordinal, but can for practical purposes often be treated as continuous outcomes. A specific issue is that quality of life scores have ceiling effects, because minimum and maximum scores apply.

### 4.1.2 Economic Outcomes

Health economics is another important field where continuous outcomes are considered, such as length of stay in hospital, or length of stay at a specific ward (e.g. the intensive care unit), or total costs for patients.<sup>88</sup>

Cost data are usually not normally distributed. Such economic data have special characteristics, such as patients without any costs (zero), and a long tail because some patients having considerable costs. We might consider the median as a good descriptor of the outcome. Interestingly, we are however always interested in the mean costs, since the expectation is what matters most from an economical perspective. Sometimes analyses have been performed to identify “high-cost” patients, after dichotomizing the outcome at some cost threshold.

### \*4.1.3 Example: Prediction of Costs

Many children in moderate climates suffer from an infection by the respiratory syncytial virus (RSV). Some children, especially premature children are at risk of a severe infection, leading to hospitalization. The mean RSV hospitalization costs were 3,110 euros in a cohort of 3,458 infants and young children hospitalized for severe RSV disease during the RSV seasons 1996–1997 to 1999–2000 in the Southwest of The Netherlands. RSV hospitalization costs were higher for some patient categories, e.g. those with lower gestational age or lower birth weight, and younger age. The linear regression model had an adjusted  $R^2$  of 8%.<sup>345</sup> This indicates a low explanatory ability for predicting hospitalization costs of individual children. However, the model could accurately estimate the anticipated mean hospitalization costs of groups of children with the same characteristics. These predicted costs were used in decision analyses of preventive strategies for severe RSV disease.<sup>46</sup>

### 4.1.4 Transforming the Outcome

An important issue in linear regression is whether we should transform the outcome variable. The residuals ( $y - \hat{y}$ ) from a linear regression should have a normal distribution with a constant spread (“homoscedasticity”). This can sometimes be achieved by,

e.g. a log transformation for cost data, but other transformations are also possible. As Harrell points out, transformations of the outcome may reduce the need to include transformations of predictor variables.<sup>174</sup> Care should be taken in backtransforming predicted mean outcomes to the original scale. Predicted medians and other quantiles are not affected by transformation. The log-normal distribution can be used for the mean on the original scale after a log transformation, but a more general, non-parametric, approach is to use “smearing” estimators.<sup>341</sup>

### 4.1.5 Performance: Explained Variation

In linear regression analysis, the total variance in  $y$  (“total sum of squares”, TSS) is the sum of variability explained by one or more predictors (“model sum of squares”, MSS) and the error (“residual sum of squares”, RSS):

$$\text{TSS} = \text{MSS} + \text{RSS}$$

$$\text{var}(\text{regression on } x_i) + \text{var}(\text{error}) = \sum (\hat{y} - \text{mean}(y))^2 + \sum (y - \hat{y})^2$$

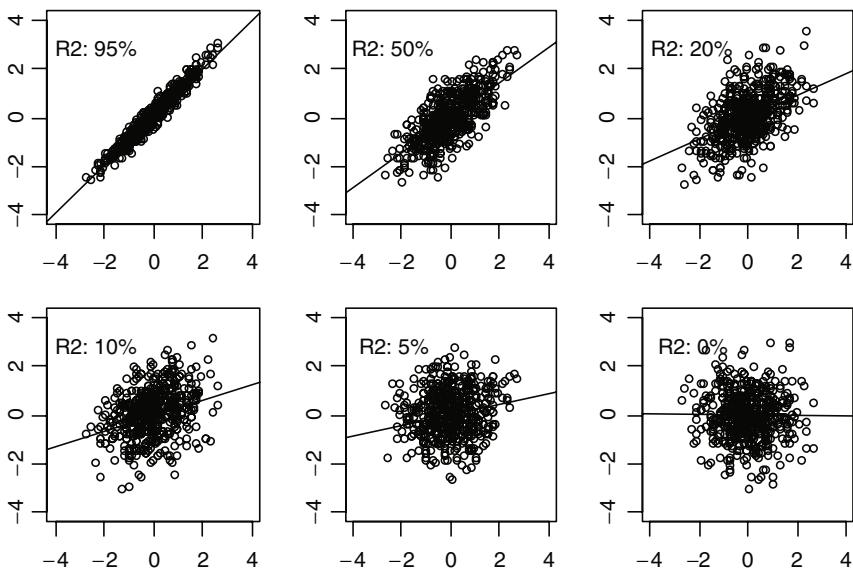
The estimates of the variance follow from the statistical fit of the model to the data, which is based on the analytical solution of a least squares formula. This fit minimizes the error term in the model, and maximizes the variance explained by  $x_i$ . Better prediction models explain more of the variance in  $y$ .  $R^2$  is defined as  $\text{MSS} / \text{TSS}$ .<sup>472</sup>

To appreciate values of  $R^2$ , we consider six hypothetical situations where we predict a continuous outcome  $y$ , which has a standard normal distribution ( $N(0,1)$ , i.e. mean 0 and standard deviation 1) with one predictor  $x$  ( $N(0,1)$ ). The regression coefficients for  $x$  are varied in simulations, such that  $R^2$  is 95%, 50%, 20%, 10%, 5%, and 0% (Fig. 4.1). We note that an  $R^2$  of 95% implies that observed outcomes are always very close to the predicted values, while gradually relatively more error occurs with lower  $R^2$  values. When  $R^2$  is 0%, no association is present.

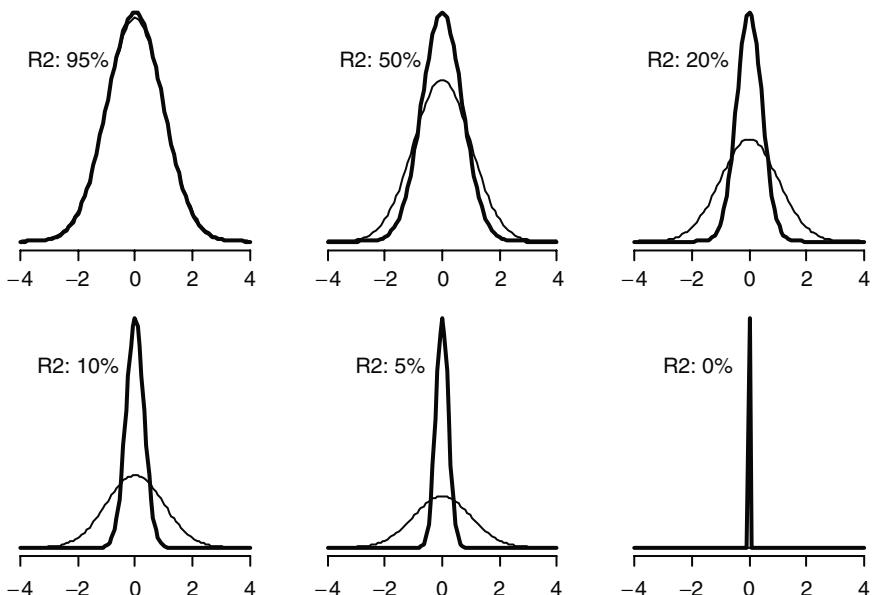
To appreciate  $R^2$  further, we plot the distributions of predicted values ( $\hat{y}$ ). The distribution of  $\hat{y}$  is wide when  $R^2$  is 95%, and very small when  $R^2$  is 5%, and near a single line when  $R^2$  is 0% (Fig. 4.2). The distribution of  $y$  is always normal with mean 0 and standard deviation 1.

### 4.1.6 More Flexible Approaches

The generalized additive model (GAM) is a more flexible variant of the linear regression model.<sup>180, 181, 472</sup> A GAM allows for more flexibility especially for continuous predictors. It replaces the usual linear combination of continuous predictors with a sum of smooth functions to capture potential non-linear effects:  $y = b_0 + f_i(x_i) + \text{error}$ , where  $f_i$  refers to functions for each predictor, e.g. loess smoothers.



**Fig. 4.1** Linear regression analysis with true regression models with  $y = \beta \times x + \text{error}$ , where  $\text{sd}(y) = \text{sd}(x) = 1$ . The outcome  $y$  is shown on the  $y$ -axis,  $x$  on the  $x$ -axis



**Fig. 4.2** Probability density functions for observed and predicted values (“fitted values”,  $\hat{y}$ ). For the first graph ( $R^2 = 95\%$ ), the distribution of predicted values (thick line) is nearly identical to the distribution of observed  $y$  values (thin line), while for the last graph all predictions are for the average of 0

Loess smoothers are based on locally weighted polynomial regression.<sup>75</sup> At each point in the data set a low-degree polynomial is fit to a subset of the data, with data values near the point where the outcome  $y$  is considered. The polynomial is fitted using weighted least squares, giving more weight to nearby points and less weight to points further away. The degree of the polynomial model and the weights can be chosen by the analyst.

The estimation of a GAM is more computationally demanding than for linear models, but this is no limitation anymore with modern computer power and software. A GAM assumes that the outcome is already appropriately transformed, and then automatically estimates the transformation of continuous predictors to optimize prediction of the outcome.

An even more flexible approach is the alternating conditional expectation method.<sup>174, 181</sup> Here,  $Y$  and  $Xs$  are simultaneously transformed to maximize the correlation between the transformed  $Y$  and the transformed  $Xs$ .

$g(y) = \alpha + f_i(x_i) + \text{error}$ , where  $g$  refers to a transformation of the outcome  $y$ , and  $f_i$  refers to functions for each predictor. For cost data, several other specific approaches have been proposed.<sup>27, 341</sup>

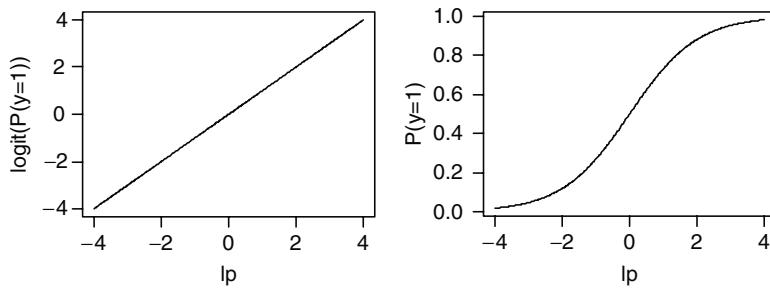
## 4.2 Binary Outcomes

For outcome prediction, we often consider diagnostic (presence of disease) or prognostic outcomes (e.g. mortality, morbidity, complications, see Chap. 2). The logistic regression model is the most widely used statistical technique nowadays for such binary medical outcomes.<sup>174, 472</sup> The model is flexible in that it can incorporate categorical and continuous predictors, non-linear transformations, and interaction terms. Many of the principles of linear regression also apply for logistic regression, which is an example of a *generalized* linear model. As in linear regression, the binary outcome  $Y$  is linked to a linear combination of a set of predictors and regression coefficients  $\beta$ . We use the logistic link function to restrict predictions to the interval  $<0,1>$ . The model is stated in terms of the probability that  $y = 1$  (“ $P(y=1)$ ”), rather than the outcome  $Y$  directly.

Specifically, we write the model as a linear function in the logistic transformation (logit), where  $\text{logit}(P(y=1)) = \log(\text{odds}(P(y=1)))$ , or  $\log([P(y=1)/(P(y=1)+1)])$ :

$\text{Logit}(P(y=1)) = \alpha_0 + \beta_i x_i = \text{lp}$ , where logit indicates the logistic transformation,  $\alpha$  the intercept,  $\beta_i$  the estimated regression coefficients,  $x_i$  the predictors, and lp linear predictor.

The coefficients  $\beta_i$  are usually estimated by maximum likelihood in a standard logistic regression approach, but this is not necessarily the case. For example, we will discuss penalized maximum likelihood methods to shrink the  $\beta_i$  for predictive purposes (Chap. 13). The interpretation of the coefficients  $\beta_i$  is as for any regression model, that the coefficient indicates the effect of a 1-unit increase in  $x_i$ , keeping the other predictors in the model constant. When we consider a single predictor in a logistic model,  $\beta_i$  is an unadjusted, or univariate effect; with multiple predictors, it is an “adjusted” effect, conditional on the values of other predictors in the model. The exponent of the regression coefficient ( $e^\beta$ ) indicates the odds ratio.



**Fig. 4.3** Logistic function. The linear predictor  $lp$  is related to the predicted probability  $P(y=1)$  as:  $\text{Logit}(P(y=1)) = lp$ , or  $P(y=1) = 1 / (1 + \exp(-lp))$

Predicted probabilities can be calculated by backtransforming:  $p(y=1) = e^{lp} / (1 + e^{lp}) = 1 / (1 + e^{-lp})$ . The quantity  $e^{lp}$  is the odds of the outcome. The logistic function has a characteristic sigmoid shape, as is bounded between 0 and 1 (Fig. 4.3). We note that a  $lp$  value of 0 corresponds to a probability of 50%. Low  $lp$  values correspond to low probabilities (e.g.  $lp -4, p < 2\%$ ), and high  $lp$  values correspond to high probabilities (e.g.  $lp +4, p > 98\%$ ).

#### 4.2.1 *R*<sup>2</sup> in Logistic Regression Analysis

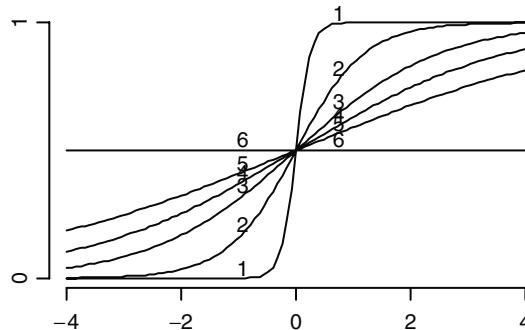
We learned from the linear regression examples that  $R^2$  is related to the relative spread in predictions. When predictions cover a wider range, the regression model better predicts the outcome. This concept also applies to dichotomous outcomes, e.g. analyzed with a logistic regression model. Better prediction models for dichotomous outcomes have a wider spread in predictions, i.e. predictions close to 0% and close to 100%.

To illustrate this concept, we use the same simulated data as for the examples of linear regression models, but we now dichotomize the outcome  $y$  (if  $y < 0$ ,  $y_d = 0$ , else  $y_d = 1$ ). The relationship between a standard normal variable  $x$  and the six  $y_d$  outcomes is shown in Fig. 4.4.

#### \*4.2.2 Calculation of $R^2$ on the Log Likelihood Scale

Where the linear model is optimized with least squares estimation, the logistic model is usually optimized with maximum likelihood techniques. The likelihood refers to the probability of the data given the model, and enables estimation of parameters in various non-linear models. The natural logarithm of the likelihood (log likelihood, LL) is usually used for convenience in numerical estimation. The LL is calculated as the sum over all subjects of the distance between the natural log of the predicted probability  $p$  for the binary outcome to the actually observed outcome  $y$ :

$$\text{LL} = \sum y \times \log(p) + (1 - y) \times \log(1 - p),$$



**Fig. 4.4** Predicted probabilities of a 0/1 outcome by six logistic models according to a normally distributed  $x$  variable. The predictive strength varied, with Nagelkerke's  $R^2$  decreasing from 87% (labelled "1") to 0% (label "6")

where  $y$  refers to the binary outcome and  $p$  the predicted probability for each subject.

If  $y = 1$ , the probability should be high (ideally 100%), such that  $\log(p)$  gets close to 0. Then the term  $(1-y)$  drops out. If  $y=0$ , the term  $(1-y) = 1$ , and  $p$  should be low (ideally 0%), such that  $\log(1-p)$  gets close to zero. A perfectly fitting model would have an LL of zero. In medical problems, perfect predictions cannot be made, unless a fully deterministic model is identified. The LL is hence usually negative for a fitted logistic regression model. A better model will have an LL closer to zero.

As reference we consider the LL of a model with average predictions:

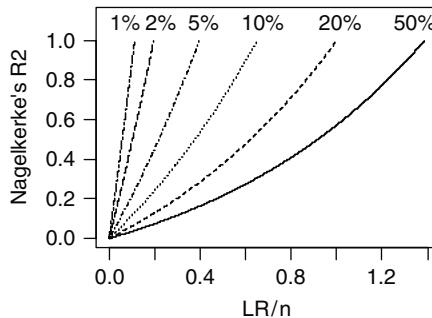
$$\text{LL}_0 = \sum y \times \log (\text{mean}(y/n)) + (1 - y) \times \log (1 - \text{mean}(y/n)),$$

where  $\text{LL}_0$  refers to the log likelihood of the Null model, and  $\text{mean}(y/n)$  is the average probability of the binary outcome  $y$ . The  $\text{LL}_0$  is negative, unless  $y/n$  is 0 or 1.

We can quantify the performance of a prognostic model by comparison with the Null model. We multiply by  $-2$ , since the difference on the  $-2$  LL scale is a Likelihood Ratio statistic (LR), which follows a  $\chi^2$  distribution:

$$\text{LR} = -2 (\text{LL}_0 - \text{LL}_1),$$

where  $\text{LL}_1$  refers to the model with predictors,  $\text{LL}_0$  to the Null model, and LR is the likelihood ratio. The LR statistic can be used for univariate analysis, but also for testing the joint importance of a larger set of predictors in the model ("global LR statistic"). We can also easily make comparisons between nested submodels, which contain a subset of the predictors in a larger model. For example, we can compare models with and without age as a predictor to determine the LR for age, or compare models with and without a block of predictors, e.g. with and without a set of patient history characteristics. Statistical testing is straightforward between such nested models.



**Fig. 4.5** Relationship between Nagelkerke's  $R^2$  and the LR statistic for incidence of the outcome of 1–50%. The LR is divided by  $n$  to make the scale independent of sample size. We note a reasonably linear relationship, especially for lower incidences. Largest LRs per subject are possible with an incidence of 50%

The absolute value of the LR depends on  $n$ , the number of patients, similar to the sum of squares in linear regression analysis. Several attempts have been made to define an  $R^2$  measure for generalized linear models, relating LR to  $-2 \text{LL}_0$ .  $R^2$  values ideally enable direct comparison across predictors, irrespective whether the predictor was categorical or continuous, and independent of the sample size. A nowadays popular definition of  $R^2$  uses the LR and  $-2 \text{LL}_0$  as follows:

$$R^2 = (1 - \exp(-\text{LR}/n)) / (1 - \exp(-2 \text{LL}_0 / n)),$$

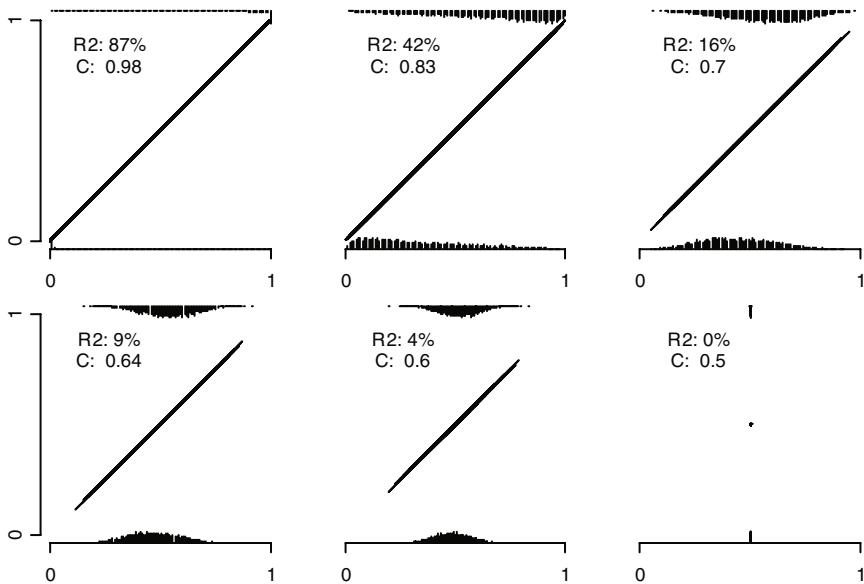
where  $n$  is the number of patients.

This definition of  $R^2$  was proposed by Nagelkerke, and has the advantage of being scaled between 0 and 100%.<sup>309</sup> For a perfect model,  $\text{LR} = +2 \text{LL}_0$ , and  $R^2 = 100\%$ . The relationship between the LR statistic and Nagelkerke's  $R^2$  is more or less linear (Fig. 4.5).

We will use the Nagelkerke definition of  $R^2$  throughout this book. The scaling between 0 and 100% makes it a natural measure to indicate how close we are with our predictions to the observed 0 and 1 outcomes (Fig. 4.6). The calculation is based on the LL scale, which is the scale used in the fitting process to optimize the model given the data. The calculation includes the LR, which is the theoretically preferred quantity for testing of significance in logistic models.

### 4.2.3 Models Related to Logistic Regression

Logistic regression can be viewed as an improvement over linear discriminant analysis, which is an older technique.<sup>170</sup> Discriminant analysis usually makes more



**Fig. 4.6** Distribution of observed outcomes (0 or 1), in relation to predicted probabilities from logistic models relating  $y$  to a predictor  $x$ . The  $y$  variable was created from the linear regression example in Fig. 4.1 by dichotomization, and had an average incidence of 50%. We note that Nagelkerke's  $R^2$  values for logistic regression are slightly smaller than the Pearson  $R^2$  values for linear regression in Fig. 4.1. Discrimination is indicated by the  $c$  statistic (equivalent to the area under the receiver operating characteristic curve, see Chap. 15)

assumptions on the underlying data, for example multivariate normality, which is not the case in logistic regression. The data need to follow a binomial distribution, which is a natural assumption for 0/1 data. However, when correlations between outcomes exist, for example because of grouping of patients within hospitals, this assumption may be violated. Generalized estimation equations (GEE) are an extension of logistic regression for correlated data.<sup>322, 472</sup>

#### 4.2.4 Bayes Rule

Bayes rule has often been used in a diagnostic context for the prediction of the likelihood of an underlying disease.<sup>331</sup> A prior probability of disease ( $p(D)$ ) is considered before information becomes available (e.g. from history taking, or from a diagnostic test, denoted as predictor  $x$ ). The information is used to calculate a posterior probability of disease ( $p(D|x)$ ).

This approach has been followed with some success in the 1970s by De Dombal in deriving diagnostic estimates for patients with abdominal pain.<sup>92</sup> Probabilities were estimated with a Bayesian approach, where the prior probability of a diagnosis was updated with information from a large database. This database contained data on the prevalence of signs and symptoms according to the outcome diagnosis. This information can efficiently be summarized with diagnostic likelihood ratios (“LR”). The diagnostic LR for a specific sign or symptom  $x$  is

$LR(x) = p(x|D) / p(x|!D)$ , where  $D$  indicates presence of the disease (determined by a reference standard), and  $!D$  no disease.

The combination of a prior probability of disease and LR is straightforward with Bayes’ formula:

$$\text{Odds}(D|x) = \text{Odds}(D) \times LR(x), \text{ where}$$

$\text{Odds}(D)$  is the prior odds of disease, calculated as  $p(D)/(1 - p(D))$ .

In logit form the formula reads as:

$$\text{Logit}(D|x) = \text{Logit}(D) + \log(LR(x))$$

This looks very similar to the logistic model shown before. The intercept  $\alpha$  is replaced by  $\text{Logit}(D)$ , the prior probability of disease, and  $\beta_1 \times x_1$  is replaced by  $\log(LR(x))$ . The term “ $\log(LR(x))$ ” has been referred to as “weight of evidence”, since it indicates how much the prior probability changes by evidence from a test.<sup>397</sup>

For a test with a positive or negative result, there is a simple relationship between LR and OR:

$$\text{OR} = LR(+) / LR(-), \text{ and}$$

$\log(\text{OR}) = \text{coefficient} = \log(LR(+)/LR(-)) = \log(LR(+)) - \log(LR(-))$ , where  $LR(+)$  and  $LR(-)$  are the LRs for positive and negative test results, respectively.

In a logistic model with one predictor representing the test (+ or – result), the intercept  $\alpha$  reflects the logit(y) when the test is negative. When the test is positive, the change in logodds is given by the coefficient, and  $\text{logit}(y) = \text{intercept} + \text{coefficient}$ .

#### \*4.2.5 Example: Calculations with Likelihood Ratios

Suppose we have a test with 80% sensitivity and 90% specificity, and a prevalence of disease of 10%. For 1,000 patients, the cross-table may look like Table 4.1.

$$\text{The } LR(+) = p(\text{Test } +|D)/p(\text{Test } +|!D) = 0.8 / 0.1 = 8.$$

The  $LR(-) = p(\text{Test } -|D)/p(\text{Test } -|!D) = 0.2 / 0.9 = 0.22$ . We can calculate the posterior probabilities of disease with the formula  $\text{Odds}(D|x) = \text{Odds}(D) \times LR(x)$ . For a positive test,  $\text{Odds}(D|x) = 100/900 \times 8 = 8/9$ . The probability is calculated as  $\text{odds}/(\text{odds}+1) = (8/9)/(8/9 + 1) = 47\%$ .

For a negative test results,  $\text{Odds}(D) \times LR(-) = 100/900 \times 0.2/0.9 = 2/81$ , or a probability of 2.4% ( $(2/81) / (2/81 + 1)$ ).

These numbers can also be calculated directly from the table: prior = 100/1,000 = 10%; posterior  $80/180=47\%$  and  $20/920=2.4\%$ . On logodds scale the change =

**Table 4.1** Cross-tabulation of a test with + or - results with presence of disease ( $D$  or  $\text{!}D$ )

	$D$	$\text{!}D$	Total
Test +	80	90	180
Test -	20	810	920
Total	100	900	1,000

**Table 4.2** Logistic regression analysis for example in Table 4.1

Variable	$b$	SE	OR [95% CI]
Intercept	-3.701	0.226	
Test	3.583	0.274	36 [21–62]

$\log(8) = +2.1$  for a positive test and  $\log(0.22) = -1.5$  for a negative test result. The odds ratio is  $8/0.22 = 36$ , and the  $\log(\text{OR}) = 2.1 - 1.5 = 3.6$ .

From a logistic regression analysis, we obtain: intercept = -3.7, coefficient for test is 3.6; OR=36 (Table 4.2). So, the linear predictor is -3.7 for a negative test and -0.1 for a positive test, which corresponds to probabilities of 2.4% and 47%; as expected this is identical to the calculations with LRs, or as directly observed from Table 4.1.

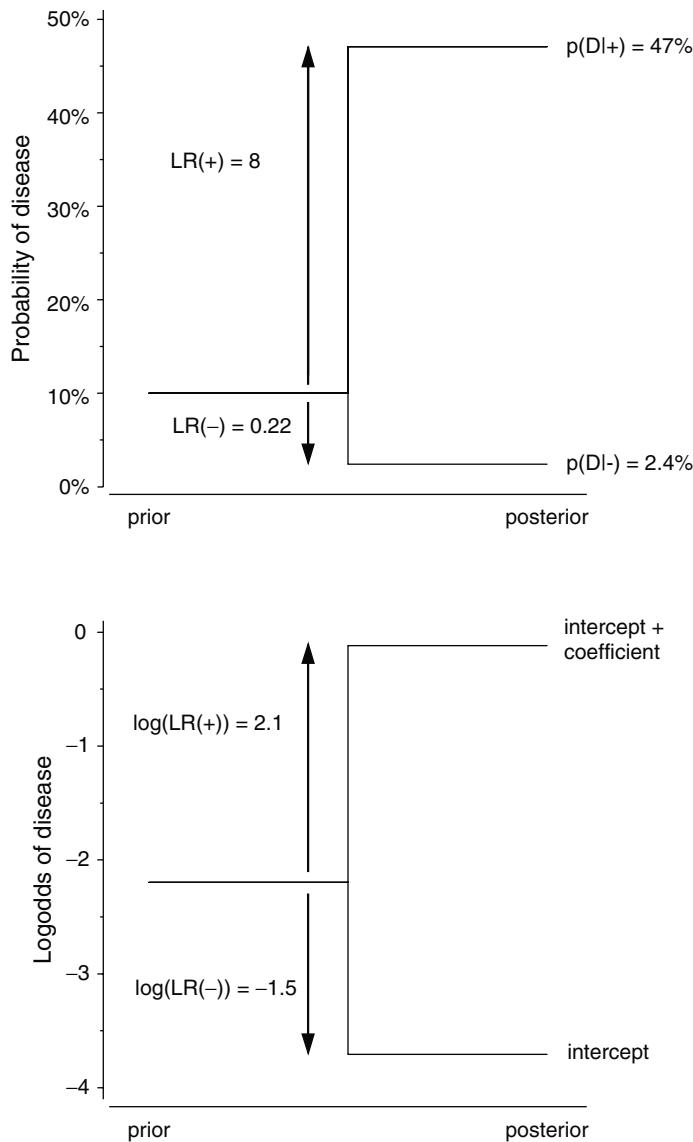
Graphically, we can well illustrate how Bayes' formula works for a positive or negative test result to obtain a posterior probability from a prior probability (Fig. 4.7).

#### 4.2.6 Prediction with Naïve Bayes

Bayes rule is a general scientific approach to handle conditional probabilities, e.g. to obtain  $p(D|x)$  from  $p(x|D)$ . The  $p(x|D)$  can sometimes easier be estimated than  $p(D|x)$ . For example, sensitivity and specificity of a dichotomous test are estimated conditional on disease status. For prediction, we are however interested in  $p(D|x)$ .

De Dombal and others have used a simple method to estimate posterior probabilities for combinations of signs and symptoms.<sup>92</sup> The posterior probability after considering  $x_1$  is used as the prior when considering  $x_2$ , etc. This approach is reasonable if the  $x_1$ ,  $x_2$ , etc. are conditionally independent. Usually positive correlations are however present which makes that the effect of  $x_2$  is smaller once  $x_1$  has already been considered, compared to considering  $x_2$  unconditionally. Such violation of conditional independence makes that  $\text{LR } x_2|x_1(x) < \text{LR } x_2(x)$ .<sup>217</sup>

This sequential application of Bayes' rule is equivalent to using the univariate logistic regression coefficients in a linear predictor. Because of its simplicity and mathematical incorrectness, Naïve Bayes is sometimes referred to as "Idiot's Bayes".



**Fig. 4.7** Graphical illustration of Bayes's formula for a prior probability of disease of 10%. Diagnostic LRs of 0.22 and 8 change the posterior probability of disease to 2.4% and 47%, respectively. The second graph shows the probabilities on the logodds scale

The linear predictor reads like

$$Lp_u = \beta_{1,u} \times x_1 + \beta_{2,u} \times x_2 + \dots + \beta_{p,u} \times x_p,$$

where the subscript  $u$  indicates univariate estimates for the logistic regression coefficients.

### \*4.2.7 Examples of Naïve Bayes

A naïve Bayes modeling approach has been studied by Spiegelhalter, who found remarkably good performance for discrimination.<sup>395</sup> Also, the method has been applied in modelling the effects of genetic markers, where robustness in modelling is required at the expense of accepting bias in coefficients.<sup>387</sup>

### \*4.2.8 Calibration and Naïve Bayes

The problem with Naïve Bayes estimation is that correlations between the predictors are ignored. In the case of positive correlations, predictions will be too extreme, since the effects of predictors are overestimated. Both too low and too high predictions arise. This is reflected in a regression coefficient for the linear predictor (“calibration slope”,  $\beta_{\text{cal}}$ ) below 1 in the model:  $y \sim \text{lp}_u$ . A simple approach hence is to correct this calibration problem with a single coefficient for the linear predictor:  $\text{Logit}(y) = \alpha + \beta_{\text{cal}} \times \text{lp}_u$ .

In terms of multivariable OR ( $\text{OR}_m$ ) or multivariable LR ( $\text{LR}_m$ ), the exponent can be used for easy of notation:  $\text{OR}_m = \text{OR}_u^{\beta_{\text{cal}}}$  or  $\text{LR}_m = \text{LR}_u^{\beta_{\text{cal}}}$ . The idea of recalibrating of the linear predictor comes back in Chap. 15 and 20.

### \*4.2.9 Logistic Regression and Bayes

The diagnostic LR can be used mathematically correct in a multivariable context. The key trick is to rescale test results. Instead of a “1” for positive and a “0” for negative, the univariate log(LR) values can be filled in for the test results.<sup>395</sup> In a multivariable model, the joined effects for the test results are subsequently estimated. Coefficients for the rescaled test results reflect to degree of correlation between test results from different tests. If there are no correlations, the coefficients of each test would be close to 1.

Multivariable diagnostic LRs can also be calculated by comparing models with and without the test of interest. The model without the test is the prior, and the model with the test included provides the posterior probabilities.<sup>217</sup> Subtracting these two equations provides the LRs.

### \*4.2.10 More Flexible Approaches to Binary Outcomes

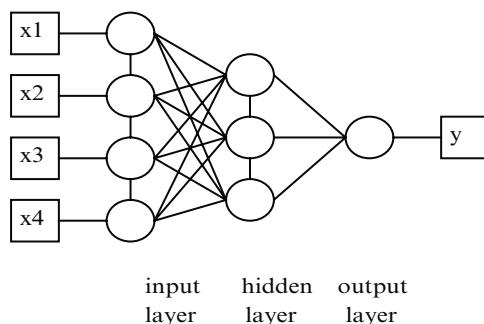
Naïve Bayes estimation is an example of a more simplistic and robust method than logistic regression. A more flexible alternative model is a generalized additive model (GAM), as was already discussed for linear regression models.<sup>180,181,472</sup>

Another alternative is to consider generalized non-linear models. Here, the outcome is no more related to a mathematically simple linear combination of estimated regression coefficients and predictor values. Instead, non-linear combinations of predictors are possible. Generalized non-linear models are currently implemented as neural networks. Neural networks are often presented as fancy tools, “that represent the way our brain works,” but it may be more useful to consider them as non-linear extensions of linear logistic models.<sup>436,438</sup>

The most common neural network model is the multilayer perceptron (Fig. 4.8). In such a network, the neurons are arranged in a layered configuration containing an input layer, usually one “hidden” layer, and an output layer. The values of input variables (patient characteristics) are imported into the network via the input layer and multiplied with the weights of the connections. These multiplied values constitute the input of the next (hidden) layer, from where the process is continued to produce the output variables (e.g. risk of mortality) in the output layer.

A neural network does not use any preliminary information about the links between the input and output variables; the relationships between input and output variables are determined by the data. It is hence not easily possible to explicitly force external knowledge into a model, e.g. that an age effect should be monotonically increasing. Neural networks learn by example; the errors from the initial prediction for the patients are fed back into the network and the weights for connections are adjusted to minimize the error; for the second time predictions are made and compared to the actual outcome. The process from input to output layer is repeated many times. However, to prevent “overtraining” the repetitions are usually stopped before the network is fully trained to the data.<sup>410,436</sup>

The hidden layer makes the network more flexible to recognize patterns in the data compared to a standard logistic regression model. The number of hidden layers and number of nodes are chosen by the analyst. A neural network without a hidden layer is equivalent to a logistic regression model.<sup>436,438</sup>



**Fig. 4.8** A simple neural network with four input variables (predictors  $x_1 - x_4$ ), one hidden layer with three nodes, and one output layer (outcome  $y$ )

### ***4.2.11 Classification and Regression Trees***

Recursive partitioning or Classification And Regression Tree (“CART”) methods have been promoted by some as strong tools for predictive modeling. Recursive partitioning is a statistical method to construct binary trees.<sup>57</sup> The method is based on statistically optimal splitting (“partitioning”) of the patients into pairs of smaller subgroups. Splits are based on cut-off levels of the predictors, which produce maximum separation among two subgroups and a minimum variability within these subgroups with respect to the outcome. The predictor causing the largest separation is situated at the top of the tree, followed by the predictor causing the next largest separation, and so on. Splitting continues until the subgroups reach a minimum size or until no improvement can be obtained. Several variants of recursive partitioning algorithms are available which use different criteria to construct a tree. Details of the statistical procedures can be found elsewhere.<sup>57</sup>

### ***\*4.2.12 Example: Mortality in Acute MI Patients***

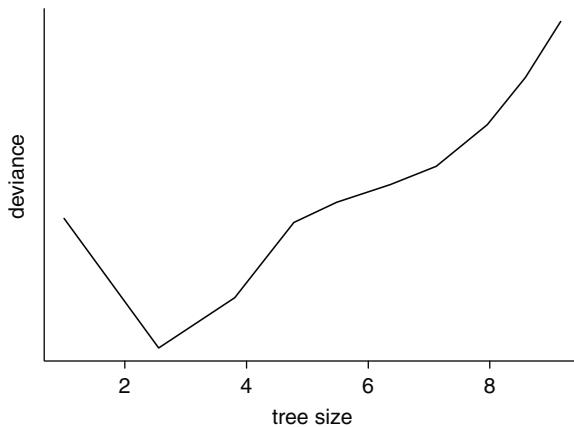
We illustrate the creation of a tree in patients with an acute myocardial infarction (MI). We use a data set from the GUSTO-I trial (see Chap. 22) which is labeled “sample5”. It contains 429 patients, of whom 24 died by 30 days. We consider the predictors age (continuous) and Killip (4 categories, Fig. 4.9). An initial tree was quite complex, with many splits, especially at many age cut-offs. Some counter-intuitive patterns arose, such as a zero mortality among older patients within sub-branches. A technique to construct better prediction trees is to prune a tree back to an “optimal” size. This can be achieved by using a cross-validation procedure (see also Chap. 17). Performance is determined in randomly drawn independent parts of the data for different tree sizes (Fig. 4.10). A pruned tree of size 3 was subsequently created (Fig. 4.11). So an enormous reduction in size was necessary to construct a more stable tree. Prediction of outcome for a new patient is accomplished by simply running that patient down the tree, according to the values of the predictors.<sup>57</sup>

### ***4.2.13 Advantages and Disadvantages of Tree Models***

An advantage of a tree is its simple presentation. Some claim that a tree represents how physicians think: starting with the most important characteristic, followed by another characteristic depending on the answer on the first, etc. Indeed, humans are remarkably quick in pattern recognition based on a few clues. However, humans have typically been outperformed by systematic prediction methods in experiments where a balanced, quantitative judgement was required, such as estimation of a probability based on a set of characteristics.<sup>265</sup> So, the fact that a tree may represent human thinking for classification does not argue in favour of the method for prediction. A true advantage may be that interaction effects are naturally incorporated in a tree, while a standard logistic regression



**Fig. 4.9** Initial tree fitted in a small subsample of GUSTO-I (“sample5”) with age and Killip class as predictors. Splits in the tree are labelled with the criterion for the split, e.g. Killip < 2.5 indicates that patients with Killip class 1 or 2 go to the left in the tree and patients with Killip class 3 or 4 go to the right. The nodes are labeled with 30-day mortality as a fraction, e.g. 0.60 indicates a 60% mortality among those with Killip class 3 or 4. Vertical distances in the tree are based on the statistical improvement between parent and children nodes



**Fig. 4.10** Cross-validated deviance in relation to tree size; optimal size is around 3

model usually starts with main effects, that is one coefficient  $\beta_i$  per predictor. When multiple, high-order interactions are expected in a huge data set, and only categorical predictors are considered, a tree might be a good choice. Such situations may be rare in medical data, but may possibly be encountered in other areas of research.

Disadvantages of trees can be noted by considering a tree as a special case of linear logistic regression. First, all continuous variables have to be categorized, which implies a loss of information. As illustrated in Figs. 4.9 and 4.11, age is con-



**Fig. 4.11** Pruned tree with size 3 for terminal nodes

sidered with different splits at different places in the tree, while the age effect could well be approximated with a single linear term in a logistic model (see Chap. 6). Moreover, these cut-points are determined from a search over all possible cut-points, which is well known to be very dangerous in a prediction context.<sup>12</sup>

Further, the tree assumes interactions between all predictors. After the first split, this interaction is of the first order, i.e.  $x_1 \times x_2$ . At the third level, second-order interactions are assumed ( $x_1 \times x_2 \times x_3$ ). In regression analysis, it is common practice to include main effects of predictors when interactions are considered; this principle is not followed in tree modelling. A higher-order interaction term is included to model the effect of a predictor in a specific branch, and simply omitted from the other branches. A predictor is typically selected in one branch of the tree and not in another. This poses a clear risk of testimation bias (Chap. 5): predictors are selectively considered when their effects are relatively large, and not if their effects are small.

#### \*4.2.14 Trees as Special Cases of Logistic Regression Modelling

From a model selection viewpoint, trees have three distinctive characteristics compared to a logistic regression model when we consider a set of potential predictors.

1. In a logistic model, a default strategy is to include all predictors as main effects. This model can be extended with interaction terms if the power to examine these is sufficient. It is rare to study interactions that are more complex than considering three variables (second order). In contrast, trees by default assume that higher-order interaction are present, and cannot model main effects.

2. Continuous variables should not be categorized in regression models.<sup>355</sup> Trees do so by necessity, which causes a loss of information.
3. One might use a stepwise selection method in a logistic model, especially in larger data sets with sufficient power to select all relevant predictors. Generally a high  $p$ -value is advisable to prevent various problems (Chap. 11).<sup>174</sup> A tree however always needs to be selective in the inclusion of predictors, and quickly runs out of cases within branches. Limited power is a major problem in the development of trees.

As an example we write the linear predictor for the tree in Fig. 4.11 as:

$$Lp = \beta_1 \times \text{Killip} > 2 + \beta_2 \times \text{Killip} \leq 2 \times \text{age} \leq 67.5 +$$

$$\beta_3 \times \text{Killip} = 1 \times \text{age} > 67.5 + \beta_4 \times \text{Killip} = 2 \times \text{age} > 67.5.$$

We estimate four parameters which identify the four terminal nodes. If we want a more standard formulation with an intercept we could write:

$$Lp = \alpha + \beta_1 \times \text{Killip} > 2 + \beta_2 \times \text{Killip} = 1 \times \text{age} > 67.5 + \beta_3 \times \text{Killip} = 2 \times \text{age} > 67.5,$$

where the intercept term refers to patients with Killip 1 or 2, and age = 67.5.

In this formulation, it is clear that age is ignored among those with Killip class  $> 2$ , and that a dichotomized age variable is used in interaction with patients in Killip class 1 or 2.

In a logistic regression model, we could combine Killip class 3 and 4 (representing “shock”), and omit the interaction of Killip with age:

$$Lp = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip} = 1 + \beta_3 \times \text{Killip} = 2 + \beta_4 \times \text{Killip} > 2.$$

Even simpler, we could include Killip as a linear rather than as a categorized predictor:

$$Lp = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip}.$$

We could extend this model to allow for  $\text{age} \times \text{Killip}$  interaction:

$$Lp = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip} + \beta_3 \times \text{age} \times \text{Killip}.$$

#### \*4.2.15 Other Methods for Binary Outcomes

Various other methods are available or under development. Such methods include multivariate additive regression splines (MARS) models. These form a kind of hybrid between generalized additive models and classification trees.<sup>129</sup> MARS models aim to find low-order additive structure as well as interactions between risk factors.

A support vector machine (SVM) performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.<sup>464</sup> Specialized texts are available that discuss these and other statistical models for binary data.<sup>181</sup>

#### **4.2.16 Summary on Binary Outcomes**

In sum, logistic regression provides a quite flexible model to derive predictions from empirical data. Interactions and nonlinearity can be incorporated. Some other models, such as GAM, neural nets (GNLM), MARS, can be seen as extensions, with the default linear logistic model as a special case. Naïve Bayes is a simplified version of logistic regression, ignoring correlations between predictors. Trees can be seen as special cases of logistic regression, requiring categorizations of continuous variables and assuming higher order interactions.

### **4.3 Categorical Outcomes**

Categorical outcomes without a clear ordering are common in diagnostic medical problems. The diagnostic process starts with considering presenting signs and symptoms of a patient. This leads the physician towards a set of differential diagnoses. Each diagnosis has a probability given the patient's clinical and nonclinical profile. Usually, one of these differential diagnoses is defined as the working diagnosis or target disease, to which the diagnostic work-up is primarily directed. Consequently, diagnostic studies commonly focus on the ability of tests to include or exclude the presence of this target disease. The alternative diagnoses (which may all direct different treatment decisions) are thus included in the outcome category "target disease absent." After dichotomization of the diagnostic outcome, we may develop diagnostic prediction rules with logistic regression analysis. However, considering only the target disease is a simplification of clinical practice.

Table 4.4 Characteristics of some statistical models for binary outcomes

Categories	Interactions	Linearity	Selection	Estimation
Linear logistic regression	Possible	Flexible	Flexible	Standard ML or penalization
Idiot's Bayes	No	Often categories for diagnostic outcome	Flexible	Univariate effects (+ calibration slope)
GAM	Possible	Highly flexible	Flexible	Nonparametric, close to penalized ML
GNLM, neural net	Assumed	Highly flexible	Flexible	Backpropagation, early stopping to prevent overfitting
Trees	Assumed	Categorization	Assumed	Various splitting methods

### 4.3.1 Polytomous Logistic Regression

Several studies discussed the use of polytomous logistic regression to accommodate simultaneous prediction of three or more unordered outcome categories.<sup>29,484</sup> The model for  $j$  outcome categories can be written as:

$$\text{Logodds}(y=j \text{ vs. } y=\text{reference}) = \alpha_j + \beta_{ij} \times x_{ij} = \text{lp}_j$$

where  $j - 1$  models are fitted each with separate sets of intercept  $\alpha_j$  and regression coefficients  $\beta_i$ . We illustrate the polytomous model for prediction of three diagnostic outcome categories in a detailed case study.

#### \*4.3.2 Example: Histology of Residual Masses

After chemotherapy, patients with nonseminomatous testicular germ cell tumor may have residual masses of metastases.<sup>425</sup> These residual masses may contain benign tissue, mature teratoma, or cancer cells. Surgery is not necessary for benign tissue. Mature teratoma can grow and hence cause problems during follow-up. The most serious diagnosis is residual cancer, where a direct benefit from surgery is plausible.

We consider three outcome categories with varying therapeutic benefit: no benefit for benign tissue, some for teratoma, and most benefit for surgical removal of residual cancer.<sup>35</sup> We have proposed to weigh the benefit as 1:3:8 based on expert estimates of the prognosis of unresected vs. resected masses.<sup>422</sup> This ordering in severity of the outcome was not used in the modeling, since biological knowledge was available that implied that prognostic relationships would be very different for the different histologies. For example, some histologies are known to produce certain tumor markers while others do not. Masses with teratoma masses are not expected to decrease substantially in size by chemotherapy, while cancer is usually responsive. Hence, a substantial decrease would make residual cancer unlikely.

Polytomous logistic regression analysis requires that one of the outcome categories is chosen as reference category. For the other outcome categories the polytomous logistic regression analysis fits simultaneously submodels that compare the outcome categories with the chosen reference. Thus, for each outcome category, different regression coefficients are estimated for each predictor. These submodels together comprise the polytomous model and can be used to estimate the probability of presence of each diagnostic outcome. In our example study, the reference diagnosis was viable cancer. Hence, we fitted a polytomous regression model, consisting of two submodels, one for benign tissue compared to viable cancer, and one for mature teratoma compared to viable cancer. These models take a similar form as the binary logistic model:

$$\begin{aligned} \text{Logit(benign vs. cancer)} &= \alpha_b + \beta_{1,b} \times x_1 + \beta_{2,b} \times x_2 + \dots + \beta_{p,b} \times x_p = \beta_{i,b} \times X = \text{lp}_b; \\ \text{Logit(teratoma vs. cancer)} &= \alpha_t + \beta_{1,t} \times x_1 + \beta_{2,t} \times x_2 + \dots + \beta_{p,t} \times x_p = \beta_{i,t} \times X = \text{lp}_t. \end{aligned}$$

The subscript  $b$  indicates that we predict the odds of benign tissue, and subscript  $t$  for teratoma with  $p$  predictors.

The interpretation of the regression coefficients is similar as for dichotomous logistic regression, i.e., the logodds of the outcome (benign tissue or mature teratoma) relative to cancer per unit change in the predictor values. The probabilities of benign and teratoma tissue can be calculated by:

$$P(\text{benign tissue}) = \exp(lp_b) / [1 + \exp(lp_b) + \exp(lp_t)]$$

$$P(\text{mature teratoma}) = \exp(lp_t) / [1 + \exp(lp_b) + \exp(lp_t)].$$

As probabilities need to sum to 1, the probability of cancer can then be calculated by:

$$P(\text{cancer}) = 1 - P(\text{benign tissue}) - P(\text{mature teratoma}).$$

We fitted a multivariable polytomous logistic regression model with six predictors to enable estimation of the probabilities of benign tissue, mature teratoma, and viable cancer. Variable selection was not applied; we simply included all six of the available predictors.

### \*4.3.3 Alternative Models

For comparison reasons, we may fit consecutive multivariable dichotomous logistic models. In our example, we make one model to predict benign tissue (vs. mature teratoma or viable cancer). The second, consecutive, model aimed to predict the odds of mature teratoma vs. viable cancer in patients who did not have benign tissue.

$$\text{Logit}(\text{benign vs. teratoma/cancer}) = \alpha_b + \beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_p \times x_p = \beta_i \times X = lp;$$

$$\text{Logit}(\text{teratoma vs. cancer}) = \alpha_t + \beta_{1,t} \times x_1 + \beta_{2,t} \times x_2 + \dots + \beta_{p,t} \times x_p = \beta_{i,t} \times X = lp_t.$$

The latter formula is identical to a previous formula for the polytomous model, but the coefficients are estimated differently. In the polytomous model, all coefficients are estimated jointly. In the consecutive logistic model, a selection of patients is made to estimate the coefficients.

With these two binary logistic models the diagnostic probabilities are calculated by:

$$P(\text{benign tissue}) = \exp(lp) / (1 + \exp(lp))$$

$$P(\text{mature teratoma}) = (1 - P(\text{benign tissue})) \times \exp(lp_t) / [1 + \exp(lp_t)]$$

$$P(\text{cancer}) = 1 - P(\text{benign tissue}) - P(\text{mature teratoma})$$

In our example, we use the same six predictors, but in principle we could select different predictors for  $lp$  and  $lp_t$ . Also, we could have considered different transformations of the continuous predictors related to LDH and mass size.

In both approaches, 14 parameters were estimated: 2 intercepts ( $\alpha$ ) and 2 sets of 6 regression coefficients ( $\beta_{i,t}$ ). The performance of the two approaches was very similar according to discrimination (area under ROC curve) and  $R^2$  measures. See Biesheuvel et al. for a more detailed description of this case study.<sup>35</sup> Further discussion of approaches to unordered outcomes is provided in other reports.<sup>352, 425</sup>

### \*4.3.4 Comparison of Modelling Approaches

We considered a total of 1,094 patients, where 425 (39%) had benign tissue, 535 (49%) mature teratoma, and 134 (12%) viable cancer. Table 4.5 shows the distributions of the six predictors across the three diagnostic outcome categories and in the total study population.

The odds ratios for the predictors are shown in Table 4.6, considering a polytomous regression model, and a consecutive logistic model. We note that the odds ratios for teratoma vs. cancer differ slightly between these modeling approaches. The odds ratios for necrosis vs. cancer are larger for most predictors than for necrosis vs. other histology.

## 4.4 Ordinal Outcomes

Ordinal outcomes are quite common in medical and epidemiological studies. Often, such scales are either simplified to binary outcomes, or treated as continuous outcomes. As an example, we consider the Glasgow Outcome Scale (GOS).<sup>430</sup> This scale has five levels (Table 4.7).

This scale has often been dichotomized as mortality vs. survival, or an unfavorable (GOS 1, 2 or 3) vs. favorable (GOS 4 or 5) outcome. However, we can also explore the use of the full GOS. A practical consideration is that the GOS 2 category is very small, and that some may debate whether vegetative state is better than death. Therefore we combine the GOS categories 1 and 2, such that an outcome with four ordered levels is formed.

Table 4.5 Distribution of predictors across outcome categories in the total study population ( $n = 1,094$ )

	Benign	Mature teratoma	Viable cancer	Total
	N (%)	N (%)	N (%)	N (%)
<i>Predictors</i>				
No teratoma in primary tumor	279 (55)	170 (34)	54 (11)	503 (46)
Normal AFP level	200 (59)	112 (33)	27 (8)	339 (31)
Normal HCG level	184 (49)	154 (41)	40 (10)	378 (35)
Standardized value of LDH*	1.5 (0.39–70)	1.2 (0.12–21)	1.8 (0.34–64)	1.4 (0.12–70)
Postchemotherapy size (mm)*	18 (2–300)	30 (2–300)	40 (2–300)	28 (2–300)
Reduction in size (%)*	60 (−150–100)	20 (−150–100)	43 (−250–100)	43 (−250–100)
<i>Outcome</i>				
Histology at resection	425 (39)	535 (49)	134 (12)	1,094 (100)

\* Median (range)

AFP Alpha-fetoprotein, HCG Human chorionic gonadotropin, LDH Lactate dehydrogenase

Table 4.6 Results of the multivariable polytomous and consecutive dichotomous logistic regression analysis. Values represent odds ratios with 95% confidence intervals

Predictor	Polytomous regression		Consecutive dichotomous regression	
	Benign vs. cancer	Teratoma vs. cancer	Benign vs. other	Teratoma vs. cancer
No teratoma in primary tumour	2.2 (1.4–3.3)	0.66 (0.44–0.99)	3.0 (2.2–4.0)	0.61 (0.40–0.92)
Normal AFP serum level	2.8 (1.7–4.6)	0.94 (0.57–1.5)	2.9 (2.1–4.0)	0.90 (0.54–1.5)
Normal HCG serum level	1.4 (0.89–2.3)	0.72 (0.46–1.1)	1.9 (1.3–2.6)	0.70 (0.44–1.1)
Log of standardized value of LDH	1.2 (0.84–1.6)	0.58 (0.42–0.78)	1.7 (1.4–2.2)	0.60 (0.44–0.81)
Square root of postchemotherapy mass size	0.79 (0.71–0.88)	0.91 (0.84–0.99)	0.85 (0.77–0.92)	0.89 (0.82–0.98)
Reduction in mass size (per 10%)	1.14 (1.06–1.22)	0.97 (0.92–1.02)	1.18 (1.12–1.24)	0.96 (0.92–1.0)

Table 4.7 Definition of the Glasgow Outcome Scale

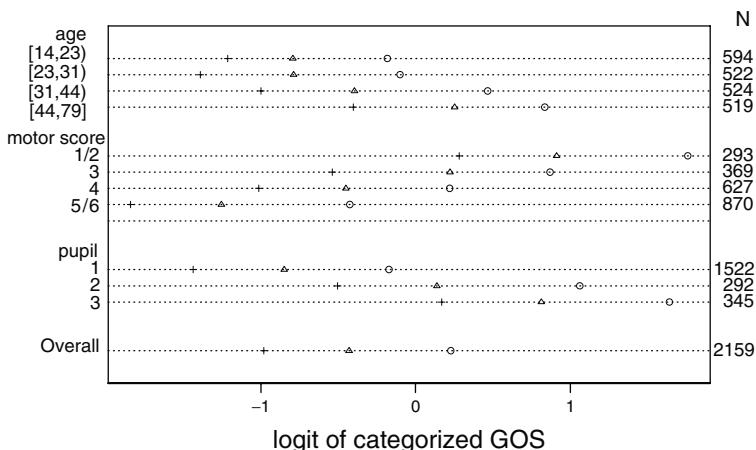
Category	Label	Definition
1	Dead	—
2	Vegetative	Unable to interact with environment; unresponsive
3	Severe disability	Conscious but dependent
4	Moderate disability	Independent, but disabled
5	Good recovery	Return to normal occupational and social activities; may have minor residual deficits

#### 4.4.1 Proportional Odds Logistic Regression

A standard logistic regression model can be used for each of the three possible dichotomous categorizations of the GOS: 12 (dead/vegetative) vs. 345, 123 vs. 45 (favorable), 1234 vs. 5 (good recovery). A straightforward extension of the logistic model is the proportional odds logistic model. Here, a common set of regression coefficients is assumed across all levels of the outcome, and intercepts are estimated for each level. So, in our example we have three intercepts  $\alpha$ , but only one set of  $\beta$ , instead of three sets of  $\beta$  coefficients when fitting a polytomous logistic model. The common set of  $\beta$  coefficients can be thought of as an average over the three separate sets of  $\beta$ s estimated at each possible dichotomization. As an example we consider a simple model with age, motor score, and pupillary reactivity in a model to predict 6-month outcome in data from two RTCs in traumatic brain injury.<sup>203</sup>

An advantage of the proportional odds model is its parsimony in dealing with an ordered outcome. The price we pay is the assumption of proportionality of the

odds. This assumption is equivalent to saying that any cut-point on the outcome scale would lead to the same logistic regression coefficient. The model further has very similar assumptions as the usual logistic model. We can graphically check the proportionality assumption in univariate analyses for each predictor (Fig. 4.12). Distances between points should be identical on the logit scale within each category of a predictor (looking horizontally), or equivalently, the effects of predictors should be the same for every point (looking vertically). The assumption of proportional odds can formally be assessed with a score test. One could also develop usual logistic models by each categorization, and check for systematic trends in the estimated odds ratios (Table 4.8). There is considerable overlap in patients in such evaluations, but clear deviations from proportional odds should become visible. In



**Fig. 4.12** Assessment of the proportional odds assumption for each of three predictors (univariate analysis) to predict for GOS at 6 months after traumatic brain injury. Data from the Tirilazad trials ( $n=2,159$ ). The circle, triangle, and plus sign correspond to the GOS categorizations 12 vs. 345, 123 vs. 45, and 1234 vs. 5. For example, the overall logit of the last categorization is  $-1$ , or a probability of 27% (589/2,159 patients). The proportional odds assumption is well satisfied, since the horizontal distance between the points is constant within each category

Table 4.8 Logistic and proportional odds models for GOS at 6 months after traumatic brain injury in 2,159 patients from the Tirilazad trials<sup>203</sup>

Categorization	12 vs. 345	123 vs. 45	1234 vs. 5	Proportional
Age (per decade)	1.36	1.47	1.45	1.43
Motor 1/2	5.88	6.50	6.18	5.86
3	2.98	3.82	3.00	3.15
4	1.95	1.95	1.62	1.82
5/6	1	1	1	1
Pupils 2 reactive	1	1	1	1
1 reactive	1.73	1.81	2.51	2.01
Nonreactive	3.26	3.53	4.23	3.55

in our example, the ORs per categorization are reasonably constant, and the proportional odds ratio provides a nice summary measure over the three categorizations.

### \*4.4.2 Alternative: Continuation Ratio Model

An alternative to the proportional odds model is the continuation ratio model. This model is related to the Cox proportional hazards model and allows predictors to have different effects on different levels of the ordinal outcome. An extensive illustration is provided by Harrell et al.<sup>174,178</sup>

## 4.5 Survival Outcomes

Survival analysis is appropriate for outcomes that occur during follow-up of patients. The outcome may for example be death or another event, such as recurrence of disease in cancer, or a complication after implantation after a heart valve. A key characteristic of survival data is that the follow-up of patients is typically incomplete. For example, some patients may have been followed for 1 year, others for 3 years, etc., while we may be interested in estimates of 5-year survival. Patients with such incomplete data are called censored observations. Because of censoring, logistic regression for the outcome (a binary variable) is inappropriate. One could think of linear regression on the survival time (a continuous outcome), but again censoring makes such an analysis usually meaningless.

### 4.5.1 Cox Proportional Hazards Regression

In medical and epidemiological studies, the Cox proportional hazard model is the most often used method for survival outcomes.<sup>85</sup> It is the natural extension of the logistic model to the survival setting. Indeed, the Cox model is equivalent to conditional logistic regression, with conditioning at times where events occur.<sup>251</sup> In the logistic model, we use an intercept in the linear predictor, while in the Cox model a baseline hazard function is used. The hazard function indicates the risk of the outcome during follow-up. The baseline hazard is nonparametric in the Cox model. As for the logistic model, simpler and more extensive methods exist, which can be seen as special cases or extensions of the Cox model.

The Cox regression model is often stated as a function of the hazard function<sup>472</sup>:

$$\lambda(t|X) = \lambda(t) e^{\beta X},$$

Where  $\lambda(t)$  is the hazard at time  $t$ , and is usually estimated at the mean values of the predictors and  $\beta X$  is the linear predictor,  $\beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_p \times x_p$ .

The linear predictor is usually centered at the mean values of the predictors, and  $e^{\beta X}$  then indicates the hazard ratio compared to the average risk profile. Note that the linear predictor relates to the log of the hazard:

$$\log(\lambda(t|X)) = \log(\lambda(t)) + \beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_p \times x_p.$$

The Cox regression model is semiparametric. It makes a parametric assumption on the effect of predictors, i.e., proportionality of effect during follow-up. The baseline hazard function  $\lambda(t)$  is nonparametric. This is an advantage of the model, especially when we focus on the effect of predictors. Regression coefficients  $\beta_i$  can readily be estimated. The quantity  $e^{\beta_i}$  is the hazard ratio, similar to the odds ratio in logistic regression.

### 4.5.2 Predicting with Cox

When we want to make predictions, we need to consider the risk over time, for example by using the cumulative hazard, or survival function. The standard formulation of the predicted survival at time  $t$ , given a set of predictors  $X$ , is as

$$S(t|X) = S(t)^{e(\beta X)},$$

Where  $S(t|X)$  denotes the predicted survival at time  $t$ , given a set of predictors  $X$ ,  $S(t)$  is the baseline survival, usually estimated at the mean values of the predictors, and  $\beta X$  is the linear predictor.

The baseline survival is estimated from the nonparametric baseline hazard function as

$$S(t) = e^{-\Lambda(t)}$$

where  $\Lambda(t)$  is the cumulative hazard at time  $t$ .

Note that  $\log(\Lambda(t))$  can range between  $[-\infty, +\infty]$ ;  $\Lambda(t) [0, \infty]$ ;  $S(t) [1, 0]$ . This is very similar to the behavior of quantities in logistic regression: logit, odds, and probability. The baseline survival in the development data determines the precise time points where we can make predictions for, which is not very natural for application of the model in new subjects.

### 4.5.3 Proportionality Assumption

The effect of predictors is assumed to be constant in time or more precisely stated: the hazards are assumed to be proportional. The proportionality assumption can be assessed in a number of ways, including graphical and analytical methods. A general approach is to calculate interval specific hazard ratios. With proportional hazards, the hazard ratio should be similar across any interval considered. Follow-up time can also be considered as a continuous variable, where assessing interaction with  $\log(\text{time})$  may be a useful approach.<sup>174</sup>

If we find that the effect of a predictor is nonproportional, we can stratify for categorical variables in the baseline hazard. For example, we could estimate baseline hazards for males and females separately. For continuous predictors, e.g., age, we could specify interactions with  $\log(\text{age})$  as the time variable. Nonproportionality can also be visualized in a more nonparametric approach, i.e., with Kaplan–Meier curves.

#### **4.5.4 Kaplan–Meier Analysis**

Kaplan–Meier analysis is a nonparametric approach to survival outcomes.<sup>224</sup> It adequately deals with censored data, and provides attractive graphs on the relationship between predictor values and the outcome over time. The method can be seen as an extension of a cross-table for survival data. More technically, it can be interpreted as a Cox model with stratification of the baseline hazard to all predictor levels. For example, we could make a Cox model with sex as a stratification variable for the baseline hazard, without any other variables, which is equivalent to a Kaplan–Meier analysis with sex as a predictor. Also, testing in a Kaplan–Meier analysis is usually done with a log-rank test, which is equivalent to the Score test in the Cox model.

Kaplan–Meier analysis often has a role in prognostic modeling at the start of the analysis, i.e., to show univariate relationships graphically or to compute survival fractions at a certain time of follow-up. Also at the end of a modeling process, Kaplan–Meier curves are often used to present the predictions from the model. It is then necessary to group patients by their predictions, since Kaplan–Meier analysis cannot handle continuous predictors. Kaplan–Meier curves are for survival analysis what cross-tables are for binary or categorical outcomes.

#### **\*4.5.5 Example: NFI After Treatment of Leprosy**

Nerve-function impairment (NFI) commonly occurs during or after chemotherapy in leprosy. It is the key pathological process leading to disability and handicap. A simple clinical prediction rule was developed with 2,510 patients who were followed-up for 2 years in Bangladesh.<sup>87</sup> In total, 166 patients developed NFI (Kaplan–Meier 2-year estimate: 7.0% [95%CI 6.0–8.0%]. A Cox regression model included two strong predictors (Table 4.10). Patients with no, one, or two unfavorable

Table 4.10 Multivariable hazard ratios from Cox proportional hazard analysis.<sup>87</sup> Three risk groups could be formed based on presence of no, one, or two unfavorable predictive characteristics, since the hazard ratios were very similar

Predictor	Hazard ratio [95% CI]
Leprosy group (MB vs. PB)	7.5 (5.3–11.0)
Nerve-function loss at registration	8.1 (5.7–12.0)

ble predictive characteristics had 1.3% (95% CI 0.8–1.8%), 16.0% (12–20%), and 65% (56–73%) risks of developing NFI within 2 years of registration, respectively.

#### **4.5.6 Parametric Survival**

Whereas Kaplan–Meier analysis represents a more nonparametric approach, parametric survival models are less flexible than Cox regression in their dealing with the baseline hazard function. Parametric models typically assume proportionality of the predictor effects, but a more smoothed hazard in time. Examples of parametric models include the exponential model (or Poisson model, using a constant hazard) and the Weibull model (two parameters to let the hazard increase or decrease monotonically over time). The exponential and Weibull model can also be seen as examples of accelerated failure time (AFT) models. Here, the effects of predictors are not viewed as multiplicative on the hazards scale, but as multiplicative on the time axis (or additive at the log-time axis). Other examples of AFT models are the log-normal and log-logistic model.<sup>174,472</sup>

Regression coefficients in exponential or Weibull models are hazard ratios after exponentiating. In AFT models, they represent a change in the log-time. The advantage of parametric survival models is their concise, parsimonious formulation, and smoothing of the underlying hazard. This makes these models especially to be considered for prediction purposes. Extrapolation is readily possible with parametric models, but not with Cox or Kaplan–Meier analysis because of their nonparametric nature. Predictions at the end of the follow-up are quite unstable with Cox or Kaplan–Meier analysis, and more robust with parametric methods. For estimation of the effect of predictors, the Cox model is often more suitable, since this model is less restrictive than an exponential or Weibull model. However, log-logistic models have been useful in situations where predictors worked especially during an early, acute phase of the hazard, which would show as non-proportional hazards in a Cox model.<sup>174</sup> Note finally that some of the more flexible methods for binary data have also been extended to survival models, but are not commonly used yet (e.g., neural networks).<sup>181</sup>

#### **\*4.5.7 Example: Replacement of Risky Heart Valves**

In Chap. 2, we presented an overview of the decision dilemma on Björk-Shiley convexo-concave (BScc) mechanical heart valves.<sup>448</sup> Poisson regression models were constructed to estimate survival and the risk of strut fracture.<sup>415</sup> Poisson regression was especially useful to disentangle the effects of increasing age of the patient during follow-up from the increasing time since implantation of the valve during follow-up. The follow-up time was divided in yearly intervals, each with an age and time since implantation. Time since implantation started at zero, and increased

Table 4.11 Common statistical models for survival outcomes

Categories	Proportionality	Baseline hazard
Cox proportional hazards	Assumed	Nonparametric
Kaplan–Meier	No	Nonparametric
Exponential and Weibull	Assumed	Parametric
Log-normal, log-logistic	No, but multiplicative in time	Parametric

in steps of 1 year during follow-up. Age started at the age at implantation, and also increased in steps of 1 year during follow-up. The Poisson model could easily estimate the effects of both predictors, which would have been more complicated in a Cox regression analysis. Moreover, extrapolation to longer time since implantation was readily possible with the Poisson model.

#### 4.5.8 Summary on Survival Outcomes

In sum, the Cox regression model provides a default framework for prediction of long-term prognostic outcomes. Kaplan–Meier analysis provides a nonparametric method, but requires categorization of all predictors. It is the equivalent of cross-tables for categorical outcomes for a survival context. Parametric survival models may be useful for predictive purposes because of their parsimony and robustness, for example at the end of follow-up, or even beyond the observed follow-up.

## 4.6 Concluding Remarks

Regression models are available for several types of outcome that we may want to predict, such as continuous, binary, unordered categorical, ordered categorical, and survival outcomes. The corresponding default regression models are the linear, logistic, polytomous, proportional odds, and Cox regression models, respectively. Both more and less flexible methods are available. Flexible methods may fit particular patterns in the data better, but may on the other hand lead to overfitting (Chap. 5). It is therefore not immediately clear what kind of model is to be preferred in a specific prediction problem (Chap. 6).

Special types of data can be encountered that require specific types of analyses. Correlated outcome data may occur by the design of a study, for example by clustering per hospital. In survival analysis, repeated and correlated events may occur, asking for extensions of the Cox model. Also, we may want to consider competing risks in estimation of actual risk instead of actuarial risks.<sup>124,158,159</sup>

## Questions

### 4.1 Explained variation

- (a) What is the difference between explained variation in linear and logistic regression models?
- (b) Is the choice of scale for explained variation natural in linear and logistic regression models?
- (c) Why are larger likelihood ratios seen with an incidence of 50% compared to 1% in Fig. 4.5?

### 4.2 Categorical and ordinal outcomes

- (a) What is the proportionality assumption in the proportional odds model?
- (b) Mention at least two ways how the proportionality assumption can be checked
- (c) Would the proportionality assumption hold in the testicular cancer case study (Table 4.6)?
- (d) We could also make two logistic regression models for the testicular cancer case study, with one model for benign vs. other and another for cancer vs. other. What would be the problem with predictions from these models?

### 4.3 Parametric survival models

- (a) Why may we label the Cox regression model “semiparametric”?
- (b) Do you agree that Kaplan–Meier analysis is a fully nonparametric model?
- (c) Why is the Weibull model attractive for making long-term predictions? At what price?

# Chapter 5

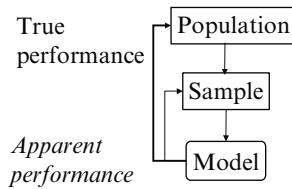
## Overfitting and Optimism in Prediction Models

**Background** If we develop a statistical model with the main aim of outcome prediction, we are primarily interested in the validity of the predictions for new subjects, outside the sample under study. A key threat to validity is overfitting, i.e. that the data under study are well described, but that predictions are not valid for new subjects. Overfitting causes optimism about a model's performance in new subjects. After introducing overfitting and optimism, we illustrate overfitting with a simple example of comparisons of mortality figures by hospital. After appreciating the natural variability of outcomes within a single centre, we turn to comparisons across centres. We find that we would exaggerate any true patterns of differences between centres, if we would use the observed average outcomes per centre as predictions of mortality.

A solution is presented, which is generally named “shrinkage.” Estimates per centre are drawn towards the average to improve the quality of predictions. We then turn to overfitting in regression models, and discuss the concepts of selection and estimation bias. Again, shrinkage is a solution, which now draws estimated regression coefficients to less extreme values. Bootstrap resampling is presented as a central technique to correct overfitting and quantify optimism in model performance.

### 5.1 Overfitting and Optimism

To derive a model, we use empirical data from a sample of subjects, drawn from a population (Fig. 5.1). The sample is considered to be drawn at random. The data from the sample are only of interest in that they represent an underlying population.<sup>13,409</sup> We use the empirical data to learn about patterns in the population, and to derive a model that can provide predictions for new subjects from this population. In learning from our data an important risk is that the data under study are well described, but that the predictions do not generalize to new subjects outside the sample. We may capitalize on specifics and idiosyncrasies of the sample. This is referred to as “overfitting.” In statistics, overfitting is sometimes defined as fitting a statistical model that has too many parameters, or as the “curse of dimensionality.”<sup>181</sup> For prediction models, we may define overfitting more precisely as fitting a statistical



**Fig. 5.1** Graphical illustration of optimism, which is defined as the difference between true performance and apparent performance. The apparent performance is determined on the sample where the model was derived from; true performance refers to the performance in the underlying population. The difference between apparent and true performance is defined as the optimism of a prediction model

model with too many degrees of freedom in the modelling process. Degrees of freedom are used by estimation of the coefficients in a regression model, but also by searching for the optimal model structure. The latter may include procedures to search for important predictors from a larger set of candidate predictors, optimal coding of predictors, and consideration of potential non-linear transformations.

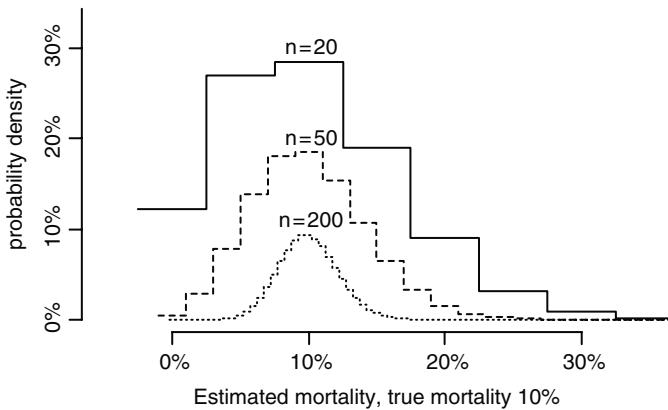
Overfitting leads to a too optimistic impression of model performance that may be achieved in new subjects from the underlying population. Optimism is defined as true performance minus apparent performance, where true performance refers to the underlying population, and apparent performance refers to the estimated performance in the sample (Fig. 5.2). Put simply: “what you see may not be what you get.”<sup>23</sup>

### 5.1.1 Example: Surgical Mortality in Oesophagectomy

Surgical resection of the oesophagus (oesophagectomy) may be performed for subjects with oesophageal cancer. It is among the surgical procedures that carry a substantial risk of 30-day mortality (see also Fig. 6.2).<sup>125,213</sup> Underlying differences in quality between hospitals may affect the 30-day mortality. A question is whether we can identify the better hospitals, and whether we can predict the mortality for a typical subject in a hospital.<sup>260</sup>

### 5.1.2 Variability within One Centre

We first illustrate the variability of mortality estimates within a single centre, according to different sample sizes. For oesophagectomy, we assume 10% as an average estimate of mortality among elderly patients, based on analyses of the



**Fig. 5.2** Estimated mortality in relation to sample size. When the true mortality is 10% in samples of  $n = 20$ , around 30% of these will contain two deaths (estimated mortality, 10%). With larger sample sizes, observed mortalities are more likely close to 10%

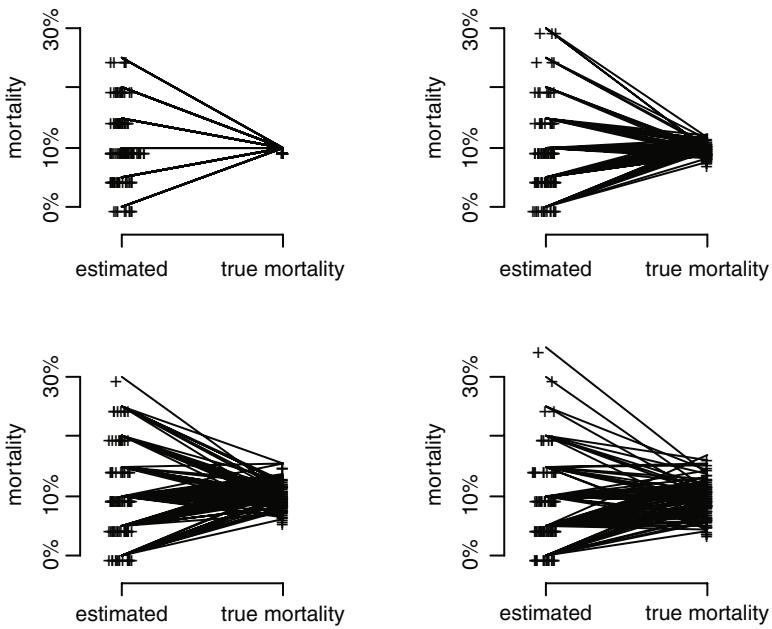
SEER-Medicare registry data, where mortality exceeded 10%: 221 of 2,031 subjects had died within 30 days after surgery, or 10.9% [95% CI, 9.6%–12.3%].<sup>423</sup>

For illustration, we assume that case-mix is irrelevant, i.e. that all patients have the same true mortality risks. The observed mortality rate in a centre may then be assumed to follow a binomial distribution (Fig. 5.2). When the true mortality is 10% in samples of  $n = 20$ , around 30% of these will contain two deaths (estimated mortality, 10%). With larger sample sizes, observed mortalities are more likely close to 10%; e.g. when  $n = 200$ , mortality is estimated between 8% and 12% in 71% of the samples.

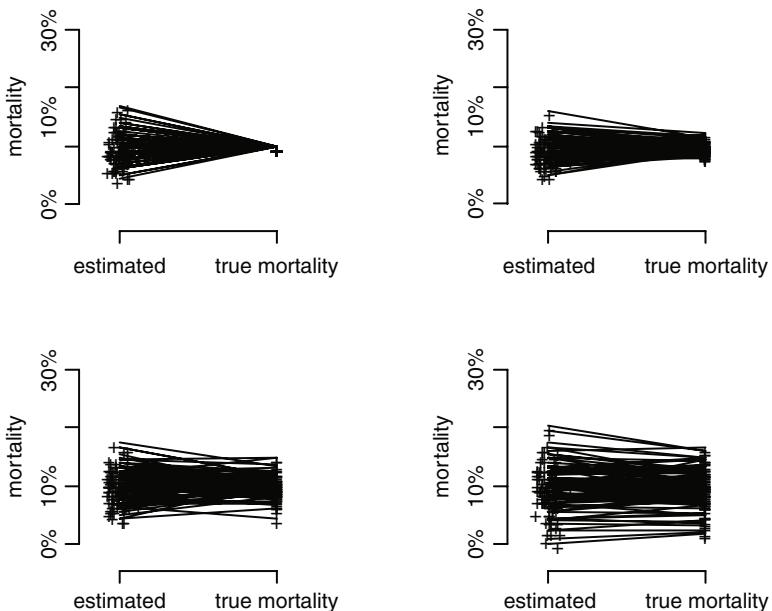
### 5.1.3 Variability between Centres: Noise vs. True Heterogeneity

We need to appreciate within centres variability when we want to make predictions of mortality by centre. For example, consider that 100 centres each reported mortality in 20 subjects, while the true mortality risk was 10% for every patient. On average two deaths are hence expected per centre (10% of 20). The expected distribution of the estimated mortality is as in Fig. 5.2: 12% of the centres will have 0% mortality, and 13% will report a 20% or higher mortality. An actual realization is shown in Fig. 5.3. A statistical test for differences between centres should be non-significant for most of such comparisons (for 95% of the cases when  $p < 0.05$  is used as criterion for statistical significance).

Of more interest is the situation that the true mortality varies by centre. This can be simulated with a heterogeneity parameter, often referred to as  $\tau$  (tau). Assuming a normal distribution for the differences across centres, we can write:



**Fig. 5.3** Estimated and true mortality for 100 centres that analyzed 20 subjects each, while the average mortality was 10% for all (*upper left panel*),  $10\% \pm 1\%$  (*upper right panel*),  $10\% \pm 2\%$  (*lower left panel*),  $10\% \pm 3\%$  (*lower right panel*)



**Fig. 5.4** Estimated and true mortality for 100 centres that had 200 subjects each, while the average was 10% for all (panel a),  $10\% \pm 1\%$  (panel b),  $10\% \pm 2\%$  (panel c), and  $10\% \pm 3\%$  (panel d)

true mortality  $\sim N(10\%, \text{sd} = \tau)$ . With  $\tau = 1\%$ , 95% of the centres have a mortality between 8% and 12%, while setting  $\tau$  to 2% and 2.5% implies that 95% of the centres have a mortality between 6% and 14%, and between 5% and 15%, respectively. This underlying heterogeneity causes the estimated mortality to have more variability than expected from the binomial distribution with a single true mortality of 10%. This is recognized in the distributions of Fig. 5.3. Differences between centres can be tested, and will be identified as significant depending on the magnitude of the heterogeneity ( $\tau$ ), and the sample size (number of centres, sample size per centre).

### 5.1.4 *Predicting Mortality by Centre: Shrinkage*

We recognize that the estimated mortalities are too extreme as predictions compared with the distribution of the true mortalities (Fig. 5.3). Predictions other than 10% are by definition too extreme when there is no heterogeneity. Too extreme predictions also occur when there is underlying variability across centres (e.g. true mortality between 6 and 14%). Per centre, the estimated mortality is an unbiased estimator of the true mortality in each centre. But the overall distribution of estimated mortality suffers from the low numbers per centre, which makes that chance severely affects our predictions.

The phenomenon in Fig. 5.3 is an example of regression to the mean.<sup>301</sup> It is a motivation for shrinkage of predictions to the average, a principle that is also important in more complex regression models.<sup>81,459</sup> We should shrink the individual centre's estimates towards the overall mean to make better predictions overall.

We can also say that predictions tend to be overfitted: They point at very low and very high risk hospitals, while the truth will be more in the middle. The identification of extreme hospitals will be unreliable with small sample size. With larger sample size, e.g. 200 subjects per centre, the overfitting problem is reduced (Fig. 5.4). Empirical Bayes and random effects methods have been proposed to make better predictions (see Chap. 21).<sup>22,458</sup>

## 5.2 Overfitting in Regression Models

### 5.2.1 *Model Uncertainty: Testimation*

Overfitting is a major problem in regression modelling. It arises from two main issues: model uncertainty and parameter uncertainty (Table 5.1). Model uncertainty is caused by specification of the structure of our model, such as which characteristics are included as predictors, or information of the data set under study. The model structure is therefore uncertain. This model uncertainty is

**Table 5.1** Causes and consequences of overfitting in prediction models

Issue	Characteristics
<i>Causes of overfitting</i>	
Model uncertainty	The structure of a model is not pre-defined, but determined by the data under study. Model uncertainty is an important cause of overfitting
Parameter uncertainty	The predictions from a model are too extreme because of uncertainty in the effects of each predictor (model parameters)
<i>Consequences of overfitting</i>	
Testimation bias	Overestimation of effects of predictors because of selection of effects that withstood a statistical test
Optimism	Decrease in model performance in new subjects compared with performance in the sample under study

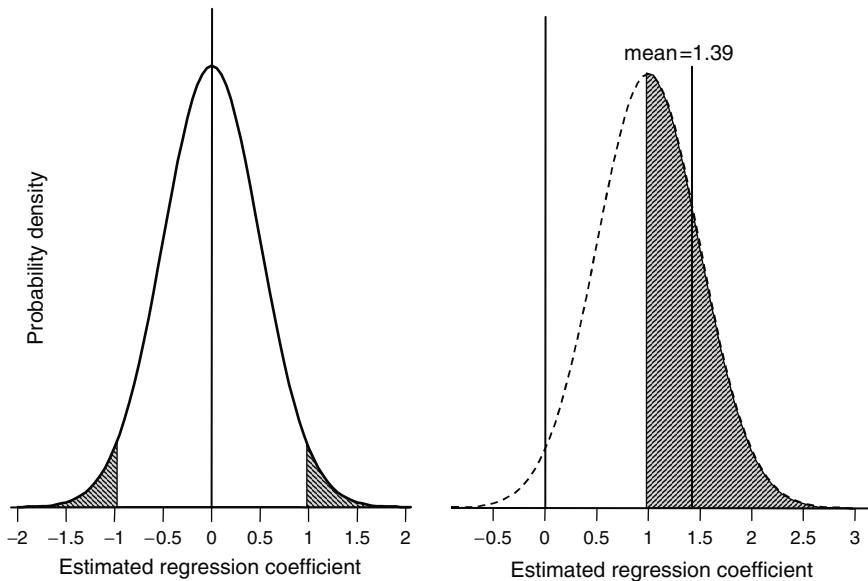
usually ignored in statistical analyses, which falsely assume that the model was pre-specified.<sup>69,101,194</sup>

The result of model uncertainty is selection bias.<sup>26,82,365,407</sup> Note that selection bias here refers to the bias caused by selection of predictors from a larger set of predictors, in contrast to the selection of subjects from an underlying population in standard epidemiological texts. Suppose that we investigate 20 potential predictors for inclusion in a prognostic model. If these are all noise variables, the true regression coefficients are zero. On average one variable will be statistically significant at the  $p<0.05$  level. The estimated effect will be relatively extreme, since otherwise the effect would not have been significant. If this one variable is included in the model, it will have a quite small or quite large effect (Fig. 5.5, left panel). On average the effect of such a noise variable is still zero.

If some of the 20 variables are true predictors, they will sometimes have a relatively small and sometimes a relatively large effect. If we only include a predictor when it has a relatively large effect in our model, we are overestimating the effect of such a predictor. This phenomenon is referred to as *testimation bias*: Because we test first, the effect estimate is biased.<sup>26,69</sup>

In the example of a predictor with true regression coefficient 1 and Standard Error (SE) 0.5, the effect will be statistically significant if estimated as lower than  $-1.96 \times SE = -0.98$ , or exceeding  $+1.96 \times SE = +0.98$  (52% of the estimated coefficients, Fig. 5.5, right panel). The average of the estimated coefficients in these 52% cases is 1.39 rather than 1. Hence, a bias of +39% occurs. In formal terms, we can state: if  $b$  is significant, then  $b=b$ , else  $b=0$ . Instead of considering the whole distribution of predictor effects, we only consider a selected part.

Testimation bias is a pervasive problem in medical statistics and predictive modelling.<sup>174</sup> The bias is large for relatively weak effects, as is common in medical research. Selection bias is not relevant if we have a huge sample size, or consider



**Fig. 5.5** Illustration of estimation bias. In case of a noise variable, the average of estimated regression coefficients is zero, and 2.5% of the coefficients is below  $-0.98 (1.96 \times \text{SE of } 0.5)$ , and 2.5% of the coefficients is larger than  $+0.98 (1.96 \times \text{SE of } 0.5)$ . In case of a true coefficient of 1, the estimated coefficients are statistically significant in 52%. For these cases, the average of estimated coefficients is 1.39 instead of 1

predictors with underlying large effects, since these predictors will anyway be selected for a prediction model. Neither does selection bias occur if we pre-specify the prediction model (“full model”).<sup>174</sup>

### 5.2.2 Other Biases

A well-known problem in prediction is bias by selection of an “optimal” cut-point for a continuous predictor.<sup>12,117,355</sup> A similar problem occurs if we examine different transformations for predictor variables as a check for linearity. For example, we may add a square term to a linear term, and omit the square term if not statistically significant.<sup>148</sup> More subtle variants occur when we less formally assess alternative model specifications. For example, we may consider different transformations of the outcome variable in a linear model, and visually judge the best transformation for use in further modelling. Or we examine different coding variants of a categorical predictor, with merging of groups with what we consider to have “similar outcomes.” These issues are discussed in more detail in Chap. 9 and 10 on coding of predictors, and Chap. 11 and 12 on selection of predictors.

### 5.2.3 *Overfitting by Parameter Uncertainty*

It appears that even when the structure of our model is fully pre-specified, predictions are too extreme when multiple predictors are considered. This is because parameters, such as regression coefficients, are estimated in the model with uncertainty. This surprising finding has been the topic of much theoretical research.<sup>81,459</sup> An intuitive explanation is related to how we create a linear predictor in regression models. Hereto, the regression coefficients of multiple predictors are multiplied with the predictor values. With default estimation methods (e.g. least squares for linear regression and maximum likelihood for logistic or Cox regression), each of the coefficients is estimated in a (nearly) unbiased way. But each coefficient is associated with uncertainty, as reflected in the estimated standard error and 95% confidence interval (CI). This uncertainty tends us to overestimate predictions at the extremes of a linear predictor, i.e. low predictions will on average be too low, and high predictions will on average be too large. This is an example of regression to the mean. We can shrink coefficients towards zero to prevent this overfitting problem.<sup>81,174,459</sup>

This phenomenon is related to “Stein’s paradox”: biased estimates rather than unbiased estimates are preferable in multivariable situations to make better predictions.<sup>107,398</sup> Shrinkage introduces bias in the multivariable regression coefficients, but if we shrink properly the gain in precision of our predictions more than offsets the bias. The issue of bias–variance trade-off is central in prediction modelling,<sup>181</sup> and will be referred to throughout this book. Estimation with shrinkage methods is discussed in more detail in Chap. 13.

### 5.2.4 *Optimism in Model Performance*

Overfitting can visually be appreciated from the distributions of estimated mortality as in Figs. 5.3 and 5.4, but also from model performance measures. For example we may calculate Nagelkerke’s  $R^2$  for a logistic model that includes 20 centres (coded as a factor variable, with 19 dummy variables indicating the effect of 19 centres against a reference hospital). If the true mortality in all hospitals was 10%, the estimated  $R^2$  was 9.4% when each hospital contained 20 subjects (Table 5.2). In fact,  $R^2$  was 0%, since no true differences between centres were present. The estimated 9.4% is based on pure noise. We refer to the difference between 9.4% and 0% as the optimism in the apparent performance (Fig. 5.1). With larger sample sizes, the optimism decreases, e.g. to 0.1% for 20 centres with 2,000 subjects each (total 40,000 subjects, 4,000 deaths on average). Statistical testing of the between centre differences was by definition not significant in 95% of the simulations. We might require statistical significance of this overall test before trying to interpret between centre differences.

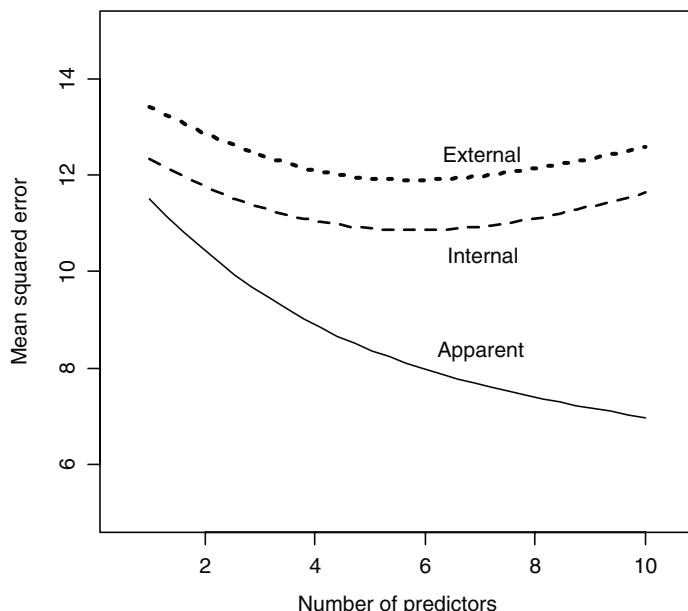
When true differences between centres were present (e.g. a range of 6–14% mortality,  $\tau = 2\%$ ), the true  $R^2$  was close to 1% ( $n = 2,000$ ). With small sizes per centre, the estimated  $R^2$  was 10.1%, which is again severely optimistic (Table 5.2).

A well-known presentation of optimism is to visualize the trade-off between model complexity and model performance.<sup>181</sup> We illustrate this trade-off in Fig. 5.6,

**Table 5.2**  $R^2$  for a logistic model predicting mortality in 20 centres. True mortality was 10% in the first series of simulations, and  $R^2$  reflects pure noise. True mortality varied between 6% and 14% ( $\tau = 2\%$ ) in the second series of simulations

True mortality	Sample size	$R^2_{app}$	$R^2_{adj}$	$R^2_{bootstrap}$
10%	$20 \times n = 20$	9.4	-0.1	NA
	$20 \times n = 200$	1.0	0	-0.5
	$20 \times n = 2,000$	0.1	0	0
$10\% \pm 2\%$	$20 \times n = 20$	10.1	0.3	NA
	$20 \times n = 200$	1.9	0.9	0.3
	$20 \times n = 2,000$	1.0	0.9	0.8

Nagelkerke's  $R^2$  calculated in logistic regression models,<sup>309</sup> averaged over 500 repetitions.  $R^2_{app}$ ,  $R^2_{adj}$ ,  $R^2_{bootstrap}$  refer to the apparent, adjusted and bootstrap-corrected estimates of  $R^2$ . The  $R^2_{adj}$  included "LR - df" instead of "LR" in the formula. Note that not all coefficients could directly be estimated, since some hospitals had 0% estimated mortality with  $n = 20$ ; for these we used 1% as the estimated mortality (adding one subject as dead, with a weight of  $1\% \times 20 = 0.2$ ). Bootstrapping with these weighted samples was not readily possible.



**Fig. 5.6** Mean squared error of predictions from models with increasing complexity (1,000 simulated samples with  $n = 50$ ). Apparent performance improves with more predictors, but internal and external performances worsen with more than five predictors

where we considered a simple linear regression model with 1 to 10 predictors. The model performance is evaluated by the mean squared error (mean  $(y - \hat{y})^2$ ) for the underlying population (internal validation), and for a population where the true regression

coefficients were slightly different (external validation). With 50 subjects per sample for estimation of the model (1,000 simulations), we note that the apparent error decreases with more predictors considered. But the internal and external performances do not improve after approximately five predictors are included. Overfitting occurs after approximately five predictors, and optimism increases from modest for one predictor to substantial for models with ten predictors.

### \*5.2.5 *Optimism-Corrected Performance*

In linear regression analysis, an adjusted version of  $R^2$  is available, which compensates for the degrees of freedom used in estimation of a model. Such an adjusted version can also be considered for Nagelkerke's  $R^2$ , which we consider e.g. for logistic and Cox models. We could subtract the degrees of freedom used to estimate the LR of the model in the calculation:

$$R^2_{\text{adjusted}} = (1 - e^{-(LR-df)/n}) / (1 - e^{(-2LL0)/n}),$$

where LR refers to the difference in  $-2 \log \text{likelihood}$  ( $-2LL$ ) of the model with and without the predictor,  $df$  are the degrees of freedom of the predictors in the model,  $N$  is the sample size, and  $LL0$  is the log likelihood of the Null model (without predictors).

This adjusted version is not standard in most current software however. When we apply this formula for the simulated centre outcome as shown in Figs. 5.3 and 5.4, the average adjusted  $R^2$  for noise differences is 0, with approximately half of the adjusted  $R^2$  values being negative (Table 5.2). The adjustment made the  $R^2$  estimates a bit conservative for small samples. For example, when true differences existed, the adjusted  $R^2$  was 0.3% rather than 0.9% (Table 5.2).

A more general optimism correction is possible with bootstrapping, which is explained in the next section. In Table 5.2, bootstrap-corrected performance was more conservative than the adjusted  $R^2$  formula, which may be caused by a not fully normal distribution of the optimism in  $R^2$ .<sup>401</sup>

## 5.3 Bootstrap Resampling

Bootstrapping alludes to a German legend about Baron Münchhausen, who was able to lift himself out of a swamp by pulling himself up by his own hair. In later versions of the legend he was using his own bootstraps to pull himself out of the sea, which gave rise to the term *bootstrapping*. A bootstrap was a loop of leather sewn onto the back of each boot to hold onto when pulling boots onto one's feet. In statistics, bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample.<sup>486</sup>

Bootstrapping mimics the process of sampling from the underlying population. Since we only have a sample from the population, this sampling is not truly

**Table 5.3** Illustration of five bootstrap samples drawn with replacement from five ages

Original sample	Bootstrap samples
20, 25, 30, 32, 35	20, 20, 30, 32, 35
	20, 25, 25, 30, 35
	20, 25, 30, 30, 32
	25, 32, 35, 35, 35
	30, 30, 32, 35, 35
	...

For easier interpretation, values were sorted per sample

possible, similar to the legend about Baron Münchhausen. Bootstrap samples are drawn with replacement from the original sample to introduce a random element. The bootstrap samples are of the same size as the original sample, which is important for the precision of estimates in each bootstrap sample.

For example, the GUSTO-I subsample 5 includes 429 subjects (Chap. 24). When we draw bootstrap samples, these each contain 429 subjects, but some subjects may not be included, others once, others twice, others three times, etc. On average, a subject has 63.2% chance of being at least once selected for a bootstrap sample.<sup>108</sup> For illustration we consider the simple case of the age of five subjects who are 20-, 25-, 30-, 32-, and 35-years old. Bootstrap samples might look like these in Table 5.3.

### 5.3.1 Applications of the Bootstrap

Bootstrapping is a widely applicable, non-parametric method. It can provide valuable insight in the empirical distribution of a summary measure from a sample. Bootstrap samples are repeatedly drawn from the data set under study, and each analyzed as if they were an original sample.<sup>108</sup>

For some measures, such as the mean of a population, we can use a statistical formula for the standard deviation ( $SD = \sqrt{var} = \sqrt{[(x_i - \text{mean}(x))^2 / (n - 1)]}$ ). We can use the SD to calculate 95% CI as  $\pm 1.96 \times SE$  or  $\pm 1.96 \times SD/\sqrt{n}$ . The bootstrap can be used to calculate the SE for any measure. For the mean, the bootstrap will usually result in a similar SE and 95% CI estimates as obtained from the standard formula. For other quantities, such as the median, no SE or 95% CI can be calculated with standard formulas, but the bootstrap can. See Harrell for an extensive illustration.<sup>174</sup>

### 5.3.2 Bootstrapping for Regression Coefficients

The bootstrap can assist in estimating distributions of regression coefficients, such as standard errors and CIs. The bootstrap can be useful in estimating distributions of related measures such as the difference between an adjusted and an unadjusted regression coefficient.<sup>472</sup> In the latter case, two regression coefficients would be estimated in each bootstrap sample. The difference would be

calculated in each sample, and the distribution over bootstrap samples would be interpreted as the sampling distribution. CIs can subsequently be calculated with three methods:

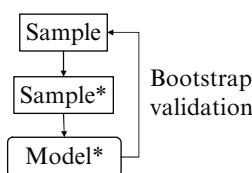
1. Normal approximation: The mean and SE are estimated from the distribution (note: the SD over bootstraps is the SE of the mean).
2. Percentile method: Quantiles are simply read from the empirical distribution. For example, 95% CIs are based on the 2.5% and 97.5% percentile, e.g. the 50th and 1,950th bootstrap estimate out of 2,000 replications.
3. Bias-corrected percentile method: Bias in estimation of the distribution is accounted for, based on the difference between the median of the bootstrap estimates and the sample estimate (“BCa”).<sup>108</sup>

For reliable estimation of distributions, large numbers of replications are advisable, e.g. at least 2,000 for method 2 and 3. Empirical  $p$  values can similarly be based on bootstrap distributions, e.g. by counting the number of estimates smaller than zero for a sample estimate larger than zero (giving a one-sided empirical  $p$  value).<sup>108</sup>

### 5.3.3 Bootstrapping for Optimism Correction

A very important application of bootstrapping is in quantifying the optimism of a prediction model.<sup>69,108,174,459</sup> With a simple bootstrap variant, one repeatedly fits a model in bootstrap samples, and evaluates the performance in the original sample (Fig. 5.7).

The average performance of the bootstrap models in the original sample can be used as the estimate of future performance in new subjects. A more accurate estimate is however obtained in a slightly more complicated way.<sup>108</sup> The bootstrap is used to estimate the optimism: The decrease between performance in the bootstrap sample (Sample\*) Fig. 5.7) and performance in the original sample. This optimism is subsequently subtracted from the original estimate to obtain an “optimism-corrected” performance estimate.<sup>174</sup>



**Fig. 5.7** Schematic representation of bootstrap validation for optimism correction of a prediction model. Sample\* refers to the bootstrap sample that is drawn with replacement from the Sample (the original sample from an underlying population). Model\* refers to the model constructed in Sample\*.

### \*5.3.4 Calculation of Optimism-Corrected Performance

Optimism-corrected performance is calculated as

Optimism-corrected performance = Apparent performance in sample – Optimism,  
where

Optimism = Bootstrap performance – Test performance.

The exact steps are as follows:

1. Construct a model in the original sample; determine the apparent performance on the data from the sample used to construct the model;
2. Draw a bootstrap sample (Sample\*) with replacement from the original sample (Sample, Fig. 5.7);
3. Construct a model (Model\*) in Sample\*, replaying every step that was done in the original Sample, especially model specification steps such as selection of predictors from a larger set of candidate predictors. Determine the bootstrap performance as the apparent performance of Model\* on Sample\*;
4. Apply Model\* to the original Sample without any modification to determine the test performance;
5. Calculate the optimism as the difference between bootstrap performance and test performance;
6. Repeat steps 1–4 many times, at least 100, to obtain a stable estimate of the optimism;
7. Subtract the optimism estimate (step 5) from the apparent performance (step 1) to obtain the optimism-corrected performance estimate.

Note that the original sample is used for testing of Model\*, while it contains largely the same subjects as the bootstrap sample (Sample\*). Although this may seem invalid, both theoretical and empirical research supports this process. Alternative bootstrap validation procedures have been proposed.<sup>1</sup> Appealing variants are the .632 and .632+ methods, where the testing of the models from the bootstrap sample is on subjects from the original sample who were not included in the bootstrap sample.<sup>109</sup> On average, 63.2% of the subjects are selected in a bootstrap sample, giving the method its name. On average 36.8% of the subjects are left for testing of the model. These .632 and .632+ variants did however not have clear advantages over the bootstrap procedure described earlier in some empirical studies.<sup>413,479</sup>

We can apply the bootstrap approach to any performance measure, including the  $R^2$ ,  $c$  statistic, and calibration measures such as calibration slope. A strong aspect of the bootstrap is that we can incorporate various complex steps from a modelling strategy. This is important since exact distributional results are virtually impossible

---

<sup>1</sup>The “simple bootstrap” compares the performance of the model from the original sample in bootstrap samples. This was less efficient than the procedure described here, where models from the bootstrap samples are tested in the original sample (see Efron).

to obtain, even for simple common selection algorithms.<sup>336</sup> The bootstrap can hence give insight in the relevance of model uncertainty, including both testimation bias and parameter uncertainty. In practice, however, it may be hard to fully validate a prediction model, including all steps made in the development of the model. For example, automated stepwise selection methods can be replayed in every bootstrap sample, leading to reasonably correct optimism-corrected performance estimates.<sup>401</sup> But more subtle modelling steps usually cannot fully be incorporated, such as choices on coding and categorization of predictors. The optimism-corrected estimate may then be an upper bound of what can be expected in future subjects. Only a fully specified modelling strategy can be replayed in every bootstrap sample.

It is often useful to calculate the optimism of a “full model,” i.e. a prediction model, including all predictors without any fine-tuning such as deleting less-important predictors. The optimism estimate of such a full model may be a guide for further modelling decisions.<sup>174</sup> If the optimism is substantial, it is a warning that we should not base our model only on the data set at hand. Using external information may improve the future performance of the model.<sup>164</sup>

### \*5.3.5 Example: Stepwise Selection in 429 Patients

As an example, we consider a sample of 429 patients from the GUSTO-I study, which studied 30-day mortality in patients with acute myocardial infarction (details in Chap. 24). We first fitted a model with eight predictors, as specified in the TIMI-II study (“full model”).<sup>302</sup> This model had a Nagelkerke  $R^2$  of 23% as apparent performance estimate. In 200 bootstrap samples, the mean apparent performance was 25% (Table 5.4). When the models from each bootstrap sample were tested in the original sample, the  $R^2$  decreased substantially (to 17%). The optimism hence was 25% – 17% = 8%, and the optimism-corrected  $R^2$ , 23% – 8% = 15%.

We can follow a backward stepwise selection procedure with  $p < 0.05$  for factors remaining in the model (Chap. 11). This leads to inclusion of only three predictors (age, hypotension, and shock). The apparent performance drops from 23% to 15% by excluding six of the eight predictors. The stepwise selection was repeated in every bootstrap sample, leading to an average apparent performance of 18%, which dropped to 12% when models were tested in the original sample (optimism, 6%; optimism-corrected  $R^2$ , 9%). When we falsely assume that the 3 predictor model was pre-specified, we would estimate the optimism as 3% rather than 6%. This discrepancy illustrates that optimism by selection bias was as important as the optimism due to parameter uncertainty in this example.

We note that the apparent performance in the bootstrap samples was higher than the apparent performance in the original sample (Table 5.4). This pattern is often noted in bootstrap model validation. It may be explained by the fact that some patients appear multiple times in the bootstrap sample. Hence, it is easier to predict the outcome, reflected in higher apparent performance. Further, we note that the

**Table 5.4** Example of bootstrap validation of model performance, as indicated by Nagelkerke's  $R^2$  in a subsample of the GUSTO-I data base (sample5, n=429)

Method	Apparent (%)	Bootstrap (%)	Test (%)	Optimism (%)	Optimism-corrected (%)
Full 8 predictor model	22.7	24.7	17.2	7.6	15.1
Stepwise, 3 predictors, p<0.05	17.6	18.7	12.7	5.9	11.7
Stepwise model falsely assumed to be pre-specified	17.6	18.2	15.4	2.9	14.7

optimism is smaller after model specification by stepwise selection (6% instead of 8%). However, the optimism-corrected performance of the stepwise model  $R^2$  12% is clearly lower than the performance of the full 8 predictor model ( $R^2$  15%). This pattern is often noted. A full model will especially perform better than a stepwise model when the stepwise selection eliminates several variables that are almost significant while they have some true predictive value. When a small set of dominant predictors is present, including only these would logically be sufficient. The bootstrap would show that these predictors are nearly always selected, and that other variables are most often excluded; the optimism would be relatively small and optimism-corrected performance similar to that of a full model. The leprosy case study is such an example (see Chap. 2). In the case that many noise variables are present in the full model, a selected submodel performs better than a full model. Careful pre-selection of candidate predictors is hence advisable, based on subject knowledge (literature, expert opinion), to prevent that pure noise variable are considered in the modelling process.

## 5.4 Cost of Data Analysis

The development of a prediction model for outcome prediction is a constant struggle in weighing better fit to the data against generalizability outside the sample. The more we incorporate from a specific data set in a model, the less the model may generalize.<sup>101</sup> This has aptly been labelled the "cost of data analysis." On the other hand, we do not want to miss important properties of the data, such as a clearly non-linear relationship of a predictor to the outcome. A prediction model where underlying model assumptions are fulfilled will provide better predictions than a model where assumptions are violated. Therefore, it is natural to assess such assumptions as linearity of continuous predictor effects and additivity of effects (Chap. 12). However, if we test all assumptions of a model and iteratively adapt the model to capture even small violations, the model will be very specific for the data analyzed.

### \*5.4.1 Example: Cost of Data Analysis in a Tree Model

An interesting concept was proposed by Ye, who determined the “generalized degrees of freedom” (GDF) of a model selection and estimation procedure.<sup>494</sup> The GDF indicate the overfitting that was associated with a modelling strategy. For example, Ye showed that a stepwise selection strategy that selected a model with five predictors (apparent  $df = 5$ ) had a GDF of 14.1. A regression tree had 19 nodes (apparent  $df = 19$ ), but GDF of 76.<sup>494</sup>

An essential part of Ye’s method is to determine the apparent performance of a model when developed with pure noise. In Table 5.2, we note that the optimism in  $R^2$  in the pure noise simulations was indeed very similar to the optimism as determined with an adjusted  $R^2$  or with bootstrapping when some true effects were present. For example, for  $n = 200$ , the optimism was 1% with pure noise or with true effects.

### 5.4.2 Practical Implications

In the development of prediction models, we have to be aware of the cost of all data analysis steps. The appropriateness of a modelling strategy is indicated by the generalizability of results to outcome prediction for new patients. Some practical issues are relevant in this respect.

- Sample size: With a small sample size we have to be prepared to make more assumptions about our data; the power to detect deviations from assumptions will anyway be small. If deviations from assumptions are detected, and the model is adapted, testimation bias will occur and the validity of predictions for new patients may not necessarily be improved (Chap. 13);
- Robust strategies: Some modelling strategies are more “data hungry” than other strategies. For example, fitting a pre-specified logistic regression model with age and sex uses only two degrees of freedom. If we test for linearity of the age effect, and interactions between age and sex, we spend more degrees of freedom. If we use a method such as regression tree analysis, we search for cut-points of age, and model interactions by default, making the method more data-hungry than logistic regression (Chap. 4). Similarly, stepwise selection asks more of the data than fitting a pre-specified model. Not only do we want to obtain estimates of coefficients, we also want to determine which variables to include as predictors (Chap. 11);
- Bootstrap validation: The bootstrap can assist in determining an appropriate level of fine-tuning of a model to the data under study. However, when many alternative modelling strategies are considered, the bootstrap results may become less reliable in determining the optimal strategy, since the optimum may again be very specific for the data under study. The bootstrap works best to determine optimism for a single, pre-defined strategy.

## 5.5 Concluding Remarks

In science in general, and in prediction modelling specifically, we need to seek a balance between curiosity and skepticism. On the one hand, we want to make discoveries and advance our knowledge, but on the other hand we must subject any discovery to stringent tests, such as validation, to make sure that chance has not fooled us.<sup>23</sup> It has been demonstrated that our scientific discoveries are often false, especially if we search hard and explore *a priori* unlikely hypotheses.<sup>210</sup> Overfitting and the resulting optimism are important concerns in prediction models.

## Questions

### 5.1 Overfitting and optimism

- (a) What is overfitting and why is it a problem?
- (b) What are the two main causes of overfitting? What is the difference and give some examples?

### 5.2 Shrinkage for prediction (Figs. 5.3 and 5.4)

A solution against the consequence of overfitting is shrinkage. For example, estimates per centre can be drawn towards the average to improve the quality of predictions in Figs. 5.3 and 5.4.

- (a) Is the required shrinkage more, or less, in Fig. 5.4 compared with Fig. 5.3?
- (b) Is the underlying true heterogeneity more, or less, in Fig. 5.4 compared with Fig. 5.3?

### 5.3 Bootstrapping (Sect. 5.3)

- (a) How can a bootstrap sample be created? How is this done with the `sample` command in R?
- (b) How can the test sample for the .632 bootstrap variant be selected in R?
- (c) How can bootstrapping be used to derive optimism-corrected estimates of model performance, addressing the two main causes of overfitting?

# Chapter 6

## Choosing Between Alternative Statistical Models

**Background** Any scientific model will have to make simplifying assumptions about reality. Nevertheless, statistical models are important tools to summarize patterns from underlying data. Statistical models can well be used to make predictions for future subjects. We consider some general issues in choosing a type of statistical model in a prediction context, with illustration in a case study on modelling age–outcome relationships in medicine. We also summarize results from some empirical comparisons of alternative statistical models.

### 6.1 Prediction with Statistical Models

In a prediction context, statistical models are merely seen as practical tools than as theories about how the world works. As long as the model predicts well, we are satisfied. This relates to the famous quote “All models are wrong, but some are useful.”<sup>51</sup> Although regression models are formulated as models of cause and effect (“y depends on x”), there need not be any causal relation at all, for example because some intermediate causal factor was not recorded. We hence simply use the terms “predictor” and “outcome.”

On the other hand, a statistical model can provide important insights in how a combination of predictors is related to an outcome. For inference and hypothesis testing, fulfillment of assumptions becomes more important than for prediction. Prediction is primarily an estimation problem, while insight in effects of predictors is related to hypothesis testing (Chap. 1). With a model, we can make predictions for future subjects, test hypotheses, and estimate the magnitude of effects of predictors. It is a philosophical question whether a true, underlying model exists. Many have argued that the notion of a “true model” is false.<sup>69</sup> Indeed, would it be imaginable that natural processes can fully be captured in a model containing relatively few variables, which are related in a mathematically simple way? Many subtle non-linear and interactive effects probably play a role. Predictors may be unobservable or not yet discovered, or predictive effects may be too small to detect

empirically. Therefore, a statistical model can only be an approximation to underlying patterns, based on the limited number of predictors that is known to us.

### 6.1.1 Testing of Model Assumptions and Prediction

If our primary aim is to make good predictions, we should not place too much emphasis on unobservable, underlying assumptions. It is a standard procedure nowadays to test model assumptions such as non-linearity and additivity, or proportionality of hazards (see also Chap. 13). Such testing may be valuable but only to the extent that adaptations to the model lead to better predictions. When assumptions are met, the model will provide a better approximation to reality and hence predict better.<sup>174, 176</sup> Statistically significant violations of underlying assumptions do not mean that a prognostic model predicts poorly.<sup>171</sup>

In a prediction context, we are lucky that we can directly measure the observed outcomes and compare these to what is predicted. This allows for direct statistical assessment of model quality with performance measures such as calibration and discrimination. Whether the underlying assumptions of the prediction model are true can never be known, since these assumptions are unobservable.

### 6.1.2 Choosing a Type of Model

Some general suggestions have been made on the type of model to be used in prognostic research.<sup>174</sup>

- The mathematical form should be reasonable for the underlying data. For example, models should not give predictions that are below 0% or above 100% for binary outcomes or survival probabilities.
- The model should use the data efficiently. Regression models need to make assumptions, but they pick up general patterns in the data better than a simple cross-tabulation approach. Cross-tables quickly run out of numbers, and hence would provide unstable predictive estimates. Similarly, survival outcomes should be analyzed with methods that use all available information.
- Robustness is preferred over flexibility in capturing idiosyncrasies. For prediction, we aim to model patterns that generalize to future subjects. Very flexible approaches will require large data sets, while medical prediction problems are often addressed with relatively small data sets. The results of the model should be transparent and presentable to the intended audience. In some fields, fully computerized models may be acceptable (e.g. neural networks), but in other fields insight in the underlying model is an advantage (e.g. effects of predictors in regression models).
- Alternative model formulations can sometimes be assessed empirically, and subject matter knowledge can assist in guiding the choice for a model. Also, practical issues play a role, such as familiarity of analysts and their readers with a method.

A major requirement of any model is of course that it adequately answers the research question, since we know that all models will miss some aspects of the underlying natural process by their relative simplicity.

We will first look at some empirical support for relatively simple regression models as tools to capture the prognostic effect of age. This is followed by a brief discussion of some head-to-head comparisons that have been made between modelling techniques.

## 6.2 Modelling Age–Outcome Relationships

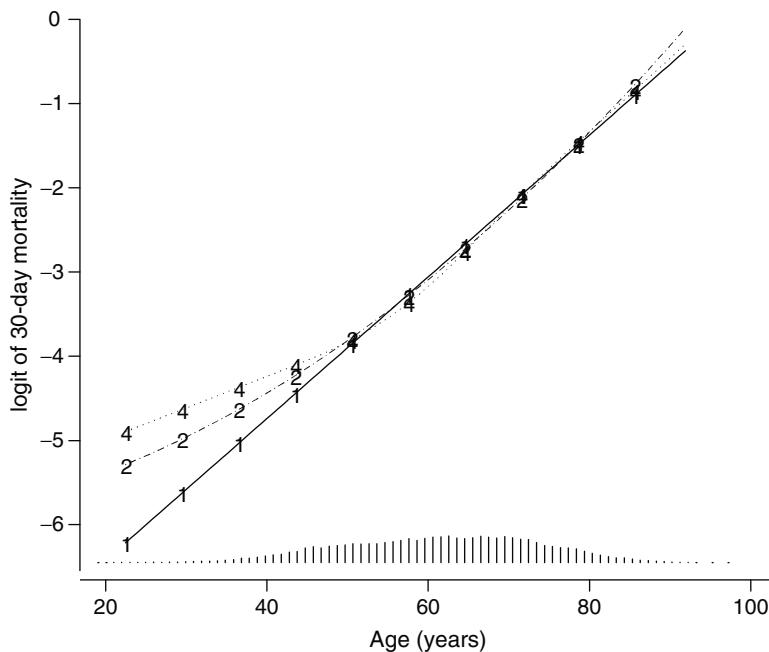
The effect of age on outcome is important in many medical prediction problems. Together with gender, age is an obvious demographic characteristic to consider in the prediction of an outcome. On the one hand, age represents the biological phenomenon of aging, with a decrease in performance of biological systems. Observed age effects do however not necessarily represent pure biological relationships, since many comorbid conditions may be present. Moreover, selection may have occurred, e.g. making that very old patients only undergo surgery when in relatively good condition. Nevertheless it is of interest to see how increasing age is related to outcome. Specifically we consider the modelling of age-related mortality with logistic regression.

### \*6.2.1 Age and Mortality After Acute MI

Within the GUSTO-I data set (details in Chap. 24), Lee et al. found that the relationship between age and 30-day mortality after an acute myocardial infarction was reasonable linear.<sup>255</sup> When we examine the relationship in detail, we see that the Likelihood Ratio (LR,  $\chi^2$ ) statistic of the linear fit is 2,099. Adding  $age^2$  increases the fit by 13 to 2,112, and a restricted cubic spline with 5 knots (4  $df$ , including the linear term, see Chap. 9) adds 23 (model  $\chi^2$  2,122). So, there is no major gain by adding non-linear transformations. The age–mortality relationships with alternative transformations are shown in Fig. 6.1. The differences between the transformations are at the lower ages (below age 50), where limited data are available. It may be that the age–mortality relationship is somewhat stronger above age 50 than below age 50. A linear spline with change point at age 50 has a  $\chi^2$  of 2,119. In sum, assuming a linear effect was quite reasonable for modelling the effect of age for mortality after acute MI.

### \*6.2.2 Age and Operative Mortality

Finlayson and Birkmeyer examined operative mortality in relation to age for 1.2 million elderly patients in the Medicare system.<sup>125</sup> They selected patients who were between 65 and 99 years old, and who were hospitalized between 1994 and 1999



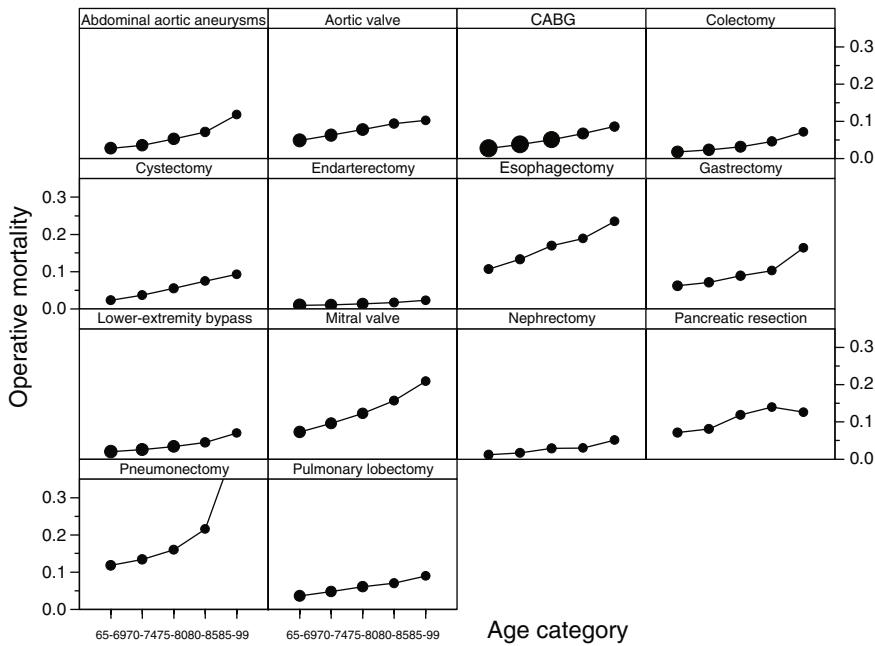
**Fig. 6.1** The relationship between age and 30-day mortality among 40,830 patients in GUSTO-I. The line marked with “1” is a linear logistic regression fit ( $1\ df$ ), the line marked with “2” is a polynomial fit ( $age + age^2, 2\ df$ ), and the line marked with “4” is a restricted cubic spline fit (5 knots,  $4\ df$ ). The distribution of the ages is shown at the bottom of the graph. Note the enormous range in mortality, since a logit of  $-6$  means a probability of 0.2% and a logit of 0 means 50%

for major elective surgery (six cardiovascular procedures and eight major cancer resections). Operative mortality was defined as death within 30 days of the operation or death before discharge, and occurred in over 38,000 patients.

The mortality risk in this huge, nationwide, representative series varied widely between procedures. Not surprisingly, it was higher than that reported in many published series from specialized centres. Operative mortality clearly increased with age. Operative mortality for patients 80 years of age and older was more than twice that for patients 65–69 years of age (Fig. 6.2).

These data can well be used to illustrate the fit of a logistic transformation for the relationship between age and mortality. The data were reported in categories. To study age as a continuous variable requires an estimate of the average age per category, which we assume to be at midpoints for the first two categories (67.5 and 72.5 years) and at 77.2, 82.0, and 90.0 years for the other three categories.

The simplest logistic regression model assumes a single age effect across categories:  $\text{mort} \sim \text{procedure} + \text{age10}$ , where mort indicates operative mortality (0/1), which is a function ( $\sim$ ) of procedure (a categorical variable for the 14 levels of procedures) and age10 (age coded per 10 years).



**Fig. 6.2** Operative mortality by surgical procedure according to age in 1.2 million elderly patients in the Medicare system<sup>125</sup>

We can test whether the age effect differs by procedure by adding the interaction term “procedure  $\times$  age10”:

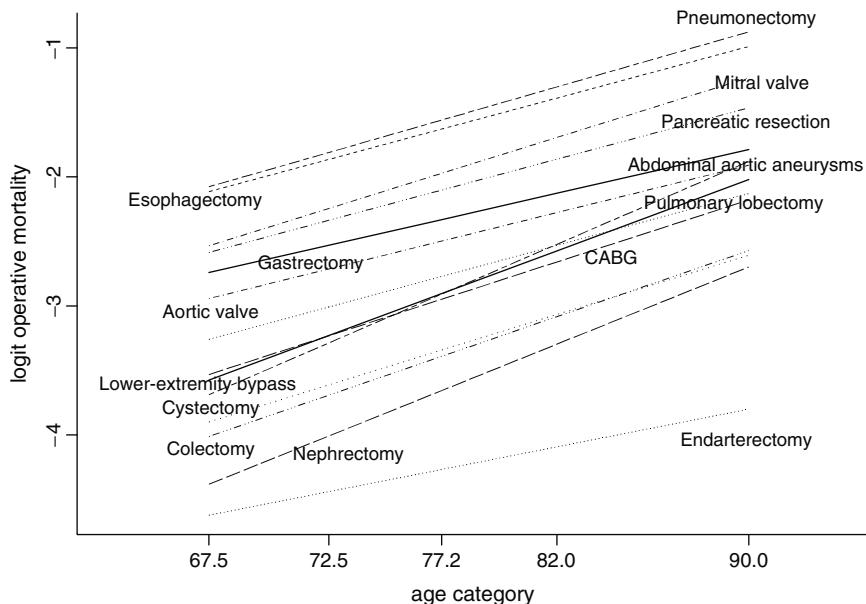
$$\text{mort} \sim \text{procedure} + \text{age10} + \text{procedure} \times \text{age10}$$

The smallest age effect was found for endarterectomy (OR 1.44 per decade), followed by gastrectomy (OR 1.53 per decade). The strongest age effects were found for nephrectomy and cystectomy (OR 2.11 and 2.23 per decade, Fig. 6.3).

The improvement in model fit obtained by this model extension was relatively small (Table 6.1). The model  $-2 \log$  likelihood decreased by 95, which was only 0.6% of the total model  $\chi^2$  including this interaction. The explained variation ( $R^2$ ) increased by 0.04%, and the area under the ROC curve (or  $c$  statistic) remained the same. See Chap. 15 for a detailed discussion of these performance measures. We can also assess non-linearity in the age effect by adding a square term ( $\text{age10}^2$ ) to the simplest model:

$$\text{mort} \sim \text{procedure} + \text{age10} + \text{age10}^2$$

Remarkably, adding such a square term made no contribution to the model fit ( $\chi^2$  increased by 2).



**Fig. 6.3** Logistic regression models for operative mortality by surgical procedure according to age, based on analysis of 1.2 million elderly patients in the Medicare system<sup>125</sup>

**Table 6.1** Age and operative mortality in 1.2 million elderly patients in the Medicare system<sup>125</sup>

Model	Age10	Model $\chi^2$	R <sup>2</sup> (%)	AUC
Procedure + age10	1.75	16,841	5.62	0.683
... + Procedure $\times$ age10	1.44–2.23	16,936	5.66	0.683
... + age10 <sup>2</sup>	1.74	16,843	5.62	0.683

In conclusion, the age–outcome relationships were linear for these surgical procedures. Moreover, we may assume that the effect is in the order of an odds ratio of 1.75 per decade, or a doubling in odds per 12.5 year. Both results may be useful as prior knowledge when we model the effect of age, especially when we are dealing with a small data set. As a starting point we may assume linearity on the logistic scale, and an age effect between 1.5 and 2.2 per decade, the latter depending on the procedure.

### \*6.2.3 Age–Outcome Relationships in Other Diseases

Many studies described the relationship between age and outcome in other diseases. We performed a meta-analysis in traumatic brain injury, where again a linear transformation was adequate, although adding the square term of age provided a

somewhat better fit.<sup>204</sup> Among seriously ill hospitalized adults, age had a linear effect that differed slightly by diagnosis, similar to the evaluation of operative operative.<sup>236</sup> Other studies also support a more or less linear association between age and outcome.

Remarkably, some studies, especially smaller ones, conclude from a statistically non-significant age effect that age was not related to outcome. This is a clear illustration of interpreting absence of evidence as evidence of absence.<sup>11</sup>

## 6.3 Head-to-Head Comparisons

Several studies have described head-to-head comparisons of alternative methods. Especially, attention has been given to alternative methods to predict binary outcomes. Main classes of statistical methods include regression modelling, trees, and neural networks.<sup>181</sup> Some comparisons were in favour of regression-based techniques, and some in favour of more modern approaches such as neural networks.

A problem in many comparisons is that one of the comparators is not developed with state-of-the-art methods, while the other is. For example, computer scientists often have been working on variants of neural networks, which were shown to do better than logistic models which were derived with simplistic, suboptimal techniques.<sup>377</sup> Methodological problems are even more severe for comparisons of methods to predict survival outcomes.<sup>376</sup> Kaplan–Meier and Cox regression adequately deal with survival data, but ad hoc approaches have usually been followed for other techniques. Moreover, there are no agreed objective statistical criteria by which to judge modelling methods. Other criteria, such as how easy a model is to develop and apply, are also very important when researchers make their choice from the currently available modelling methods.

We note that the quality of predictions obtained with a statistical method may depend on three factors:<sup>288</sup>

1. The essential quality and appropriateness of the method
2. The actual implementation of the method as a computer program
3. The skill of the user

### 6.3.1 StatLog Results

An important example of a systematic comparison of statistical modelling approaches is the StatLog project.<sup>288</sup> Different approaches to classification were studied. Table 6.2 summarizes some results for data sets with a binary outcome, both from medical and non-medical applications.<sup>288</sup> It appears that logistic regression performs quite well across all examples according to error rates. More flexible techniques such as trees and neural networks only have advantages in larger data sets. In the

**Table 6.2** Error rates for problems with binary outcomes in the StatLog project<sup>288</sup>

Data set	<i>N</i> dev	Predictors	Logistic	Naïve Bayes	Tree (CART)	Neural network <sup>a</sup>
<i>Non-medical</i>						
Credit management	15,000	7	0.030	0.043	NA	0.023
Australian credit	690	14	0.141	0.151	0.145	0.154
German credit	1000	24	0.538	0.703	0.613	0.772
Cut (letters in text)	11,220	20	0.046	0.077	NA	0.043
	11,220	50	0.037	0.112	NA	0.041
Belgian Power	1250	28	0.007	0.062	0.034	0.017
Instability	2000	57	0.028	0.089	0.022	0.022
<i>Medical</i>						
Heart disease	270	13	0.396	0.374	0.452	0.574
Diabetes	768	8	0.223	0.262	0.255	0.248
Tsetse	3500	14	0.117	0.120	0.041	0.065

NA: Not available

<sup>a</sup>Backpropagation algorithm

medical context, data sets are often relatively small, especially with respect to number of events, and the predictive information is relatively limited, leading to an unfavourable noise to signal ratio.<sup>171,181</sup>

### \*6.3.2 GUSTO-I Modelling Comparisons

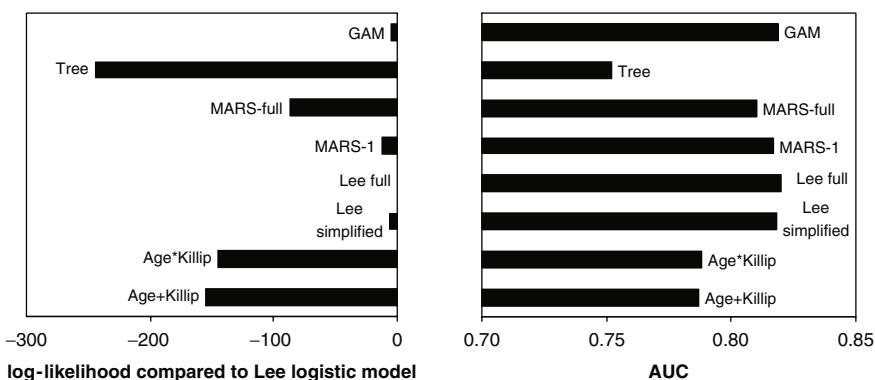
Several simulation studies have been performed with the GUSTO-I database. Ennis et al. compared a variety of modern learning methods, including logistic regression, Tree, GAM, and MARS methods.<sup>115</sup> Logistic regression can be considered as a classic prediction method. The other methods have more flexibility in capturing interaction terms or non-linear terms, and may be referred to as adaptive non-linear methods. These methods require data sets of substantial size, which is the case in GUSTO-I ( $n=40,830$ ). Because of the huge size, a large independent test set could be kept separate from the development set.

- Four different logistic regression models were considered,<sup>115</sup> containing
  1. Age and Killip class;
  2. Age, Killip class, and interactions between age and Killip class;
  3. All covariates as in Lee et al.'s model,<sup>255</sup> but no interactions and no non-linear (spline) terms;
  4. Lee et al.'s model, including the interactions and non-linear (spline) terms.
- Classification trees stratify the population in a binary tree form, and are especially good at finding interactions between risk factors.<sup>57</sup> A classification tree was constructed using state-of-the-art methods considering all 17 predictors.

- A generalized additive logistic regression model (“GAM”) is a non-linear generalization of the usual linear logistic model. It used smooth spline functions in place of linear risk terms.<sup>180</sup> The model contained smoothing splines with 4 degrees of freedom for the variables age, height, weight, pulse rate, systolic blood pressure, and time to treatment. No interaction terms were included.
- MARS stands for “multivariate additive regression splines,” and is a kind of hybrid between generalized additive models and classification tree.<sup>129</sup> It is designed to find low-order additive structure as well as interactions between risk factors. MARS models of degree 1 (additive) and 17 (all interactions allowed) were considered.

### \*6.3.3 GUSTO-I Results

The performance in the test set of 13,610 patients was remarkable (Fig. 6.4). The most basic logistic model had an AUC, or ( $c$  statistic), of 0.787, which improved substantially when more predictors were considered (Lee et al.’s logistic model variant 3 and 4,  $c$  around 0.82). The performance of Lee et al.’s traditional logistic model<sup>255</sup> could not be improved by any other method. A similar performance was found for the GAM and additive MARS model. The more flexible variant of the MARS model (with all interactions allowed) had a  $c$  statistic of 0.81 (0.01 lower). The tree performed worst, with a  $c$  statistic of 0.75. Results were similar when the log-likelihood was used as a measure for predictive performance (Fig. 6.4).



**Fig. 6.4** Performance of alternative prediction models in a test part of the GUSTO-I data set (13,610 patients with acute MI).<sup>115</sup> Results are shown for four logistic regression models, two variants of MARS models, a classification tree, and a GAM. In the *left panel*, the log-likelihood is compared to the Lee full model; in the *right panel* the area under the ROC curve (or  $c$  statistic) is shown. Age+Killip: main effects; Age  $\times$  Killip: main and interaction effects; Lee simplified: a simplified version of Lee model; Lee full: Lee et al. model.<sup>255</sup> We note that the Lee full model performed best according to both performance criteria

The authors also examined various variants of multi-layer neural networks using advanced backpropagation algorithms and various approaches to prevent overfitting (weight decay, early stopping, bagging).<sup>115</sup> None of these methods led to a better predictive performance than Lee et al.'s traditional logistic regression model.

## 6.4 Concluding Remarks

We recognize that true models do not exist, and that any model only approximates relationships between predictors and outcome. A model will only reflect underlying patterns, and hence should not be confused with reality. This is also shown in René Magritte's famous painting "La trahison des images" (The Treachery Of Images, [[http://en.wikipedia.org/wiki/The\\_Treachery\\_of\\_Images](http://en.wikipedia.org/wiki/The_Treachery_of_Images)]). This painting shows a pipe, with the words "Ceci n'est pas une pipe" (This is not a pipe) painted below the pipe. Indeed, the painting is not a pipe, it is an image of a pipe.

Nevertheless, statistical models that approximate reality closer are better for predictive purposes, as well as for inference on predictive effects of predictors. If we derive models from empirical data, the sample size needs to be sufficient for the complexity of the model that is fitted. For flexible models, the GUSTO-I results illustrate that non-linear effects and interactions may need to be quite strong before an advantage is obtained over relatively simple regression models. In medical prediction problems, the signal to noise ratio may often be relatively low. This makes regression analysis an appropriate default approach in clinical prediction models.

## Questions

### 6.1 Reasonable modeling approaches

In traumatic brain injury, the Glasgow Outcome Scale is a 5 point, ordered scale. It is common to determine the GOS at 6 months post injury. A researcher proposes to use linear regression analysis to analyse relationships for predictors with this scale. What are pros and cons of this approach for estimation of predictive effects, and for making predictions?

### 6.2 Predictions from cross-tabulations (Fig. 6.2)

A researcher might argue that the observed mortality as shown in Fig. 6.2 can directly be used for predictive purposes, similar to the cross-tabulation provided in an analysis of genetic mutation risks among 10,000 women.<sup>127</sup> What are pros and cons of this approach?

### 6.3 GUSTO-I results (Sect. 6.3)

More flexible methods performed worse than a logistic regression model in the GUSTO-I case study. What results would you expect for the comparison in Fig. 6.4 with (a) smaller and (b) larger sample sizes for model development (e.g. 1,000 and 1,000,000 patients)?

## **Part II**

# **Developing Valid Prediction Models**

In part I, we presented a number of issues that are relevant to the context of prediction model development and application. We summarize these issues as preliminaries for model development in a proposed checklist. In part II, we focus on the development of prediction models that are valid for the population from which the sample originates. Generalizability to other, plausibly related, populations is discussed in Part III. We will discuss seven steps of model development in the following chapters (Chaps. 7–18).

## Checklist for Developing Valid Prediction Models

Step	Specific issues	Chapter
<i>General considerations</i>		
Research question	Aim: predictors/prediction?	1
Intended application?	Clinical practice/research?	2
Outcome	Clinically relevant?	3
Predictors	Reliable measurement? Comprehensiveness	3
Study design	Retrospective/prospective? Cohort; case–control	3
Statistical model	Appropriate for research question and type of outcome?	4 and 6
Sample size	Sufficient for aim?	2–6
<i>Seven modelling steps</i>		
Data inspection	Missing values	7 and 8
Coding of predictors	Continuous predictors Combining categorical predictors Restrictions on candidate predictors	9 and 10
Model specification	Appropriate selection of main effects? Assessment of assumptions (distributional, linearity, and additivity)?	11 and 12
Model estimation	Shrinkage included? External information used?	13 and 14
Model performance	Appropriate statistical measures used? Clinical usefulness considered?	15 16
Model validation	Internal validation, including model specification and estimation? External validation?	17
Model presentation	Format appropriate for audience?	18
<i>Validity</i>		
Internal: Overfitting	Sufficient attempts to limit and correct for overfitting?	4–18
External: Generalizability	Predictions valid for plausibly related populations?	19–21

# Chapter 7

## Dealing with Missing Values

**Background** Missing data are a common problem in prediction research. We concentrate on missing values of predictor values ( $X$ ), in the context of a prediction model for a single outcome ( $Y$ ). Traditional complete case analysis suffers from inefficiency, selection bias of subjects, and other limitations. We briefly review the theoretical background on mechanisms of missingness of predictor values and how these may affect prognostic modelling. We further concentrate on imputation methods as a solution, where a completed data set is created by filling in missing values for the statistical analysis. Special attention is given to the specification of an imputation model, which is the essential step in imputation. A sophisticated method is to generate completed data sets multiple times (“multiple imputation”), but single imputation is more straightforward and may sometimes be sufficient. Several examples are provided. Chapter 8 presents a case study of dealing with missing values in a meta-analysis of individual patient data on prognosis in traumatic brain injury. Tentative guidelines are provided on how to deal with missing data in relation to the research question.

### 7.1 Missing Values in Predictors

Missing data are a common, but as yet underappreciated problem in medical scientific research. In this chapter, we concentrate on missing values of predictors, assuming that true predictor values are hidden by the missing values.<sup>263</sup> Standard statistical software for regression analysis deletes subjects with any missing predictor value from the analysis. With such a complete case analysis, all subjects with a missing value for any potential predictor are excluded.<sup>155,263</sup> An “available case analysis” will consider subjects with complete data for a specific predictor, but who may have missing values for other covariates that are not considered in the specific model. With such an analysis, numbers may therefore vary per analysis. Both complete case and available case analysis discard information from subjects who have information on some (but not all) predictors. They are hence statistically inefficient, as further illustrated below. For simplicity, we use the term complete case (CC) analysis further onwards.

### 7.1.1 Inefficiency of Complete Case Analysis

As a hypothetical example, we consider a data set with 500 subjects. Among these, 100 suffer the event that we want to predict (e.g. 30-day mortality). We aim to estimate regression coefficients for a prognostic model consisting of five predictors. In case of complete data, we have 20 events per variable. Such a situation is commonly thought to be sufficient for reliable estimation of the regression coefficients in a model. Suppose, however, that each predictor has 10% missing data and that each patient has at most 1 missing value. Hence, each patient has at least four values of the predictors recorded (Table 7.1). A CC analysis will ignore  $5 \times 10\% = 50\%$  of the subjects, and will leave only 250 subjects for estimation of the regression model. The number of events per variable drops to 10:1, which is commonly thought of as a minimum for reliable modelling.

The information available is 250 complete cases ( $250 \times 5 = 1,250$  predictor values) + 250 incomplete cases ( $250 \times 4 = 1,000$  predictor values). Of the required  $500 \times 5 = 2,500$  predictor values, 2,250 or 90% are available. The approach of using only 50% of the information instead of 90% hence is quite inefficient: 10% of the required values are missing, but 50% of the subjects are discarded. The inefficiency occurs to a lesser extent if multiple missing values may occur within the same patient, which is a more realistic situation.

**Table 7.1** Hypothetical missing data pattern: 250 subjects have partially complete data (missing data indicated with .), and 250 have fully complete data (indicated with X)

ID	X1	X2	X3	X4	X5	Y
1	.	X	X	X	X	X
...	.	X	X	X	X	X
50	.	X	X	X	X	X
51	X	.	X	X	X	X
...	X	.	X	X	X	X
100	X	.	X	X	X	X
101	X	X	.	X	X	X
...	X	X	.	X	X	X
150	X	X	.	X	X	X
151	X	X	X	.	X	X
...	X	X	X	.	X	X
200	X	X	X	.	X	X
201	X	X	X	X	.	X
...	X	X	X	X	.	X
250	X	X	X	X	.	X
251	X	X	X	X	X	X
...	X	X	X	X	X	X
...	X	X	X	X	X	X
500	X	X	X	X	X	X
Total	450	450	450	450	450	500

### 7.1.2 Interpretation of Analyses with Missing Data

Further concerns with missing data include problems in interpretation of results from analyses. When different models are compared, it is difficult to interpret results when the numbers of subjects vary across the analyses. For example, when a univariate odds ratio is based on 450 subjects for each  $X$  variable in Table 7.1, we cannot interpret the change in odds ratio of an adjusted analysis performed on 250 subjects, due to missing values for in total 250 subjects. It is then impossible to infer whether differences arose between univariate and adjusted analysis because of correlation between the predictors or because of a selection of subjects due to missing values. Other problems include cumbersome comparison of  $p$ -values and of the performance of two models, when they are based on different numbers of subjects. This problem would not occur if we would analyse 250 patients in both univariate and adjusted analysis in a true complete case analysis.

Another concern is that bias may arise due to systematic differences between subjects with complete data and subjects with missing data. It appears that bias will especially occur in the estimated regression coefficient for a predictor when missingness is associated in some way with the outcome.<sup>445</sup> This issue will be discussed in more detail later.

### 7.1.3 Missing Data Mechanisms

Different mechanisms may lead to missing data (Table 7.2).<sup>263,357</sup> It is important to consider these, since approaches to handle missing data in the statistical analysis rely on assumptions on the mechanism. The nomenclature of the mechanisms is puzzling to many applied researchers, but has been adopted uniformly in the scientific literature.

Missing values can occur completely at random (MCAR). Examples of MCAR mechanisms include administrative errors that occur at random, such as accidents in laboratories (e.g. spilling of material, handling errors, breakdown of equipment), or postal mail that is lost. MCAR is a strict assumption, and can be tested. With an MCAR mechanism, the subjects with missings are representative of the population with complete data. The incomplete population is a random sample from the complete population.

**Table 7.2** Three types of missing data mechanisms

Label	Missing mechanism	Description
MCAR	Missing completely at random	Administrative errors, accidents
MAR	Missing at random	Missingness related to known patient characteristics, time or place ("MAR on $x$ "), or to the outcome ("MAR on $y$ ")
MNAR	Missing not at random	Missingness related to the value of the predictor, or to characteristics not available in the analysis

In medical data, missing values often occur specifically in certain types of subjects. If we can observe the variables that are associated with missingness, we have an MAR situation (missing at random). This means that the probability of a missing value on a predictor (“missingness”) is independent of the values of the predictor itself, but depends on the observed values of other variables. The MAR assumption is fulfilled if missingness is only related to measured values in the data set but not to unmeasured values. MAR examples include more missing values in older subjects, subjects from a certain region, or from an earlier calendar time. Also, the design of a study may intentionally leave values missing for some type of subjects, which is by definition an MAR mechanism. For example, we may choose not to measure a lab value in younger patients.

With an MAR mechanism, the subjects with complete predictor and outcome values are not representative anymore for the population where we want to generalize to. We will use various examples to illustrate how a CC analysis affects estimates of the regression coefficients and the estimated performance.

A problematic situation arises when data are MNAR (missing not at random). An MNAR mechanism arises when the missingness depends on the values that are missing, or on other predictors that are not observed. Examples include selective non-response on certain questions (e.g. sexual orientation, income), or clinical condition (e.g. missing if a severe condition is present, which is not measured accurately).

### 7.1.4 Summary Points

Missing data lead to

- Inefficient analyses of research questions
- Difficulties in interpretation when analyses differ in numbers of subjects
- Possible bias in regression coefficients

Missing data mechanisms can be described as MCAR, MAR, and MNAR (Table 7.2).

## 7.2 Regression Coefficients Under MCAR, MAR, and MNAR

For illustration we consider a simple linear regression model where an outcome  $Y$  depends on  $X_1$  and  $X_2$ :

$$Y = \beta_1 \times X_1 + \beta_2 \times X_2 + \text{error, with}$$

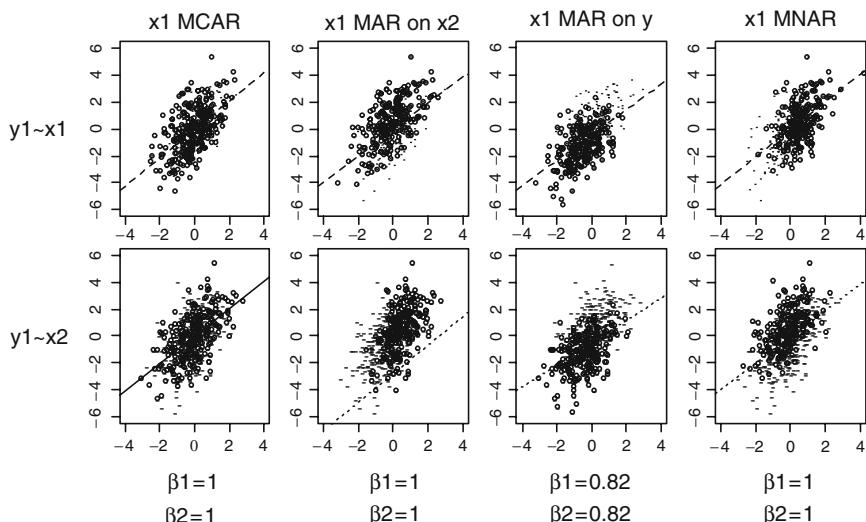
$X_1$  and  $X_2$  independent standard normal variables (distributed  $N(0,1)$ );  
 regression coefficients  $\beta_1$  and  $\beta_2$  both 1,  
 and the error distributed  $N(0,1)$ .

When we fit a linear regression model, the estimated regression coefficients  $\beta_1$  and  $\beta_2$  are on average 1 for both  $X_1$  and  $X_2$ . When we create missing values in  $X_1$  fully at random, we simulate an MCAR situation, and the estimated regression coefficients are on average 1 again for both  $X_1$  and  $X_2$  in the subjects with complete data (Fig. 7.1).

Of more interest is the situation of “MAR on  $x$ .” First, we consider the situation that missingness of  $X_1$  depends on  $X_2$ , with  $X_1$  only known with higher values of  $X_2$  (variable “ $x_1\text{MAR}_x$ ”). When we estimate the regression model  $Y \sim x_1\text{MAR}_x + X_2$  in the subjects with complete data, the estimated  $\beta_1$  and  $\beta_2$  remain unaffected (both 1, Fig. 7.1).

Second, we consider the MAR situation that missingness of  $X_1$  depends on  $Y$  (“MAR on  $y$ ”), with  $X_1$  only known with lower values of  $Y$  (variable “ $x_1\text{MAR}_y$ ”). When we estimate the regression model  $Y \sim x_1\text{MAR}_y + X_2$  in the subjects with complete data, the estimated  $\beta_1$  and  $\beta_2$  are both biased (Fig. 7.1). We selectively generated missings in  $X_1$  for the upper range of  $Y$  values. This makes that both  $\beta_1$  and  $\beta_2$  now have an expected value of 0.82 instead of 1. So, MAR of  $X_1$  on  $Y$  not only leads to bias in  $\beta_1$ , but also in  $\beta_2$ .

A correlation between missingness of a predictor and the outcome hence poses a serious problem in predictive modelling. Note, however, that if we measure all predictors prospectively, before the outcome is known, such a dependency cannot occur in a direct way. We register the predictors before the outcome.<sup>445</sup>



**Fig. 7.1** Effect of missing values on estimated regression coefficients  $\beta_1$  and  $\beta_2$  in the model  $y \sim X_1 + X_2$ . Original data are marked as “dot” and “dash” for  $X_1$  and  $X_2$ , respectively. Complete data under MCAR, MAR, and MNAR are marked with a circle. Plots show results for  $n = 500$ ; expected values for  $\beta_1$  and  $\beta_2$  are shown under the graphs (based on  $n = 100,000$ )

This holds both for diagnostic and prognostic problems. If an association between missingness of predictors  $X$  and outcome  $Y$  is noted in a prospective study, the explanation must be through other predictors, including predictors that are further down the causal pathway. If these predictors are not measured, we have an MNAR rather than an MAR situation.

Finally, we consider MNAR. Missingness of  $X1$  depends on the values of  $X1$ , with  $X1$  only known with higher values of  $X1$  (variable “ $x1MNAR$ ”). When we estimate the regression model  $Y \sim x1MNAR + X2$  in the subjects with complete data, the estimated  $\beta_1$  and  $\beta_2$  remain unaffected (both 1). This may be somewhat surprising at first sight, but is in line with the principle of conditioning in regression modelling: estimates of  $\beta_1$  and  $\beta_2$  are conditional on  $X1$ , and hence selection on  $X1$  does not affect these regression coefficients.

In sum, regression coefficients in this simple example remained unbiased under various missing data generating mechanisms. Bias only arose in the situation of an MAR on  $y$ , in this example this was that  $X1$  was only known for lower values of  $Y$ .

### \*7.2.1 R Code

In R, the command for MAR on  $x$  was:

```
x1MARx <- ifelse (rnorm(n=n, sd=.8) < x2, x1, NA)
```

Here `rnorm` generates random numbers from the normal distribution with SD=0.8. If the random value is smaller than  $x2$ ,  $x1MARx$  gets the value of  $X1$ , otherwise  $x1MARx$  is set to missing (“NA”). For low values of  $X2$ ,  $X1$  will more often be set to missing, reflecting an MAR on  $x$  situation. This simulation makes that missingness of  $X1$  has a smooth relationship with values of  $X2$ . Further, missingness of  $X1$  has an  $R^2$  of approximately 50% with values  $X2$ . When we estimate the regression model  $Y \sim x1MARx + X2$  in the subjects with complete data, the estimated  $\beta_1$  and  $\beta_2$  remain unaffected (both 1, Fig. 7.1).

MAR on  $y$  is simulated as:

```
x1MARY <- ifelse (rnorm(n=n, sd=1.5) >y, x1, NA)
```

Here `rnorm` generates random numbers from the normal distribution with SD=1.5, which makes that missingness of  $X1$  has an  $R^2$  of approximately 50% with  $Y$  (more missings for higher  $Y$  values).

An MNAR mechanism is simulated as:

```
x1MNAR <- ifelse (rnorm(n=n, sd=.8) < x1, x1, NA)
```

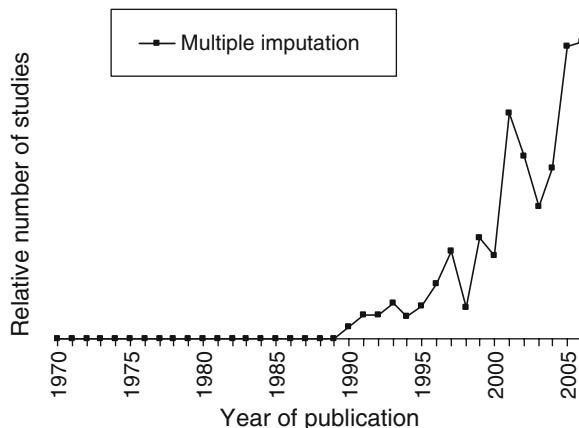
Again, `rnorm` generates random numbers from the normal distribution with SD=.8, which makes that missingness of  $X1$  has an  $R^2$  of approximately 50% with  $X1$ .

## 7.3 Missing Values in Regression Analysis

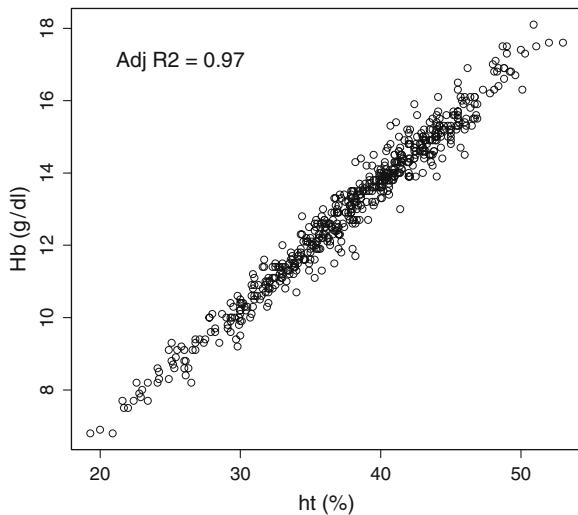
Most statisticians nowadays agree that we may opt for two sophisticated statistical approaches to deal with missing values in predictive regression models. The first is a maximum likelihood (ML) approach and the second is multiple imputation (MI). MI is a specific imputation method, where missing values are filled in (“imputed”) multiple times. MI methods make efficient use of all available data and take into account information implied by the available data. Hence, these methods are generally preferred over a CC analysis. ML methods have not yet become that popular in medical applications, but discussions on their use and merits are available. Both methods have theoretical and empirical support.<sup>263,357,370</sup> Further focus here is on imputation methods as a practical approach to missing values in prediction research (Fig. 7.2).

### 7.3.1 Imputation Principle

Imputation methods substitute the missing values with plausible values so that the completed data can then be analysed with standard statistical techniques. In some data sets, we may find a characteristic or combination of characteristics that closely defines the predictor with missing values, for example when variables are strongly related to same underlying phenomenon. For example haematocrit (“ht”) and haemoglobin (“Hb”) are both red blood cell indices. If we aim to include Hb in a prognostic model, it is useful to estimate Hb from ht for patients that have both



**Fig. 7.2** Studies in PubMed with the term “multiple imputation,” published between 1970 and 2005. We note a remarkable increase since 1990, with, for example, 41 publications in 2005, on a total of 676,000 PubMed publications. Many earlier publications on multiple imputation can be found in the methodological literature



**Fig. 7.3** Correlation between haematocrit (ht) and haemoglobin (Hb) in 566 patients with traumatic brain injury. The final imputation model included ht ( $p < 0.001$ ) and gender ( $p = 0.01$ ), with  $R^2$  of 0.97

**Table 7.3** Approaches to dealing with missing values, including imputation methods

Label	X/Y used?	Approach
CC	—	Complete case analysis; subjects with missing values are excluded from the analysis
CM	X	Single imputation with the conditional mean. The conditional mean can, e.g. be estimated with a regression model
SI	X+Y	Single imputation with a random draw from the predictive distribution from an imputation model (“stochastic regression imputation”)
MI	X+Y	Multiple imputation with a random draw from the predictive distribution from an imputation model, repeated, e.g. 5 times

measurements (Fig. 7.3). The predicted Hb can subsequently be filled in for those patients with ht available but Hb missing. This is an example of a regression imputation approach (Table 7.3). In this example, it appears that the correlation between Hb and ht is very strong.

### 7.3.2 Simple and More Advanced Single Imputation Methods

Simple imputation methods include substitution of a missing value of a continuous predictor with the mean, or the most frequent category for a categorical predictor. Such simple methods ignore potential correlation of the values of predictors among

each other, and are hence suboptimal. Further they lead to an underestimation of variability in the predictor values among subjects.

Regression imputation,<sup>114</sup> or “conditional mean imputation,”<sup>263</sup> does consider the correlation among predictors. An imputation model is made to predict the missing values (see for example, Fig. 7.3). Expected values can then be imputed reflecting the correlations in the data. An alternative is to take a random draw from the distribution of predicted values (“stochastic regression imputation”<sup>114</sup>). The random element reflects that the imputed values are not certain, which is especially important in the case of relatively uncertain predicted values.

Simple, conditional mean, and stochastic regression imputation methods are examples of *single* imputation methods. In contrast, Rubin proposed a *multiple* imputation method for handling of missing data.<sup>357</sup>

### 7.3.3 *Multiple Imputation*

With multiple imputation,  $m$  completed data sets are created instead of a single completed data set. Missing values are imputed  $m$  times using  $m$  independent draws from an imputation model. As with (stochastic) regression imputation, the imputation model aims to reasonably approximate the true distributional relationship between the missing data and the available information. This means that for each variable with missing data, a conditional distribution for the missing data can be specified given other data.<sup>100</sup>

A problem with imputation models is that we may want to predict missing values for one predictor, using other predictors which also have missing values. Fortunately, this may well be solved with data augmentation methods, which follow an iterative process of an imputation step, which imputes values for the missing data, and a posterior step, which draws new estimates for the model parameters based on the previously imputed values.<sup>370</sup> This process continues until convergence. The final imputed values are used as the first imputed data set. The whole process is repeated with different starting values to obtain the  $m$  imputed data sets. The variation among the  $m$  imputations reflects the uncertainty with which the missing values can be predicted from the observed data. Further details of this procedure are, e.g. provided by Schafer.<sup>370</sup>

After creating  $m$  completed data sets,  $m$  analyses are performed by treating each completed data set as a real complete data set. Standard procedures and software can be used, as we would for a data set without any missings.

Finally, the results from the  $m$  complete-data analyses are combined, for example to obtain the estimates of regression coefficients and performance estimates, while properly taking into account the uncertainty in the imputed values. Point estimates are simply the average over the  $m$  imputed data sets (and could in principle also have been obtained in a large stacked data set instead of  $m$  separate data sets). The variance of the estimates (e.g. regression coefficients, performance measures) is the average of the variance as estimated within the  $m$  imputed data

sets plus the variance between these  $m$  data sets. The latter element is the essential difference between single or stochastic regression imputation, and multiple imputation. MI takes the uncertainty into account that is caused by having to estimate an imputation model. The formula for MI results is relatively straightforward. For an estimated regression coefficient  $\beta$ , the variance over  $M$  imputed data sets is

$$\begin{aligned}\text{Var}(\beta) &= \text{var}(\beta) \text{ within } m + (1 + 1/M) \text{ var}(\beta) \text{ between } m \\ &= \text{mean} (\text{var}(\beta_m)) + (1 + 1/M)(1/(M-1)) \sum (\beta_m - \text{mean} (\beta))^2,\end{aligned}$$

where  $m = 1 \dots M$  imputed data sets

This formula (or closely related variants) are implemented in many software packages that can perform MI. The number  $M$  for the imputed data sets is usually set to 5 or 10. If  $M = 10$ , the mean variance estimates within the  $m$  imputed data sets are the dominant factors in the formula, since the term  $(1+1/M)$  becomes 1.1 and the “between imputation” variance is usually much smaller than the “within imputation” variance. Rubin found that when 50% of the data are missing for a predictor, an estimate based on five imputations has a standard deviation (SD,  $\sqrt{\text{var}}$ ) that is only 5% wider than a variance estimate based on an infinite number of imputations.<sup>358</sup> Other simulation studies have shown that  $M$  can be as low as 3 for data with 20% missingness.<sup>447</sup> Setting  $M$  to 1 makes MI a single stochastic regression imputation procedure (SI).

The most important step in any imputation procedure is the definition of the imputation model. We therefore discuss this step in more detail, largely following guidelines provided by Van Buuren et al.<sup>447</sup>

## 7.4 Defining the Imputation Model

The imputation model aims to approximate the true distributional relationship between the unobserved data and the available information. The imputation model is an explicit attempt to model the MAR process. Imputation models can be specified for each potential predictor with missing data, irrespective of the quantity of missing data. Two modelling choices usually have to be made: the form of the model (e.g. linear, logistic, polytomous) and the set of variables that enter the model, including potential transformations of predictors.

For binary predictor variables (e.g. the presence or absence of a patient characteristic) it is convenient to use a logistic model, for categorical variables with three or more levels a polytomous logistic model, and for continuous variables a linear regression model. A problem may arise when imputations are constructed that are outside the observed range of values. In such cases, it may be reasonable to truncate imputed values, so that they remain within a plausible range.

The variables in the imputation model can be differentiated in various categories. All predictors that appear in the prediction model should be included in the

imputation models. Failure to do so may bias the analysis. Next, some variables that do not appear in the prediction model may serve as auxiliary variables. For example, calendar time or geographic site may be associated with missingness and should be considered for the imputation model. Finally, we need to include the outcome that we consider in the prediction model. This may appear a bit circular, since the aim of a prediction model is to predict the outcome. However, not including the outcome in the imputation model may cause substantial bias in the MI analysis of prognostic effects, even in the MCAR situation. Simulation studies have shown that a severe dilution of the predictive effects may occur if the outcome is not included in the MI procedure. If 50% of the data is MCAR, omitting the outcome approximately halves the estimated regression coefficient. We can even consider to include other types of outcomes if these are available.

It has been observed that including many variables in the imputation model tends to make the MAR assumptions more plausible. Putting in noise variables does not harm the imputation process,<sup>77</sup> unless computational problems arise because of multicollinearity and inclusion of many predictors with missing data.<sup>73</sup> It is therefore generally convenient to include all predictors, some auxiliary variables, and the outcome in the imputation model for an MI procedure.

### \*7.4.1 *Transformations of Variables*

A difficult topic is how transformations among  $X$  variables and between  $X$  and  $Y$  should be handled. Current software for MI deals with this issue in different ways. The `mice` function assumes linearity of associations among  $X$  variables and between  $X$  and  $Y$  in the default setting. Specific forms of imputation models can be specified by the user, using for example  $x + x^2$  for some  $X$  variables. In contrast, the advanced `aregImpute` function may search for transformations among variables such that the correlations are maximized. If non-linear associations are present, this may be of advantage. However, it is inefficient in the case of linear associations. The default settings can be changed such that `aregImpute` resembles `mice`. Indeed, very similar results in simulations between `mice` and `aregImpute` have been found when linearity was enforced (using the identity function ("I") in `aregImpute`).

### 7.4.2 *Imputation Models for SI*

For single imputation with the conditional mean (e.g. from a regression model), only the  $X$  variables should be used in the imputation model.<sup>174,445</sup> If the outcome  $Y$  is also used in the imputation model, we exaggerate the strength of relationships between predictors and outcome in the prediction model. In contrast, stochastic regression imputation should be performed with the outcome. This is

because a random element is added to the predicted values from the imputation model, similar to an MI procedure (see Table 7.3).

### 7.4.3 Summary Points

Imputation models for a multiple imputation (MI) procedure need to include

- All predictors considered in the prediction model
- Auxiliary variables, related to the predictors, but not included in the prediction model
- The outcome considered in the prediction model

For single conditional mean imputation, the outcome  $Y$  should not be included in the imputation model (Table 7.3).

## \*7.5 Simulations of Imputation Under MCAR, MAR, and MNAR

Some simulations are presented at the book's website to illustrate benefits and limitations of imputation for predictive regression models. We consider estimates of regression coefficients (bias and precision) and estimates of predictive performance.

A first simple situation considered predictors  $X_1$  and  $X_2$  for a continuous outcome  $Y$ , with missing value mechanisms (MCAR, MAR, and MNAR as in Fig. 7.1).  $X_1$  and  $X_2$  were generated as uncorrelated and correlated predictors. Simulations confirm that a CC analysis gives quite reasonable estimates of the two regression coefficients under most missing value mechanisms (MCAR, MAR on  $x$ , MNAR). The problematic situation was MAR on  $y$ , where stochastic SI or MI had clear advantages. Both use  $X + Y$  for imputation of missing values. In the MNAR situation, CC analysis led to unbiased regression coefficients, in contrast to all other approaches.<sup>445</sup> CC analysis did however not give a good impression of the predictive performance of the regression model in the original population. This occurs since the performance of the model with CC analysis is assessed on a selection of subjects in the MAR on  $x$ , MAR on  $y$ , and MNAR situations. The limited spectrum of subjects in the CC analysis led to lower estimates of the predictive performance as quantified by an adjusted  $R^2$  statistic.

Overall, MI gave good results in this simple simulation study: regression coefficients were at least as well estimated as a CC analysis, and the estimated SEs were quite correct. As a next best, stochastic SI was reasonable, with slightly poorer estimation of regression coefficients, but good estimation of predictive performance. Both SI and MI rely on MCAR, MAR on  $x$ , or MAR on  $y$ . Under MNAR, CC analysis is unbiased for the regression coefficients, but underestimates predictive performance for the original, complete data.

### \*7.5.1 Multiple Predictors

In prognostic analyses, we usually study more than two predictors. We re-consider the situation of Table 7.1, where 250 subjects have 1 missing value, and 250 have fully complete data for 5 predictors. Regression coefficients were set to 1 for all 5 predictors.

A CC analysis uses only 250 subjects. Regression coefficients are unbiased, but have considerably more variability than the estimates from an MI procedure (Table 7.4). The conditional mean (CM) and stochastic SI procedures perform quite similar to MI. All approaches correctly estimate the predictive performance as an adjusted  $R^2$  around 35%.

**Table 7.4** Regression under different missing value mechanisms, and the effect of imputation procedures. Results are means over 1,000 repetitions of samples with 500 subjects. The square root of the mean squared error is highlighted in bold for the strategy with the best result in dealing with missing values

	Table 7.1 $\beta \pm$ SE; sqrt(MSE)	Adj $R^2$	Mix of mechanisms $\beta \pm$ SE; sqrt(MSE)	Adj $R^2$
<b>X1-X5 correlated</b>	<b>10% missing (total 50%)</b>		<b>20% missing (total 75%)</b>	
Original data, no missings		35%		35%
$\beta_1$	1.00±0.18; 0.18		1.00±0.18; 0.18	
$\beta_2$	1.01±0.19; 0.19		1.00±0.19; 0.19	
$\beta_3$	1.00±0.19; 0.20		1.00±0.19; 0.19	
$\beta_4$	0.99±0.19; 0.20		1.00±0.19; 0.20	
$\beta_5$	1.00±0.20; 0.20		1.00±0.20; 0.20	
Complete case analysis		35%		19%
$\beta_1$	1.00±0.26; 0.26		0.66±0.36; 0.49	
$\beta_2$	1.03±0.27; 0.27		0.66±0.38; 0.51	
$\beta_3$	0.98±0.27; 0.27		0.69±0.33; 0.45	
$\beta_4$	1.00±0.28; 0.28		0.68±0.33; 0.47	
$\beta_5$	1.00±0.28; 0.28		0.67±0.39; 0.52	
Conditional mean with $X$		33%		27%
$\beta_1$	0.99±0.19; <b>0.19</b>		1.08±0.21; 0.23	
$\beta_2$	1.01±0.20; <b>0.21</b>		0.75±0.23; 0.32	
$\beta_3$	0.99±0.21; <b>0.21</b>		1.05±0.23; 0.24	
$\beta_4$	0.98±0.21; <b>0.21</b>		1.07±0.24; 0.25	
$\beta_5$	0.99±0.21; <b>0.22</b>		1.03±0.28; 0.29	

(continued)

**Table 7.4** (continued)

	Table 7.1 $\beta \pm$ SE; sqrt(MSE)	Adj $R^2$	Mix of mechanisms $\beta \pm$ SE; sqrt(MSE)	Adj $R^2$
SI with $X+Y$			36%	35%
$\beta_1$	1.00±0.18; 0.22		1.03±0.18; 0.25	
$\beta_2$	1.01±0.19; <b>0.21</b>		1.02±0.19; 0.30	
$\beta_3$	1.00±0.19; 0.23		1.03±0.19; 0.27	
$\beta_4$	1.01±0.19; 0.22		1.03±0.20; 0.27	
$\beta_5$	1.00±0.20; 0.23		1.02±0.23; 0.34	
Mice with $X+Y$		35%		35%
$\beta_1$	1.00±0.19; <b>0.19</b>		1.03±0.22; <b>0.22</b>	
$\beta_2$	1.02±0.20; <b>0.21</b>		1.02±0.26; <b>0.26</b>	
$\beta_3$	1.00±0.21; <b>0.21</b>		1.02±0.23; <b>0.24</b>	
$\beta_4$	0.99±0.21; <b>0.21</b>		1.03±0.23; <b>0.24</b>	
$\beta_5$	0.99±0.21; <b>0.22</b>		1.02±0.28; <b>0.29</b>	

Model:  $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \text{error}$  For the 10% missing example, all  $X$  variables were independent standard normal, and error  $N(0,4)$ . 10% MCAR per variable were created as in Table 7.1. For the second example,  $X$  variables were correlated:  $X_1 \sim N(0,1)$ ;  $X_2 \sim 0.2 \times X_1 + N(0, 0.98)$ ;  $X_3 \sim 0.2 \times X_1 + 0.16 \times X_2 + N(0, 0.97)$ ;  $X_4 \sim 0.2 \times X_1 + 0.16 \times X_2 + 0.14 \times X_3 + N(0, 0.96)$ ;  $X_5 \sim 0.2 \times X_1 + 0.16 \times X_2 + 0.14 \times X_3 + 0.12 \times X_4 + N(0, 0.95)$ ; error  $\sim N(0,4)$ . For each  $x$  variable, 20% missings were created, with MCAR for  $X_1$ ; MAR on  $y$  for  $X_2$ ; MAR on  $X_1$  for  $X_3$ ; MAR on  $X_2$  for  $X_4$ ; and MNAR for  $X_5$ . Covariances of missingness were set at 50%

A more complex situation was also simulated. More missing values were created (20% vs. 10%), with more complex missing value mechanisms for correlated  $X_1$ – $X_5$  (covariance 0.2 for all). MCAR was used for  $X_1$ , MAR on  $y$  for  $X_2$ , MAR on  $x$  for  $X_3$  and  $X_4$ , and MNAR for  $X_5$ . A CC analysis led to biased estimates for all regression coefficients, which can be attributed to the MAR on  $Y$  mechanism for  $X_2$ . Hence, MAR on  $y$  for only one of the five predictors was sufficient to bias all coefficients. Also, the variability was considerable, since only 25% of the subjects were included in the CC analysis. MI did quite well overall. SI was a next best, with slightly poorer estimation of the regression coefficients. In the conditional mean (CM) analysis, the effect of  $\beta_2$  was underestimated, but less so than with a CC analysis. Coefficients  $\beta_3$  to  $\beta_5$  were well estimated. Both the CC and CM analyses underestimate the predictive performance (adjusted  $R^2$  19% and 27% instead of 35%).

## 7.6 Imputation of Missing Outcomes

The outcome for a patient can be missing in several situations. A common situation is that follow-up time is insufficient to observe the outcome for all subjects. Survival analysis techniques deal with this situation by considering these incomplete

observations as “censored.” Also, sometimes an outcome is measured multiple times (repeated measures) at different time points. For example, in TBI studies measurements may be done at 1, 3, 6, and 12 months, while 6-month outcome may be our primary outcome measure (see Chap. 8). In this case, missing 6-month outcomes can possibly be imputed based on measurements at other time points, exploiting the correlations between outcome measurements.<sup>203</sup> Sometimes imputation is based on extrapolation (“last observation carried forward”).<sup>489</sup> This is a simple method to impute missing outcome values when repeated measurements are available, but the method has many problems. It leads to biased regression coefficients and underestimated variability. Multiple imputation or other methods are preferred that make use of the correlation between predictors and outcomes.<sup>440</sup>

A specific situation is that we are interested in a single outcome at a specific point in time, and this outcome is missing in some subjects, e.g. a diagnosis. Missingness makes that we cannot analyse the relationship between predictors and this outcome, while this relationship is of primary interest. In principle it is possible to impute the outcome, similar to imputation of predictors. An imputation model does not “know” what is  $X$  and what is  $Y$ . The distribution of the outcome will reflect the relationships of the predictors with the outcome. Imputation will hence not provide new information on these relationships, and is therefore generally not useful for the purpose of better estimation of regression coefficients.

On the other hand, we can imagine that certain parameters are of interest that benefit from imputation of the outcome. For example, we may be interested in the mean prevalence of a diagnosis, while the diagnosis is not available for all subjects. The reference standard may selectively be not determined, which leads to verification bias. If the diagnosis was missing especially for low risk subjects, a better prevalence estimate is obtained after imputation of the missing diagnoses. Finally, performance measures such as  $R^2$  or ROC area may better be estimated after completing the outcome for the whole spectrum of subjects. In prediction research, subjects with missing outcome data are generally discarded. We have therefore focused on dealing with missing values in predictors.

## 7.7 Guidance to Missing Values in Prediction Research

We provide some guidance for dealing with missing values and imputation in prediction research, based on previous research, findings in simulations (Table 7.4) and practical considerations.

### 7.7.1 Patterns of Missingness

As a preliminary step, it is recommended to investigate the missing data patterns.<sup>174,444</sup>

1. We need to examine how many missings occur for each potential predictor; this examination is part of the basic approach to any data analysis. Missing values are easily noted when examining frequency distributions of the predictors.
2. We want to know whether predictor values are correlated with missingness of other predictors; this determines how well we may be able to impute a missing value, and how useful the remaining information on subjects without missing values is. We may also study associations with auxiliary variables, such as calendar time and site. Patterns of missing data can be visualized with cluster analysis methods, e.g. the `naclus` function.<sup>174</sup>
3. Regression tree and logistic regression analysis can be used to assess associations between predictors and missingness of the predictor as the outcome. When associations are identified, the MCAR assumption is violated.
4. As an extension of point 3, it is especially important to assess whether missingness was associated with the outcome. This can easily be assessed by examining outcome, e.g. mortality, by missingness of the predictor (value available/missing). Often we may note a poorer outcome in those with missing values. The first question is whether this association can be explained by observed predictors. Hereto, logistic regression analysis can be helpful, with missingness as the dependent variable, and the outcome  $Y$  and other predictors as covariates. If the study was truly prospective, a missing  $X - Y$  association can only occur through other characteristics; it is logically impossible to have selective missingness on the outcome when the data were collected before the outcome was known. The other characteristics that mediate the observed missingness-outcome association may be known; this is an MAR on  $x$  situation. If some of the mediating predictors are not known, or measured imprecisely, some kind of residual confounding occurs, leading to an MNAR situation. Imputation with  $Y$  may at least partly resolve this situation.
5. Subject matter knowledge should be used to judge plausible mechanisms for the missing values, for example whether MNAR is plausible. The MCAR assumption can be tested, and may often be rejected in medical research. But the MAR assumption cannot be tested, and MNAR hence always remains a possibility.

### 7.7.2 *Simple Approaches*

A historically popular method in epidemiological research was to create a category “missing” for missing values in the regression analysis. Such a “missing indicator method” is especially straightforward for categorical predictors. For example, we can recode a predictor that was incompletely recorded as “absent,” “present,” and “missing.” However, such a procedure ignores correlation of the values of predictors among each other. Simulations have shown that the procedure may lead to severe bias in estimated regression coefficients.<sup>155,294</sup> The missing indicator should hence generally not be used. An alternative in such a situation might be to change

the definition of the predictor, i.e. by assuming that if no value is available from a patient chart, the characteristic is absent rather than missing.

If the missing values are among many predictors, or if we aim to include a predictor with many missing values anyway, MI may be considered as generally preferable to a simple CC analysis. When using an imputation model, we have to assume an MAR mechanism. The MAR mechanism becomes more reasonable when more detailed characteristics are included in the imputation model. Also, the form of associations (e.g. linearity for continuous predictors) in the imputation model has to be adequate.

### ***7.7.3 Maximum Fraction of Missing Values Before Omitting a Predictor***

When we are interested in the specific effect of a predictor, the “face validity” of an analysis is higher with fewer missing values. If a substantial number of missing values occur specifically in one predictor, it may be convenient to omit this predictor from the analysis. Especially when the predictor is of primary interest, it would not be natural to impute the missing values. For example, when we had missing treatment allocation for some patients in a randomized controlled trial (RCT), we would never impute these missing values.

It is difficult to provide a guide to what is still an acceptable number of missing values. Evidence for selective missingness (e.g. MAR on  $y$ ) may already make a CC analysis of a predictor with 10% missings suspect; in other cases 20% missingness may be quite acceptable (e.g. MCAR assumed).

Theoretically, MI solves any missing data problem, as long as we correctly model the missing data mechanism, and do not have an MNAR situation. So, the effect of a predictor with 90% missing values could still be estimated, but with relatively large uncertainty.

In practice, we can only approximate the missing data mechanism. Effects of predictors with more than 50% missings in a specific data set will generally be distrusted. Such predictors might hence be discarded. Other considerations may include the reasons for missingness. If missings occur because of the study design, we may be less worried in interpreting findings based on a relatively limited set of known values. For example, in the TBI case study (Chap. 8), missing values occurred especially because some studies included in the meta-analysis did simply not record the predictor.

### ***7.7.4 Single or Multiple Imputation for Predictor Effects?***

In prediction research, we may generally think of studying effects of predictors of specific interest (univariate and adjusted analyses) and of studying predictions

**Table 7.5** Dealing with missing values to estimate predictor effects

Analysis	Predictor of interest $X_1$ (e.g. motor score)	Confounders $X_2$ – $X_i$ (other predictors)
Univariate analysis	Ignore	–
Adjusted analysis	Ignore	SI/MI

(deriving prognostic equations, evaluating model performance). We usually start with a univariate analysis of predictive effects, e.g. a cross-tabulation of a predictor with a binary outcome or with time-to-event in a Kaplan–Meier survival analysis. Equivalently, we can calculate the regression coefficients in a univariate logistic or Cox regression to obtain estimates of predictor effects. A complete case analysis is the most obvious approach. In Table 7.5 this is indicated as ignoring incomplete records for variable  $X_1$ . An example may be that we are interested in the prognostic effect of the motor score from the Glasgow outcome scale in traumatic brain injury (see Chap. 8).

Next, we are often interested in adjusted effects, i.e. the effect of  $X_1$  corrected for correlation with other variables ( $X_2$  to  $X_i$ ). The variables  $X_2$  to  $X_i$  are considered as confounders, since they may be associated with the outcome and with  $X_1$ . Such an adjusted analysis may well be done with imputation of missing data for the confounders ( $X_2$  to  $X_i$ ), but without imputation of  $X_1$ . This ensures comparability with the univariate analysis, because numbers will be the same in univariate and adjusted analyses. SI will underestimate the variability in the adjusted regression coefficient. An MI procedure for the confounding variables results in better estimates of the variability in the adjusted regression coefficient. This issue will be illustrated in Chap. 8.

An alternative is to perform univariate and adjusted analyses with imputed data, both for the predictor of interest  $X_1$  and the confounders  $X_2$  to  $X_i$ . Many medical researchers will however appreciate univariate analyses that stay closer to the observed data, at least as an initial analysis. In general we should be careful in interpreting a univariate effect. Confounding may cause the effect of a predictor to appear too extreme (because of positive associations with other predictors) or too small (because of negative associations with other predictors).

### 7.7.5 Single or Multiple Imputation for Predictions?

To derive predictions, MI may be the best approach. However, given that not all analysts are familiar with these methods, some next best strategies can be envisioned. Especially, single imputation may be a good alternative, if a good imputation model is used. A stochastic SI data set can easily be created as the first of a series of MI data sets. Every investigator can easily work with such a SI data set, and does not have to bother with the combination of results over different MI data sets. More experienced data analysts may consider this advantage trivial. The conditional

mean approach could also be followed, but has some problems in an MAR on y situation. The GUSTO-I data set, which is used as an example throughout this book, is a CM data set, with at most 8% imputed values for some of the predictors.<sup>255</sup>

The primary disadvantage of stochastic SI is the underestimation of the uncertainty associated with imputed values. A second disadvantage is less stability in the point estimates, because of the random element in stochastic SI. These disadvantages are less relevant with relatively few missing values, and in large data sets. MI may be preferable with relatively small data sets (for example with less than 100 events), since imputations will vary considerably from imputation to imputation.

To derive predictions for individual subjects in a data set, it is often advisable to impute missing data for all predictors. An exception is the situation that we know that we cannot obtain complete data in future applications. It may then be reasonable to develop the prediction model in a selection of subjects where the data will be available in the future. SI may often be sufficient to provide reasonable point estimates of predictor effects in a multivariable analysis, and hence SI is also sufficient to obtain reasonable predictions. The estimated predictions are of primary interest, and the underestimation of uncertainty by SI is less relevant. Performance measures such as discrimination and calibration can readily be calculated after SI.

Finally, we may want to present the prognostic model in a simple form for practical application (Chap. 18). A score chart based on rounded coefficients is well obtainable with SI. MI will provide better estimates of variability of the scores, but variability is only of secondary interest, if presented at all. MI may therefore have only a minor advantage over SI for model presentation. In summary, multivariable analysis, performance estimation, and model presentation can all be done with SI or MI approaches.

### **7.7.6 Reporting of Missing Values in Prediction Research**

A review was performed by Burton and Altman of 100 prognostic studies in cancer that were published in 2002.<sup>61</sup> Missing values were present in 81 studies, and led to the exclusion of >10% of the patients available for multivariable analyses in about 50% of the studies. The most common technique was complete or available case analysis. Some studies omitted a predictor because of many missing data, or included a separate category for missingness. Three papers applied some form of single imputation, and only one applied multiple imputation. Hence, dealing with missing data was suboptimal in many studies, and (multiple) imputation was only infrequently used yet.

Suggested reporting guidelines include three major issues<sup>61</sup>:

1. Quantification of the completeness of predictor data
2. Approaches to dealing with missing predictor data (including imputation methods)
3. Exploration of the missing data (including results for complete case and completed case analysis, Table 7.6)

**Table 7.6** Guidelines for reporting of prognostic studies with missing predictor data<sup>61</sup>

Issue	Aspect
Quantification of completeness	If completeness of data is an inclusion criterion, specify numbers excluded Provide total $n$ and $n$ with complete data Report frequency of missingness for every predictor
Approaches to dealing with missing data	Provide sufficient details on the methods used, including references if imputation was done Specify the $n$ of patients and number of events for all analyses
Exploration of missing data	Discuss reasons for missingness Present comparisons of characteristics between cases with and without missing data

**Table 7.7** Imputation methods as applied in some examples

Method	Characteristics	Example
Simple imputation	Mean or most frequent category	Guillain-Barré: few missing <sup>461</sup>
Conditional mean imputation	Estimate predicted value based on correlations between predictors	Historical examples: GUSTO-I, <sup>255</sup> ReHiT study <sup>417</sup>
Stochastic regression imputation	Draw imputed value from distribution of predicted values	Adjusted analysis in IMPACT study <sup>305</sup>
Multiple imputation	Develop imputation model and draw imputed value from distribution of predicted values; combine estimates over $m$ imputed data sets	Ovarian cancer, <sup>73</sup> testicular cancer <sup>453</sup>

We list some examples on dealing with missing values in Table 7.7. Methods include single imputation (simple, conditional mean, stochastic regression), or multiple imputation. More details of these studies are provided at the book's website.

## 7.8 Concluding Remarks

Missing values pose important challenges in prediction research. Straightforward methods such as complete case analysis are often oversimplistic from a methodological point of view. Simulations support the view that imputation methods are superior to complete case analysis, which is currently still the dominant approach in the medical literature. Some case studies also illustrate benefits of imputation.<sup>73,453</sup> Possibly, many clinicians are unaware of the problems with CC analyses, and of modern developments in this area, especially on (multiple) imputation methods. The inclusion of the outcome in a multiple imputation process may be met with skepticism, while it is clearly necessary.

The best solution for missing values is of course to ensure that no data are missing. It may sometimes be possible to retrieve missing data by going back to medical charts. In some settings, it may be reasonable to define missing as “No.” If characteristics are measured multiple times, we may sometimes use a measurement from another time point. If missing values do occur, they have to be dealt with in a reasonable way, i.e. such that the research questions are addressed efficiently.

The research question is not to estimate the missing values but to estimate model parameters (univariate effects, adjusted effects, multivariable effects, prognostic equations, performance). These parameters should be valid for the population where the model will be applied in the future. The sample serves to learn for this future application, and we should try to use all available information. Imputation of some missing values prevents that we throw away useful information recorded for other predictors. The primary benefit of imputation is hence an increase in power to detect prognostic effects, and in deriving better predictions. A second benefit of imputation is comparability of results over analyses. The price we pay for these benefits is making the assumption of MAR; we need to include all variables (predictors, outcome, and auxiliary variables) that are potentially correlated with the missingness of the predictor.

As in any statistical analysis, the sensible judgment of the analyst is important, based on subject knowledge and the research question. Comparing results of complete case and completed case analyses may be informative, and together with a judgment about the plausibility of assumptions in a particular situation we can decide on which is the primary analysis.

### 7.8.1 *Summary Statements*

- Missing values in predictors are common, and lead to inefficiency, difficulties in comparing results between analyses with different numbers, and potentially biased regression coefficients and predictions.
- Theoretical analyses and simulations conclude that imputation methods, especially multiple imputation, are superior to complete case analyses.
- Advanced stochastic single imputation methods, based on the first data set of a multiple imputation sequence, are also reasonable to address prediction questions.
- Imputation methods make the assumption of MAR; more specifically, MAR given the information used in the imputation process.
- The MAR assumption is not testable, but becomes more reasonable with imputation models that include a wide range of characteristics, including predictors, the outcome, and auxiliary variables.

### \*7.8.2 *Currently Available Software and Challenges*

Multiple imputation software is widely available nowadays, and further improvements may be expected during the coming years. For R and S+, the `mice` library is freely available (developed by Van Buuren et al.). It includes state of the art functions, has flexible settings, but the computation time can be substantial. An interesting alternative is the `aregImpute` function developed by Harrell, which performed well in a number of assessments. Stata has the sophisticated `ice` and `ice2` functions, which were developed by Royston. Some packages, such as the MVA module in SPSS v11 and `proc MI` in SAS v8 have limitations, since they make stronger parametric assumptions about the multivariate distribution of the data. With any imputations, we should as a minimum check distributions of the observed and imputed values, e.g. by histograms.

Several methodological challenges may require further study:

- If some sort of selection process is done, e.g. stepwise selection, how can this be combined with imputation? (Chap. 11)<sup>488</sup>
- How should we perform internal validation, e.g. bootstrapping, when missing values are imputed? We propose to validate the modelling process within each imputed data sets (Chap. 17).
- How should we perform external validation, when some predictors are missing? Impute based on an imputation model from the development setting, impute based on the validation data? (Chap. 19)
- How large are the advantages of MI over SI in prognostic research?
- How tenable is the MAR assumption in practical examples? What is the influence of non-linear relationships between predictors and/or outcome?

## Questions

### 7.1 Missing values vs. incomplete cases

- (a) How many values are missing from the required values for a model with 3 predictors, estimated in 1000 subjects, where predictor 1 has 100, predictor 2 has 200, and predictor 3 has 400 missing values?
- (b) If the missing values occur completely at random, how many subjects would approximately be discarded in a complete case (CC) analysis?

### 7.2 MCAR, MAR, or MNAR?

Consider a prognostic study among patients undergoing heart valve surgery aiming to quantify the predictive value of intra-operative characteristics (e.g. intra-operative blood pressure and complications) for mortality after 30 days (outcome). What type of missingness pattern do we have in the following two situations:

- (a) Among patients who actually developed an intra-operative complication, the intra-operative data are often missing?
  - (b) Among patients with a less severe indication for surgery based on presurgical data, the intra-operative data are missing?
- Suppose that clinicians do not perform a diagnostic test if their impression is that the patient does not have the diagnosis of interest. This impression may partly be captured by clinical variables that are observed, but also depend on some predictors that are not registered in the data.
- (c) Is this an MAR or MNAR situation?

### 7.3 MAR on $y$ in Fig. 7.1

Why does a missing value mechanism of MAR on  $y$  in  $x_1$  result in bias both for  $\beta_1$  and  $\beta_2$  in Fig. 7.1?

### 7.4 Problems of overall mean imputation

What is the effect of performing overall mean imputation (i.e. imputing the mean of the observed values for the missing values) on estimated regression coefficients and standard errors?

### 7.5 Imputation with outcome (“ $Y$ ”, Table 7.3)

Consider 50% missingness for a predictor  $X_1$  which is not related to other predictors. We recommend to perform SI or MI with the outcome as one of the variables in the imputation model (Table 7.3). What would happen to the univariate regression coefficient of  $X_1$  if a completed data set was analysed, where values were imputed without using  $Y$ ?

### 7.6 Complete case analysis or imputation?

- (a) For what missingness patterns is complete case analysis a reasonable solution?
- (b) In what respect is multiple imputation preferable above single imputation?

# Chapter 8

## Case Study on Dealing with Missing Values

**Background** A case study is presented on prognostic modelling in patients with moderate and severe traumatic brain injury (TBI). Individual patient data from several studies were available to quantify predictor effects and to develop and validate prognostic models. Missing values were a key issue, since few studies recorded all predictors of interest. The use of single and multiple imputation methods is illustrated with a detailed description of the analyses in R software.

### 8.1 Introduction

#### 8.1.1 Aim

Randomized controlled trials (RCTs) in TBI are complex due to the heterogeneity of the population. None of the multicentre RCTs conducted in this field over the past decades have convincingly shown benefit of new therapies in the overall population.<sup>273,310</sup> The overall aim of the study was to optimize the methodology of randomized clinical trials in the field of TBI, such that chances of demonstrating benefit with an effective new therapy or therapeutic agent would be maximized. This NIH sponsored project was labelled IMPACT: International Mission on Prognosis and Analysis of Clinical Trials in TBI.<sup>271</sup> Individual patient data from recent trials and observational studies were available.

Prognosis was central to the aims of the project. For example, prognostic models can be used for the efficient selection of patients (excluding those with an extreme prognosis, either very poor or very good) and for covariate adjustment of the treatment effect (with several advantages as described in Chap. 2).<sup>189</sup> In TBI, outcome is commonly assessed with the Glasgow outcome scale (GOS), which is an ordinal scale (Table 8.1).<sup>218</sup> The scale ranges from dead, through vegetative state, severe disability to moderate disability, and good recovery. In conventional analyses, the GOS is often dichotomized as mortality vs. survival (category 1 vs. 2–5), or as

**Table 8.1** Definition of the Glasgow outcome scale<sup>218,491</sup>

Category	Label	Definition
1	Dead	Mortality from any cause
2	Vegetative	Unable to interact with environment; unresponsive
3	Severe disability	Conscious but dependent
4	Moderate disability	Independent, but disabled
5	Good recovery	Return to normal occupational and social activities; may have minor residual deficits

unfavourable vs. favourable (category 1, 2, 3 vs. category 4, 5), although it is preferable to exploit the ordinal nature of this scale. One approach is the “sliding dichotomy” analysis, in which the split for dichotomization of the GOS is differentiated according to the baseline prognosis established prior to randomization.<sup>304</sup> Another approach is to use a proportional odds model for the GOS as an ordered outcome (see Chap. 4).

We aimed to predict the dichotomized 6-month GOS. Missing data were a key problem in the prognostic analysis.<sup>283</sup> We focus on approaches for dealing with missing data.

### 8.1.2 Patient Selection

Our focus was on patients with severe TBI (Glasgow coma score, GCS 3–8), but cohorts that included patients with moderate TBI (GCS 9–12) were also considered. The GCS is a measure for the level of consciousness. An individual patient data meta-analysis of 11 studies was performed, including 8 randomized controlled trials (RCTs), and 3 relatively unselected prospective surveys, with the potential for analysing data on 9,205 patients. Complete outcome data were available for 8,719 of the 9,205 patients (95%). We further excluded children, leaving 8,530 patients for analysis. The studies are arbitrarily designated as A to K in Table 8.2. The meta-analysis was a continuation of analyses of two related RCTs (Tirilazad, Table 8.2: study ID A and B).<sup>203</sup>

### 8.1.3 Selection of Potential Predictors

Extensive univariate analyses were performed within the IMPACT study of potential predictors. In combination with a review of the literature we identified predictors for further multivariable analyses.<sup>305</sup> These predictors included demographic characteristics (age),<sup>306</sup> injury details (cause of injury),<sup>62</sup> secondary insults (hypoxia and hypotension),<sup>284</sup> clinical measures of injury severity (Glasgow coma scale and pupillary reactivity),<sup>276</sup> characteristics of the admission CT scan,<sup>272</sup> and laboratory values.<sup>446</sup> For prognostic modelling, a core set of three strong predictors emerges from the literature since the 1970s, consisting of age, motor score, and pupillary

**Table 8.2** Availability of predictor values by study (A–K), as included in the IMPACT study ( $n=8,530$ )<sup>271</sup>

Study	A	B	C	D	E	F	G	H	I	J	K	Total
N	1,118	1,041	409	919	1,510	350	812	604	126	822	819	8,530
<i>Core predictors</i>												
Age (%)	100	100	100	100	100	100	100	100	100	100	100	100
Motor score (%)	100	100	100	100	100	100	100	100	100	100	100	100
Pupils (%)	93	95	97	0	98	100	92	100	0	96	99	85
<i>Secondary insults</i>												
Hypoxia (%)	88	89	100	93	0	0	98	100	67	99	0	64
Hypotension (%)	97	97	0	93	0	98	99	100	83	99	100	75
<i>CT</i>												
CT class (%)	99	99	100	99	0	0	0	0	100	98	99	61
tSAH (%)	97	95	99	99	100	73	0	87	100	95	100	87
EDH (%)	98	99	0	99	100	100	95	0	100	100	100	87
Cisterns (%)	89	87	99	99	0	0	0	86	100	0	0	45
Shift (%)	89	88	99	99	100	0	0	89	100	0	100	73
<i>Laboratory values</i>												
Glucose (%)	96	99	0	95	96	85	0	0	98	0	0	57
Sodium (%)	98	96	0	96	96	95	0	64	98	0	0	62
Hb (%)	99	98	0	90	30	97	0	0	93	0	0	45
Platelets (%)	0	0	0	90	29	0	0	40	93	0	0	19
Prothrombin time (%)	0	0	0	0	29	0	0	48	91	0	0	10

reactivity. We subsequently expanded this core model to a 7-predictor model by including secondary insults and CT characteristics (CT classification, traumatic subarachnoid haemorrhage).<sup>203</sup> Further modelling studies were performed with inclusion of more predictors, but are omitted here.

#### \*8.1.4 Coding and Time Dependency of Predictors

An important issue was the definition of predictors across the 11 studies. Definitions varied between data sets. The data extraction was guided by a data dictionary and original study documentation, which standardized the format of variables entered into the pooled data set. A consistent set of categories for coding was sought for each variable by collapsing more extensive codings into a simpler format. For example, the presence of hypoxia on admission was collapsed into a binary coding present/absent, although some data sets contained a more detailed coding as No/Suspect/Definite. “Cause of injury” raised this same issue but in a more complex form, since many and different categories were considered per study.<sup>276</sup>

A further issue was related to the time of measurement of a predictor. We aimed to consider predictors that would be available when patients were to be enrolled in an RCT, in line with the overall aim of the project. An interesting example is the

motor score, which is the prognostically most important element of the GCS. Four time points for assessment were defined: pre-hospital, first hospital (in case of secondary referral), admission, and post stabilization. Most data sets had data for at least two of these time points. For prognostic analysis we aimed to select the latest reliable assessment on admission to correspond with a baseline assessment prior to randomization, i.e. the post-stabilization score. If this was missing we used the next reliable value going back in time (admission, first in-hospital, pre-hospital). However, sometimes the motor score is not clinically obtainable because of early sedation or paralysis, required for artificial ventilation. The motor score was then coded as a separate category (“9,” untestable), rather than considered as a missing value. This approach made the motor score available for all patients.

It can be debated whether a more formal analysis should have been used for defining the baseline motor score; e.g. a multiple imputation procedure might have considered all four time points of the motor score, providing a formally imputed post-stabilization motor score. MI might also have provided estimates for the untestable patients (“category 9”). However, the necessity for sedation and paralysis is related to the severity of injuries. In this specific case, missingness in the sense of “untestable” may possibly be of prognostic relevance, and imputation of a virtual motor score for “untestable” patients was hence not considered appropriate.

## 8.2 Missing Values in the IMPACT Study

Missing values were present in the outcome and in predictors. We discuss dealing with both below.

### 8.2.1 Missing Values in Outcome

Data on 6-month outcome were available for 10 of the 11 studies. For one however, only the 3-month GOS was measured (study E). Since the GOS is assumed to be relatively stable between 3 and 6 months, we imputed missing 6-month GOS with the 3-month GOS. This approach is consistent with the way in which missing outcome had been imputed in a small number of patients in the individual studies (Last Value Carried Forward approach). We chose not to further attempt imputation of the 6-month GOS in the 5% of patients in whom outcome remained missing, as not to compromise the interpretation of our outcome measure.

A more formal MI procedure could have been followed, incorporating the GOS patterns over time as available in some of the studies (e.g. 1, 3, 6, 12 months), and correlations with predictors.

### 8.2.2 *Quantification of Missingness of Predictors*

Table 8.2 summarizes the availability of predictors within the 11 studies of the IMPACT database. The main reason for missingness was absence of a predictor within a given dataset. If the dataset included a predictor, availability was generally high. Data for age and motor score (including the untestable category) were complete, but some studies had no data for pupils (studies D and I, Table 8.2). If pupils were recorded, data were complete in >90% in most studies. Secondary insults (hypoxia and hypotension) had not been recorded in some studies, but if recorded, data were quite complete.

CT scans are usually performed within hours after admission, after stabilization of the patient. CT scans provide important diagnostic information, and are often classified according to the Marshall classification.<sup>280</sup> This classification was available in 7 of the 11 studies, for 61% of the 8,530 patients. Other important CT characteristics, such as traumatic subarachnoid haemorrhage (tSAH) and the presence of an epidural haematoma (EDH) were available in slightly higher numbers of patients. The presence of EDH is illustrated in Fig. 8.1.

Laboratory values were available for only few studies (Table 8.2). Glucose, pH, sodium, and Hb levels were available for around 50% of the patients, but platelets and prothrombin time (which are related to blood clotting), were available for less



**Fig. 8.1** Example of an epidural haematoma (EDH). An EDH is located directly under the skull and mainly causes brain damage due to compression. Consequently, prognosis is more favourable if it can be evacuated rapidly. A developing EDH is one of the greatest emergencies in neurosurgery

than 20% (Table 8.2). The latter percentages were so low that we did not consider these predictors for a prediction model; admittedly this judgment is arbitrary. A series of models was developed, with different selections of studies, based on availability of predictors per study.

### 8.2.3 Patterns of Missingness

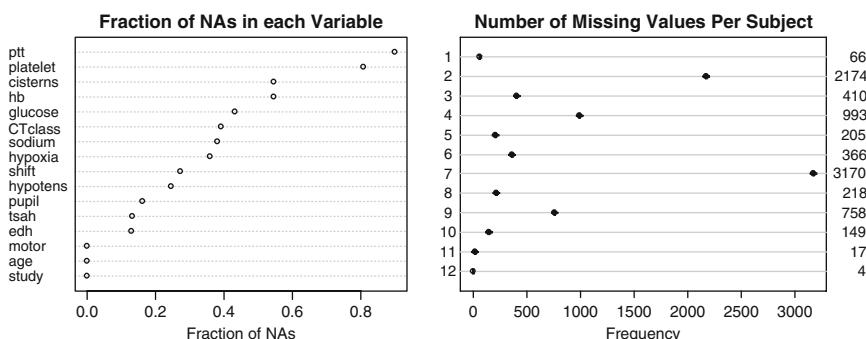
We further examined patterns of missingness, following the steps discussed in Chap. 7.

#### a. How many missings occur for each potential predictor?

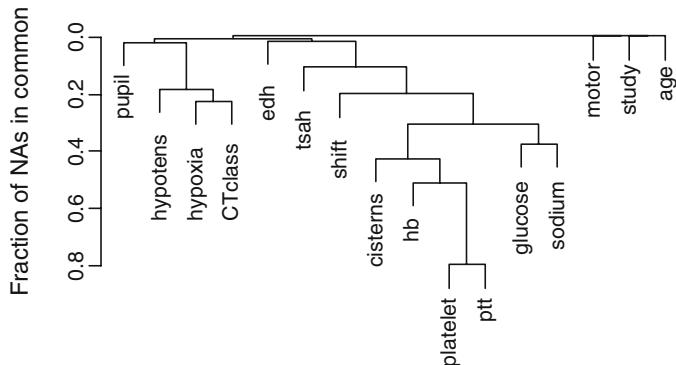
We used the `naclus` and `naptot` function to visualize missing value patterns. As was also noted in Table 8.2, missing values were most frequent for laboratory parameters and some CT characteristics (Fig. 8.2, left panel). Many patients had multiple missing values, e.g. 3,170 patients had 7 missing values, and 4 patients even had 12 missing values among the 15 predictors considered (Fig. 8.2, right panel).

#### b. Missing value mechanisms

For analysis of the mechanism of missingness we examined combinations of missing predictors, associations between predictors and missingness, and associations between outcome and missingness. As proposed by Harrell, we used the `naclus` function to visualize missing value patterns (Fig. 8.3).<sup>174</sup> We note that platelets and prothrombin time are often jointly missing, as also noted in Table 8.2. Characteristics of CT scans, such as shift and cisterns are often missing in combination, while also laboratory values are missing in such patients (hb, glucose, sodium, platelet, ptt).



**Fig. 8.2** Fraction of missing values per potential predictor (left panel), and number of missing values per subject (right panel)



**Fig. 8.3** Combinations of missing values in predictors (“NAs”), based on a hierarchical cluster analysis of missingness combinations

### c. Associations between predictors and missingness

Table 8.2 demonstrates that missingness of most predictors strongly depends on study. We explored in detail whether there were other determinants of missingness for pupils, hypoxia, hypotension, CT class, tSAH, or EDH but no clear patterns were found (Fig. 8.4). Hence, no MAR on  $x$  patterns were evident.

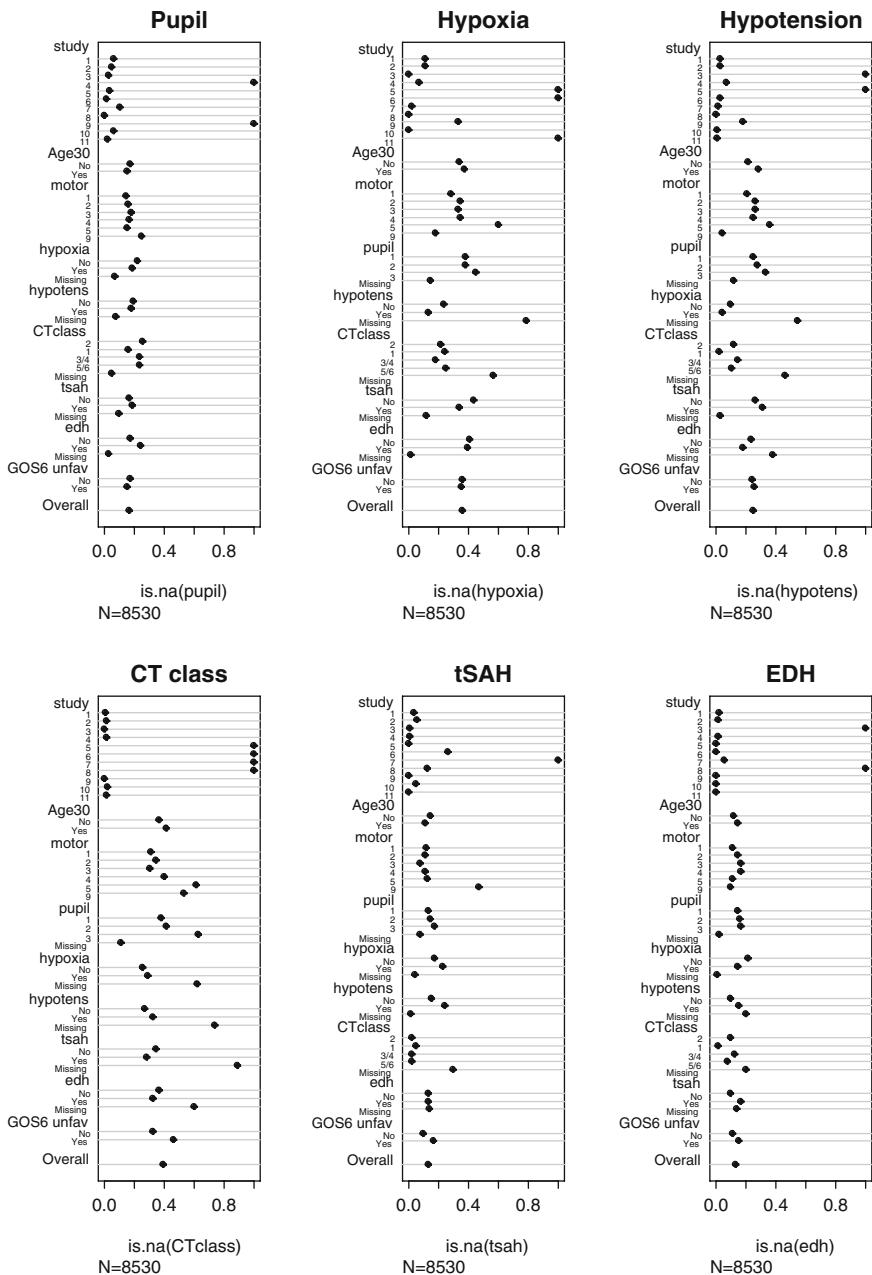
### d. Associations between outcome and missingness

Fig. 8.4 further demonstrates no clear associations between missingness and an unfavourable 6-month GOS outcome. To explore the relation between missingness and outcome in more detail, logistic regression models were constructed, but again no clear patterns were noted. Hence, there were no indications of an MAR on  $y$  mechanism.

### e. Plausible mechanisms for missingness

The most plausible mechanism for missingness was that a predictor was simply not recorded for some studies. Within studies, a mechanism close to MCAR had occurred. We conclude that missingness was essentially MCAR, conditional on the study. Hence, we would like to stratify on study when making imputations. This is however logically impossible in situations that predictor values are 100% missing in a study, as study specific estimates cannot be derived.

We hence imputed values conditional on values of the other predictors, but not conditional on study. On the other hand, we excluded some studies from analyses if we judged that too many predictors were 100% missing in a study.



**Fig. 8.4** Missingness in relation to study (numbered 1–11), other predictors (age to EDH), and outcome (GOS6). Study was the main determinant of missingness. Only weak associations were observed with other predictors, and no relationship with the 6-month outcome (GOS6)

## 8.3 Imputation of Missing Predictor Values

### 8.3.1 Correlations Between Predictors

In Chap. 7, we noted that multiple imputation became more relevant when predictors were correlated. Table 8.3 shows that the correlations between variables were generally modest, implying that both single and multiple imputation procedures may be considered. Some more substantial correlations ( $r>0.4$ ) were noted among CT scan characteristics and between some laboratory values. The associations between cisterns/shift and the CT classification are to be expected, as these characteristics are used in the definition of the CT classification. Hb and platelets are correlated, as both will decrease following blood loss.

### \*8.3.2 Imputation Model

An imputation model was considered that included all relevant potential predictors and the outcome (6-month GOS, in five categories). No auxiliary variables were used. The imputation model was fitted using the mice library and aregImpute from the Hmisc library in R. We show the commands below for illustration, with more details on the web site.

```
# mice imputation model for pmat as predictor matrix, with default
# settings
gm <- mice (TBIallR2, m = 10,
  imputationMethod =c("polyreg", "polyreg", "pmm", "polyreg",
  "polyreg", "polyreg", "logreg", "logreg", "logreg", "pmm",
  "logreg", "logreg", "logreg", "logreg", "pmm", "pmm", "pmm",
  "pmm"), predictorMatrix = pmat, seed=1)
# aregImpute for data set TBIallR2, with default settings
g <- aregImpute (formula = ~d.gos+as.factor(trial) + age +
  as.factor(motorrr) + as.factor(pupil) + as.factor(CTclass) +
  tsah + cisterns + shift + size + sdh + edh +
  hypoxia + hypotens + d.sysbpt + hbt + glucoset + sodiummt,
  n.impute = 10, data=TBIallR2)
```

Here, d.gos is the derived 6-month GOS; trial is the study; age is age in years; motorrr is the Motor score; pupil is pupillary reactivity; CTclass is CT classification; tsah is presence of tSAH; cisterns is presence of compressed cisterns on CT; shift is shift  $\geq 5$  mm on CT; size is shift in mm; sdh and edh refer to subdural and epidural haematomas; hypoxia and hypotens refer to secondary insults; d.sysbpt is derived systolic blood pressure; hbt is truncated Hb; glucoset is truncated glucose; sodiummt is truncated sodium.

**Table 8.3** Rank correlations between predictors, with correlations > 0.4 in bold

	Study	Age	Motor	Pupil	Hypoxia	Hypotens	CTclass	TSAH	EDH	Cisterns	Shift	Gluco	Sodiu	Hb	Platelet	Pt
Study <sup>a</sup>	1	0.17	0.28	0.22	0.19	0.16	0.08	0.22	0.12	0.34	0.26	0.08	0.31	<b>0.52</b>	0.29	
Age	0.17	1	-0.01	0.03	0.00	0.05	0.20	0.14	0.01	0.03	0.14	0.06	-0.07	-0.05	-0.17	0.03
Motor	0.28	-0.01	1	0.37	0.15	0.14	0.11	0.05	-0.02	0.21	0.11	0.11	-0.03	-0.06	-0.01	0.21
Pupil	0.22	0.03	0.37	1	0.14	0.17	0.22	0.11	-0.02	0.23	0.16	0.18	-0.06	-0.05	-0.03	0.11
Hypoxia	0.19	0.00	0.15	0.14	1	0.29	0.02	0.01	-0.05	0.02	-0.02	0.12	-0.01	-0.03	0.02	0.13
Hypotens	0.16	0.05	0.14	0.17	0.29	1	-0.02	0.05	-0.06	0.02	-0.03	0.15	0.02	-0.23	-0.14	0.31
Ctclass	0.08	0.20	0.11	0.22	0.02	-0.02	1	0.14	0.31	<b>0.44</b>	<b>0.48</b>	0.15	-0.06	-0.04	-0.01	-0.05
tSah	0.22	0.14	0.05	0.11	0.01	0.05	0.14	1	-0.04	0.13	0.07	0.10	-0.02	0.01	-0.04	0.12
EDH	0.12	0.01	-0.02	-0.02	-0.05	-0.06	0.31	-0.04	1	0.06	0.15	0.00	0.01	-0.05	-0.05	0.03
Cisterns	0.34	0.03	0.21	0.23	0.02	0.02	<b>0.44</b>	0.13	0.06	1	<b>0.51</b>	0.09	-0.03	-0.08	0.07	0.09
Shift	0.26	0.14	0.11	0.16	-0.02	-0.03	<b>0.48</b>	0.07	0.15	<b>0.51</b>	1	0.03	-0.01	-0.11	-0.10	0.08
Glucose	0.24	0.06	0.11	0.18	0.12	0.15	0.15	0.10	0.00	0.09	0.03	1	-0.13	-0.04	0.21	0.11
Sodium	0.08	-0.07	-0.03	-0.06	-0.01	0.02	-0.06	-0.02	0.01	-0.03	-0.01	-0.13	1	0.04	-0.08	0.06
Hb	0.31	-0.05	-0.06	-0.05	-0.03	-0.23	-0.04	0.01	-0.05	-0.08	-0.11	-0.04	0.04	1	<b>0.46</b>	-0.21
Platelet	<b>0.52</b>	-0.17	-0.01	-0.03	0.02	-0.14	-0.01	-0.04	-0.05	0.07	-0.10	0.21	-0.08	<b>0.46</b>	1	-0.34
Pt	0.29	0.03	0.21	0.11	0.13	0.31	-0.05	0.12	0.03	0.09	0.08	0.11	0.06	-0.21	-0.34	1

<sup>a</sup>Based on generalized Spearman rank correlation as calculated with the `spearman2` function; other correlations based in standard Spearman rank correlation. All correlations were calculated with pairwise available patients

The `gm` and `g` objects each consist of ten imputed data sets of the IMPACT database. In total 18 variables were considered in the imputation model. Data were complete for the outcome (`d.gos`), `trial`, `age`, and motor score. With `aregImpute`,  $R^2$  values are given to indicate how well each variable can be predicted from the other variables.  $R^2$  values were very high for shift coded as a binary variable and size of shift in millimetres, which are by definition strongly correlated (shift defined as size  $\geq 5$  mm). Similarly, details of the imputations by `mice` can be inspected.

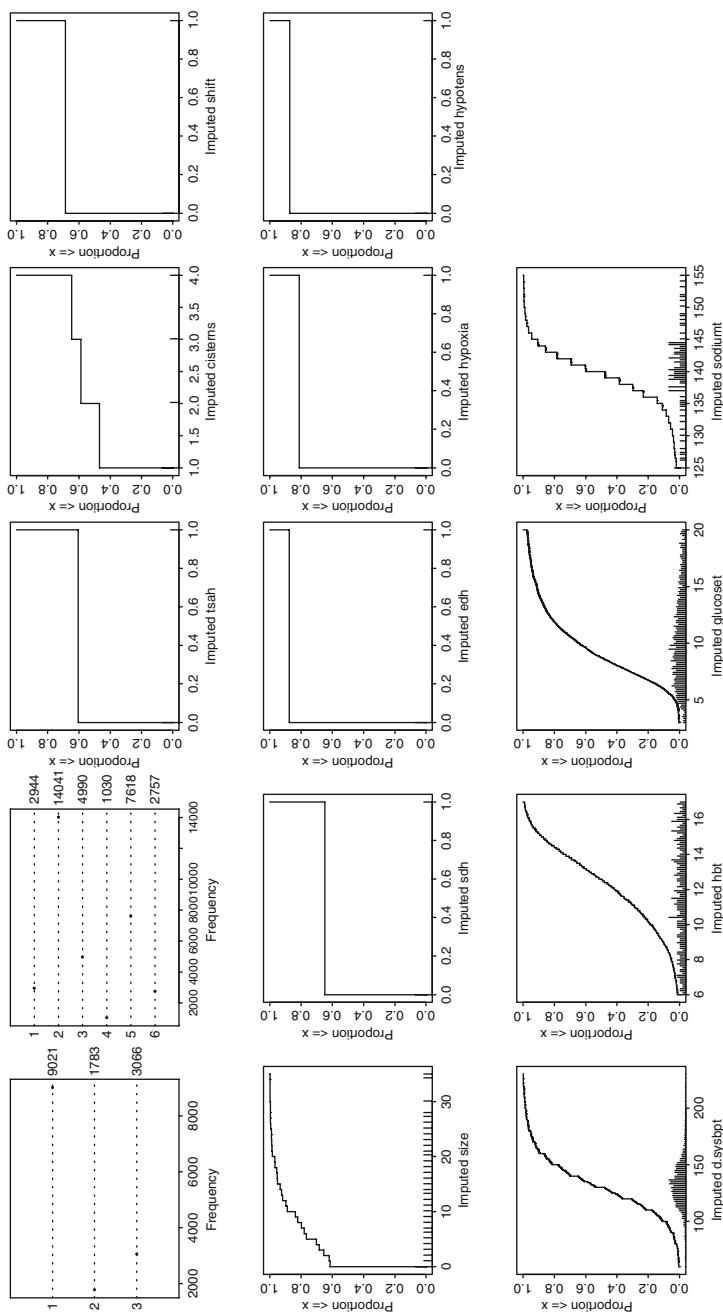
### 8.3.3 *Distributions of Imputed Values*

The distributions of imputed values in object  $g$  were checked for the plausibility of imputations (e.g. within a plausible range, no strange peaks, Fig. 8.5). The frequencies of categorical variables are shown as dot charts. For example, the first graph shows the imputations over ten sets for “pupil” (values 1, 2, 3), and the second for CTclass (values 1–6). For predictors that are treated as linear variables, the cumulative distribution is shown. For example, the third graph shows that imputed tSAH values were 0 in 60%, and 1 in 40%. Although size was considered as a linear variable, this does not imply that normality was assumed for the distribution; many values for “Imputed size” were zero. The before last graph shows imputations for glucose, which are truncated at 2 and 20, as in the original predictor definition.

## 8.4 Estimating Adjusted Effects

After imputation, we estimated the adjusted effects of each predictor of interest in turn, using imputed versions of other predictors. These other predictors are hence considered as potential confounders. We present all results for `aregImpute` for adjusted analyses; results with `mice` are only presented for the multivariable models. As confounders we considered seven predictors that had also shown convincing effects in previous TBI studies. These include the three core predictors (age, motor score, pupils), two secondary insults (hypoxia, hypotension), and two CT characteristics (CT classification and tSAH). The outcome was GOS at 6 months, dichotomized as unfavourable vs. favourable in logistic regression models. For illustration, we show the adjusted logistic regression coefficients of each of these predictors in turn (Table 8.4). We estimate adjusted effects in the complete cases (CC), as well as in completed data sets with single (SI) or multiple imputation (MI). Odds ratios can be calculated as  $e^{\text{coefficient}}$ .

Numbers of patients differ dramatically between the univariate and CC analyses, since only 2,428 patients had complete values for all 7 predictors considered. Per predictor, values were complete for some (age, motor score). Values were



**Fig. 8.5** Distribution of imputed values with aregImpute in the IMPACT study. The 14 imputations shown are for the predictors pupil, CTclass, tasah, ciysters, shift, sdh, hypoxia, hypolens, systbp, hb.glucose, and sodium, respectively

**Table 8.4** Logistic regression coefficients of predictors in univariate and adjusted analyses. Numbers are coefficients (SE).

	N	Univar		Adjusted	
		N	N = 5,192–8,530	CC, n = 2,428	SI, n=5,192–
					8,530
Age (per decade)	8,530		0.32 (0.015)	0.36 (0.033)	0.33 (0.018)
Motor score	8,530				
1 or 2			1.87 (0.065)	1.65 (0.160)	1.48 (0.074)
3			1.38 (0.077)	1.36 (0.157)	1.14 (0.086)
4			0.69 (0.065)	0.71 (0.128)	0.57 (0.071)
5 or 6			Zero (ref)	Zero (ref)	Zero (ref)
9			0.91 (0.112)	1.06 (0.259)	0.82 (0.127)
Pupillary reactivity	7,143				
Both pupil reactive			Zero (ref)	Zero (ref)	Zero (ref)
One non-reactive			0.97 (0.076)	0.51 (0.149)	0.56 (0.085)
Both non-reactive			1.77 (0.067)	0.94 (0.144)	1.18 (0.076)
Hypoxia	5,473		0.80 (0.072)	0.49 (0.125)	0.38 (0.085)
Hypotension	6,440		0.99 (0.070)	0.68 (0.133)	0.68 (0.084)
CT class	5,192				
1 or 2			Zero (ref)	Zero (ref)	Zero (ref)
3 or 4			1.08 (0.079)	0.77 (0.134)	0.78 (0.089)
5 or 6			0.96 (0.066)	0.67 (0.115)	0.55 (0.075)
Traumatic SAH	7,393		0.99 (0.050)	0.84 (0.101)	0.74 (0.057)
					0.73 (0.058)

most incomplete for CT class ( $n = 5,192$ ). The coefficients of most of the predictors were largest in univariate analyses, and smaller in adjusted analyses. This reflects the positive correlations between predictors (see Table 8.3). The estimates in adjusted analyses were largely similar for SI or MI, but were sometimes quite different from the CC analyses, e.g. smaller for motor score. The SEs in the CC analyses are higher than in the imputed analyses, reflecting smaller numbers. The MI analyses showed larger SEs than SI analyses, but differences were minor (3rd decimal).

Technical details of the model fitting are further discussed with detailed code for R programs. We first describe the modelling for complete predictors (age, motor), followed by the approach for predictors with missing values, such as pupils.

#### \*8.4.1 *Adjusted Analysis for Complete Predictors: Age and Motor Score*

Age and motor score were completely available ( $n=8,530$ ). Univariate effects can easily be estimated with logistic models:

```
lrm(d.unfav~as.factor(trial)+age, data=TBIall)
lrm(d.unfav~as.factor(trial)+as.factor(motorr),
data=TBIall)
```

The estimated regression coefficients are shown in Table 8.4.

Here, d.unfav refers to unfavourable GOS at 6 months, trial is the study indicator, such that analyses are stratified by study.

A CC model with adjustment for confounders included only 2,428 patients, due to exclusion of patients with any missing value for the other predictors (pupil, hypoxia, hypotens, CTclass, tsah). Only patients from studies A, B, and J are included:

### CC model:

```
lrm(formula = d.unfav ~ as.factor(trial) + age + as.factor(motorr) + as.factor(pupil)
+ hypoxia + hypotens + CTclass34 + CTclass56 + tsah, data = TBIall)
```

Frequencies of Missing Values Due to Each Variable

	d.unfav	trial	age	motorr	pupil	hypoxia	hypotens	CTclass34	CTclass56	tsah
0	0	0	0	1387	3057	2090	3338	3338	3338	1137

Obs	Max Deriv	Model	L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R2	Brier
2428	6e-010		840	14	0	0.823	0.645	0.646	0.315	0.393	0168
			Coef		S.E.		Wald Z		P		
Intercept			-3.59911		0.186332		-19.32		0.0000		
trial=A			-0.14818		0.121132		-1.22		0.2212		
trial=J			0.07172		0.137206		0.52		0.6012		
age			0.03571		0.003348		10.67		0.0000		
motorr=1/2			1.64538		0.159743		10.30		0.0000		
motorr=3			1.35782		0.156498		8.68		0.0000		
motorr=4			0.71459		0.128421		5.56		0.0000		
motorr=9			1.06208		0.258924		4.10		0.0000		
pupil=2			0.51432		0.148866		3.45		0.0006		
pupil=3			0.94368		0.143710		6.57		0.0000		
hypoxia			0.49115		0.124781		3.94		0.0001		
hypotens			0.67864		0.133171		5.10		0.0000		
CTclass34			0.76777		0.134252		5.72		0.0000		
CTclass56			0.67493		0.114807		5.88		0.0000		
tsah			0.84091		0.101395		8.29		0.0000		

In this specific case with complete data on age and motor score, fitting age and motor score with imputed data (SI or MI) is identical to fitting a model in the fully imputed data set ( $n=8,530$ ). For SI, we create imputed data from the first MI data set in the g object, for example:

```
TBIall$pupil.i <- TBIall$pupil
TBIall$pupil.i[is.na(TBIall$pupil)] <- g$imputed$pupil[,1]
```

This is done for all predictors with missing values, with the extension “.i” added to indicate that we consider imputed data for a predictor.

### SI model:

```
lrm (formula = d.unfav ~ as.factor(trial) + age + as.factor(motorr) + as.factor(pupil.i)
+ hypoxia.i + hypotens.i + CTclass34.i + CTclass56.i + tsah.i, data = TBIall)
```

Obs	Max	Deriv	Model	L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R2	Brier
8530	2e-009			2678	22	0	0.805	0.609	0.61	0.304	0.36	0.18
				Coef	S.E.	Wald Z		P				
	Intercept	-3.193738	0.111591		-28.62	0.0000						
	...											
	age	0.032630	0.001774		18.39	0.0000						
	motorr=1/2	1.475716	0.074192		19.89	0.0000						
	motorr=3	1.169648	0.085977		13.60	0.0000						
	motorr=4	0.574532	0.071067		8.08	0.0000						
	motorr=9	0.820593	0.126781		6.47	0.0000						
	pupil.i=2	0.588143	0.076883		7.65	0.0000						
	pupil.i=3	1.103948	0.068252		16.17	0.0000						
	hypoxia.i	0.264818	0.068488		3.87	0.0001						
	hypotens.i	0.670742	0.073482		9.13	0.0000						
	CTclass34.i	0.570787	0.069579		8.20	0.0000						
	CTclass56.i	0.491745	0.059293		8.29	0.0000						
	tsah.i	0.723821	0.053876		13.43	0.0000						

The MI model for age and motor score is fitted using the `fit.mult.impute` function, which automatically combines results over imputed data sets.

### MI model:

```
fit.mult.impute(d.unfav ~ as.factor(trial) + age + as.factor(motorr) + as.factor(pupil)
+ hypoxia + hypotens + as.factor(CTclass == 3 | CTclass == 4) + as.factor (CTclass
==5 | CTclass == 6) + tsah, lrm, xtrans = g, data = TBIall)
```

Variance Inflation Factors Due to Imputation:

Intercept	trial=B	trial=C	trial=D	trial=E	trial=F	trial=G	trial=H	trial=I				
1.07	1.01	1.01	1.06	1.03	1.04	1.07	1.02	1.05				
trial=J	trial=K	age	motorr=1/2	motorr=3	motorr=4	motorr=9	pupil=2	pupil=3				
1.01	1.05	1.02	1.03	1.02	1.03	1.02	1.53	1.56				
hypotens	CTclass=3/4	CTclass=5/6	tsah=TRUE					1.76				
1.15	1.59		1.23		1.1							
Obs	Max	Deriv	Model	L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R2	Brier
8530	2e-009			2688	22	0	0.805	0.61	0.611	0.305	0.361	0.179
				Coef	S.E.	Wald Z		P				
	Intercept	-3.22374	0.116132		-27.76	0.0000						
	...											
	age	0.03321	0.001799		18.46	0.0000						
	motorr=1/2	1.46459	0.075307		19.45	0.0000						
	motorr=3	1.15620	0.086893		13.31	0.0000						
	motorr=4	0.57085	0.072146		7.91	0.0000						
	motorr=9	0.82014	0.128497		6.38	0.0000						
	pupil=2	0.57979	0.095312		6.08	0.0000						
	pupil=3	1.15770	0.085706		13.51	0.0000						
	hypoxia	0.35117	0.090620		3.88	0.0001						
	hypotens	0.63300	0.079237		7.99	0.0000						
	CTclass=3/4	0.55875	0.088365		6.32	0.0000						
	CTclass=5/6	0.47454	0.066028		7.19	0.0000						
	tsah	0.73805	0.056594		13.04	0.0000						

<sup>1</sup>These statistics are from the last fit with imputed data, in this case the tenth imputed data set

In conclusion, single and multiple imputation yielded very comparable results in this example: model statistics were similar (LR statistic,  $c$  statistic,  $R^2$  estimate), as well as regression coefficients and standard errors.

### \*8.4.2 Adjusted Analysis for Incomplete Predictors: Pupils

Pupillary reactivity was recorded for 7,143 patients. This selection of patients was used in univariate and adjusted analyses.

#### Univariate analysis:

```
lrm(d.unfav ~ as.factor(trial) + as.factor(pupil), data = TBIall)
Frequencies of Missing Values Due to Each Variable
d.unfav trial pupil
0          0      1387
Obs Max Deriv Model L.R. d.f. P C      Dxy    Gamma Tau-a R2      Brier
7143 1e-008     1097 10   0 0.708 0.417 0.441 0.208 0.19 0.213

      Coef      S.E.  Wald Z      P
Intercept -0.73594 0.06787 -10.84 0.0000
...
pupil=2   0.96801 0.07590  12.75 0.0000
pupil=3   1.77194 0.06670  26.56 0.0000
```

#### Adjusted analysis following single imputation:

```
lrm(d.unfav ~ as.factor(trial) + age + as.factor(motorr) + as.factor(pupil) +
hypoxia.i + hypotens.i + CTclass34.i + CTclass56.i + tsah.i, data = TBIall)
Obs Max Deriv Model L.R. d.f. P C      Dxy    Gamma Tau-a R2      Brier
7143 1e-008     2403 20   0 0.814 0.628 0.629 0.314 0.381 0.175

      Coef      S.E.  Wald Z      P
Intercept -3.25518 0.120631 -26.98 0.0000
...
pupil=2   0.55594 0.085164   6.53 0.0000
pupil=3   1.17557 0.076062  15.46 0.0000
...
```

For adjusted analyses, we can also use multiple imputations, e.g. from `aregImpute`. We first rename the predictor of interest (e.g. “.`o`” for “original”) such that this predictor is not imputed:

```
TBIall$pupil.o <- TBIall$pupil
fit.mult.impute(d.unfav ~ as.factor(trial) + age + as.factor(motorr) + as.factor(pupil.o) +
hypoxia + hypotens + as.factor(CTclass==3|CTclass==4) + as.factor(CTclass==5|Ctclass
==6) + tsah, lrm, xtrans = g2, data = TBIall)
```

This original version of the pupil variable remains missing in 1,387 patients:

```
Frequencies of Missing Values Due to Each Variable
d.unfav trial age motorr pupil.o hypoxia hypotens CTclass      tsah
0          0      0      0      1387      0      0      0      0
Obs Max Deriv Model L.R. d.f. P C      Dxy    Gamma Tau-a R2      Brier
7143 9e-009     2386 20   0 0.813 0.626 0.627 0.313 0.379 0.176

      Coef      S.E.  Wald Z      P
Intercept -3.28470 0.12572 -26.13 0.0000
...
pupil.o=2  0.56648 0.08624   6.57 0.0000
pupil.o=3  1.17570 0.07670  15.33 0.0000
...
```

Again, the results obtained with single or multiple imputation procedures were very similar. Analyses for the other predictors with missing values were performed in a similar way. A series of papers presents further results for the other predictors with missing values.<sup>62,272,276,284,306,446</sup>

## 8.5 Multivariable Analyses

After studying adjusted effects per predictor, we are further interested in the multivariable effects of all predictors combined. We start with a core model, consisting of three predictors age, motor score, and pupils. All studies could reasonably be considered for this model, since they had age and motor score completely available ( $n=8,530$ ). A CC analysis included 7,143 patients, because of 1,387 missing values for pupils. These 1,387 values led to exclusion of  $1,387/8,530=16\%$  of the patients, while they represented 5.4% of the required values for the three predictors.

Next, we considered a more extended model, including the seven predictors that were also used as confounders before: three core predictors plus secondary insults plus CT characteristics. It was not considered reasonable to include study #E in this analysis, since secondary insults and CT classification was not recorded in the database for this study. We hence considered 10 studies, with a total of 7,020 patients. These were included in SI and MI procedures. A CC analysis was possible with only 2,428 patients, representing a loss of 4,592 patients (65%), while only 13% of the required values were missing ( $6,426$  of  $7 \times 7,020 = 49,140$ ).

The multivariable coefficients are shown in Table 8.5, together with rounded prognostic scores. Scores were based on multiplying coefficients by 10, and rounding to whole numbers ("round(10 \* fit\$coef)"). We note that the SI and MI coefficients and prognostic scores were largely similar. Scores never differed by more than 2 points. The CC analysis gave quite different estimates compared to SI or MI, demonstrating the substantial limitations of CC analyses. Prognostic scores with MI were lower for motor scores, larger for pupils, lower for hypoxia, and similar for CT characteristics.

## 8.6 Concluding Remarks

This case study illustrates how we may deal with missing values in assessing predictor effects (univariate and adjusted effects), and in multivariable modelling to derive prediction models. The difference in numbers of patients was dramatic between complete case and single or multiple imputed data. Since a reasonable imputation model could be constructed, we should have more confidence in the results after imputation (either SI or MI) than the CC results. The presented R code is available at the book's web site, and may be useful in implementing MI in other case studies.

**Table 8.5** Multivariable regression coefficients and rounded prognostic scores for a 7-predictor model in the IMPACT study.

	CC, n=2,428			SI, n=7020			MI, n=7020		
	Coeff	Score	aregImpute	Coeff	Score	mice	Coeff	Score	aregImpute
Age (coef per decade score per 3 year)	0.36	1	0.32	1	0.33	1	0.31	1	0.34
Motor score									
1 or 2	1.65	17	1.44	14	1.54	14	1.42	14	1.57
3	1.36	14	1.12	11	1.13	11	1.11	11	1.18
4	0.71	7	0.54	5	0.54	5	0.53	5	0.57
5 or 6	Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)
9	1.06	11	0.82	8	1.04	10	0.79	8	0.83
Pupillary reactivity									
Both pupil reactive	Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)
0.51	5	0.62	6	0.48	5	0.60	6	0.48	5
0.94	9	1.14	11	1.09	11	1.22	12	1.01	10
0.77	8	0.30	3	0.37	4	0.38	4	0.33	3
Hypotension	0.67	7	0.68	7	0.61	6	0.64	6	0.58
CT Class									
1 or 2	Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)		Zero (ref)
3 or 4	0.84	8	0.62	6	0.62	6	0.63	6	0.59
5 or 6	0.49	5	0.51	5	0.45	5	0.49	5	0.43
Traumatic SAH	0.68	7	0.72	7	0.64	6	0.74	7	0.62

## Questions

### 8.1 Missingness mechanisms

We state that most predictors were missing complete at random (MCAR), conditional on study (Sect. 8.2.3 e).

- (a) Does Table 8.2 support an MCAR mechanism?
- (b) What do we learn from Fig. 8.4 with respect to MAR on  $x$ , or MAR on  $y$  mechanisms?
- (c) Can we exclude a MNAR mechanism from the presented tables and figures?
- (d) The imputation models did not include “study” as a variable. Why was this desirable, but not possible?

### 8.2 Imputation results (Sect. 8.4.1)

- (a) For the MI model, the `aregImpute` imputation procedure lists “Variance Inflation Factors Due to Imputation.” What do these factors refer to? When are they larger than 1? Which predictor has the largest VIF?
- (b) Compare the predictor effects of age between the CC, SI, and MI models. When is the standard error estimated as the smallest?

### 8.3 Numbers in adjusted vs. multivariable analyses (Sect. 8.4.2 and 8.5)

The adjusted analysis for the predictor pupillary reactivity (“`pupil`”) was performed with 7,143 patients (Sect. 8.4.2), while the multivariable analysis included 7,020 patients (Table 8.5).

- (a) How did this difference arise?
- (b) Do you agree with this approach? Or explain alternatives.

# Chapter 9

## Coding of Categorical and Continuous Predictors

**Background** When developing a prediction model, an important consideration is how we code the predictors. Raw data from a study are often not in a form appropriate for entering in regression models and must first be manipulated. This is known as “coding.” As in any data analysis, we will usually start with obtaining an impression of the data under study, such as occurrence of missing values and the distribution of predictors. Descriptive analyses, such as frequency tables are useful to this aim. We will consider various issues in coding of unordered and ordered categorical predictors. For continuous predictors, we specifically discuss how we can limit the influence of outliers and interpret regression coefficients.

### 9.1 Categorical Predictors

Categorical predictors can be unordered, for example a diagnostic category, or site of treatment. Categorical predictors are usually coded as “factor” variables, with coding as dummy variables. For example, smoking was coded originally as 1 for never, 2 for past, and 3 for current smoker in GUSTO-I. For analysis as a factor, we might create two dummy variables for category 2 vs. 1 and 3 vs. 1. Logistic regression coefficients for these dummies refer to the comparison of past vs. never smokers and current vs. never smokers. Dummy coding may often be convenient in prediction research. Specific attention should be paid to the choice of reference category (here: never smokers). By default, the lowest or highest numbered category is used as reference in many statistical packages. If this category is relatively small, comparisons with this reference category may show statistically non-significant and unstable results, while the factor has an important predictive effect overall. The predictions from a model are not affected by the choice of reference category.

It may be convenient to combine categories if these are relatively small. For example, a cancer study might list a very large number of stages (e.g. T1a, T1b, T1c, T2a, T2x, etc.) that might be converted into a smaller number of groups (e.g. T1, T2, T3, and T4). In other situations, some categories might be combined in an “other” category. If small categories are kept, some sort of penalized estimation or shrinkage is required to obtain reliable estimates.<sup>391,468</sup> When a combination of cate-

**Table 9.1** Impact of various codings of categorical predictors in GUSTO-I ( $n=40,830$ )

Predictor	Coding	<i>df</i>	Model $\chi^2$ <sup>a</sup>
<i>Unordered</i>			
Location of infarct	Anterior vs. other	1	343
	Ant/Inf/Other	2	361
<i>Ordered</i>			
Killip class	Shock (3/4 vs. 1/2)	1	861
	Linear (1–4)	1	1388
	Linear + square	2	1388
	Factor	3	1389
Smoking	Never/past/current	2	483
	Linear (1–3)	1	482

<sup>a</sup>Model  $\chi^2$  was calculated as the difference between a model with and without the predictor on the  $-2 \log$  likelihood scale

gories is based on the similarity of the relationship with the outcome, overfitting may occur and the apparent model performance will be optimistic. In practice, a balance has to be sought between combining categories blinded to the outcome (e.g. based on frequency distributions) and adequately capturing patterns of outcome by category. Using the coding from previous studies may often be helpful in smaller sized data sets.

Ordered predictors are also common in prediction research. Often, a small number of categories is made, for example by dichotomization. Table 9.1 illustrates that ignoring ordering in predictors may cause a substantial loss of predictive information. Simply assuming linearity of predictor effects may sometimes work well. Some advanced estimation techniques might also be considered that force monotonicity of the effect but are more flexible than a linear coding.<sup>468</sup>

### \*9.1.1 Examples of Categorical Coding

In patients with an acute MI, location of infarction is an important predictor of 30-day mortality. In GUSTO-I, the categorization was as anterior vs. inferior vs. other. The other location category contained only 3% of the patients.<sup>255</sup> Such a refined coding is only possible in large studies; in smaller-sized studies we might combine the inferior and other categories.

The refined coding with three categories led to a slightly better predictive performance than the combined coding with two categories. The  $\chi^2$  statistics were calculated as the differences between a model with and without location of infarction on the  $-2 \log$  likelihood scale and were 361 vs. 343 (Table 9.1), at the expense of 1 *df* extra.

An example of an ordered predictor is Killip class, a measure for left ventricular function ranging from I to IV. It can be recoded as shock (Killip 3/4 vs. 1/2).<sup>302</sup>

Alternatively, we can analyse ordered predictors as continuous variables, possibly with a check for non-linearity by adding a square term. Ignoring the ordinal nature of a variable such as Killip class causes a major loss in predictive ability. A simple linear coding captures much more of the predictive information ( $\chi^2$  861 vs. 1,388). For a less clearly ordered variable such as smoking (never/past/current), linear coding had the same performance as a factor variable, using 1 instead of 2 *df* (Table 9.1).

## 9.2 Continuous Predictors

Continuous variables formally should be measured on a interval or ratio scale, and should be able to take any value in a range. We however noted in Table 9.1 that treating ordered variables as linear was sometimes reasonable for prediction, at least for some variables considered in GUSTO-I.

### 9.2.1 Examples of Continuous Predictors

Age is a good example of a continuous predictor variable. We already found that the age effect could often quite well be captured with a linear term (Chap. 6). Remarkably, age has often been considered as a categorical variable in prognostic studies, for example in traumatic brain injury (TBI).<sup>204</sup> In GUSTO-I, a dichotomy at 65 years leads to a  $\chi^2$  of 1,463 instead of 2,099 (Table 9.2). Considering three categories limits the loss in information somewhat ( $\chi^2$  1,775, 85% of the information of age as a linear variable).

The predictor “number of leads with ST elevation” ranges from 0 to 11 in the GUSTO-I data (Chap. 22). The number of categories is large for consideration as a

**Table 9.2** Impact of various codings of continuous predictors in GUSTO-I ( $n=40,830$ )

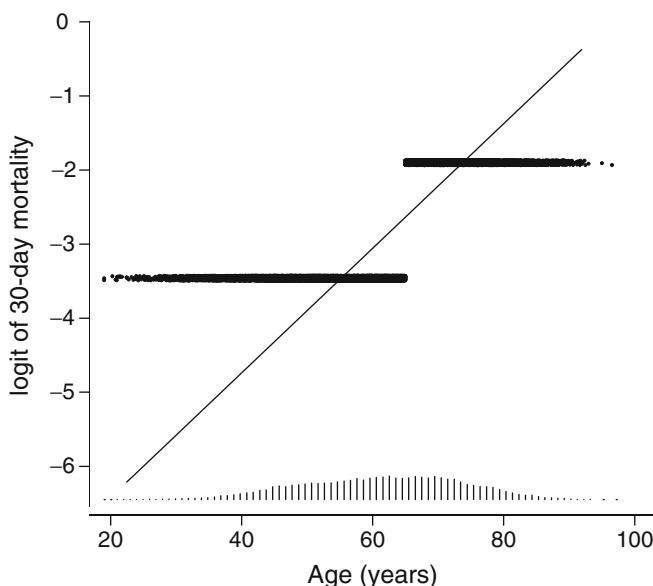
Predictor	Coding	<i>df</i>	Model $\chi^2$
Age	<=65 vs. >65 years	1	1,463
	<=60, 61–70, >=71	2	1,775
	Linear	1	2,099
	Linear + square	2	2,112
	RCS, 5 knots <sup>a</sup>	4	2,122
ST elevation	>4 vs. <=4	1	259
	Linear (0–11)	1	281
	Linear + square	2	306
	Linear + square + cubic	3	339
	RCS, 5 knots <sup>a</sup>	4	350
	Factor	11	367

<sup>a</sup>RCS denotes restricted cubic spline function; 5 knots lead to 4 *df* for the transformation of the predictor (see Sect. 9.3)<sup>177</sup>

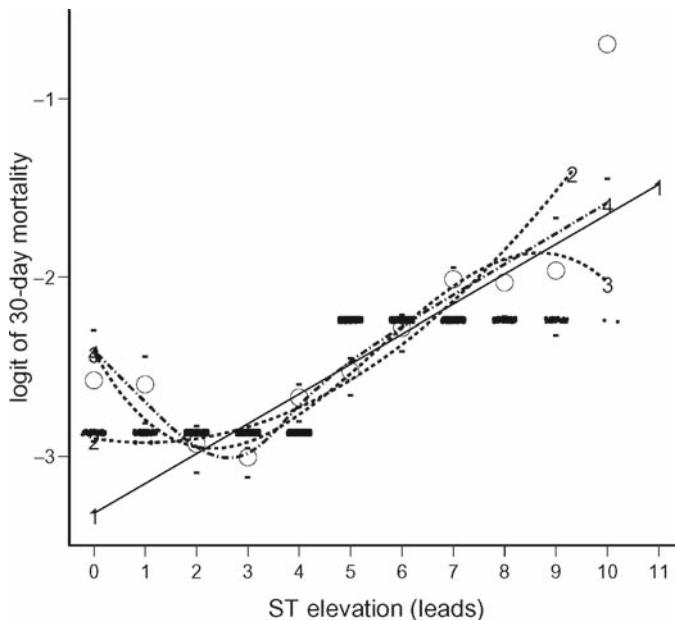
factor variable, but this can technically still be done. Simple linear coding leads to a better performance than a dichotomy at 4 or more leads ( $\chi^2 281$  vs. 259). Adding a square term led to further improvement in fit, but a restricted cubic spline function with 4 *df* made an even better approximation ( $\chi^2 350$ ).<sup>174,177</sup>

### 9.2.2 *Categorization of Continuous Predictors*

Dichotomization of a continuous predictor has many disadvantages.<sup>355</sup> The first unnatural aspect is the step in predictions, as illustrated for age  $\leq 65$  vs.  $> 65$  years in Fig. 9.1. Would risks be really very different for patients who had their 65th birthday yesterday compared to patients who had their 65th birthday today? Similarly, the assumption of a constant risk below or above a threshold is unnatural. A patient of age 40 has lower risks of mortality than a patient of age 64; and a patient of age 90 is different from a patient of age 66. There are only two points where the dichotomized version of age is adequate, that is around the intersections of predicted risks with the predicted risks according to the continuous variable (either linear or transformed). However, if there had been a different distribution of ages, e.g. no patients older than 70 years old, the step function in Fig. 9.1 would have been much different. The continuous model, which conditions on all values of age, would remain relatively unchanged.



**Fig. 9.1** Relationship between age and 30-day mortality in GUSTO-I ( $n=40,830$ ). Age is modelled as a linear variable and dichotomized at age 65 (see Table 9.2). The distribution of ages is shown at the bottom of the graph.



**Fig. 9.2** Relationship between number of leads with ST elevation and 30-day mortality in GUSTO-I ( $n=40,830$ ). ST elevation is modelled as a linear variable (“1”), and with extension with a square term (“2”), and square + cubic terms (“3”). A restricted cubic spline with 4  $df$  (“4”) is shown as well as a dichotomized version of ST elevation ( $>4$  leads, see Table 9.2). The observed risk for each number of leads with ST elevation is shown with *circles* (Table 9.3)

A similar problem arises with treating ST elevations as a dichotomous variable. A reasonable fit is achieved with a linear + square coding (Table 9.2, Fig. 9.2). The risk associated with a low number of elevated leads (0–4) could well be captured with a category “ $\leq 4$  leads.” But the risk rises steeply with increasing numbers of elevated leads, and this risk is poorly estimated with a constant risk for all patients with  $>4$  leads elevated (Table 9.3). The relationship of ST elevations with mortality is complex, as further illustrated in Fig. 9.2, but dichotomization does poorest of all transformations considered.

In epidemiological research, continuous variables are often divided in four or five categories. This may be attractive as an exploratory step for predictor–outcome relationships, but should not be used in a final prediction model.<sup>298</sup> Jumps in predictions are unnatural, and smooth relationships are biologically far more plausible.

### 9.3 Non-Linear Functions for Continuous Predictors

When we consider a continuous predictor as a linear term in a prediction model, we assume that the effect is the same at each part of range of the predictor. For example, in Fig. 9.1 we assume that the effect of being 10 years older is the same at age 30

**Table 9.3** Number of leads with ST elevation and 30-day mortality: univariate analysis in GUSTO-I ( $n=40,830$ )

Predictor	Number of leads elevated	N	30-day mortality (%)
ST elevation	0	607	(7.1)
	1	1,702	(6.9)
	2	4,594	(5.1)
	3	12,744	(4.7)
	4	5,774	(6.5)
	5	5,162	(7.4)
	6	4,573	(9.3)
	7	3,848	(12)
	8	1,456	(12)
	9	333	(12)
	10	33	(33)
	11	4	(0)

(40 vs. 30 years) and 70 years (80 vs. 70 years) for patients with an acute MI. If a non-linear function is expected, various options can readily be considered in regression models. Below we discuss non-linear modelling of continuous predictors with polynomials, fractional polynomials, and spline functions.

### 9.3.1 Polynomials

A general approach to continuous predictors in regression analysis is to add polynomial terms as extensions to a model with a linear term. Commonly, square and cubic terms are considered.<sup>148</sup> For example, we can examine models with  $X$ ,  $X+X^2$  and  $X + X^2 + X^3$ , where  $X$  is a continuous predictor. This results in nested models, and we can statistically test each extension. From a pragmatic point of view, there is no objection to considering a model such as  $X + X^3$ , but it is more common to consider sequential extensions with terms of increasingly higher order. Other common transformations to consider are the inverse ( $X^{-1}$ ) and square root ( $X^{0.5}$ ), and logarithmic ( $\log(X)$ ,  $\exp(X)$ ). We may use these terms as replacement of the linear term  $X$ , or as extension to a model with  $X$  as a linear term included. Polynomials are limited in the shapes they can take. We therefore consider wider families of models.

### 9.3.2 Fractional Polynomials

Fractional polynomials (FPs) have been advocated recently to model continuous predictors. FPs are an extension of earlier proposals on transformation of predictors.<sup>52,354</sup> FPs allow for smooth and flexible transformation of continuous predictors by combining polynomials. FPs extend ordinary polynomials by including non-positive and fractional powers from the set  $-2, -1, -0.5, 0, 0.5, 1, 2, 3$ . This defines eight transformations, including inverse ( $X^{-1}$ ), log ( $X^0$ ), square root ( $X^{0.5}$ ), linear ( $X^1$ ),

squared ( $X^2$ ), and cubic transformations ( $X^3$ ). In addition to these 8 “FP1” functions, 28 “FP2” functions can be considered of the form  $X^{p1} + X^{p2}$ ; when  $p1 = p2$  one defines another 8 FP2 functions as  $X^p + X^p \log X$ , for a total of 36 FP2 functions.<sup>367</sup> The  $df$  used by these functions are 2 and 4, respectively. This includes 1 or 2 degrees of freedom for searching the power transformation. The width of confidence intervals may however be too small if we ignore such model uncertainty.

For medical problems, two terms (FP2 transformations) have been suggested as sufficient to describe non-linear relationships, e.g.  $\text{age}^2$  and  $\text{age}^{0.5}$ . Such parametric combinations can be written down easily. Procedures have been proposed to select FP transformations in multivariable models.<sup>15,354</sup>

A disadvantage of FPs is distortion caused by values at the tails of the predictor distribution. The influence of extreme values can be prevented by a type of truncation, but the global shape of fractional polynomials remains influenced by the values at the tails. Furthermore, fractional polynomial functions are not invariant to a change of origin of the covariate, and negative values cannot be handled. A pragmatic approach to these issues has been proposed to improve the robustness of FP models.<sup>356</sup>

### 9.3.3 Splines

Very flexible transformations are provided by spline functions. Various types of spline functions can be considered, such as natural splines. These can well be fitted with generalized additive models (GAM).<sup>180,1</sup> The extreme flexibility leads sometimes to wiggly patterns of predictions, which are unlikely to be reproduced in new data. Smoothness can be enforced by parameters in the model fitting process, e.g. penalty terms in the likelihood function (see Chap. 13).<sup>181</sup> Without such penalty, splines may easily overfit patterns in the data.

Restricted cubic spline (RCS) functions have been proposed for a more stable approach for prediction models.<sup>174,177</sup> RCSs are cubic splines (containing  $X^3$  terms) that are restricted to be linear in the tails. These splines are still very flexible, and can take more forms than a parametric transformation with the same  $df$  in the model. For example, adding  $X^2$  restricts the relationship to be parabolic, while an RCS with 2  $df$  (3 knots) incorporates a wider family of functions. See Harrell for many illustrations of the form that an RCS can take.<sup>174</sup>

A spline function requires the specification of knots. The spline will bend around these knots. Fortunately, the exact position of the knots is usually not critical to the shape that the spline will take. It is common to specify the location from the distribution of the predictor variable.<sup>174</sup> More difficult is the choice of the number of knots. Empirical illustrations have shown that 5 knots is sufficient to capture many non-linear patterns. In smaller data sets, it may often be reasonable to use linear terms or splines with 3 knots (2  $df$ ), especially if no strong prior information suggests that a non-linear function is necessary.<sup>174</sup> If a large data set is available, 4 or 5 knots are reasonable, especially if we anticipate a non-linear function.

---

<sup>1</sup>GAMs are also often used for nonparametric regression functions, such as lowess.

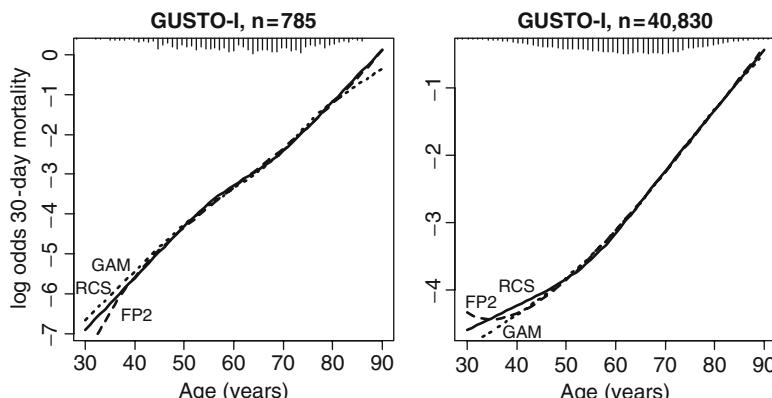
RCS of increasing complexity are not nested functions, so testing of higher-order transformations to simplify a complex non-linear model in a stepwise manner is not formally correct. It is, however, possible to study the increase in model likelihood ratio (LR) while taking the extra degrees of freedom into account, e.g. as  $\chi^2 - 2 \times df$  (Akaike's Information Criterion, AIC).

### \*9.3.4 Example: Functional Forms with RCS or FP

We examine the transformations for continuous predictors in the GUSTO-I study; both in a large subsample ( $n=785$ ) and the full data set ( $n=40,830$ , Fig. 9.3). In the subsample, we first fit a second-order fractional polynomial (FP2); the chosen model is  $AGE^{-2} + AGE^3$ . We compare the shape to an RCS function. An FP2 function uses 4  $df$ , but the shape can not have more than 2 bendings, which corresponds to an RCS with 4 knots (3  $df$ ). For age, weight, and height, FP2 functions were explored in univariate and multivariable logistic regression analysis; no statistically significant non-linearity was identified in the subsample. In the full GUSTO-I data set,  $AGE^2$  was chosen as the optimal transformation. For weight and height non-linearity was not statistically significant.

### 9.3.5 Extrapolation and Robustness

Extrapolation beyond the range of observed data is always dangerous, but this is possible with RCS functions. Essentially linear extrapolation will take place. In Fig. 9.2, the RCS with 4  $df$  (5 knots) draws a straight line for STE 9–11, while the cubic transformation with 3  $df$  curves downwards at STE 10.



**Fig. 9.3** FP, RCS (4 knots, 3  $df$ ), and GAM (3  $df$ ) functions for age in a subsample of GUSTO-I ( $n=795$ ) and in the full GUSTO-I study ( $n=40,830$ )

**Table 9.4** Options for dealing with continuous predictors in prediction models

Procedure	Characteristics	Recommendation
Dichotomization	Simple, easy interpretation	Bad idea
More categories	Categories capture prognostic information, better but are not smooth, sensitive to choice of cut-points and hence instable	Primarily for illustration
Linear	Simple	Often reasonable as a start
Polynomials	Square, cubic terms added; tails may behave unstable	Reasonable as checks for non-linearity
Transformations	Log, square root, inverse, exponent, etc.	May provide robust summaries of non-linearity
Fractional polynomials	Flexible combinations of polynomials; tails may behave unstable	Flexible descriptions of non-linearity
Restricted cubic splines	Flexible functions with robust behaviour at the tails of predictor distributions	Flexible descriptions of non-linearity
Splines in GAM	Highly flexible functions with smoothness set by penalty terms	Highly flexible descriptions of non-linearity

An interesting intermediate is to aim for a parametric transformation that captures most of the non-linearity. Adequacy of the fit can be tested by adding RCS functions based on the transformed variable.<sup>177</sup> For example, in a prostate cancer prediction problem, PSA values were linearly related to outcome after a log transformation,<sup>424</sup> while the original model in this prediction problem was constructed with RCS functions with 5 knots.<sup>227</sup> The log transformation often performs well for laboratory measurements, such as hormone concentrations. Restricting a continuous predictor to a parametric transformation may seem to harm the apparent performance somewhat. But it will limit optimism in performance, and increase a model's robustness. Care should always be given to predictions at the tails of a distribution.<sup>353</sup>

Empirical comparisons between FPs and RCSs have not yet been made. The main differences will occur at the tails of the distribution, exactly where the RCS was restricted to have better behaviour for prediction (see Fig. 9.3). If we have a predictor where a true curvature does occur at the tails, this will be captured by the FP and less by the RCS. If such curvatures are spurious, RCS will do better. In practice, both approaches may perform similarly in fitting a non-linear relationship given the same number of *df*. A number of options for dealing with continuous predictors in prediction models is summarized in Table 9.4.

## 9.4 Outliers and Truncation

Outliers are an important concern in statistical analyses.<sup>315</sup> Outliers are values that are outside the typical range for a variable. In box plots, a box is usually shown with the median and the interquartile range (IQR, 25–75 percentile). Outliers are defined by Tukey as values at least 3 times the IQR above the third quartile or at least 3

**Table 9.5** Dealing with outliers and extreme values of continuous predictors in prognostic research

Procedure	Method	Recommendation
Outlier detection	Box plot	Verify correctness (data entry error?) and biological plausibility (missing if implausible value)
Truncation	Shift low and high values to the middle	Shift approximately 0.5–1% of values to lower and upper ends of range

times the IQR below the first quartile.<sup>193</sup> We consider outliers as any values that potentially have a large influence in a regression model (Table 9.5).

The first question to address for an outlier is whether the value is realistic. For example, does the value reflect a data entry error? The records of a patient could be checked for that, e.g. the hospital chart or the case report form (CRF) when the patient participated in a trial.

Another check is on biological plausibility. This judgment requires expert opinion, and depends on the setting. For example, a systolic blood pressure of 250 mmHg is biologically plausible in the acute care situation for traumatic brain injury patients, but may not be plausible in an ambulatory care situation. Implausible values may best be considered as errors and hence set to missing.<sup>315</sup>

For biologically possible values, various statistical approaches are subsequently possible. To reduce the influence on the regression coefficients (“leverage”), we may consider to transform the variable by “truncation.” Very high and very low values are shifted to truncation points:

$$\begin{aligned} \text{If } X > X_{\max} \text{ then } X = X_{\max}; \\ \text{If } X < X_{\min} \text{ then } X = X_{\min}; \\ \text{else } X = X \end{aligned}$$

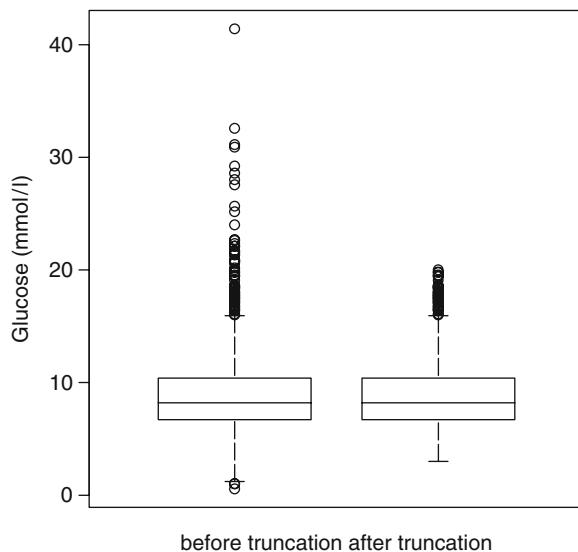
Here,  $x_{\max}$  and  $x_{\min}$  are the upper and lower truncation points. These may be defined from examining distributions, e.g. with box plots and histograms, and the predictor-outcome relationship.

#### \*9.4.1 Example: Glucose Values and Outcome of TBI

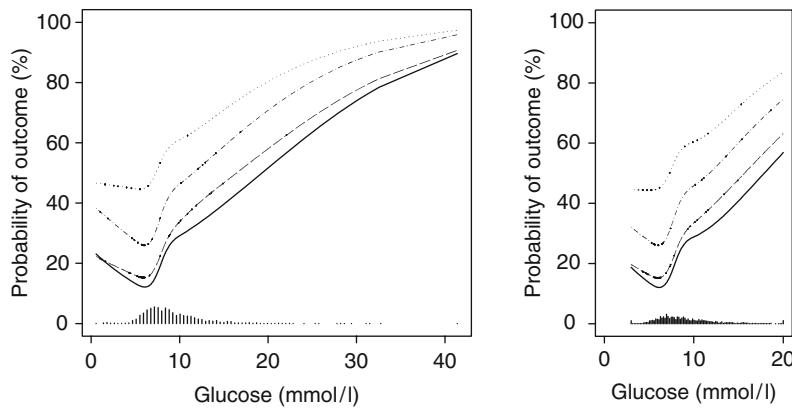
We consider glucose values measured at admission to predict 6-month outcome of patients with TBI. Outcome is measured with the Glasgow outcome scale (GOS), which has 5 levels (dead to good recovery, Chap. 8).

First, we consider an upper threshold for biologically possible glucose values at 100 mmol l<sup>-1</sup>. Among 4,831 values, 3 were above this threshold and set to missing. For further illustration, we consider 2,096 patients from the Tirilazad trials, who had glucose values and outcome available.

Second, we truncate glucose values to the interval 3–20 mmol l<sup>-1</sup> to limit the influence of extreme values (Fig. 9.4). The glucose – outcome relationship becomes slightly more linear after truncation (Fig. 9.5).



**Fig. 9.4** Distribution of glucose values for 2,096 TBI patients before and after truncation



**Fig. 9.5** Relationship of glucose to outcome at 6 months after TBI before and after truncation to the interval [3–20 mmol l]<sup>-1</sup>. The lowest line (*solid*) indicates the probability of mortality (GOS 1), the second the combination of mortality and vegetative (GOS 1 or 2), the third unfavourable outcome (GOS 1, 2, or 3), and the fourth line the probability of less than good outcome (GOS < 5). Relationships were analysed with restricted cubic spline functions in logistic regression models. The glucose–outcome relationship becomes slightly more linear after truncation. The distribution of glucose values is indicated at the bottom of each graph

## 9.5 Interpretation of Effects of Continuous Predictors

Effects of predictors can be interpreted through various presentations. A general way is to examine the predictions by predictor values, for example as the predicted probabilities from a logistic model. A graph is often very useful, especially for non-linear effects of predictors.

We can interpret the coefficients from a regression model by converting them to odds ratios (logistic regression) or hazard ratios (survival models, e.g. Cox). For binary variables such as gender, scaling is not a problem: the OR will refer to the comparison of males females if vs. males are coded one unit higher than females (e.g. 0/1). The OR will refer to females vs. males if the coding is reversed.

For linear, continuous variables, the scaling is very important for interpretability and comparability of effects. For example, the predictive effect is usually small for age coded in years. In GUSTO-I, the univariate logistic regression coefficient is 0.084, or an OR of 1.088 per year older, in the analysis of 30-day mortality. A simple improvement is to divide the age variable by 10 before estimating the model, such that the age effect is interpreted by decade. We can also multiply the coefficient that was estimated by year. In GUSTO-I, the univariate logistic regression coefficient becomes  $10 \times 0.084$ , and the OR  $1.088^{10} = 2.32$ . Also for other variables with a wide range in units, e.g. laboratory measurements, division by 10, 100, or 1000 may help. Comparability of effects of different continuous variables is still difficult then.

Another approach is to standardize linear, continuous predictors by dividing them by their standard deviation.<sup>2</sup> A variant on this approach was proposed by Harrell, i.e. to compare effects of predictors at the 75 vs. 25 percentiles.<sup>174</sup> For linear, continuous variables, this can be achieved by dividing the values by the interquartile range. Note that such rescaling does not affect *p*-values or predictions in any way.

For non-linear codings of continuous variables we can compare the predicted outcomes at the 75 vs. 25 percentile, but interpretability is difficult for parabolic relationships (e.g. a quadratic form); an OR near 1 may be found when comparing the 75 vs. 25 percentile predictions. A somewhat related, simple alternative is to code a non-linear variable with two dummy variables: one indicating values below the 25 percentile and one indicating values above the 75 percentile. The middle category is defined by the 25–75 percentile and serves as a reference category for both dummy variables. Such categorized coding implies a loss of information. Moreover, the effects in the dummy coding depend on the distribution of the predictor, similar to the dichotomized coding of age in Fig 9.1.<sup>355</sup> Dummy coding is therefore more useful for illustration of a predictor effect than for making predictions.

---

<sup>2</sup> Note that standardization does not work for categorical variables or non-linear transformations such as polynomials

### \*9.5.1 Example: Predictor Effects in TBI

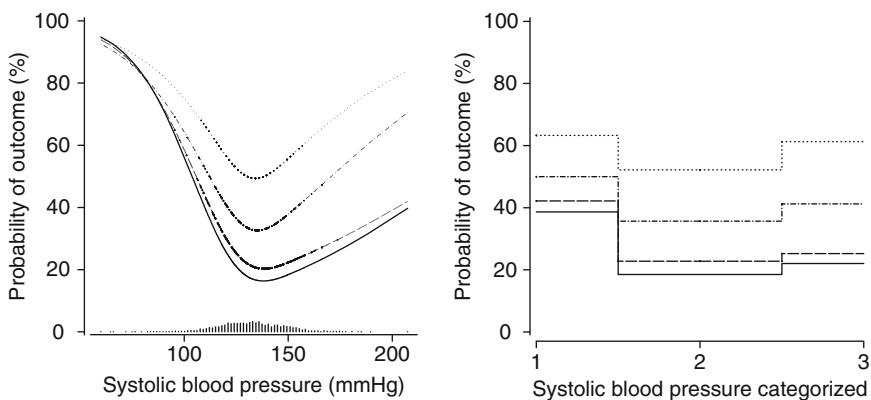
Various continuous predictors were measured at admission to predict 6-month outcome of patients with TBI. The relationships of age and glucose to outcome were reasonably linear. Effects were presented for the interquartile range (IQR, Table 9.6).

The relationship of systolic blood pressure with outcome was non-linear: low blood pressure was especially associated with a poor GOS, and GOS was also poorer at higher blood pressure values. This relation was modelled with an RCS with 3 knots ( $2\ df$ , Fig. 9.6).

The 75 percentile is a pressure of 141; the 25 percentile 121 mmHg. The OR for the comparison of predictions at these points is 1.39 [1.25–1.54]. For illustration, we categorize blood pressure at 120 and 150 mmHg (chosen because of clinical

**Table 9.6** Examples of coding of continuous predictors in predicting outcome of TBI

Predictor	Procedure	Interpretation
Age (linear)	Compare predictions at age 40–30 years Coding: Divide by 10	Age by decade
Systolic blood pressure (quadratic relationship)	Illustrate non-linear effect by making three categories, with dummy variables for <120 mmHg, > 150 mmHg	Effects for relatively low and high blood pressure



**Fig. 9.6** Relationship of systolic blood pressure to outcome at 6-months after TBI before and after categorization as <120, 120–150, and >150 mmHg. The lowest line (solid) indicates the probability of mortality (GOS 1), the second the combination of mortality and vegetative (GOS 1 or 2), the third unfavourable outcome (GOS 1, 2, or 3), and the fourth line the probability of less than good outcome (GOS<5).

relevance). This leads to ORs of 0.50 and 0.76 for blood pressure < 120 and > 150 vs. 120–150 mmHg, respectively. So, the 6-month outcome is twice as likely to be poor with a relatively low blood pressure, and 1.3 times with a relatively high pressure. This categorized coding has an  $R^2$  of only 2.4%, compared to 4.8% for the continuous coding, arguing against use of dummy coding for predictive purposes. The categorized coding is only intended to illustrate effects rather than to replace the non-linear continuous variable in a predictive model.

## 9.6 Concluding Remarks

We have seen that some decisions on coding can be made while we are still blinded to the relationship of the predictor to the outcome in our sample. Such blinding limits overfitting. Another general strategy is to use codings of predictors as used in other studies.

Special attention is required for continuous predictors. Most natural processes have a more or less smooth association with an outcome. Simple extensions of linear terms, such as the square and square root, can be useful, as well as more flexible functions such as restricted cubic splines and fractional polynomials.

Here, we focused on the effects of single predictors, which are usually first considered in a univariate analysis. The effects may also be studied with adjustment for other predictors (“confounders”). The aim may also be to derive a prediction model, with less interest in the specific forms of the relationships of each predictor with the outcome. A detailed modelling strategy has been proposed for the simultaneous selection of predictors for a prediction model and their FP transformations (“multivariable fractional polynomial (MFP) modeling”<sup>367</sup>). Harrell has suggested to determine the number of degrees of freedom with RCSs in univariate analyses, and use the chosen level of complexity in further multivariable analyses, irrespective of statistical significance of higher-order terms.<sup>174</sup> But another reasonable strategy might be to fit a model with all continuous predictors in flexible forms, e.g. with 5  $df$ , and then decide on reducing  $df$  per predictor based on their contribution to the model, e.g. according to partial  $R^2$ . Stronger and weaker predictors are then given more or less flexibility, respectively, without considering the degree of non-linearity. Further discussion of non-linearity in multivariable analysis is provided Chap. 12.

### 9.6.1 Software

RCS and FP functions can be used with any statistical program, but may require some programming. RCS functions are very easy to use with modern software such as R (Courier function) and Stata. Algorithms for fractional polynomials are available for R, Stata, and SAS.

## Questions

### 9.1 Dichotomization: a bad idea<sup>355</sup>

- (a) What are the problems of dichotomization when studying the effect of one specific predictor, such as age?
- (b) What are the problems of dichotomization when studying the effect of gender (male vs. female) and potential confounders, such as age, are dichotomized in an adjusted analysis?
- (c) What are the problems of dichotomization of predictors, such as age, when we aim to make individualized predictions?

### 9.2 Categorization of continuous predictors<sup>268</sup>

In an analysis of BNP, the authors of a paper state: “To produce odds ratios, cut points were used for age (65 years) and BNP (62 pg/mL) to reduce them to nominal variables.”<sup>268</sup>

- (a) Why should continuous predictors not be categorized?
- (b) For what purpose they could?
- (c) What suggestion would you have for the authors if they want to calculate interpretable odds ratios for continuous predictors?

### 9.3 Truncation of extreme values

- (a) Why are extreme values a problem in regression analysis?
- (b) How could you define truncation in one simple statement in R software for a continuous predictor “ $x$ ”?

### 9.4 Flexible continuous functions

What are advantages and disadvantages of flexible modelling of continuous predictors, e.g. using spline functions?

# Chapter 10

## Restrictions on Candidate Predictors

**Background** A major problem in predictive modelling is that we often have many candidate predictors available for the analysis, while the data set available for analysis is relatively small. A small sample size leads to problems as discussed in Chap. 5, such as limited power to test main effects of potential predictors, and too extreme predictions when predictions are based on the standard regression coefficients (overfitting). We discuss some procedures to increase the robustness and validity of a prediction model, including restriction of the number of candidate predictors, considering distributions of predictors, combining similar variables, and averaging the effects of similar variables. We provide a detailed description of a case study of modelling similar effects of aspects of family history for robust prediction of the presence of a genetic mutation.

### 10.1 Selection Before Studying the Predictor–Outcome Relationship

Ideally, candidate predictors are selected without studying the predictor–outcome relationship in the data under study. Two approaches are to use subject knowledge, and to study the distribution of predictors in the data under study.

#### 10.1.1 Selection Based on Subject Knowledge

The list of candidate predictors can often be reduced based on a review of the literature on the specific topic, combined with consulting experts in the field. The development of a prognostic model in situations without such subject knowledge on at least some predictors is nearly impossible, unless huge sample sizes are available. In many cases, a list in the order of 5–20 candidate predictors is reasonable to develop an adequate predictive model. Even in genetic research, it has been suggested that at most 20 genes should be included in a prediction model (although many more are usually considered, necessitating large sample sizes).<sup>258</sup> On the other hand, simulations show that many genes are needed if effects per gene are small.<sup>216</sup>

### \*10.1.2 Example: Too Many Candidate Predictors

In predicting the underlying diagnosis in children presenting with fever without obvious cause, models were developed that considered 57 candidate predictors.<sup>41</sup> The sample size was relatively small, with 231 patients and 58 having the diagnosis of interest (severe bacterial infection). The model was developed with stepwise methods after a univariate screening for statistically significant predictors. The model seemed to perform reasonably but external validation showed poor results.<sup>39</sup> On further analysis, bootstrapping of the full modelling process indicated a substantial decrease in model performance. A large part of the poor performance in new patients could be attributed to the modelling strategy with too many candidate predictors.<sup>401</sup>

### 10.1.3 Meta-Analysis for Candidate Predictors

We may consider to perform a systematic literature review or even a formal meta-analysis to identify candidate predictors. Some objections can be made against meta-analysis of univariate effects of predictors. Correlations between variables make that their effects are different in multivariable analyses. In the case of negative correlations, the univariate effects are suppressed. This results in no relation between the predictor and outcome in univariate analysis, while multivariable analysis does show a relationship. This situation may be relatively rare, but if this is suspected, the univariate effect of the predictor from previous studies should not be used as guidance to whether the candidate predictor is considered. In medical applications, most correlations between variables are however positive, making univariate effects larger than multivariable effects.

Another question is whether we should only count the number of times that a predictor was identified as “important,” or perform a formal meta-analysis. Counting may be sufficient for identification of the key predictors in a prediction problem. Meta-analysis is desired if we want to use the univariate effects of previous studies as a kind of prior estimate in our model (see Chap. 15). Publication bias is an important objection to meta-analysis of prognostic factors. Many studies will not report the effect of a predictor if not statistically significant; this biases the reported effects to more extreme values. One approach is to consider only studies that report the results for all predictors considered, but this may severely limit the numbers of studies in the meta-analysis.

### \*10.1.4 Example: Predictors in Testicular Cancer

We reviewed the prognostic value of a core set of prognostic factors for the histology of residual masses in testicular cancer.<sup>420</sup> The predictors that emerged as most relevant in the review were subsequently used in the prediction model. Some further

fine-tuning was done.<sup>425</sup> This fine-tuning included searching for good transformations of continuous variables, and choosing between three variables related to mass size: pre-chemotherapy mass size, post-chemotherapy mass size, and reduction in mass size (calculated as [presize–postszie]/presize).

### ***10.1.5 Selection Based on Distributions***

After restricting of the list of potential predictors, we should consider the distributions of predictors for missing values and width of the distribution. We may choose to eliminate variables that have a large number of missing values, especially if

- The predictor is relatively important in the problem, such that imputation of missing values will be suspect to many readers
- The predictor will be missing in applications of the model

We may choose to eliminate variables that have a narrow distribution especially if the variable is not expected to be important, this may be reasonable. For this reason, 6 of 49 potential predictors were eliminated in a study that aimed to predict the outcome of stroke.<sup>480</sup>

The situation is more difficult when a predictor has a very skewed distribution, but is known to be highly predictive. For example, in GUSTO-I, shock occurred in 2% of the patients but had a large effect (univariate odds ratio 10.9). Several options are available to deal with such a variable, such as

1. Include the variable as a predictor, since the effect is substantial.
2. Omit the variable from the model, since the effect cannot be estimated reliably; the model might be presented with a warning that specific conditions, such as shock are not included in the model.
3. Omit patients with shock, making the model applicable only to patients without shock.

As a default strategy we might prefer option 1, i.e. to include important variables, even though they are infrequent. The second option in fact holds for all variables that are not included in a model: predictions only consider information on variables that were included. The third option may only be defendable when we postulate that patients with shock are different with respect to prognostic relationships of other variables, i.e. we assume interaction between shock and other predictors.

## **10.2 Combining Similar Variables**

Sometimes variables can be grouped based on subject knowledge, or based on statistical clustering techniques.<sup>174</sup> For example, atherosclerosis is a systemic disease which is reflected in many symptoms. These symptoms can hence be considered as one group reflecting the underlying concept of “presence of atherosclerotic disease”

**Table 10.1** Illustrations of simple summary variables based on combinations of different predictors

Concept	Variables	Range
Atherosclerotic disease in predicting renal artery stenosis <sup>243</sup>	Any femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or had vascular surgery	0–1
Comorbidity in predicting surgical mortality in oesophageal cancer <sup>423</sup>	Count of chronic pulmonary disease, cardiovascular disease, diabetes, liver disease, renal disease	0–5
Family history in predicting a genetic mutation <sup>25</sup>	Sum of # affected first-degree family members plus 0.5 * # affected second-degree family members	0–3

(Table 10.1).<sup>243</sup> We could code “presence of atherosclerotic disease” as 0 or 1, depending on the presence of any symptom. We could also make a simple unweighted sum, by counting the number of symptoms. For 6 symptoms, the sum ranges from 0 to 6. In coding, we could truncate such an unweighted sum as 0, 1, 2, 3+, depending on the distribution, and start modelling with this sum as a linear, continuous predictor.

### \*10.2.1 Example: Coding of Comorbidity

Concomitant diseases are important in many prediction problems. These are commonly referred to as “comorbidity.” Various systems have been proposed to measure comorbidity. Weighted sumscores can be used such as proposed by Charlson<sup>68</sup> or ACE-27.<sup>337</sup> Note that these scores were derived from specific populations. Subject matter knowledge hence needs to support that it is reasonable to apply such as a pre-defined weighted score in another setting.

Alternative codings may be considered, depending on sample size. In very large data sets, e.g. using >100,000 records, a detailed coding can be imagined, which considers study-specific regression coefficients for each comorbidity. Also, a simple score can be attractive, for example the sum of a number of comorbidities.<sup>120</sup> Such an unweighted sum may be rather robust and generalize well to new patients. Such a simple sum was applied in a prediction model for surgical mortality after oesophagectomy (Table 10.1).<sup>423</sup>

### \*10.2.2 Assessing the Equal Weights Assumption

Simple sums of predictors make the assumption of equal weights for each predictor. This assumption can be assessed by adding the conditions one by one in a regression model that already contains the sumscore. The coefficient of the

**Table 10.2** Illustration of testing deviations for each condition in a sum score. Data from oesophageal cancer patients who underwent surgery (2,041 patients from SEER-Medicare data, 221 died by 30- days<sup>423)</sup>

Model	Logistic regression coefficient	P-value
Comorbidity sumscore	0.44 ( $\pm 0.13$ )	<0.001
+ chronic pulmonary disease	-0.22 ( $\pm 0.31$ )	0.48
+ cardiovascular disease	-0.13 ( $\pm 0.33$ )	0.69
+ diabetes	+0.32 ( $\pm 0.29$ )	0.27
+ liver disease	+1.31 ( $\pm 1.03$ )	0.20
+ renal disease	-1.09 ( $\pm 1.11$ )	0.33
		0.46 (overall, 4 df)

condition added in a model indicates the deviation from the common effect based on the other conditions. We can use an overall test for the decision whether a simple sum is reasonable, or that a more refined coding is required. In the example of comorbidity, we may consider the sum of five comorbidity conditions (Table 10.2). We may assess the effect of each of the five conditions by fitting five logistic regression models, with a separate coefficient for the deviation from the common effect for each of the five conditions in turn. We note that the deviations from the common effect are relatively small, except for liver disease and renal disease. Renal disease even seemed to have a protective effect. Both effects were based on small numbers. The standard errors were large, and the effects were statistically non-significant. The overall test for deviations from the simple sum had a  $\chi^2$  statistic of 3.6, 4 df, and a p-value of 0.46, in a model with the simple sum and four comorbidities added (chronic pulmonary disease, cardiovascular disease, diabetes, renal disease). We hence stick to our assumption of a similar effect for all comorbidities.

### \*10.2.3 Logical Weighting

Instead of equal weights we can sometimes base weights on a logical relationship. For example, when we model family history, we know that the genetic distance between family members is 0.5 between second and first-degree relatives, and 0.25 between third and first-degree relatives. This relationship can be used to define a variable for family history (Table 10.1).

Such a coding was used in a model to predict the likelihood of a genetic mutation in patients suspected of Lynch syndrome.<sup>25</sup> A proband with one affected first-degree family member gets a similar score for family history (1) as a proband with two affected second-degree family members. An implicit assumption here is that the numbers at risk are similar for 1st and second-degree family members, e.g. with similar numbers and similar age distributions.

### \*10.2.4 Statistical Combination

Harrell proposed to use principal component analysis to summarize the information from all candidate predictors.<sup>175</sup> This clustering does not use information on the predictor-outcome relationship, and has shown promising results in empirical evaluations. However, some theoretical and practical objections can be made. For example, the interpretability of regression coefficients is lost, and all predictor values have to be filled in to calculate predictions. Few modelling studies used principal components analysis of the predictor variables, but the concept should be kept in mind. There is some similarity with the clustering analysis applied in some studies of genetic markers.<sup>387</sup>

## 10.3 Averaging Effects

In regression modelling, we usually start with modelling main effects of variables. We may subsequently assess interaction effects as tests for additivity of effects. Conceptually, main effects average over underlying subgroup effects. This averaging may be reasonable as long as no strong interactions exist, and adds to the robustness of the model. This issue has a parallel with how we study treatment effects in RCTs. The main question is on the average treatment effect, and subgroup effects are commonly considered as secondary analyses.<sup>18,339,477</sup>

### 10.3.1 Example: *Chlamydia Trachomatis* Infection Risks

The starting point for modelling determines how our final model may look. For example, prediction of *Chlamydia trachomatis* infections has traditionally focused on infection prevalences in women. However, when we have a data set which contains infection status for both men and women, we may debate how to view model development. On the one hand, we may develop a male model fully independent of the female model. This is equivalent to assuming interactions between all predictors and sex. The models for males and females may adequately fit risk patterns for both sexes separately, but the predictions will be less reliable because of the reduced sample size. The obvious alternative is to start with a model of the combined data, which assumes similar effects in males and females. This assumption can specifically be tested by interaction terms of sex  $\times$  predictor. In this example, only the effect of urogenital symptoms clearly differed between the sexes.<sup>145</sup>

### \*10.3.2 Example: Acute Surgery Risk Relevant for Elective Patients?

In the *Chlamydia trachomatis* example, we were interested in prediction for both males and females. In another case, we were specifically interested in patients undergoing elective replacement of a heart valve.<sup>454</sup> In our data set, we also had

information on patients undergoing acute valve replacement. Should these patients be excluded? We decided to include these patients in the modelling, of course with a main effect for acute vs. elective surgery. We tested whether predictive effects were different between these types of patients, and found no such indication for any predictor separately nor in an overall test for interaction. Hence, it might be reasonable to assume that increasing the sample size by adding these high-risk acute surgery patients helped to improve our predictions for elective patients. More precisely stated, we assume that the relevance of any bias is smaller than the increase in precision by increasing sample size. This assumption seems reasonable from the data, but paradoxically sample size limits the power to detect differential effects. So, subject matter knowledge is the main guidance whether effects would be too different to model in the total group.

## \*10.4 Case study: Family History for Prediction of a Genetic Mutation

We consider the case study of diagnosing mutations in patients suspected of Lynch syndrome, or “hereditary nonpolyposis colorectal cancer” (HNPCC).<sup>25</sup> Mutations can be diagnosed with a genetic test, which is costly. Therefore, some selection of patients for definite testing is required. Family history has traditionally been used for such selection. Age at diagnosis is an important predictor, with higher likelihood of a mutation with younger age. Also, the number of affected first- and second-degree relatives is important, with more affected family members making a mutation as cause of the cancer more likely.

### 10.4.1 Clinical Background and Patient Data

Lynch syndrome is the most common hereditary colorectal cancer syndrome in western countries, accounting for 2–5% of all colorectal cancers (CRC).<sup>270</sup> Lynch syndrome is associated with underlying mutations in the mismatch repair system, most commonly in the *MLH1* and *MSH2* genes. Several guidelines have been developed to identify Lynch syndrome families, including the Amsterdam Criteria<sup>32</sup> and Bethesda Guidelines.<sup>349,442</sup> Such guidelines intend to support health care providers to select subjects for mutation testing. More recently, empirically derived prediction models have been developed for the likelihood of mutations in individual patients or families, enabling a more refined selection of subjects. Some models use logistic regression,<sup>25,28,262,485</sup> while others use Bayesian methods.<sup>71</sup> Several aspects of family history are considered in these models, related to the presence and age at diagnosis of cancer in the proband (the index person who is first being tested in a family), and the presence and age at diagnosis of cancer in his/her relatives. Modelling family history is complex, since the spectrum of cancers associated with *MLH1* and *MSH2* mutations is diverse. Mutation carriers are mainly at risk of

developing colorectal and endometrial cancer.<sup>270</sup> Young age at diagnosis is a risk factor of being a mutation carrier, and family members with various degrees of genetic relationship to the proband need to be considered.

We consider a development sample of 898 patients who were tested for presence of mutations (130 with mutation). Patients usually had one or more of various cancer diagnoses, including CRC ( $n=536$ ), women with endometrial cancer ( $n=91$ ), and other HNPCC-related cancer ( $n=100$ ). Of the 898 patients, 118 had multiple cancers. Details on predictor and outcome definitions are described elsewhere.<sup>25</sup>

### **10.4.2 Similarity of Effects**

Colorectal cancer (CRC) at a young age is a well-known predictor of a mutation. Especially if multiple CRCs occur in the same patient, this is very suspect for underlying genetic cause. Further, CRC in the family history points at HNPCC. We illustrate the modelling of CRC effects for the prediction of the presence of a mutation.

#### **10.4.2.1 CRC in a Proband Before Age 50**

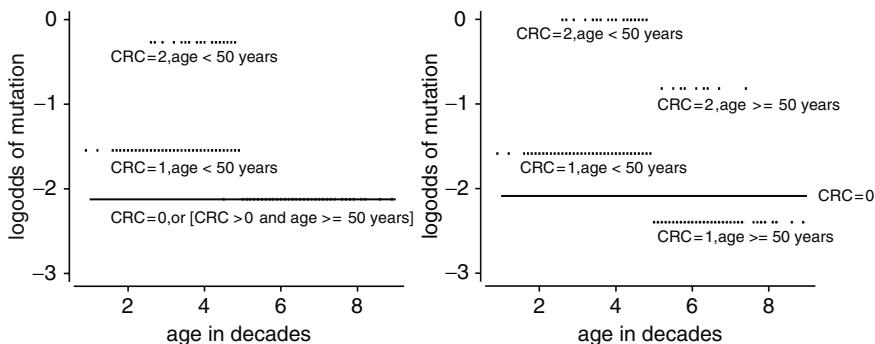
We study the effect of CRC below 50. We make 2 dummy variables: 1 for having 1 CRC below age 50 years ( $\text{CRC1} < 50$ ) and another for having 2 CRC diagnoses with the first diagnosis made below age 50 years ( $\text{CRC2} < 50$ ). The model is:  $\text{Mutation} \sim \text{CRC1} < 50 + \text{CRC2} < 50$ , where Mutation indicates the presence of a mutation (0/1), and  $\sim$  indicates the logistic regression link.

We can also use 2 terms for each, reflecting probabilities of mutation below and over 50 years:  $\text{Mutation} \sim \text{CRC1} < 50 + \text{CRC1} \geq 50 + \text{CRC2} < 50 + \text{CRC2} \geq 50$ .

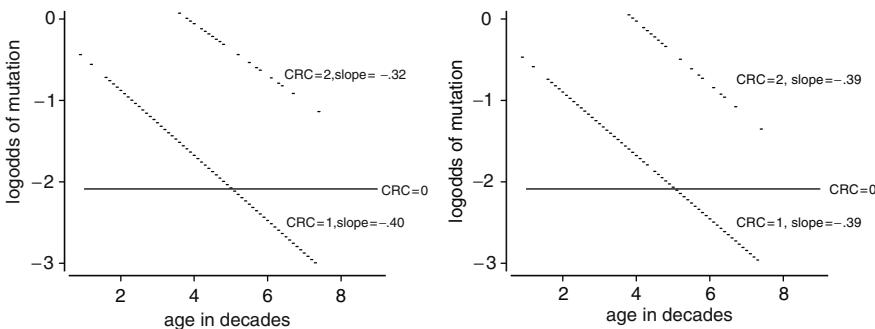
In the first formula, 2 coefficients are estimated for those with CRC at age  $< 50$  years (“ $\text{CRC} < 50$ ”). All other patients form the reference category. Estimated coefficients were 0.58 and 1.86. In the second formula, 4 coefficients are estimated for those with CRC, and patients without CRC are the reference (Fig. 10.1). Coefficients for  $\text{CRC1}$  were 0.54 and  $-0.50$ , and for  $\text{CRC2}$  2.09 and 1.82 (age  $< 50$  and age  $\geq 50$  years, respectively). So patients with 1 CRC, diagnosed after age 50, had a lower estimated probability of mutation compared to patients without a CRC.

#### **10.4.2.2 CRC in a Proband and Age Continuous**

To analyse age of diagnosis as a continuous predictor, we need to insert an age for those without CRC. A simple strategy would be to impute “0” for patients without CRC. An indicator variable would then be used for “CRC,” referring to the difference in probability of mutation at age zero between those with and without CRC. To obtain a more interpretable effect of the indicator variable for CRC, we set age



**Fig. 10.1** Mutation prevalence in relation to presence of a single or multiple CRCs diagnosed before age 50 in the proband (*left*), or before or after age 50 (*right*)



**Fig. 10.2** Mutation prevalence in relation to presence of a single or multiple CRCs, with age at diagnosis as a linear term (*left*, assuming separate age effects; *right*, assuming identical age effects)

at 45 years, since 45 years is around the average age of patients with CRC diagnoses. Further, we scaled age per decade ( $\text{CRCAge10} = \text{CRCAge}/10$ ). The interpretation of the indicator variables  $\text{CRC1}$  and  $\text{CRC2}$  is then as the presence of one or two CRCs vs. no CRC at the age of 45 years. For calculation of the linear predictor and graphical display (Fig. 10.2), the specific coding is irrelevant.

We may analyse the effect of the age at diagnosis of CRC with separate coefficients for  $\text{CRC1}$  and  $\text{CRC2}$  patients:

$$\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRC1age} + \text{CRC2age},$$

where  $\text{CRC1age}$  and  $\text{CRC2age}$  indicate the age of CRC diagnosis; in those without CRC the age is arbitrarily set close to the mean age of diagnosis (45 years). The main effects  $\text{CRC1}$  and  $\text{CRC2}$  are interpretable as the effect at age 45 of having a CRC diagnosis (one or multiple CRCs).

We may also assume a single  $\text{CRCAge}$  effect for both  $\text{CRC1}$  and  $\text{CRC2}$  patients:

$$\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRCage}$$

A test of whether the more complex model is better than the simpler one is provided by a likelihood ratio test (comparison of the  $\chi^2$  statistics, 1  $df$ ).

The age effects were very similar in both groups (CRC1age coefficient  $-0.40$ , CRC2age coefficient  $-0.32$ ), and the difference in effects was far from significant ( $p=0.80$ ). Hence, it is reasonable to assume a single age effect for patients with one or two CRCs; performance remained identical (Table 10.3).

We may subsequently test for non-linearity in the age effect. A linear coding was reasonable, since we found no improvement in fit by adding a square term ( $p = 0.50$ ) or considering restricted cubic splines (3 knots, non-linearity  $p=0.76$ ; 4 knots, non-linearity  $p=0.94$ ).

#### 10.4.3 CRC and Adenoma in a Proband

Adenoma polyps can be considered as precursors of CRC. They hence occur on average before the age of diagnosis of CRC. But the predictive effect for, e.g. a 10 years younger diagnosis of adenoma is *a priori* expected to be similar to the age–outcome relationship for CRC. Let us first consider the CRC and adenoma effects plus their age effects:

$$\text{Logit}(\text{Mutation}) = \text{CRC1} + \text{CRC2} + \text{CRCage} + \text{Adenoma} + \text{AdenomaAge}$$

The coefficients for the age effects are  $-0.38$  for CRC and  $-0.36$  for adenoma. It is tempting to estimate only 1 coefficient for these two effects. However, among a total of 141 patients with adenomas, only 100 had *only* adenomas as their diagnosis. CRC and adenoma are hence not mutually exclusive. How can we force the CRCage and AdenomaAge effects to be identical? In other words, we want to estimate one  $\beta_{\text{CRCAdenoma}}$  instead of  $\beta_{\text{CRC}}$  and  $\beta_{\text{Adenoma}}$  in a regression equation as

$$\text{Logit}(\text{Mutation}) = \beta_{\text{CRC}} \times \text{CRCage} + \beta_{\text{Adenoma}} \times \text{AdenomaAge} + \dots$$

The requirement is that  $\beta_{\text{CRC}} = \beta_{\text{Adenoma}}$ . This is achieved quite simply:

$$\text{Logit}(\text{Mutation}) = \beta_{\text{CRCAdenoma}} \times (\text{CRCage} + \text{AdenomaAge}) + \dots$$

**Table 10.3** Performance of alternative modes for the predictive effect of CRC and its age of diagnosis in patients tested for mutations in HNPCC (898 patients, 130 mutations). The third coding is preferred (3  $df$ ), with a single, linear term for the continuous variable “CRCage”

Model	<i>df</i>	$R^2$	<i>C</i>
CRC1<50+CRC2<50	2	4.6%	0.602
CRC1<50+CRC2<50+CRC1>=50+CRC2>=50	4	6.9%	0.634
CRC1+CRC2+CRCage	3	7.6%	0.651
CRC1+CRC2+CRC1age+CRC2age	4	7.6%	0.649

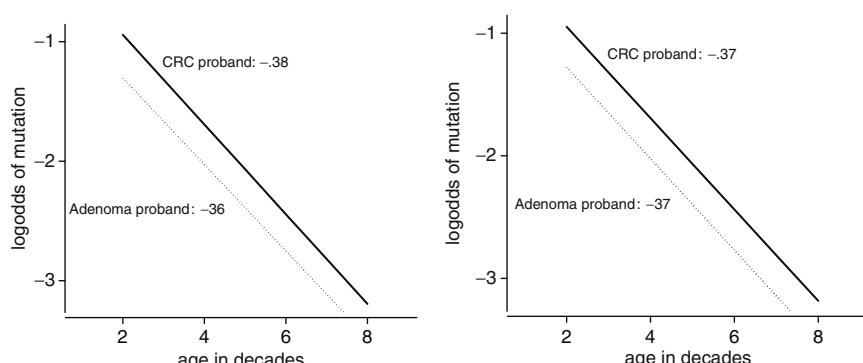
Again we include the indicator variables CRC1, CRC2, and adenoma in such a model. CRCage and AdenomaAge are set to 45 years for those with missing diagnoses, and recoded per decade for better interpretability of effects. The value of  $\beta_{\text{CRCAdenoma}}$  was  $-0.37$  per decade: in between the effects for the two separate coefficients  $\beta_{\text{CRC}}$  and  $\beta_{\text{Adenoma}}$  (Fig. 10.3).

#### 10.4.4 Age of CRC in Family History

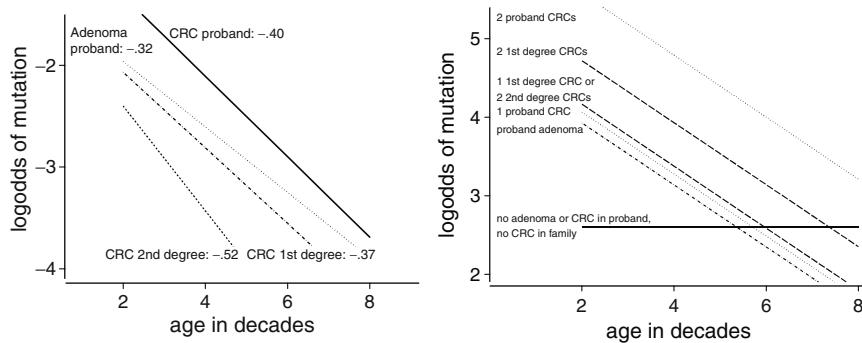
A further extension is to consider the effects of age at CRC diagnosis in first and second-degree relatives (Fig. 10.4). A CRC diagnosis at young age in a relative is more suspect for HNPCC than a CRC diagnosis at a more advanced age. We can again assume that the age effects should be similar, and add indicator variables for the presence of first or second-degree relatives. Four separate age effects are fitted with the formula:

$$\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRCage} + \text{Adenoma} + \text{AdenomaAge} + \text{CRC1st} \\ + \text{CRCage1st} + \text{CRC2nd} + \text{CRCage2nd},$$

where CRCage and AdenomaAge indicate age at diagnosis of CRC and adenoma in the proband, respectively, and CRCage1st and CRCage2nd indicate age at diagnosis of CRC in first and second-degree relatives, respectively. CRC1 and CRC2 refer to 1 or 2 CRCs in the proband, Adenoma to adenoma in the proband, CRC1st and CRC2nd to the number of CRC affected first and second-degree relatives.



**Fig. 10.3** Mutation prevalence in relation to age at diagnosis of CRC and age at diagnosis of adenoma as a linear term (left, assuming separate age effects; right, assuming identical age effects)



**Fig. 10.4** Mutation prevalence in relation to age at diagnosis of CRC in the proband, first, or second-degree relatives, and age at diagnosis of adenoma. Separate logistic regression coefficients were estimated for the four age effects in model 5a (*left*). In the *right panel*, one single age effect is estimated by considering the sum of all four ages, and family history is summarized in a single weighted score instead of four separate family history effects (1 or 2 first-degree relatives with CRC, 1 or 2 second-degree relatives with CRC, model 5c). The predictive performance of model 5c was similar to that of model 5a while using 5 instead of 11 degrees of freedom (Table 10.4)

A single effect for the four age variables is estimated by calculating the sum of all four ages, to force the four age coefficients to be the same:

$$\text{CRC.Adenoma.age} = \text{CRCAge} + \text{AdenomaAge} + \text{CRCAge1st} + \text{CRCAge2nd}.$$

We reduce a model with four age effects to a model with a single common age effect for CRC and adenoma.

Moreover, we can combine the family history of first and second degree relatives as  $\text{CRCfam} = \text{CRC1st} + 0.5 \times \text{CRC2nd}$ , instead of considering indicator variables for having 1 or 2 first degree relatives, and 1 or 2 second degree relatives with CRC. So, we reduce a concept with 4 to 1 *df*.

The chosen coding for CRC and adenoma effects hence is as

$$\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{adenoma} + \text{CRCfam} + \text{CRC.Adenoma.age}$$

In total, we reduce a model with 11 *df* to 5 *df*. We find that the performance of both model variants is similar (Table 10.4). Importantly, more stability is expected, and better generalizability to future patients.

#### 10.4.5 Full Prediction Model for Mutations

A final predictive model was constructed where other diagnoses were treated in a similar way. For endometrial cancer, we create an indicator variable with as reference category females without endometrial cancer and all males ("endo"). Age at diagnosis in the proband was combined with age at diagnosis in first and second degree

**Table 10.4** Performance of alternative modes for the predictive effect of age of diagnosis for CRC and adenoma in the proband, and CRC in first and second-degree relatives. Models created with data from patients tested for mutations in HNPCC (898 patients, 130 mutations). The fourth coding is preferred, with a single, linear term for a continuous age effect.

Model	<i>df</i>	<i>R</i> <sup>2</sup>	<i>C</i>
CRCage + AdenomaAge; adenoma; CRC1+CRC2;	5	8.4%	0.662
CRC.Adenoma.Age; adenoma; CRC1 + CRC2	4	8.4%	0.662
CRCage + AdenomaAge + CRC1stAge + CRC2ndAge + adenoma; CRC1 + CRC2; CRC 1st (0,1,2), CRC2nd (0,1,2)	11	19.4%	0.767
1 age effect; adenoma; CRC1+CRC2; CRCfam	5	18.3%	0.757

relatives, and family history was coded as for CRC: # affected first degree relatives + 0.5 # affected second-degree relatives.

For other HNPCC related cancers, indicator variables were created for the proband (“other”) and relatives (“rother”, scored as for CRC and endo). No age effect was identified. The final model hence was:

$$\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{Adenoma} + \text{CRCfam} + \text{CRC.Adenoma.age} + \text{Endo} \\ + \text{EndoFam} + \text{EndoAge} + \text{Other} + \text{OtherFam}$$

This model incorporates information on CRC, adenoma, endometrial cancer, and other cancer diagnoses from the proband and from first and second-degree relatives with only 10 *df*. The *R*<sup>2</sup> was 24.9%, and *c* 0.81. External validation was performed with 1,016 new patients from the same setting.<sup>25</sup> The *R*<sup>2</sup> was 24.0% and *c* 0.80.

This case study illustrates how predictors related to the same underlying phenomenon can be combined for parsimonious and robust modelling. Such a strategy may especially be useful in relatively small data sets, where specification of complex models would not be reasonable, and lead to unstable estimation of regression coefficients. Further statistical detail is provided elsewhere.<sup>400</sup>

## 10.5 Concluding Remarks

Model specification is the most difficult step in prediction modelling. We considered several steps to develop more robust models for prediction purposes by reducing the degrees of freedom considered in the modelling process.

- 1 The first step obviously is to match the number of candidate predictors with the available effective sample size. If we have only a small sample for modelling, a more restricted set of candidate predictors is necessary compared to the situation

of a large sample for modelling. Subject matter knowledge may assist in limiting the selection, such as literature review and consultation of experts.

- 2 Second, we may consider distributions of predictors. We may exclude candidate predictors based on number of missing values and skewness of distributions.
- 3 Related predictors can sometimes be combined in summary scores, such as illustrated for comorbidity.
- 4 Finally, we may want to average effects of predictors across subgroups for more stability, exploit logical relationships, and estimate single effects for combinations of predictors. As illustrated for the case study on mutation prevalence, this may lead to a parsimonious, robust model, which still captures most of the predictive information.

The risk of such restrictions is that we may exclude certain predictors and specific predictor effects from a model; the specific circumstances should guide us in what strategy is most reasonable.

## Questions

### 10.1 Data reduction

- (a) What is meant with candidate predictors, in contrast to included predictors in a model?
- (b) What problems can occur when considering many candidate predictors for inclusion in a prediction model?
- (c) What kind of strategies do not use the predictor–outcome relationship in reducing the number and degrees of freedom of the candidate predictors?
- (d) What kind of strategies do use the predictor–outcome relationship while attempting to reduce the number and degrees of freedom of the candidate predictors?

### 10.2 Combining similar variables

What objections can be made against the combination of similar variables in summary predictors (e.g. comorbidity scores), or the combination of effects of similar predictors (e.g. age effects in the family)?

### 10.3 Interpretation of case study (Sect. 10.4)

The case study illustrates robust coding of family history for prediction of an underlying mutation.

- (a) CRCage may be coded as  $(45 - \text{years})/10$ . How can we then interpret the coefficients for CRCage, CRC1 and CRC2 in the regression formula  $\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRCage}$ ?
- (b) The model for aspects of CRC in the family was  $\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{adenoma} + \text{CRCfam} + \text{CRC.Adenoma.age}$  What would the values of the predictors be for someone with no CRC or adenoma, and no CRC in the family?
- (c) How can we test for deviations of the age effects in first and second degree relatives in the variable  $\text{CRC.Adenoma.age}$  (2 age effects vs. 1 age effect for relatives)?
- (d) Endometrial cancer can only occur in females. How do we code the predictors Endo and EndoFam to obtain interpretable coefficients?

### 10.4 Splitting analyses

A researcher considers to analyse males and females separately, and proposes to split the files for such analyses. A colleague says there is no need to do so. How can effects for males and females be analysed in one dataset?

# Chapter 11

## Selection of Main Effects

**Background** Model specification is the most difficult part of prediction modelling.<sup>472</sup> Especially in smaller data sets it is virtually impossible to obtain a reliable answer to the question: which predictors are important and which are not? In this chapter, we focus on the advantages and problems that are associated with model reduction techniques such as stepwise selection, including overfitting and the quality of predictions from a model. Specific issues include instability of selection, biased estimation of coefficients, and exaggeration of  $p$ -values. We explore the influence of including noise variables as predictors in a model, and find that their influence is not so detrimental to legitimize widespread use of stepwise methods. Alternative approaches include making a list of a limited number of candidate predictors to consider for the prediction model, e.g. based on a meta-analysis of available literature, and some more modern selection methods.

### 11.1 Predictor Selection

#### 11.1.1 *Reduction Before Modelling*

In the previous chapters, we have discussed several approaches to limit the degrees of freedom that are considered in the modelling process. Use of subject matter knowledge is essential to preselect candidate predictors, e.g. from a systematic review of the literature, and from discussions with experts in the field. We should also consider strategies for robust coding of predictors (Chap. 10). These steps may reduce the chance that there are noise variables among the candidate predictors. Predictive modelling then turns into an estimation problem rather than a testing problem. Ideally, we end up with a limited list of candidate predictors, which can all be entered in a “full model,” which contains the main effects of all candidate predictors.<sup>174</sup> Model specification is then restricted to consideration of model assumptions such as additivity (with interaction terms) and non-linearity (with non-linear terms, see Chap. 12).<sup>14</sup>

### ***11.1.2 Reduction While Modelling***

We may consider to reduce the set of candidates predictors for various reasons. One reason is that it is not practical to use a large set of predictors in medical practice. Formally, this is only an argument if variables are not all available in future patients, or have a cost associated with their collection. Also, some predictors may have very small or implausible effects, which makes it questionable why they are included in a model. In some circumstances, we may also have a list of new predictors, where some are expected to have no true relationship to the outcome at all. For example, when predicting valve fracture with production characteristics, it was unclear which specific aspects would be important.<sup>42</sup> Also in genetic and proteomic research, identification of which characteristics are predictive from a very large set of candidate predictors is the main goal. This makes such analyses quite exploratory in nature, more aimed at biological knowledge discovery than prediction.

### ***\*11.1.3 Collinearity***

Another argument in favour of model reduction includes collinearity, which refers to the issue that predictors may be strongly correlated with each other. Collinearity is reflected in “variance inflation factors” (VIF), which measure the degree to which collinearity among the predictors degrades the precision of estimate coefficients. Collinearity hampers reliable estimation of regression coefficients of the correlated variables, especially if correlations are very strong (say correlation coefficient  $r>0.8$ , or  $VIF>10$ ).<sup>472</sup>

Is collinearity relevant for clinical prediction models?

- Correlations do of course exist between predictors. We perform multivariable analysis to consider the joint effects of predictors which cannot be inferred from a univariate analysis.
- In many practical examples, correlations are less than 0.5. For example, the strongest correlation in the GUSTO-I study is between height and weight with  $r=0.5$ . All other correlations are weaker, typically with  $r$  around 0.1–0.2.
- Sometimes we create highly correlated variables, e.g. age and  $age^2$  ( $r>0.9$ ), but we can estimate their coefficients quite reliably.

If predictors are relatively strongly correlated, it may be wiser to combine them in a single combined variable. For example, a strong correlation generally exists between diastolic blood pressure (DBP) and systolic blood pressure (SBP), with  $r$  of 0.62 in one study.<sup>14</sup> When choosing between DBP and SBP, “mean blood pressure” may be a better choice ( $2\times DBP + 1\times SBP$ ) than choosing either one of them. But again, subject matter knowledge is important. For example, systolic pressure is known to be the more relevant predictor for cardiovascular risk.<sup>428</sup>

### 11.1.4 *Parsimony*

Another argument in favour of smaller models is made by referring to the principle of parsimony (“Occam’s razor”). This principle states that simpler explanations are preferred over more complex explanations. Better predictive abilities can be expected from a simpler model. This is an appealing philosophical principle when judging two alternative theories. It is, however, not obvious how this principle translates to prediction models. The traditional reasoning is that a model where some less significant variables are eliminated is more parsimonious than a full model with more predictors, and is hence to be preferred. When we consider how predictive regression models are created we however come to the opposite conclusion:

- A full model does not ask more from the data than estimating regression coefficients
- A reduced model asks two questions:
  - (a) which variables can be eliminated?
  - (b) what are the coefficients of the remaining predictors, given that the other variables are eliminated?

So, a reduced model reflects the answer to two questions rather than the answer to one, which is arguably more complex.

A practical issue may be that smaller models are easier to interpret and use in practice. For example, prediction rules with a few, simple predictors may be easy to remember for clinicians. This “parsimony” comes at a price: such smaller models are conditional on selecting the right predictors from the candidate predictors.

### 11.1.5 *Should Non-Significant Variables Be Removed?*

Finally, some may argue that statistically non-significant variables should not be included as predictors in a model, since their effects are not proven. This belief may result from mixing the fundamental statistical concepts of hypothesis testing and estimation. Prediction is about estimation; hence it is quite reasonable to include a predictor with a  $p$ -value higher than the magical value of 5%. Especially, this is reasonable if the data set is relatively small, the predictor uncommon (a rare but strong predictor such as “shock” in GUSTO-I), or when the predictor is well known from previous research to be predictive. Non-significance does not mean that there is evidence for a zero effect of a predictor; as always absence of evidence is not evidence of absence.<sup>11</sup> Finally, simulation studies with true noise variables in a model do show only a limited decrease in predictive ability<sup>410</sup>.

### **11.1.6 Summary Points**

In sum, some arguments can be put forward in favour of predictor selection based on findings in our data:

- Larger models are less practical to work with
- Some predictors may have very small or implausible effects

False arguments include

- “Statistically non-significant variables should be excluded”; for estimation, significance testing is not relevant, especially if estimated effects are supported by subject knowledge
- “Collinearity precludes obtaining reliable predictions”; although collinearity makes estimates of individual coefficients unstable, reliable predictions can still be obtained
- “Referring to the parsimony principle”; this may hold when pre-specified models are compared, not when models are selected by studying patterns in the data.

## **11.2 Stepwise Selection**

We will first consider traditional approaches such as stepwise selection of predictors in a model, followed by some promising alternative approaches to model selection.

### **11.2.1 Stepwise Selection Variants**

Currently, stepwise selection methods are probably the most widely used in medical applications. These automated methods aim to include only the most significant predictors in a model. Significance is determined with a selection criterion: the  $F$  test in linear regression; a likelihood ratio (LR), Wald, or Score statistic in logistic or Cox regression models. Forward selection starts with inclusion of the most significant candidate predictor to a model that does not contain any predictor. Backward selection starts with elimination of the least significant candidate predictor from a full model including all candidate predictors. Forward and backward selection may also be combined, such that an iterative procedure is followed.

A backward selection approach is generally preferred if stepwise selection is attempted. First, the modeller is forced to consider the full model with a backward approach, and can judge the effects of all candidate predictors simultaneously.<sup>174</sup> Second, correlated variables may remain in the model, while none of them might enter the model with a forward approach.<sup>96</sup>

An extension of stepwise selection strategies is “all possible subsets regression.” Here, every possible combination of predictors is examined to find a best

fitting model.<sup>289</sup> All possible subsets regression can identify combinations of predictors not found by the more standard forward or backward procedures. This comes at a price: we examine many models, with multiple testing, easily resulting in overfitted models.<sup>96</sup>

### \*11.2.2 Stopping Rules in Stepwise Selection

The stopping rule for inclusion or exclusion of predictors is a central issue in stepwise selection methods. It is far more important than the specific variant of the stepwise selection method (e.g. forward, backward, combined, all possible subsets). Usually, one applies the standard significance level for testing of hypotheses ( $\alpha=0.05$ ), but the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) are also often used. In all possible subset selection, the stopping criterion often is to maximize Mallow's  $C_p$ , which is similar to optimizing AIC. Stopping rules are usually applied for testing contributions of individual predictors, but may also be applied to the pooled degrees of freedom of unselected predictors.<sup>174</sup>

AIC and BIC compare models based on their fit to the data, but penalize for the complexity of the model. With AIC, we require that the increase in model  $\chi^2$  has to be larger than two times the degrees of freedom:  $\chi^2 > 2 \text{ df}$ . When considering a predictor with 1  $\text{df}$ , such as gender, this implies that  $\chi^2$  has to exceed 2, equivalent to  $p<0.157$ . With 2  $\text{df}$ ,  $\chi^2 > 4$ , or  $p<0.135$ ; and with 4  $\text{df}$ ,  $p<0.092$  (Table 11.1).

With BIC, we penalize the model fit such that  $\chi^2$  has to exceed  $\log(n)$ . The effective sample size should be used for  $n$ , e.g. the number of events in Cox regression for survival data.<sup>474</sup> With small sample size, e.g.  $n=20$ , BIC is equivalent to  $p<0.083$  for selection. With larger sample sizes, the  $p$ -values are much lower (Table 11.2). Hence, selection with BIC will generally lead to smaller models than selection with AIC.<sup>1</sup> The theory behind AIC and BIC criteria can be found elsewhere in detail<sup>181</sup>; for the applied researcher, the  $p$ -value that is effectively used as a stopping criterion is most relevant.

**Table 11.1**  $P$ -value associated with Akaike's Information Criterion (AIC) for selection of candidate predictors with different degrees of freedom ( $\text{df}$ )

Df	$\chi^2$ has to exceed 2 times $\text{df}$	P-value
1	2	0.157
2	4	0.135
3	6	0.112
4	8	0.092
5	10	0.075

<sup>1</sup> Selection with BIC may lead to “underfitting,” since many predictors are excluded

**Table 11.2**  $P$ -value associated with Bayesian Information Criterion (BIC) for selection of candidate predictors

N	$\chi^2$ has to exceed $\log(n)$	$P$ -value
20	3.0	0.083
50	3.9	0.048
100	4.6	0.032
200	5.3	0.021
500	6.2	0.013
1000	6.9	0.009
2000	7.6	0.006

There is no specific reason to stick to a  $p$ -value of 0.05, or low  $p$ -values as implied by applying BIC. Using AIC has been recommended.<sup>14</sup> Using even higher  $p$ -values ( $p < 0.20$  or  $p < 0.50$ ) have been found to provide more power for the selection of predictors with relatively weak effects,<sup>253</sup> and to provide better predictions in small data sets with a set of established candidate predictors.<sup>409</sup>

### 11.3 Advantages of Stepwise Methods

Stepwise selection methods have a number of advantages. They are usually relatively straightforward to apply in modern statistical packages. Some care should be taken with missing values; if we start with a full model, the number of available cases is restricted by the combination of missing values in any of the candidate predictors. It is therefore important to use imputed data set to deal with missing values. Multiple imputation (MI) poses some complexities if we would select predictors per imputed data set, where predictors may be selected in some replicates of the data set and not in other replicates. The preferable approach is to perform selection based on the results from the combined data sets. For example, with a backward procedure, we first obtain  $p$ -values for each predictor in a full model, fitted on MI data sets. We then eliminate the least significant predictor, provided that the  $p$ -value is higher than our stopping criterion. We refit the model in the MI data, and eliminate the next predictor. We stop when all predictors have  $p$ -values less than the stopping criterion.

Stepwise methods are also relatively objective. When another analyst is provided with the same data set and the same list of candidate predictors, the resulting selection should be very similar. The objectivity of stepwise selection makes it possible to replay this model reduction strategy in validation procedures such as the bootstrap (Chap. 16). Optimism can hence be estimated including model uncertainty.<sup>69,401</sup>

Stepwise methods usually reach their goal of making a model smaller. In larger data sets, such as GUSTO-I, all variables that are important for prediction will have small  $p$ -values. Sometimes  $p < 0.01$  is therefore chosen in large samples. In small data sets, only few variables may have such small  $p$ -values, resulting in small models (sometimes referred to as “underfitting”). This argues for the use of AIC or an even higher effective  $p$ -value.

## 11.4 Disadvantages of Stepwise Methods

Stepwise methods have various disadvantages, including

1. Instability of the selection
2. Biased estimation of coefficients
3. Misspecification of variability and exaggeration of  $p$ -values
4. Provision of predictions of worse quality than from a full model

These issues are explained and illustrated below.

### 11.4.1 Instability of selection

Stepwise selection considers a high number of combinations of predictors. Some of these combinations may actually be rather similar in how they fit the data. This instability may be illustrated by the observation that the selection of predictors may change when we consider a slightly different selection of patients for a model.<sup>20</sup>

The instability of selection can well be illustrated with subsamples of the GUSTO-I case study; very different selections arise (Table 11.3). For example, the selected predictors were age (A65), hypotension (HYP), and shock (SHO) in sample5 ( $n=429$  patients). We also considered the selection in the other 110 small subsamples where a logistic regression model could technically be fitted. The predictors A65, HYP, and SHO were among the predictors most often selected, with A65 in 80% of the 111 subsamples, HYP in 47%, and SHO in 53%. The candidate predictors TTR and DIA were selected in only 11% and 13%, respectively. The specific selection of these 3 predictors was however replicated in only 7 of the 110 other small subsamples.

The conclusion from this case study is similar to what was found in other studies: the specific selection of predictors is unstable and should be interpreted with much caution. Statements such as “the only independent predictors in this prediction problem were age, hypotension, and shock” are overinterpretations unless the sample size was huge, and the effects of the other predictors were much smaller compared to the predictors selected. Even worse overinterpretations are related to the order of entry of a predictor in a forward stepwise procedure, or rank order of the  $p$ -value in the selected model.<sup>96</sup>

The instability of selection depends on a number of factors. One crucial aspect is the sample size. In a large sample, more stability is to be expected, since we have more power to detect truly important effects. Table 11.4 illustrates that more predictors were selected in larger subsamples than in smaller subsamples. When considering 16 large regions in GUSTO-I of at least 2,000 patients (178 events on average), around 6–7 predictors had statistically significant effects, eliminating predictors such as TTR and DIA which had minor predictive effects.

Although a larger sample size helps in many ways, we are usually tempted to study more candidate predictors in such situations. This introduces instability

**Table 11.3** Illustration of variability in selection with backward selection with  $p < 0.05$ . The first 25 small subsamples are shown from GUSTO-I. The 8 predictor model could technically be fitted in 111 of the 121 subsamples

#	A65	Sex	DIA	HYP	HRT	HIG	SHO	TTR
1	*					*	*	
3	*	*				*	*	
4	*			*			*	
5	*			*			*	
6	*	*						
7					*	*	*	
8	*	*			*			
9		*				*	*	
11	*			*			*	
13	*				*			
14	*						*	
15	*				*		*	
16	*					*	*	
17	*						*	
18	*					*		
19	*				*			
20		*	*	*	*	*		
22	*	*					*	
23	*			*			*	*
24	*							
26	*					*	*	
27	*			*		*	*	
28	*			*			*	
...								
121	*	*			*			*
Selected	80%	22%	13%	47%	23%	29%	53%	11%

**Table 11.4** Summary of number of predictors selected with different selection strategies in subsamples from the GUSTO-I data set. Numbers are mean  $\pm$  SD

Samples	Ful	$P < 0.05$	AIC	$P < 0.5$
Small subsamples	8	$2.8 \pm 1.1$	$4.2 \pm 1.1$	$6.3 \pm 1.2$
Large subsamples	8	$4.8 \pm 1.1$	$6.0 \pm 1.0$	$7.0 \pm 0.9$
Regions	8	$6.6 \pm 1.0$	$7.1 \pm 0.8$	$7.8 \pm 0.4$

again: more candidate predictors implies more potential combinations of predictors. So, a crucial aspect is the ratio between number of candidate predictors and the effective sample size. Sometimes, a ratio of 10 events per variable (EPV) is advocated; this is however only a reasonable lower bound for prespecified models. For reliable selection among candidate predictors, an EPV of 50 may be better.<sup>410</sup> So, if we consider 8 candidate predictors, at least 400 events should preferably be analysed in a logistic regression model when we want to make firm statements on which predictors are important and which are not. The total GUSTO-I model easily fulfills the 1 in 50 criterion with 2,851 events in 40,830 patients, but this is exceptional.

The instability of selection procedures can well be studied for one specific data set with bootstrapping procedures (Chaps. 5 and 16). For larger data sets, the instability will not show up as extreme as with the small subsamples in Table 11.3. Also, when a few predictors have strong effects, and others have weak effects, this should be apparent from the selection pattern over bootstrap samples.

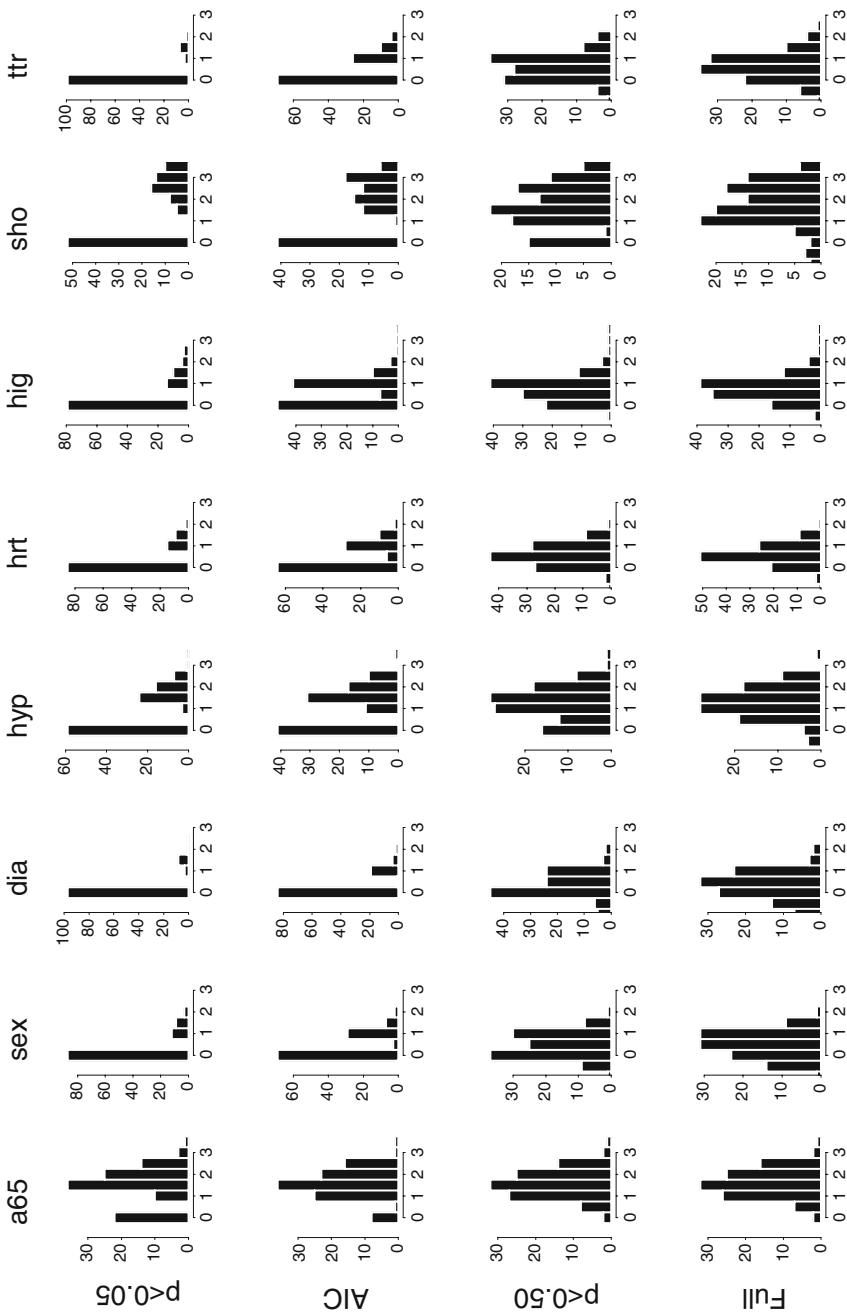
### ***11.4.2 Biased Estimation of Coefficients***

The problem of estimation after testing (“testimation”) was already discussed in Chap. 5. It clearly shows up in stepwise selected coefficients. The distributions of coefficients is biased away from zero, as illustrated in Fig 11.1 for the small subsamples in GUSTO-I. We note that many coefficients are set at zero; the predictor was not selected for the model. Between zero and 1 there is a gap; small estimated coefficients were not statistically significant and were set to zero. This gap is smaller when we select with a higher  $p$ -value (AIC or  $p < 0.50$ ), and fewer coefficients were set to zero. This is explained by the fact that smaller estimates of coefficients are included with a more relaxed stopping rule. The testimation problem is smaller in larger samples (see [www.clinicalpredictionmodels.org](http://www.clinicalpredictionmodels.org) for graphs).

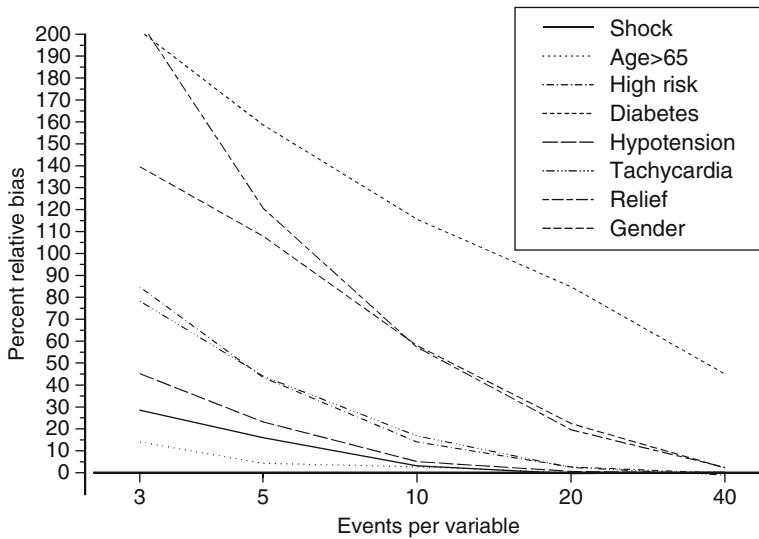
### ***11.4.3 Bias of Stepwise Selection and Events Per Variable***

We simulated small subsamples fully at random from GUSTO-I in a study on bias of estimated logistic regression coefficients in stepwise selected models.<sup>407</sup> Testimation bias was substantial when we studied coefficients of  $p < 0.05$  selected variables (Fig. 11.2). When we use a higher  $p$ -value, such as  $p < 0.50$ , the bias was much smaller (Fig. 11.3). For example, with Events per Variable (EPV) 10, the bias exceeds 50% for 3 predictors when selected with  $p < 0.05$ , but such a bias is not seen with  $p < 0.50$  selection. With a full model, the bias is small, especially for reasonable sample sizes (EPV  $\geq 10$ , Fig. 11.4). With an extreme as EPV 3, the absolute bias is  $< 20\%$  for all predictors.

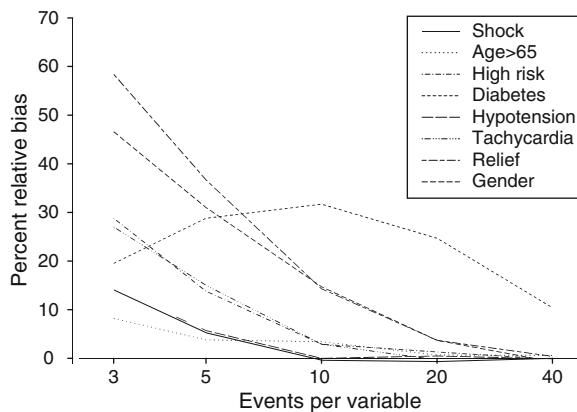
There is a direct relationship between the bias in an estimated effect and the frequency of selection. A strong predictor such as age is selected in many models and has limited bias. A weak predictor such as diabetes is selected in only a few models, and if selected, the coefficient is biased upwards by more than 100% with low EPV. There is also a relation with the underlying strength of a predictor. The mathematical relationship between strength of predictive effect and bias is shown in Fig. 11.5. The strength of effect is expressed as the ratio of the true coefficient value to the true standard error. Sample size influences the standard error, which decreases with increasing sample size. The bias appears to be at a maximum when the true coefficient to standard error ratio is around 0.6.



**Fig. 11.1** Distribution of logistic regression coefficients in 111 small subsamples within GUSTO-I. *First row:*  $p < 0.05$  selection; *second row:* AIC selection; *third row:*  $p < 0.5$  selection; *fourth row:* full model with all 8 predictors included. a65: age  $> 65$ ; dia: diabetes; hyp: hypotension; hrt: heart rate  $> 80$ ; hig: high risk (anterior infarction or previous MI); sho: shock; ttr: time to relief  $> 1$  h. Note that the coefficients in the stepwise selected models should be interpreted with caution: they are based on different sets of selected predictors. The general pattern is however that the coefficients in the stepwise selected models are zero or a value clearly above zero, since predictors with accidentally small effects are not selected. The coefficients follow an approximately normal distribution in the full models



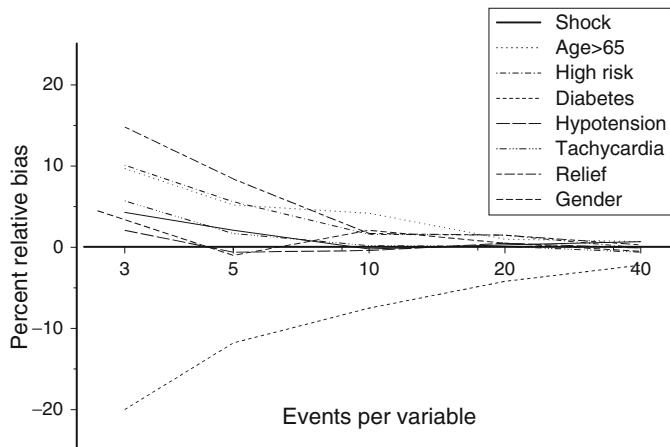
**Fig. 11.2** Average percent conditional relative bias in relation to the number of events per variable after backward stepwise selection with  $p<0.05$  from the 8 predictor model in random subsamples from GUSTO-I



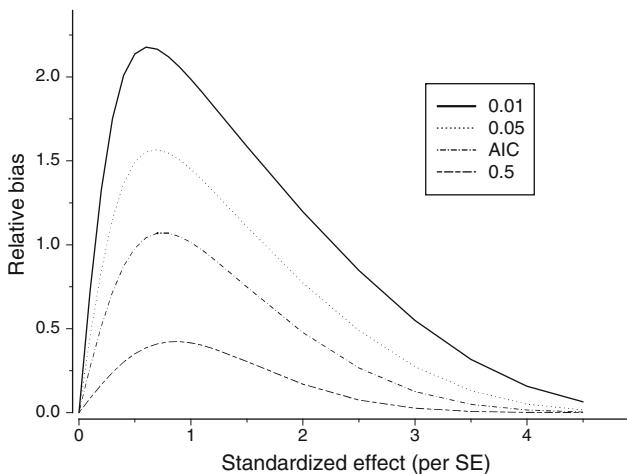
**Fig. 11.3** Average percent conditional relative bias in relation to the number of events per variable after backward stepwise selection with  $p<0.50$  from the 8 predictor model in random subsamples from GUSTO-I

#### 11.4.4 Misspecification of Variability

As noted in Fig. 11.1, the distribution of coefficients from stepwise selected models has a strange shape. From an unconditional perspective, coefficients are set to zero when the predictor was not selected. From a conditional perspective, only the values of coefficients of selected predictors are considered (Fig. 11.6).



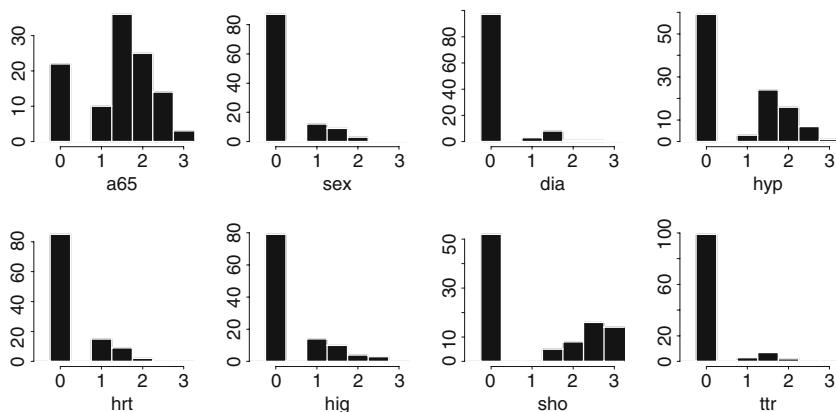
**Fig. 11.4** Average percent relative bias in relation to the number of events per variable in the 8 predictor model, without any selection, in random subsamples from GUSTO-I



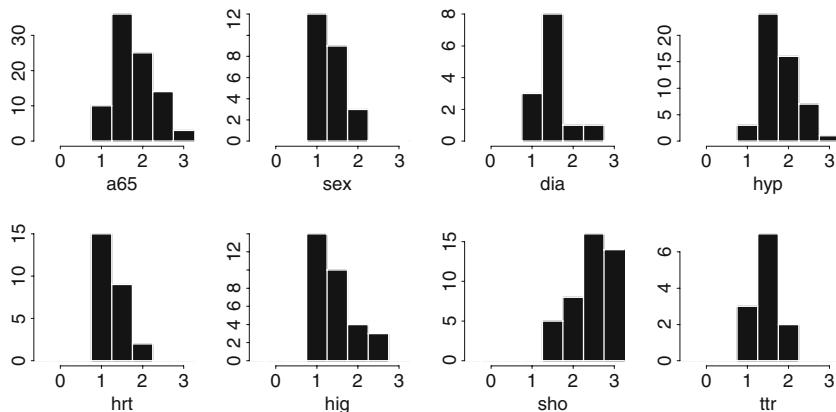
**Fig. 11.5** Mathematical relationship between bias and strength of predictive effect, expressed as coefficient/SE

The asymptotic standard error (SE) of the selected coefficient is estimated as if the model was pre-specified. The means of these asymptotic SEs were somewhat larger than the empirical SEs of the conditional coefficients for each of the 8 predictors, but smaller than the unconditional SEs (Table 11.5). The latter SEs reflect

A



B



**Fig. 11.6** Distribution of logistic regression coefficients in small subsamples within GUSTO-I, with  $p < 0.05$  for selection. Panel(a) unconditional perspective (if a predictor was not selected the coefficient is assumed to be set to zero). Panel(b) conditional perspective (only studying the distributions of selected predictors)

**Table 11.5** Standard errors of estimated coefficients in  $p < 0.05$  selected models, calculated from an unconditional or conditional perspective (standard deviation in Fig. 11.6), and as asymptotically estimated in the models (average SE)

Perspective	A65	Sex	DIA	HYP	HRT	HIG	SHO	TTR
<b>Empirical SE</b>								
Unconditional	0.87	0.55	0.53	0.98	0.55	0.77	1.51	0.46
Conditional	0.52	0.27	0.36	0.47	0.29	0.62	0.75	0.27
<b>Asymptotic SE</b>								
Conditional	0.57	0.48	0.55	0.61	0.48	0.59	0.88	0.62

that the coefficient might have been set to zero, but the interpretation of this SE is difficult, if not impossible. In sum, the distribution of coefficients is less straightforward to interpret or quantify when a stepwise selection procedure has been followed. Some may hence consider reporting of 95% confidence intervals for coefficients in a stepwise model rather meaningless.

Another consideration is the variability of predictions (rather than predictor effects), given covariate patterns. This has been studied with bootstrapping techniques. Predictions were far more variable than expected from estimates which were made as if the model was pre-specified.<sup>10,19</sup>

### ***11.4.5 Exaggeration of P-Values***

The testimation bias in coefficients and misspecification of variability leads to an exaggeration of *p*-values. The *p*-value of predictors in a stepwise model should generally not be trusted; the *p*-value is calculated as if the model was pre-specified.

### ***11.4.6 Predictions of Worse Quality Than from a Full Model***

For the studied sample, stepwise selection does not decrease model performance that much by omitting some variables. The eliminated variables have by definition relatively weak effects, otherwise they would not have been omitted.

Of more interest is the validity of the predictions outside the studied sample. We can assess the validity with internal validation techniques such as bootstrapping, and with external validation, i.e. evaluation in completely new patients. Both types of validation have shown that the performance of stepwise selected models is usually worse than that of a full model, without selection.<sup>175</sup> Table 11.6 provides an illustration of bootstrap validation of sample5 from the GUSTO-I study. This same pattern was found in a large simulation study considering all small subsamples within the training part of GUSTO-I, and evaluating them on an independent validation part of GUSTO-I.<sup>409</sup>

**Table 11.6** Illustration of bootstrap validation of model performance, as indicated by  $R^2$  in sub-sample #5 of the GUSTO-I data base ( $n=429$ , see also Table 5.4)

Method	Apparent	Bootstrap	Test	Optimism	Optimism-corrected
Full 8 predictor model	22.7	24.7	17.2	7.6	15.1
Stepwise, 3 predictors $p<0.05$	17.6	18.7	12.7	5.9	11.7

## 11.5 Influence of Noise Variables

An argument for stepwise methods is that it helps to eliminate variables that have no true relationship to the outcome (noise variables, with true regression coefficient of zero). As discussed before, the likelihood of having such noise variables in our model can be reduced by considering only predictors with external knowledge on their relevance (from literature, expert opinion). Various simulation studies have considered the behaviour of stepwise selection in the presence of noise variables.

Derkzen and Keselman found that stepwise selection produced models in which 30–70% of the selected predictors were not related to the outcome, i.e. were pure noise, when candidate predictors consisted of a mix of noise and true predictors.<sup>96</sup> The frequency of inclusion of noise and true predictors depended on the number of noise variables among the candidate predictors and on the correlations between candidate predictors. Stepwise methods are hence no guarantee for exclusion of noise variables.

Ambler et al. performed simulations in two data sets, where a mix of true and noise predictors was considered. They focused on the predictive performance of the models. Stepwise selection with AIC was optimal in their study.<sup>14</sup>

We added 9 noise variables to the 8 predictors considered thus far in GUSTO-I simulations.<sup>409,410</sup> Performance was evaluated in an independent test part with 20,318 patients (see Chap. 22). As expected, we note that discriminative ability (*c* statistics) for the full model was worse by adding noise variables, compared to a model including 17 true predictors (Table 11.7). The *c* statistic decreased from 0.784 with true predictors to 0.753 with 8 true and 9 noise predictors. The stepwise models succeeded in removing noise variables: with  $p < 0.05$  selection only 1 in 20 was retained in the model, which is approximately 1 in every 2 models (9 per model considered). The exclusion of noise variables comes at the price of, at the same time, excluding true predictors. For example, the  $p < 0.05$  selected models contained on average 0.5 of the 9 noise predictors and 4.8 of the 8 true predictors (Table 11.7).

Hence, the performance of stepwise models was worse when only true predictors were considered, but also when more than half of the candidate predictors were in fact noise (Table 11.7). Apparently, the  $p < 0.05$  stopping rule led to a suboptimal balance between elimination of noise variables and the inclusion of a sufficient number of true predictors in this case study. This case study suggests that the omission of a true predictor may be far worse than the inclusion of a noise variable.<sup>410</sup>

**Table 11.7** Selected predictors and performance of models with 17 true predictors or 9 noise variables and 8 true predictors in 23 large subsamples from GUSTO-I.<sup>410</sup> Models were evaluated in an independent test part of GUSTO-I (part B,  $n=20,318$ )

# Predictors	True predictors 17 predictor model	$P < 0.05$	Noise variables added 8+9 model	$P < 0.05$
Noise	—	—	9	0.5
True	17	5.9	8	4.8
<i>c</i> statistic	0.784	0.762	0.753	0.746

## 11.6 Univariate Analyses and Model Specification

A common way to select predictor variables for a regression model is to first study the univariate relation between each variable and the outcome. When a variable meets a univariate criterion, e.g.  $p < 0.05$ ,  $p < 0.1$ ,  $p < 0.2$ , or  $p < 0.5$ , the variable is considered further for multivariable modelling (Table 11.8). This strategy may seem advantageous, and seems to reduce problems of overfitting and stepwise selection. However, univariate pre-selection is just a variant of stepwise selection. All candidate predictors are considered in the first step, but only those meeting the univariate criterion are considered in the following steps (minus the one predictor that entered the model in the first step). This is in contrast to standard forward (or backward) selection, where all candidate predictors are considered in each step as long as they have not entered the model (or are not removed from the model). The difference between univariate pre-screening and standard backward selection is shown in Tables 11.8 and 11.9 for a hypothetical example.

**Table 11.8** Hypothetical example of univariate screening of candidate predictors, followed by stepwise backward selection. We note that 3 candidate predictors are omitted from further consideration based on univariate insignificance (#6, #7, #8), and 2 because of multivariable insignificance (#4, #5). The final model includes 3 predictors (#1, #2, #3)

**Table 11.9** Hypothetical example of backward stepwise selection of candidate predictors #1–#8. We note that the final model includes 1 of the 3 candidate predictors which were insignificant in univariate analysis (#6)

	1	2	3	4	5	6	7	8
Multivariable modeling								
Omit #8								
Omit #7								
Omit #5								
Omit #4								
Selected model								

### \*11.6.1 Pros and Cons of Univariate Pre-Selection

Univariate pre-selection has some practical advantages:

- Predictors are eliminated at an early stage if no regression coefficient can be estimated with standard fitting algorithms, e.g. for “shock” in the small GUSTO-I subsamples. A model can be developed with the remaining predictors;
- In a large data set, with many predictors, the computational burden is lower when starting with a smaller set of predictors in a “reduced full model.”

On the other hand, univariate screening of candidate predictors does not reduce the problems as noted for stepwise methods (Sect. 11.4). Other variants of univariate pre-selection are eye-balling relationships between continuous predictors and outcome, and inspection of exploratory cross-tables. In these informal inspections, the relationship between a predictor and the outcome is used. Such informal data inspections may hence contribute to overfitting.

### \*11.6.2 Testing of Predictors within Domains

A variant of univariate screening is to test the relevance of predictors within a cluster of related variables, representing a disease domain. For example, we may consider pre-selection of 1 or more predictors from variables related to hypertension: diastolic blood pressure, systolic blood pressure and treatment for high blood pressure. Such an approach has some attractiveness, but problems of stepwise selection apply here too. Some increase in power can be obtained by requiring that all domains have to be included in the final model, even when not statistically significant after the pre-selection. Alternative approaches are to combine variables within such a cluster, e.g. as mean blood pressure, or pre-selection based on prior information, e.g. evidence from other studies.

## 11.7 Modern Selection Methods

A number of more modern selection methods have emerged over the past decades. Generally these methods are quite computer intensive, and are still infrequently encountered in medical applications. Some methods use resampling methods such as the bootstrap to identify important variables. Others use principles of Bayesian analysis, such as Bayesian model averaging (BMA). Some methods use shrinkage of regression coefficients to zero as a method of selection. Finally, many methods are under consideration by computer scientists and statisticians that may prove valuable in the future, but are not discussed here.<sup>181</sup>

### \*11.7.1 Bootstrapping for Selection

Several authors have proposed to define prediction models based on selection in bootstrap samples.<sup>21,70,368</sup> For example, one may apply backward stepwise selection in bootstrap samples drawn from the original sample. Candidate predictors are ranked according to their frequency of selection in the bootstrap samples. A cut-off is then applied for selection of predictors in the model that is fitted in the original sample, e.g. all predictors selected in >50% of bootstrap samples. Evidence for the advantages of this method is still unconvincing. Models constructed with this procedure will generally be very similar to the stepwise model in the original sample, provided that the stopping rule is similar (e.g. selection in over 50% of bootstrap samples). Predictors with low  $p$ -values in the original sample tend to be selected with high frequency in bootstrap samples.

### \*11.7.2 Bagging and Boosting

Bagging (for “bootstrap aggregating”) was proposed by Breiman.<sup>59</sup> Bagging is a method for generating multiple versions of a linear predictor and using these to get an aggregated linear predictor. The multiple versions are formed by making bootstrap replicates of the sample and using these as new model development sets. The aggregation averages over multiple versions of a predictor to make predictions. If perturbing the development set can cause substantial changes in the predictor constructed, bagging can improve accuracy.<sup>59</sup>

Bagging is somewhat related to “boosting,” which is a general method for improving the performance of any learning algorithm.<sup>371</sup> Bagging works by taking a bootstrap sample from the training set. Boosting works by changing the weights on the training set. Greater weights are given to observations that were difficult to classify, and lower weights to those that were easy to classify.

### \*11.7.3 Bayesian Model Averaging (BMA)

Researchers usually ignore the uncertainty associated with modelling procedures such as stepwise selection. BMA aims to appropriately consider this uncertainty. This method selects a subset of all possible models (up to  $K = 2^p$ , where  $p$  is the number of predictors, ignoring interactions) and uses the posterior probabilities of the models to perform hypothesis testing and prediction. Equations relating to the problem of optimal model selection have been developed.<sup>194</sup> Here  $M = \{M_1, M_2, \dots, M_k\}$  is used to denote the set of all possible models to be considered and  $\Delta$  is used to identify the quantity of interest. For example,  $\Delta$  can indicate the regression coefficient in a logistic regression model. Then the posterior distribution of  $\Delta$ , given the data  $D$  is

$$\Pr(\Delta | D) = \sum_{k=1}^K \Pr(\Delta | M_k, D) \Pr(M_k | D)$$

This is an average of the posterior distributions under each model  $M_k$  ( $\Pr(\Delta | M_k, D)$ ), weighted by the corresponding posterior model probabilities ( $\Pr(M_k | D)$ ) ( $k = 1, 2, \dots, K$ ) given the data.

Hereto, we need to estimate how likely each coefficient is given a particular model and how likely each model is. This estimation requires two prior probabilities: one for the coefficient values and one for the likelihood of each model  $M_k$ . For the coefficients, a multivariate normal prior with mean at the maximum likelihood estimate and variance equal to the expected information matrix for one observation has been suggested. This can be thought of a prior distribution that contains the same amount of information as a single, typical observation. Essentially, this prior distribution is non-informative. When there is little information about the relative plausibility of the models considered, taking them all to be equally likely a priori is believed by many to be a reasonable choice.

For an analysis with  $p$  potential predictors, the number of models,  $K$ , can be enormous. To get around this problem, we may exclude models that are far less probable than the best model. This strategy is also known as “Occam’s window” approach.<sup>342</sup> For example, we may choose to discard models that are 20 times less likely as posterior models based on the data than the most likely model. This approach makes the BMA procedure computationally better feasible.

Alternatively, a two-step bootstrap model averaging approach can be applied, which consists of a screening step to eliminate covariates thought to have no influence on the response, and a model averaging step. This procedure increases practical usefulness by eliminating unimportant factors in the screening step.<sup>196</sup>

Software is increasingly available that calculates a posterior model probability, parameter estimates, and standard errors of those estimates (for example for R). This enables the testing of hypotheses, such as that the effect of a predictor is zero. Also the regression coefficient can be estimated (as the posterior mean) with a standard error (based on the posterior standard deviation). Essentially, each estimated regression coefficient  $\beta_i$  from a potential model is weighted with the posterior likelihood that this model is the final model given the data  $D$ :

$$E(\beta_i | D) = \sum_{M_k \in A} \beta_i \Pr(M_k | D)$$

Similarly, we can make predictions for future patients with all models with a posterior likelihood larger than zero, and then weight each prediction with the posterior likelihood that this model is the final model.

### 11.7.4 Practical Advantages of BMA

Simulation studies have shown that the BMA procedure may especially guard against false positive findings. When 25 noise variables were considered and no

true predictors, standard stepwise selection methods sometimes included two or more of these variables as predictors, while BMA always indicated that no predictors were identified in the data. With 2 relatively strong true predictors and 23 noise predictors, BMA again outperformed standard stepwise methods, with the latter including up to 5 noise variables in addition to the 2 true predictors.<sup>342</sup>

BMA was also applied to the Framingham data to predict coronary events. Within the data set, 12 exploratory variables were available. Using Occam's window, 6 out of 8,192 (2<sup>13</sup>) models were selected, reflecting model uncertainty. The 6 models from the BMA procedure contained 5 or 6 predictors, with posterior probability between 3% and 65%. The model selected by the backward stepwise procedure with  $p < 0.05$  contained eight variables. Overall, the posterior regression coefficients from the BMA method were similar to the maximum likelihood estimates for a model selected using the backward stepwise procedure with a significance level of 1%. The predictive performance of BMA was however somewhat better than that of the stepwise model.<sup>475</sup>

### ***11.7.5 Shrinkage of Regression Coefficients to Zero***

Shrinkage is the principle of reducing the regression coefficients to improve the quality of predictions. Several variants of shrinkage will be discussed in Chap. 13. Some variants of shrinkage methods lead to regression coefficients which are set to zero. Hence, model reduction is achieved, since variables with zero coefficients can be dropped. Examples of these methods include the “Garotte”<sup>58</sup> and the “least absolute shrinkage and selection operator” (Lasso).<sup>434</sup> This approach minimizes the log likelihood subject to the sum of the absolute values of the parameters being bounded by a constant. This constraint shrinks some coefficients to zero. The Lasso showed promising results in simulation studies and in predicting 30-day mortality in subsamples from GUSTO-I (see Chap. 13).

## **11.8 Concluding Remarks**

The problem of overfitting already starts with considering too many candidate predictors in a data set. This problem is difficult to solve with standard statistical techniques which are used by default in medical research nowadays, such as stepwise selection. Faraway has labelled the issues discussed here as “the cost of data analysis.”<sup>119</sup> Ye has proposed methods to estimate the “effective degrees of freedom” of a multi-step modelling procedure.<sup>494</sup>

Improvements in model selection can be sought in various directions. This first is to limit the necessity for selection by using subject matter knowledge, especially in relatively small data sets. Another strategy is to use better algorithms to discover patterns in the data, including better fitting algorithms (such as the “Lasso”), or by

bootstrapping and following Bayesian estimation methods.<sup>154</sup> The uncertainty of model selection is an important source of overfitting, which needs to be prevented if possible, e.g. by analysing larger data sets, and a limited use of stepwise methods. The Lasso and variants of such a method are promising techniques when prediction and parsimony are goals of predictive modelling.<sup>179,434</sup>

## Questions

### 11.1 Stepwise selection methods

Stepwise methods are abundant in the medical literature, both in the context of addressing epidemiological questions on predictive effects and in the context of deriving prediction models.

- (a) What decisions need to be made when one wants to use stepwise selection methods?
- (b) What are the major advantages and disadvantages of stepwise selection?

### 11.2 Models considered in all subset regression

Suppose we consider ten candidate predictors, and use a variant of stepwise selection that considers all combinations of predictors in selecting a model (“all possible subset regression,” Sect. 11.2.1).

- (a) How many models do we consider?
- (b) And how many if we pre-specify that four predictors have to be included?

### 11.3 Bias by stepwise methods (Fig. 11.5)

We found that the bias was at a maximum when the true coefficient to standard error ratio was around 0.6 (Fig. 11.5). Logistic regression coefficients for a binary outcome have an SE of 0.5 if the prevalence of a binary predictor such as gender has a prevalence of 50% and the incidence of the outcome is also 50%, in a total of 64 patients (32 with the outcome).

- (a) What does this imply for the bias in regression coefficients, where the SE is around 0.5?
- (b) And for predictors with a true regression coefficient around 0.5 (odds ratio 1.6), or a true coefficient around 2 (odds ratio 7.4)?

### 11.4 Application of stepwise methods<sup>363</sup>

Consider the paper by Sanada et al. published in 2007.

- (a) How many subjects were studied?
- (b) How many predictors were considered?
- (c) How many were selected by stepwise selection?
- (d) What alternatives might have been used for model specification?
- (e) Consider the Letter to the editor from Malek et al., who is very critical with respect to stepwise selection.<sup>275</sup> They propose an alternative selection strategy, called “hierarchical analysis.” What is your opinion of this strategy?

# Chapter 12

## Assumptions in Regression Models: Additivity and Linearity

**Background** In this chapter, we discuss assessment of assumptions in multivariable regression models. Specifically, we consider the additivity assumption, which can be assessed with interaction terms. We also consider the linearity assumption of continuous predictors in a multivariable regression model, where multiple non-linear terms can be included to allow for non-linear relationships between predictors and outcome. Throughout we stress parsimony in strategies to extend a prediction model with interactions and non-linear terms, since better fulfillment of assumptions in a particular sample does not necessarily imply better predictive performance for future subjects. We consider several case studies for illustration of various strategies to deal with additivity and linearity.

### 12.1 Additivity and Interaction Terms

The generalized linear regression models discussed in this book all have a linear predictor at their core:  $lp = \beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_i \times x_i$ , for models with  $i$  predictors.

The  $\beta_1$  to  $\beta_i$  are the regression coefficients, referring to the main effects of predictors  $x_1$  to  $x_i$ . This formulation implies additivity of effects. For a logistic regression model, we can calculate odds ratios as  $e^{\beta_i}$ ; the odds ratios are multiplied to obtain the odds of the outcome. Hence, effects of predictors are assumed to be multiplicative on the odds scale. For a Cox regression model,  $e^{\beta_i}$  is the hazard ratio; the assumption is that these hazard ratios can be multiplied on the hazard scale.

The scale is essential for consideration of additivity. If a treatment reduces risk from 20 to 10% in one risk stratum, and from 10 to 5% in another risk stratum, the relative risk is 0.5 in both. The odds ratios are also quite similar (0.44 and 0.47, respectively). Hence, we could say that there is a consistent halving of the risk. But on an absolute scale, the benefit is clearly dependent on the risk (10% vs. 5% reduction).<sup>237</sup>

The most common regression modelling procedure is to start model specification with main effects of predictors only. Some epidemiological text books advice to consider interactions early in the modelling process, with main effects included for

all variables that have a relevant interaction term.<sup>234</sup> Interactions between predictors can be considered by multiplicative terms of the form  $x_1 \times x_2$  (two-way or first-order interactions), and  $x_1 \times x_2 \times x_3$  (three-way, or second-order interactions); higher-order interactions are uncommon to consider for regression models. The interpretation of a two-way interaction is that the effect of one predictor depends on that of another predictor. The effect is different, depending on the value of another predictor. The effect of a predictor cannot be interpreted alone; we need to know the value of another predictor to interpret its effect.

### **12.1.1 Potential Interaction Terms to Consider**

As for main effects, prior subject knowledge may help to guide us to select interaction terms. For example, interaction terms that were identified in previous studies could be assessed. Clinical insights, e.g. on pathophysiology, are difficult to use, because using main effects in a model is assuming that predictors act in a multiplicative way on the risk scale (e.g. odds ratios and hazard ratios are multiplied). Reasoning why a certain combination of predictors would not act in an additive way on, e.g. the logodds scale, is quite difficult to imagine. Some researchers are motivated to study an interaction term when two predictors are correlated. But correlation does not imply anything on the effects of predictors conditional on each other. Two predictors may not have any correlation, but have interacting effects. Some types of interactions have been suggested that warrant consideration in prediction models (Table 12.1).<sup>174</sup>

### **12.1.2 Interactions with Treatment**

Various interactions with treatment can be considered. The benefit of treatment may depend on the severity of disease, with less relative benefit for those with less severe disease. The reverse may also be true, especially in oncology, where less

**Table 12.1** Examples of interactions to consider in clinical prediction models (based on Harrell<sup>174</sup>)

Interaction	Effect
Severity of disease $\times$ treatment	Less benefit with less severe disease
Place $\times$ treatment	Benefit varies by treatment centre
Place $\times$ predictors	Predictor effects vary by centre/region
Calender time $\times$ treatment	Learning curves for some treatments
Calender time $\times$ predictors	Increasing or decreasing impact of predictors over the years
Age $\times$ predictors	Older subjects less affected by risk factors; or more affected by certain types of disease
Follow-up time $\times$ predictors	Non-proportionality of survival effects, often a decreasing effect over time
Season $\times$ predictors	Seasonal effect of predictors

relative benefit occurs for those with more severe disease. For example, surgery in oesophageal cancer can be curative, but only for patients without distant metastases. Note that absolute benefit will anyway depend on the severity of disease, even when the relative benefit is constant. For example, Califf modelled the absolute benefit of tPA treatment for acute myocardial infarction patients in the GUSTO-I trial in relation to predictors. Benefit depended strongly on the risk profile, while it might be assumed that the relative effect of treatment was constant.<sup>63</sup> In addition to severity of disease, a treatment effect may depend on the setting, e.g. the centre where a patient was treated. This is especially the case when specific skills and facilities are required for the treatment. For example, surgical mortality is known to vary widely between centres for some procedures, such as resection of oesophageal cancer. Similarly, some treatments have a learning curve, which can be modelled by including a treatment  $\times$  calendar time interaction term, with calendar time reflecting cumulative experience.

In randomized controlled trials, subgroup effects for treatment effects are often performed, e.g. whether treatment works better for older than younger patients. Such subgroup effects should be supported by an interaction test for difference in effect; not with one  $p$ -value for older and one  $p$ -value for younger patients.<sup>339</sup> Even when subgroup analyses are pre-specified, results should be cautiously interpreted because of multiple testing of the treatment effect. Multiple testing inflates the risk of false positive conclusions. Subgroup analyses are therefore best interpreted as secondary analyses which motivate further study. This is often not the case in current practice.<sup>18</sup>

### \*12.1.3 *Other Potential Interactions*

Predictor effects may differ by place and time, which would limit their generalizability (see Part III). Basic issues to consider are whether predictor definitions were consistent across centres and during time. In some individual patient data analyses, predictor effects were however surprisingly consistent, even when definitions varied over studies (e.g. studies in traumatic brain injury<sup>271,277</sup>). As might be expected, interactions of predictors by place of treatment were small within the GUSTO-I trial, where data were collected in a highly standardized and controlled way.<sup>405</sup>

Various aspects of “time” can interact with predictor effects: calendar time (e.g. patients treated during years 1980–2005), age (e.g. 30–90 years), follow-up time (e.g. 0–10 years), and season (months January to December). The effects of predictors may change over the years because of improvements in treatment, or changing definitions. The effects of risk factors for developing cardiovascular disease are known to decrease with aging. Predictors having less effect in the elderly might be explained as that older subjects have proven to survive with the risk factors. For survival analysis, predictors are usually assumed to have proportional effects during follow-up, e.g. in the Cox proportional hazards model, but also in a Weibull model. Such proportionality of effects may not be tenable in the follow-up of

oncological patients, where relative risks of predictors for early events decrease with time, while others may increase. For example, non-proportional effects have been noted in breast cancer survival, with no effect of stage of disease after 10 years of follow-up.<sup>308</sup> The proportionality assumption is equivalent to assuming no interaction effects between predictors and follow-up time.

Furthermore, some predictors may have a different impact during the season, e.g. for infectious and respiratory diseases (Table 12.1). Other interactions may be relevant to consider in specific prediction problems. For example, sex-specific effects of predictors are commonly modelled in cardiovascular disease.

#### \*12.1.4 Example: Time and Survival After Valve Replacement

A follow-up study was done spanning over 25 years for survival of patients after aortic valve replacement.<sup>195</sup> Various changes had taken place in case-mix between the first valve replacement (in 1967) and the latest replacement analysed (in 1994). During the 25+ years period, 1,449 mechanical valves were implanted. Overall early mortality (<30 days) was 5%, and was analysed with logistic regression. Overall survival rates at 5, 10, and 15 years were 80%, 63% and 49%, respectively. Poisson regression analysis was used to disentangle the effects of calendar time, age, and follow-up. All three aspects of time appeared to be important. A substantial drop in both early and late mortality was identified around the introduction of cardioplegia (in 1997), but no strong interactions with calendar time were found. A changing, non-proportional effect was observed for several prognostic factors during follow-up. For example, increasing effects during follow-up were found for older age ( $p<0.05$ ), urgency (urgent operations and acute endocarditis) ( $p<0.05$ ), and ascending aorta surgery ( $p=0.12$ ). Early year of operation, male gender, and previous cardiac surgery (all  $p<0.05$ ) were more important during early years of follow-up. The effects of concomitant coronary bypass surgery and concomitant mitral valve surgery were more or less constant during follow-up. This study illustrated that a Poisson regression model could be used to disentangle different aspects of time in a survival analysis. This model was more easily to work with compared to the Cox regression model.<sup>195</sup>

## 12.2 Selection, Estimation and Performance with Interaction Terms

In clinical prediction models with a typical number of predictors, say 5–10, the number of potential interactions is substantial. If interactions are considered, it has been suggested to first perform an overall interaction test.<sup>174</sup> We can also obtain partial overall  $p$ -values, e.g. for all interactions with age. If this  $p$ -value is low, we may consider proceeding with studying specific interactions for inclusion in the

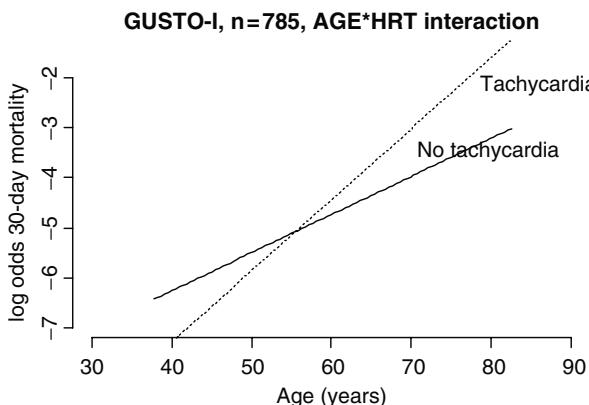
model. This approach limits the multiple testing problem, at the price of lower power for including specific interactions. An alternative is to perform interaction tests for individual combinations of predictors, but use a rather stringent  $p$ -value, such as 0.01 for inclusion. We illustrate the problems with selection of interaction terms with a small subsample from the GUSTO-I study.

### 12.2.1 Example: Age Interactions in GUSTO-I

We study interaction with age in the relatively large subsample from GUSTO-I (sample5,  $n=785$ , 52 deaths). We first fit all interactions, and then perform an overall test based on the Wald statistics. The overall test has a  $p$ -value of 0.14; but the interaction AGE $\times$ HRT is statistically significant ( $p=0.03$ , not adjusted for multiple testing). Some might be tempted to include this interaction in the model. It appears that tachycardia (HRT) has a stronger effect at higher age (a positive interaction). Equivalently, we can state that age has more effect in strength in those with tachycardia (Fig. 12.1).

### 12.2.2 Estimation of Interaction Terms

A first distinction that epidemiologists like to make is between “qualitative” and “quantitative” interactions. A qualitative interaction means that a predictor has an opposite effect in one group vs. another group of patients. Quantitative interaction means that the effect of a predictor is in the same direction, but different in strength

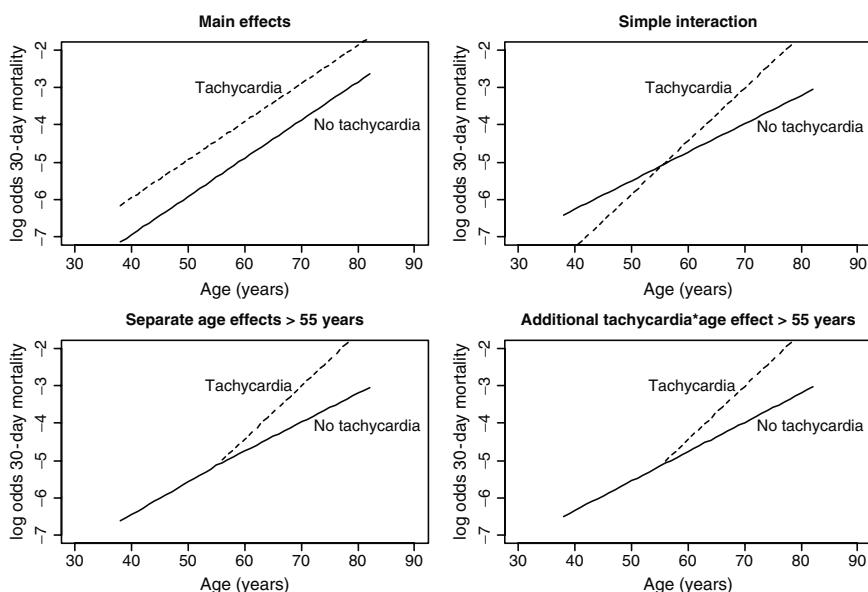


**Fig. 12.1** Age by tachycardia interactions in a subsample of GUSTO-I ( $n=785$ , 52 deaths), revealing a positive interaction

in one group than another group of patients (see e.g. Fig. 12.1). This distinction is especially important when we aim to interpret the effects of predictors; we will more be tempted to include a qualitative interaction than a quantitative interaction. For predictive performance, the distinction between qualitative and quantitative interaction is less relevant.

Another issue is that we can have somewhat counterintuitive effects of interactions. For example, Fig. 12.1 suggests that the presence of tachycardia is protective for 30-day mortality at ages younger than 55. If we consider this implausible, we can code the interaction such that no effect of tachycardia is present below age 55 (Fig. 12.2). Admittedly, the age cut-point of 55 years is data-driven. But the general idea is that we incorporate subject-specific knowledge to prevent incorporation of random noise in the model.

More generally, we should use a smart coding for interaction terms once we decide to include an interaction term in a model. This is especially useful when we want to readily obtain standard errors and confidence intervals for predictors in interaction with other predictors.<sup>122</sup> The approach is to test for interactions in models with standard multiplicative terms of the form  $x_1 \times x_2$ . But we can estimate effects with a smarter coding of the form  $x_1 + (1 - x_1) \times x_2 + x_1 \times x_2$  instead of  $x_1 + x_2 + x_1 \times x_2$ . More details are on the book's web page.



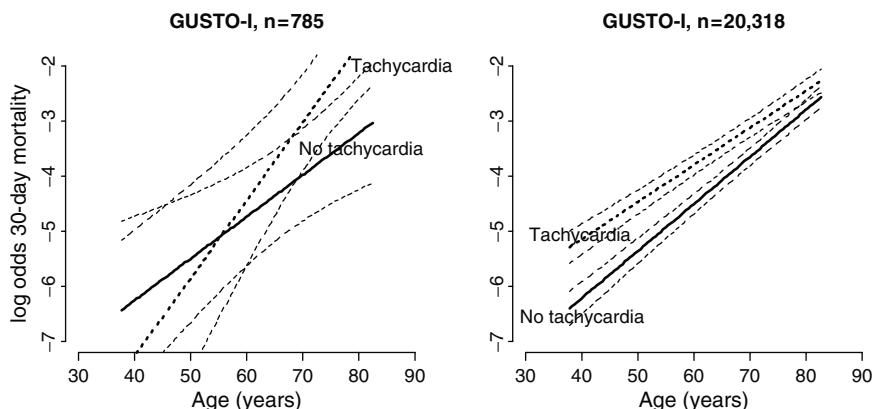
**Fig. 12.2** Age by tachycardia relationships to 30-day mortality in a large subsample of GUSTO-I ( $n=785$ , 52 deaths). Panel (a) main effects only; panel (b) simple positive interaction; panel (c) separate effects for (no) tachycardia over age 55; panel (d) one age effect and an additional effect of tachycardia over age 55 years. The difference between panel c and d is barely notable, but in panel c, three age effects are estimated, while in panel d two age effects are estimated

### 12.2.3 Better Prediction with Interaction Terms?

We may wonder we predict better with the AGE×HRT interaction (Table 12.2). We hereto test the models as shown in Fig. 12.2 in a large, independent part of GUSTO-I ( $n=20,318$ ). Surprisingly, we find that a model with the AGE×HRT interaction (Fig. 12.2b), performs worse than a model without this interaction term. The models without the counterintuitive effect of tachycardia below age 55 perform similar, both at apparent validation and at external validation in  $n=20,318$ . The explanation for this remarkable finding is in Fig. 12.3: the interaction between tachycardia and age was positive in the subsample, but negative in the independent validation part of GUSTO-I (less effect of tachycardia at older ages). This example illustrates that considering interaction in an unstructured way can damage predictive ability of a model.

**Table 12.2** Performance of models developed in a subsample of GUSTO-I ( $n=785$ ) in an independent part of GUSTO-I ( $n=20,318$ ). The model with main effects contained eight dichotomized predictors

Model	df	Apparent ( $n = 785$ )	Validation ( $n=20,318$ )
Main effects	8	0.828	0.805
Main effects+AGE×HRT interaction	9	0.831	0.796
One age effect $<55$ , 2 age effects $\geq 55$	9	0.832	0.798
HRT effect only for age $>55$ years	8	0.832	0.798



**Fig. 12.3** AGE×HRT interactions in GUSTO-I. *Left panel:* positive interaction in a subsample ( $n=785$ , 52 deaths,  $p$ -value for interaction 0.10), negative interaction in an independent validation part of the GUSTO-I data set ( $n=20,318$ , 1,428 deaths,  $p$ -value for interaction 0.002). 95% confidence intervals are given around each line

### 12.2.4 Summary Points

- An interaction term indicates that the effect of a predictor depends on values of another predictor
- Interaction terms to consider in a prediction model depend on the context, but some types of interactions may warrant specific consideration
- For better interpretation, we may use a smart coding of interactions, and eliminate counterintuitive effects, e.g. that a predictor becomes protective for some patients
- The performance of a prediction model does not necessarily increase by including an interaction term
- Pre-specification of some interaction terms for a model may be preferable to exploratory determination of which terms to include

## 12.3 Non-linearity in Multivariable Analysis

We discussed the assessment of continuous predictor variables in Chap. 9 for the univariate case, where each predictor is considered separately. Harrell advocates to use restricted cubic spline functions to define transformations of continuous variables.<sup>174,177</sup> An RCS function consists of pieced-together cubic splines (containing  $x^3$  terms) that are restricted to be linear in the tails. These functions have many favourable properties, such as appropriate flexibility combined with stability at the tails of the function. We can also consider multivariable modelling with fractional polynomials,<sup>367</sup> and with smoothing spline transformations (in multivariable generalized additive models (“GAM”), Table 12.3). The flexibility of a smoothing spline transformation in a GAM is determined by penalty terms, which relate to the effective degrees of freedom ( $df$ ). There are presently two variants of GAM available with respect to choosing the effective  $df$  in a multivariable context. One variant is that the effective  $df$  are set by the analyst.<sup>180</sup> Alternatively, a generalized cross-validation (GCV) procedure can be used to define statistically optimal transformations for multiple continuous predictors in a GAM.<sup>490</sup> We discuss these approaches in more detail below.

### 12.3.1 Multivariable Restricted Cubic Splines (RCS)

An RCS requires the specification of knots, which can well be based on the distribution of the predictor variable.<sup>174</sup> The key issue is the choice of the number of knots: 5 knots implies a function with 4  $df$ , 4 knots 3  $df$ , and 3 knots 2  $df$ . Although 5 knots are sufficient to capture many non-linear patterns, it may not be wise to include 5 knots for each continuous predictor in a multivariable model. Too much

**Table 12.3** Approaches to non-linearity in multivariable clinical prediction models

Approach	Characteristic	Multivariable strategy	R implementation
Restricted cubic splines	Cubic splines, with restriction in shape at the ends of the predictor distribution	Keep complexity as defined a priori or based on findings in univariate/multivariable analysis	<code>rccs</code> in <code>Design</code> package
Fractional polynomials	Combine one or two polynomials	Search iteratively for optimal transformations	<code>fp</code> and <code>mfp</code> in <code>mfp</code> package
Splines in GAM	Spline functions with smoothing depending on effective degrees of freedom	Degrees of freedom set by analyst or from a generalized cross-validation (GCV) procedure	<code>gam</code> and <code>mgcv</code> package

flexibility would lead to overfitting. One strategy is to define a priori how much flexibility will be allowed for each predictor, i.e. how many  $df$  will be spent. In smaller data sets, we may for example choose to use only linear terms or splines with 3 knots (2  $df$ ), especially if no strong prior information suggests that a non-linear function is necessary.<sup>174</sup> Alternatively, we might examine different RCS transformations (5, 4, 3 knots) in univariate and/or multivariable analysis, and choose an appropriate number of knots for each predictor based on the findings in the data. It might be reasonable to choose the complexity of non-linear functions based on the  $\chi^2$  statistic of each predictor, with more flexibility for stronger predictors.

### 12.3.2 Multivariable Fractional Polynomials (FP)

As discussed in Chap. 9, fractional polynomials are formulated as a power transformation of a predictor  $x$ :  $x^p$ , where  $p$  is chosen from the set  $-2, -1, -0.5, 0, 0.5, 1, 2, 3$ . This defines 8 transformations, including inverse ( $x^{-1}$ ), log ( $x^0$ ), square root ( $x^{0.5}$ ), linear ( $x^1$ ), squared ( $x^2$ ) and cubic transformations ( $x^3$ ). In addition to these 8 FP1 functions, 28 FP2 functions can be considered of the form  $x^{p1} + x^{p2}$ ; when  $p1=p2$  one defines another 8 FP2 functions as  $x^p + x^p \log x$ , for a total of 36 FP2 functions.<sup>367</sup> FP1 and FP2 have 2 and 4  $df$ , respectively.

Estimation algorithms have been developed for various software packages, including R.<sup>366</sup> The `mfp` algorithm applies a special type of backward stepwise selection procedure for the determination of reasonable functional forms for each continuous predictor. The algorithm starts with a full model including all predictors, with all continuous predictors in linear form. The predictors are considered in order of decreasing statistical significance, such that relatively important predictors are considered before unimportant ones.

For a particular continuous predictor, we may search within the 44 FP2 transformations for a best fitting function. The best transformation is compared to deleting the predictor. This procedure uses  $4\text{df}$  to test for inclusion of the continuous predictor, as having “any effect.” If this test is significant, we may continue with a test for non-linearity: FP2 vs. linear, using  $3\text{df}$ . Finally, we tests an FP2 vs. FP1 transformation as a test of a more complex function against simpler one ( $2\text{ df}$  test for “simplification”). The functional form for this predictor is kept, and the process is repeated for each other predictor. The first iteration concludes when all the variables have been processed. The next cycle is similar, except that the functional forms from the initial cycle are retained for all variables excepting the one currently being processed. Updating of FP functions and selection of variables continues until the functions and variables included in the model do not change.<sup>367</sup>

This test procedure aims to preserve the overall type I error. The price is that we are slightly conservative if the true predictor–outcome relationship is linear, i.e. a straight line. This is because in step 1, we test for overall effect with  $4\text{df}$ , leading to lower statistical significance in case of a true linear relationship.

### ***12.3.3 Multivariable Splines in GAM***

In a GAM, flexible, smooth functions are defined for continuous predictors. The smooth functions can be defined by splines or other “basis functions.”<sup>490</sup> To avoid overfitting we statistically penalize lack of smoothness (“wiggliness”) using a smoothing parameter. The penalization reduces the effective degrees of freedom used by each continuous predictor. The optimal smoothness can be determined with prediction error criteria, e.g. in a Generalized cross-validation (GCV) procedure. Further details are provided by Wood<sup>490</sup> and Hastie.<sup>181</sup>

In multivariable modelling, splines in a GAM may well serve as a reference standard for comparison of simpler, parametric transformations, such as FP (or RCS) functions.<sup>353</sup> We compare several approaches in a case study below. In practice, one would not have to perform all of these transformations but choose one approach that one is familiar with.

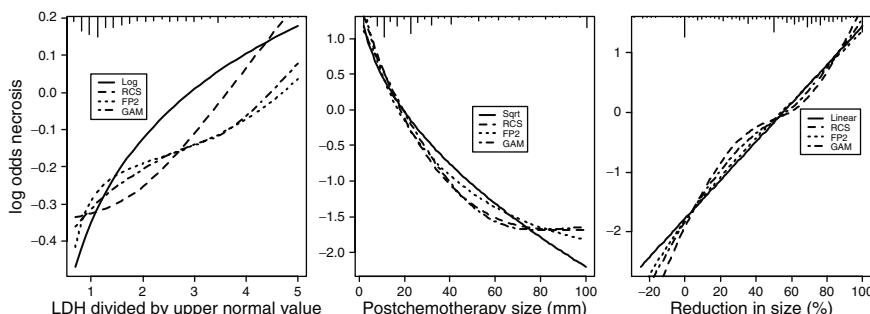
## **\*12.4 Example: Non-Linearity in Testicular Cancer Case Study**

We aim to predict the presence of benign tissue only (“necrosis”) in patients treated with chemotherapy for testicular cancer. We consider six predictors, of which three are binary (Teratoma, pre-chemotherapy elevated AFP, pre-chemotherapy elevated HCG), and three continuous (pre-chemotherapy LDH, reduction in mass size during chemotherapy, post-chemotherapy size). The LDH values were standardized by dividing by the upper limit of the local upper normal value (“LDHst” variable).

In initial univariate analyses, we used RCS functions to study non-linearity in the effects of the continuous predictors. Subsequently, we used simple parametric transformations, mainly based on visual assessment of the univariate RCS functions.<sup>425</sup> The chosen transformations were logarithmic for LDHst; linear for reduction in size; and square root for post-chemotherapy size (Fig. 12.4). We now explore the transformations chosen with other modelling strategies, including fractional polynomials and smoothing splines in generalized additive models.

We compare RCS, FP, and GAM functions with two bendings: FP2 transformations, RCS with 4 knots ( $3\ df$ ), and GAM splines with 3 effective  $df$ . For LDH, the transformations vary to quite some extent. The relationship of LDH to necrosis is rather different for a logarithmic transformation compared to other transformations. A simple linear term might also have been reasonable. This is supported by the FP procedure (Table 12.4). LDH has an effect ( $p$ -value for “any effect” = 0.02), but non-linearity was non-significant ( $p=0.48$ ). For postchemotherapy size, the RCS, FP2, and GAM transformations agree much better visually (Fig. 12.4), and the square root transformation looks reasonable. The FP procedure indicates significant non-linearity ( $p=0.0002$ ), and non-significant improvement by an FP2 function over an FP1 function ( $p=0.46$ ). The chosen FP1 function is logarithmic rather than the square root. Finally, reduction in mass size seems to be fit adequately with a linear term. The RCS, FP2, and GAM transformations fluctuate around this straight line, with the most wiggly pattern for the GAM. The FP procedure confirms that there is no reason to include non-linear terms ( $p=0.64$ ). The R code for these analyses is available at the book’s web site.

Fractional polynomials were considered in univariate logistic regression analysis, and subsequently in three multivariable logistic regression models. A full model included three binary predictors (teratoma (yes/no, 1  $df$ ), elevated AFP (yes/no, 1  $df$ ), elevated HCG (yes/no, 1  $df$ )), and three continuous predictors with FP2 functions (LDH standardized, reduction in size, post-chemotherapy size).



**Fig. 12.4** Non-linearity in univariate analysis of LDH, post-chemotherapy size, and reduction in mass size. Curves are shown for a parametric approximation (log, sqrt, linear), restricted cubic spline functions with 4 knots ( $3\ df$ ), a fractional polynomial ( $4\ df$ ), and a generalized additive model with spline smoother ( $3\ df$ ). The distributions of values are shown at the top of the graphs

### \*12.4.1 Details of Multivariable FP and GAM Analyses

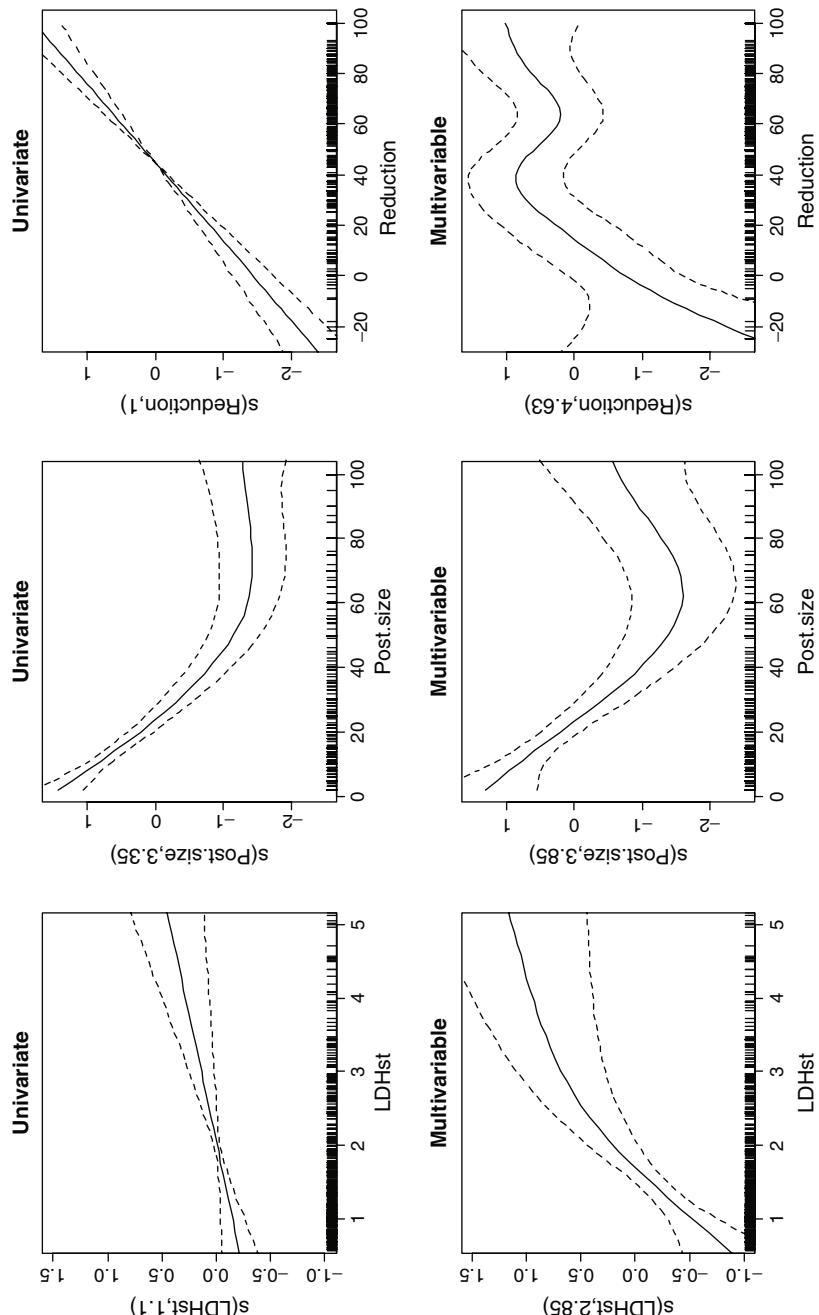
Multivariable fractional polynomials were fitted without selection (“full model,”  $3\ df$  for dichotomous +  $3 \times 4 = 12\ df$  for continuous predictors, in total  $15\ df$ ), and with a variant of a backward stepwise selection algorithm (Table 12.4). The FP2 transformations were  $\log(LDHst) + LDHst^3$ ;  $1/reduction + 1/\sqrt{reduction}$ ; and  $\sqrt{postsize} + \sqrt{postsize} \times \log(postsize)$ . A multivariable FP procedure with  $p < 0.05$  for selection led to a model with linear terms for the three continuous predictors and three binary predictors (each of the six predictors  $p < 0.01$ ). All tests for non-linearity were non-significant (Table 12.4). Selection with  $p < 0.20$  led to a linear term for  $LDHst$ ,  $1/reduction$ , and  $\log(postsize)$  in FP1 transformations. Post-chemotherapy size and reduction in size had  $p$ -values for non-linearity of 0.03 and 0.08, but FP2 transformations were not much better than FP1 transformations ( $p$ -values 0.46 and 0.27 respectively, Table 12.4).

### \*12.4.2 GAM in Univariate and Multivariable Analysis

For comparison, we examine the smooth functions selected as optimal with a GCV procedure (Fig 12.5). In univariate analysis, a (near) linear term is optimal for LDH and reduction in size (1.1 and 1 effective  $df$ ). Post-chemotherapy size

**Table 12.4** Fractional polynomial analysis of three continuous predictors in the testicular cancer data set ( $n=544$ )

	Predictor	P-value “any effect” (FP2 vs. no effect, 4 df)	P-value “non-linearity” (FP2 vs. linear, 3 df)	P-value “FP2” (FP2 vs. FP1, 2 df)	FP1	FP2
Univariate	LDH (standardized)	0.021	0.48	0.59	2	-2, 3
Full model		<0.0001	0.18	0.73	0 (=log)	<b>0 (=log), 3</b>
Stepwise $p < 0.05$		0.0003	0.46	0.62	0.5	0 (=log), 3
Stepwise $p < 0.20$		<0.0001	0.28	0.66	0 (=log)	0 (=log), 3
Univariate	Post-chemotherapy	<0.0001	0.0002	0.46	0 (=log)	0.5, 1
Full model		0.0004	0.004	0.45	0 (=log)	<b>0.5, 0.5</b>
Stepwise $p < 0.05$	size (mm)	0.012	0.086	0.30	0 (=log)	-0.5, -0.5
Stepwise $p < 0.20$		0.0005	0.034	0.46	<b>0 (=log)</b>	-0.5, -0.5
Univariate	Reduction in size (%)	<0.0001	0.64	0.63	0 (=log)	-1, 3
Full model		0.0005	0.06	0.16	-1	<b>-1, -0.5</b>
Stepwise $p < 0.05$		0.0002	0.64	0.78	-1	-1, 3
Stepwise $p < 0.20$		0.0009	0.08	0.27	<b>-1</b>	-1, -0.5



**Fig. 12.5** Generalized additive models with optimal smoothing spline transformations according to a generalized cross-validation procedure in the testicular cancer example ( $n=544$ ). *Top row*: optimal transformation in univariate logistic regression analysis; *bottom row*: multivariable logistic regression analysis with six predictors. The degrees of freedom of the optimal smoothing spline transformation are shown in each y-axis label. The distribution of values if shown at the bottom of the graphs

is modelled with a non-linear function using 3.35 effective  $df$ . In multivariable analyses, non-linear functions are used for all three continuous predictors, using 2.85, 3.85, and 4.63 effective  $df$  for LDHst, post-chemotherapy size and reduction, respectively (Fig. 12.5). Hence, more complex transformations were chosen in multivariable than in univariate analyses. The multivariable function for LDH looks much like a log transformation, as chosen previously. For post-chemotherapy size, we note an implausible increase in logodds of necrosis with very large mass sizes, and for reduction in size we note a wiggly shape between 20% and 100%. Hence, the smooth functions might not be smooth enough from a pathophysiological perspective. Further external validation might indicate whether the chosen “optimal” transformations are merely examples of overfitting.

### \*12.4.3 Predictive Performance

Finally, we study the predictive performance of the alternative non-linear transformations (Table 12.5). With linear terms only, we use 6  $df$ , and achieve a model  $\chi^2$  of 205 (apparent  $R^2$  41.9%, internally validated  $R^2$  40.3%). We find the same model by applying a multivariable FP procedure with  $p<0.05$  for selection; in fact we used more than 6 degrees of freedom in this approach, since we allowed for non-linear terms to be included in the model. If we fit a full FP2 model without selection, we use 15  $df$ , and achieve a model  $\chi^2$  of 222. The increase by 17 (from 205 to 222) with 9  $df$  is of borderline statistical significance (overall  $\chi^2$  test,  $p=0.049$ ). If we apply a more liberal  $p$ -value for non-linearity, we use two FP1 transformations (1/reduction and log(postsize)) for a model  $\chi^2$  of 213. Using RCS functions with each 4 knots leads to a better fit than the FP2 functions (231 vs. 222). The increase in model LR (from 205 to 231, +26) is statistically significant (overall  $\chi^2$  test, 6  $df$ ,  $p<0.001$ ). Our previous visual approximation of non-linearity

**Table 12.5** Logistic Regression models with alternative codings of three continuous predictors

Strategy	Model	df	Model $\chi^2$
Assume linearity (same as FP2, bw p<0.05 selection)	All linear	3+3	205
FP2, no selection	Full FP2	3+12	222
FP2, bw p<0.20 selection	LDHst, 1/reduction, log(postsize)	3+>3	213
RCS, no selection	3 RCS functions, each 4 knots	3+9	231
Visual approximation	log(LDHst) + reduction + sqrt(postsize)	3+>3	212
GAM, pre-specify $df$	3 smooth functions, each 3 $df$	3+9	232
GAM, GCV	3 optimally smoothed functions	3+(2.8+4.6+3.9)	240

in LDH and postsize led to a similar fit as the FP1 functions (model  $\chi^2$  212 vs. 213). Smoothing splines were similar in performance as the RCS model when 9 *df* were spent on the continuous predictors (model  $\chi^2$  232 vs. 231). With “optimal” transformations (GAM, GCV), more effective degrees of freedom were spent, and the highest model  $\chi^2$  or model LR was achieved (240). All model LRs indicate apparent performance. Rigorous internal validation, including all model selection steps, would be desired to indicate any true increase in performance, after correction for optimism. If inclusion of all modelling decisions is impossible, validation in a fully independent validation set may be required (split-sample, or external validation, see Chap. 17).

#### \*12.4.4 R code for Non-Linear Modelling

```
# RCS: multivariable logistic regression with 3 rcs functions,
# each 4 knots
library(Hmisc)
library(Design)
lrm(NEC ~ Teratoma + Pre.AFP + Pre.HCG + rcs(LDHst, 4) +
    rcs(Post.size, 4) + rcs(Reduction, 4), data=n544)
# FP: multivariable fractional polynomial
library(mfp)
mfp(NEC ~ Teratoma + Pre.AFP+Pre.HCG + fp(LDHst) +
    fp(Post.size) + fp(Reduction), alpha=1, data=n544)
# GAM: multivariable gam, 3 effective df for each continuous predictor
library(gam)
gam(NEC~Teratoma+Pre.AFP+Pre.HCG+s(LDHst, 3)+s(Post.size, 3) +
    s(Reduction, 3), data=n544, family=binomial)
# multivariable gam, optimal effective df for each continuous predictor
# based on generalized cross-validation (GCV)
library(mgcv)
gam (NEC ~ Teratoma + Pre.AFP+Pre.HCG + s(LDHst) + s(Post.size) +
    s(Reduction), data=n544, family=binomial)
```

## 12.5 Concluding Remarks

On the one hand, one may see the additivity and linearity assumptions as essential aspects of a regression model. Hence one might argue that we should assess these assumptions thoroughly. When we are interested in the effect of a specific predictor, this may make sense. On the other hand, a thorough assessment of assumptions increases the risk of overfitting if we are primarily interested in obtaining predictions from a model. We will be tempted to adapt the model specification based on findings in the data, i.e. extend the model with interaction terms and/or non-linear terms. The price of striving for such perfection is that we may end up with a model that performs worse for future patients than a parsimonious model without interac-

tion terms or non-linear terms. Instead, we might strive for a “wrong, but useful” model.<sup>51</sup> Such a model should provide well-calibrated and discriminating predictions, despite possibly violating some underlying model assumptions.

In the examples in this chapter, model performance did not increase impressively. Of course, results may be different in other situations, but strong qualitative interaction or U-shaped non-linearity may be relatively rare. In general, it may be sobering to assess the increase in predictive performance by inclusion of interaction terms and non-linear terms; this is often quite modest in medical examples.

Note that prediction modelling techniques deal with interactions differently. A procedure such as Naïve Bayes estimation uses univariate effects of predictors in a multivariable prediction context; additivity is assumed and interactions are not studied. In contrast, tree models assume high-order interaction by default. Similarly, neural networks assume high-order interactions, allowing for their flexibility to fit specific patterns in a data set. To explore interactions we might hence also use a tree model, since it assumes interaction by default. Interactions that stand out could subsequently be considered in a regression model, and assessed for their significance. Shrinkage or penalized estimation may be particularly valuable to reduce interaction effects that were identified among a large set of potential interactions. Penalized ML is discussed in more detail in the next chapter.

### 12.5.1 *Recommendations*

Several measures can be taken to prevent the overfitting that may occur by considering additivity and linearity assumptions. First we should balance the number of interaction and non-linear terms to be considered with the effective sample size in the analysis (Table 12.6). We might only consider interactions in studies with relatively large sample sizes, i.e. many events compared to the number of terms considered. In smaller data sets, we may simply have to rely on the additivity assumption to be reasonable. We can also say that we estimate average (or “marginal”) effects of predictors across subgroups; we know that we will never be able to exclude that we missed a relevant high-order interaction. For the linearity assumption, we might consider non-linear terms only for predictors with a presumed strong, and likely non-linear, effect. If previous studies have used a non-linear transformation for a predictor, we could also consider this transformation. Subject knowledge should also support the choice for a transformation; plotting the effect of a transformed predictor is essential (e.g. Figs. 12.1–12.5).

The second measure to prevent overfitting is to use overall tests, rather than focus on separate tests for interaction and non-linear terms. Note that based on an overall test, we would not have continued estimation of interaction of age and tachycardia in the GUSTO-I subsample (Sect. 12.2). We should also note that interaction terms make life a bit more difficult for model presentation, arguing against their inclusion in a model unless their relevance is substantial for the specific prediction problem.

**Table 12.6** Approaches to limit overfitting by assessing additivity and linearity assumptions

Approach	Description
Limited number of interaction/non-linear terms	Consider interaction term that are a priori plausible (Table 12.1); Consider non-linear terms only for predictors with a presumed strong, and likely non-linear, effect
Overall testing	Perform overall tests per interacting predictor (e.g. all age interactions)
Compare flexible vs. simple model	Compare the validated performance of a flexible model (e.g. including interactions and non-linearities) with a simple model without interaction and assuming linearity
Extra shrinkage of interaction/non-linear terms	Use a stronger shrinkage factor (<1) or more penalty in a penalized maximum likelihood procedure for interaction and non-linear terms

Third, an extension of this overall testing approach is to compare the performance of a flexible model to a simple model without interaction and non-linear effects (e.g. Table 12.5). The flexible model may for example be a neural network, or a GAM. Both the simple model and the flexible model should be validated, e.g. with bootstrapping, to see the validated rather than apparent improvement that might be achieved with inclusion of interaction and non-linear terms.

Finally, we may use shrinkage techniques to reduce the regression coefficients of selected interaction or non-linear terms. Some extra shrinkage may try to compensate for the “testimation bias” (see Chaps. 5 and 11), which is expected when terms were included in a model because they were relatively large.<sup>174</sup> The search for interactions and non-linear terms makes the effective degrees of freedom of a flexible model larger than the final degrees of freedom of a fitted model. This is recognized by FP transformations, where FP1 is tested with 2  $df$ , and FP2 with 4  $df$ . It is not included in  $p$ -values for optimal GAM transformations (according to GCV).  $P$ -values are then only approximate as a result of ignoring uncertainty in the model specification (e.g. searching for a smoothing parameter in GAM).

## Questions

### 12.1 Additivity and interaction

- (a) Explain the additivity assumption in your own words, and the relevance of the scale for assessing additivity?
- (b) Explain interaction terms in your own words?
- (c) How many interaction terms can be assessed in a model with ten binary predictors?
- (d) How many of these would be expected to be statistically significant at the  $p<0.05$  level?

### 12.2 Assumptions and model performance

- (a) Why would you consider testing of the additivity assumption with interaction terms?
- (b) What key problem can occur when interactions and non-linearities are included in the model? How can this be prevented?
- (c) Model performance increases with more flexible non-linear functions. In Table 12.5, the maximum Model LR is 240. Is this model hence preferred for predicting outcome, or do you think other considerations are also relevant?

# Chapter 13

## Modern Estimation Methods

**Background** In this chapter we discuss methods to estimate biased regression coefficients, which lead to better predictions than those obtained with traditional methods. These modern estimation methods include uniform shrinkage methods (heuristic or bootstrap based) and penalized maximum likelihood methods (with various forms of penalty, including the “Lasso”). We illustrate the application of these methods with a data set of 785 patients from the GUSTO-I trial. It appears that rather advanced procedures can now readily be performed with modern software.

### 13.1 Predictions from Regression and Other Models

In linear regression, we aim to minimize the mean squared error, which is calculated as the square distance between observed outcome  $Y$  and prediction  $\hat{Y}$ . The prediction  $\hat{Y}$  can be based on a single predictor, e.g. age predicts blood pressure, or a multivariable combination of predictors, e.g. age, sex, smoking, and salt intake are used to predict blood pressure. As discussed in previous chapters, we can improve predictions from multivariable models for future subjects if the predictions are shrunk towards the average. Statistically speaking, we can reduce the mean squared error for future subjects by using slightly biased regression coefficients.<sup>81,459</sup> This is because predictions will be slightly biased, but have lower variance. The challenge is to find the optimal balance between increasing bias and decreasing variance. This “bias–variance” trade-off underlies the problem of overfitting, and is essential in all predictive modelling (Chap. 5).

In generalized linear regression models, such as logistic or Cox models, maximum likelihood methods are the classical methods for estimation of regression coefficients. Similar to linear regression, the estimated coefficients can be considered as optimal for the sample under study. But again, introducing some bias in the coefficients may lead to better predictions for future subjects.

Neural networks are examples of generalized non-linear models. Their estimation can be done with various techniques. One popular estimation technique is minimizing the Kullback–Leibler divergence, which can be considered as a

distance between two probability densities. One density is provided by the observed outcomes, another by the estimates from the model. Minimizing the Kullback–Leibler divergence is similar to maximizing the likelihood in a generalized linear regression model. Neural networks are quite flexible, and will hence be severely overfitted when they are fully optimized to fit the data. Therefore a common procedure is “early stopping”: the model is not fully trained for maximum fit to the data, but training is stopped at the point where predictive ability is expected to be best. Commonly, the optimal number of iterations to train the model is determined from a cross-validation procedure, where the model is trained on part of the data and tested on an independent part. The optimal number of iterations is then used in the full training part to develop the neural network.

## 13.2 Shrinkage

Shrinkage of regression coefficients towards zero is one way to improve predictions from a regression model.<sup>81,459</sup> We label this method *shrinkage after estimation*, since the shrinkage is applied to regression coefficients after the model has been fitted initially with traditional methods.

Penalized estimation is an alternative method, which uses a penalty factor in the estimation of the regression coefficients: Larger values of regression coefficients are penalized in the fitting procedure, leading to smaller values being preferred. We refer to this method as *shrinkage during estimation*. Although one single penalty factor is used, the degree of shrinkage varies by predictor. A variant of penalized estimation is the Lasso (“least absolute shrinkage and selection operator”).<sup>434</sup> This approach penalizes the sum of the absolute values of the regression coefficients. This leads to some coefficients becoming zero. A predictor with a coefficient of zero can be excluded from the model, which means that the Lasso implies *shrinkage for selection* (Table 13.1).

**Table 13.1** Characteristics of three shrinkage methods

Name	Label	Characteristics
Uniform shrinkage	Shrinkage after estimation	Application of a shrinkage factor to the regression coefficients. The shrinkage factor is determined with a heuristic formula, or by bootstrapping
Penalized maximum likelihood	Shrinkage during estimation	Regression coefficients are estimated with penalized maximum likelihood. The optimal penalty factor can be determined by AIC
Lasso	Shrinkage for selection	Regression coefficients are estimated with penalized maximum likelihood with a restriction on the sum of the coefficients (“Lasso”). The optimal penalty factor can be determined by a cross-validation procedure, or AIC

### 13.2.1 Uniform Shrinkage

A simple and straightforward approach is to apply a uniform (or *linear*) shrinkage factor for the regression coefficients. Shrunk regression coefficients are calculated as  $s \times \beta_i$ , where  $s$  is a uniform shrinkage factor, and  $\beta_i$  are the estimated regression coefficients. The shrinkage factor  $s$  may be based on a heuristic formula<sup>81,459</sup>:

$$s = (\text{model } \chi^2 - df) / \text{model } \chi^2,$$

where *model*  $\chi^2$  is the likelihood ratio  $\chi^2$  of the fitted model (i.e., the difference in  $-2\log$  likelihood between the model with and without predictors), and  $df$  indicates the degrees of freedom of the number of candidate predictors considered for the model. The required shrinkage increases when larger numbers of predictors are considered (more  $df$ ), or when the sample size is smaller (smaller *model*  $\chi^2$ ).

We can also calculate the uniform shrinkage factor  $s$  with bootstrapping.<sup>459,174</sup>

1. Take a random bootstrap sample of the same size as the original sample, drawn with replacement.
2. Select the predictors according to the selection procedure and estimate the logistic regression coefficients in the bootstrap sample.
3. Calculate the value of the linear predictor for each patient in the original sample. The linear predictor is the linear combination of the regression coefficients as estimated in the bootstrap sample with the values of the predictors in the original sample.
4. Estimate the slope of the linear predictor, using the outcomes of the patients in the original sample.

Steps 1–4 need to be repeated many times to obtain a stable estimate of the shrinkage factor as the mean of the slopes in step 4. For example, we may use 200 bootstrap samples, although a fully stable estimate may require 500 bootstrap repetitions.<sup>401</sup> The shrinkage factor may take values between 0 and 1.

### \*13.2.2 Uniform Shrinkage in GUSTO-1

As an example, we consider sample4 from the GUSTO-I study of patients with an acute myocardial infarction (see Chap. 22). The data set consists of 785 patients, of whom 52 had died by 30 days. We consider 2 models for prediction of 30-day mortality after an acute MI: an 8 predictor model, and a 17 predictor model. For estimation of the heuristic shrinkage factor, we need the *model*  $\chi^2$  of each model. These were 62.6 and 73.5. The heuristic shrinkage estimate  $s$  was hence  $(62.6 - 8) / 62.6 = 0.87$ . The larger model required more shrinkage, with  $s = (73.5 - 17) / 73.5 = 0.77$ .

Next, a bootstrap procedure was performed with 200 replications. This resulted in identical estimates of the slope of the linear predictor (0.87 and 0.77, respectively). The regression coefficients are shown in Table 13.2.

**Table 13.2** Logistic regression coefficients estimated with standard maximum likelihood (“original”), uniform shrinkage, penalized maximum likelihood, and the Lasso, for sample4 (795 patients with acute MI, 52 deaths by 30 days)

Predictor	Original	Shrunk	Penalized	Lasso
SHO	1.12	0.97	1.17	1.09
A65	1.49	1.30	1.21	1.36
HIG	0.84	0.74	0.72	0.73
DIA	0.43	0.38	0.36	0.35
HYP	0.99	0.86	0.83	0.87
HRT	0.96	0.84	0.84	0.87
TTR	0.59	0.51	0.49	0.46
SEX	0.07	0.06	0.11	0.00
Shrinkage parameter	NA	s=0.87	penalty=8	s=0.88
Effective shrinkage	1	0.87	0.81–1.49	0–0.97

### 13.3 Penalized Estimation

Penalized maximum likelihood estimation is a generalization of the ridge regression method, which can be used to obtain more stable parameters for linear regression models.<sup>102</sup> Instead of maximizing the log likelihood in generalized linear models, a penalized version of the log likelihood is maximized, in which a penalty factor  $\lambda$  is used:

$$\text{PML} = \log L - 0.5 \lambda \sum (s_i \beta_i)^2,$$

where PML is penalized maximum likelihood,  $L$  is the maximum likelihood of the fitted model,  $\lambda$  a penalty factor,  $\beta$  the estimated regression coefficient for each predictor  $i$  in the model, and  $s_i$  is a scaling factor for each  $\beta_i$  to make  $s_i \beta_i$  unitless.<sup>174,468</sup> It is convenient to use the standard deviation of each predictor for the scaling factor  $s_i$ .<sup>174</sup>

#### 13.3.1 Penalized Maximum Likelihood Estimation

The PML can also be formulated as  $\text{PML} = \log L - 0.5 \lambda \beta' P \beta$ , where  $\lambda$  is a penalty factor,  $\beta'$  denotes the transpose of the vector of estimated regression coefficients (excluding the intercept), and  $P$  is a non-negative, symmetric penalty matrix. For penalized estimation, the diagonal of  $P$  consists of the variances of the predictors and all other values of  $P$  are set to 0.<sup>174</sup> If  $P$  is defined as  $\text{cov}(\beta)^{-1}$  (i.e., the inverse of the covariance matrix of the regression coefficients  $\beta$ ), shrinkage of the regression coefficients is achieved, which is identical to the use of a uniform shrinkage factor as determined by leave-one out cross-validation.<sup>468</sup> If  $P$  is equal to the matrix of second derivatives of the likelihood function, PML is similar to applying a uniform shrinkage factor  $s = 1/(1 + \lambda)$ .

The main problem in penalized estimation is how to choose the optimal penalty factor  $\lambda_{\text{opt}}$ . Maximizing a modified Akaike's Information Criterion (AIC) is an

efficient method.<sup>149</sup> Traditionally, the AIC is defined as  $-2 \log L + 2p$ , where  $L$  is the maximum likelihood of the fitted model and  $p$  is the degrees of freedom equal to the number of fitted predictors. A more convenient formulation is as

$$\text{AIC}_{\text{model}} = \text{model } \chi^2 - 2p,$$

where  $\text{model } \chi^2$  is the likelihood ratio  $\chi^2$  of the fitted model (i.e., the difference in  $-2 \log$  likelihood between the model with and without predictors). For penalized maximum likelihood estimation we use a modified AIC, defined as

$$\text{AIC}_{\text{penalized}} = \text{model } \chi^2_{\text{penalized}} - 2 df_{\text{effective}},$$

where  $\text{model } \chi^2_{\text{penalized}}$  refers to likelihood ratio  $\chi^2$  of the penalized model, and  $df_{\text{effective}}$  is the degrees of freedom after penalizing the fitted predictors. In standard logistic regression, the degrees of freedom are equal to the number of predictors in the model; the higher the number of predictors, the higher the degrees of freedom and the more likely the model is overfitted. Because of the penalization, the degrees of freedom effectively used in penalized estimation are lower than the actual number of predictors. More technically,  $df_{\text{effective}}$  is derived from the reduction in variance of penalized parameter estimates in comparison to the variance of ordinary parameter estimates<sup>149</sup>:

$$df_{\text{effective}} = \text{trace} [I(\beta) \text{cov}(\beta)],$$

where  $I(\beta)$  is the information matrix as computed without the penalty function, and  $\text{cov}(\beta)$  is the covariance matrix as computed by inverting the information matrix calculated with the penalty function. If both the  $I(\beta)$  and  $\text{cov}(\beta)$  are estimated without penalty,  $I(\beta) \text{cov}(\beta)$  is the identity matrix and  $\text{trace}[I(\beta)\text{cov}(\beta)]$  is equal to the number of estimated parameters in the model (excluding the intercept). With a positive penalty function, the  $\text{cov}(\beta)$  becomes smaller and the effective degrees of freedom decrease. With higher penalty values, the model  $\chi^2_{\text{penalized}}$  decreases (poorer fit to the data), but so does the  $df_{\text{effective}}$ . The maximum of  $\text{AIC}_{\text{penalized}}$  ( $\text{model } \chi^2_{\text{penalized}} - 2 df_{\text{effective}}$ ) is sought by varying the values of  $\lambda$  in a trial and error process. For example, we may vary  $\lambda$  over a grid such as 0, 1, 2, 4, 6, 8, 12, 16, 24, 32, 48. Larger values of  $\lambda$  are required for more complex models and larger data sets. The optimal penalty factor  $\lambda_{\text{opt}}$  is the value of  $\lambda$  that maximizes  $\text{AIC}_{\text{penalized}}$ . With this optimal  $\lambda$ , the final model is estimated. An alternative is to use cross-validation or bootstrapping to find the optimal  $\lambda$ , which is more computer intensive compared with finding the maximum of  $\text{AIC}_{\text{penalized}}$ .

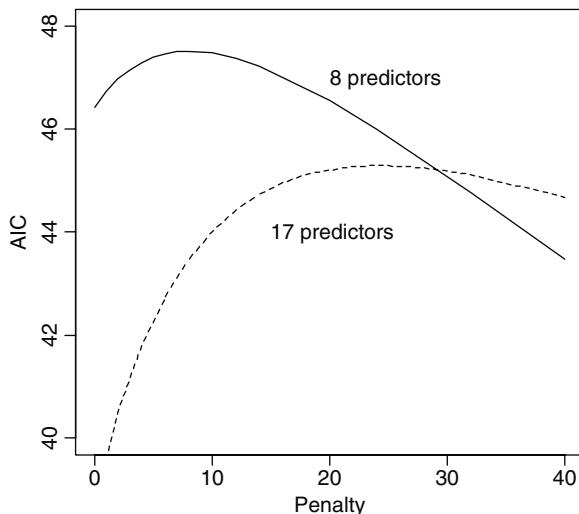
### \*13.3.2 Penalized ML in Sample4

We searched for a penalty factor  $\lambda$  over a grid using the `pentrace` function. The fitting for the 8 predictor model is as follows:

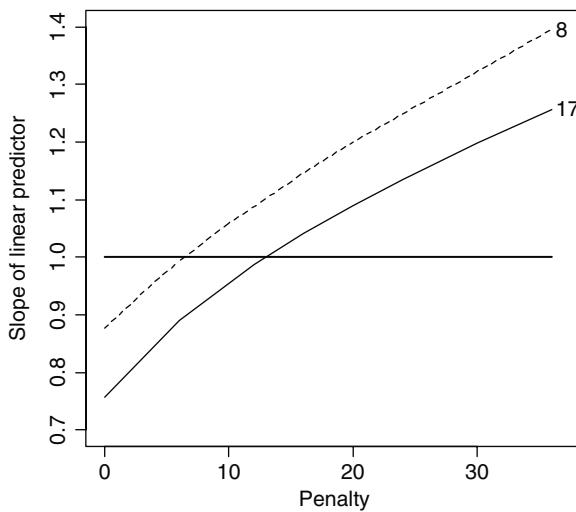
```
# logistic regression model with 8 predictors
full8<- lrm(DAY30~SHO+A65+HIG+DIA+HYP+HRT+TTR+SEX, data=gustos)
# determine performance over range of penalties
p8 <- pentrace(full8, 0:20)
# fit penalized model
full8.pen <- update(full8, penalty=p8$penalty)
```

The  $AIC_{\text{penalized}}$  is calculated with the effective degrees of freedom, and is plotted in Fig. 13.1. The optimum penalty factors were 8 for the 8 predictor model, and 24 for the 17 predictor model. The effective degrees of freedom were 6.9 (instead of 8) and 10.8. (instead of 17). Note that the  $AIC_{\text{penalized}}$  was worse for the 17 predictor model compared with the 8 predictor model, over all penalties considered. The 17 predictor model was hence actually overfitted with only 52 events in the data set.

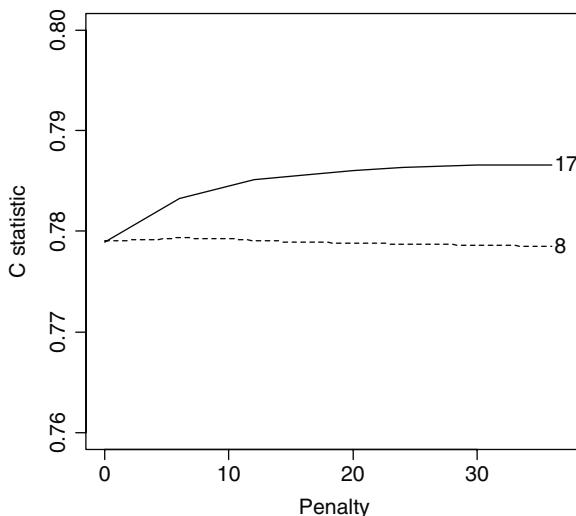
For comparison we also performed a bootstrap procedure to find the optimal penalty factor  $\lambda$ . We created logistic regression models with a range of penalty factors in bootstrap samples drawn with replacement. The models were tested in the original sample. A linear predictor was calculated with the penalized coefficients from the bootstrap sample and the predictor values in the original sample:  $lp = X_{\text{original}} \% \times \% \text{ coef}_{\text{penalized, bootstrap}}$ . Various performance measures can be calculated for this linear predictor. We focus on the slope of the linear predictor, since the primary objective of shrinkage methods is to improve calibration. As expected, the slope is below 1 when no shrinkage is applied (Fig. 13.2). It appears that the slope is 1 if we apply a penalty factor of 7 for the 8 predictor model, and 12 for the 17 predictor model. These values are slightly lower than those obtained from



**Fig. 13.1**  $AIC_{\text{penalized}}$  in relation to the penalty factor. Optimum values are 8 and 24 for the 8 and 17 predictor models, respectively



**Fig. 13.2** Slope of the linear predictor in relation to the penalty factor according to a bootstrap procedure. Optimum values are 7 and 12 for the 8 and 17 predictor models, respectively



**Fig. 13.3**  $c$  statistic in relation to the penalty factor according to a bootstrap procedure. Optimum values are 8 and 30 for the 8 and 17 predictor models, respectively

maximizing the  $AIC_{\text{penalized}}$ . This is explained by the fact that AIC considers the model  $\chi^2$  as criterion rather than the slope of the linear predictor. The model  $\chi^2$  also reflects the discriminative ability, which was higher with larger penalty values (Fig. 13.3).

### 13.3.3 Shrinkage, Penalization, and Model Selection

Uniform shrinkage and penalized estimation methods are defined for pre-specified models. If we apply a selection strategy such as stepwise selection, fewer predictors are included in the selected model, and we might expect less need for shrinkage of coefficients. However, we know that a “testimation” problem arises, i.e. coefficients of selected predictors are overestimated. This selection bias should be taken into account when calculating a shrinkage factor. This may be achieved by considering the number of candidate predictors in the heuristic formula (instead of the number of selected predictors).<sup>459</sup> In a bootstrap procedure, we can include the selection process in step 2.<sup>174</sup> Empirical research suggests that the required shrinkage is more or less similar in pre-specified or selected models.<sup>409</sup> For penalized estimates of the regression coefficients after selection, we can apply the penalty factor that was identified as optimal for the full model, before selection took place.

A specific situation is when a substantial number of interaction terms is tested, and one or more are included in the final model. For shrinkage, we could still use the original  $df$  of the model with main effects and all interactions considered. A more elegant solution was suggested by Harrell for penalized ML estimation, i.e. to penalize the interaction terms more than the main effects, for example with twice the penalty of the main effects. Similarly, non-linear and nonlinear interaction terms might be penalized by twice and 4 times the penalty of the main effects.<sup>174</sup>

## 13.4 Lasso

A formal method to achieve model selection through shrinkage is the Lasso (least absolute shrinkage and selection operator).<sup>434</sup> The Lasso can efficiently be applied to linear regression models using “least angle regression.”<sup>106</sup> The Lasso can also be used for generalized linear models such as the logistic or Cox model.<sup>435</sup> The Lasso preferentially shrinks some predictors to zero.

### 13.4.1 Estimation of Lasso Model

The Lasso estimates the regression coefficients of standardized predictors by minimizing the log-likelihood subject to  $\sum |\beta| \leq t$ , where  $t$  determines the shrinkage in the model. We may vary  $s = t / |\sum \beta_0|$  over a grid between 0 and 1, where  $\beta_0$  indicates the standard ML regression coefficients and  $s$  may be interpreted as a standardized shrinkage factor. We may estimate the final  $\beta$  with the value of  $t$  that gives the lowest mean-squared error in a generalized cross-validation procedure.<sup>435</sup> We may also aim to optimize AIC or use bootstrapping to find the optimal value for  $t$ .<sup>320</sup>

### \*13.4.2 Lasso in GUSTO-I

We used the `glmpath` package for *R* to perform lasso analyses, but other packages are nowadays available. This is a path-following algorithm for L1 regularized generalized linear models and Cox proportional hazards model.<sup>320</sup> The logistic regression coefficients were estimated given a bound (“L1”) to the sum of absolute  $\beta$ ,  $|\beta|$ . The predictors are standardized such that sum  $|\beta|$  does not depend on coding of predictors.

```
# make list of predictors in matrix x, outcome in y
gustosd <- list(x=full8$x, y=full8$y)
# fit logistic models over a range of L1
gustopath <- glmpath(data=gustosd)
# plot results: Fig 13.4
plot.glmpath(gustopath, type="coefficients")
plot.glmpath(gustopath, type="aic")
```

With a low *L1* bound, small coefficients were estimated for the predictors A65 (age>65 years), SHO (Shock), and HRT (Tachycardia). This occurred both in the 8 and 17 predictor models (Fig. 13.4). The other predictors had coefficients set to zero. With larger bounds, non-zero coefficients were estimated for these predictors as well. With a bound over 0.6 (8 predictor model) or over 0.9 (17 predictor model), the original, unshrunk logistic model was estimated.

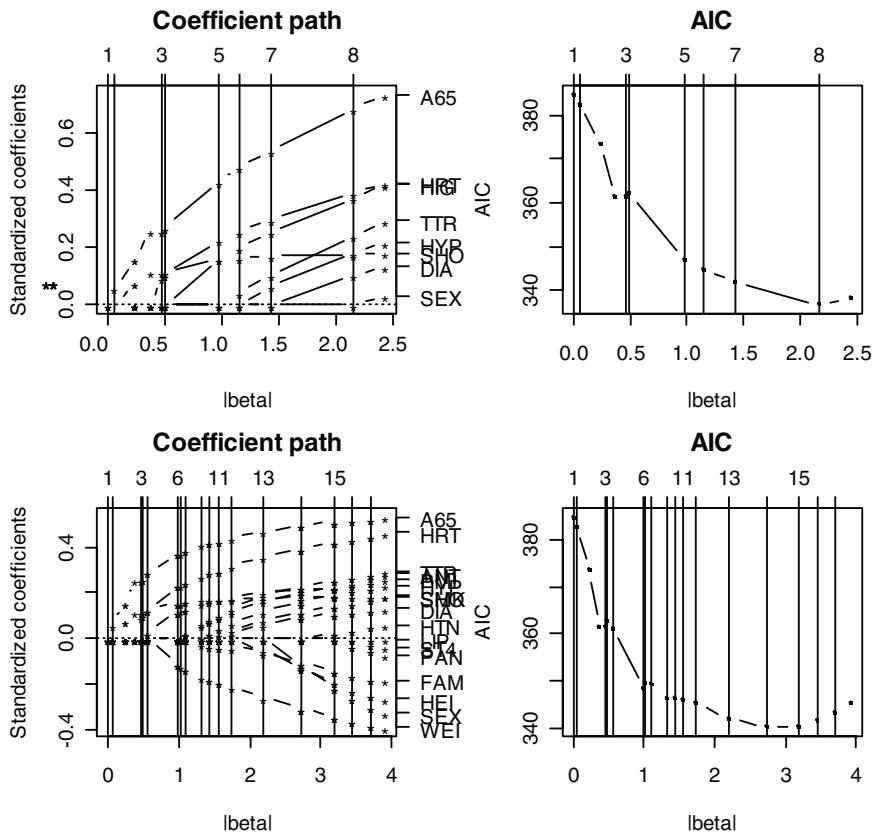
The optimum penalty can be estimated by studying the AIC (Fig. 13.4). This suggests an optimal selection of seven predictors in the 8 predictor model (*L1* = 2.1), and a selection of 14 predictors for the 17 predictor model (*L1* = 3.0). We can validate the selection and estimated coefficients through a bootstrap analysis (see the book’s website). The coefficients for the final model are chosen at the lowest AIC value. The effect of SEX was set to zero, and the coefficient of DIA was small (standardized coefficient, 0.10).

```
# coefficients at lowest AIC: Table 13.2
gustopath$b.predictor[gustopath$aic==min(gustopath$aic),]
Intercept   SHO    A65    HIG    DIA    HYP    HRT    TTR    SEX
-4.55       1.09   1.36   0.73   0.35   0.87   0.87   0.46   0.00

# linear predictor with Lasso model, step 12 has lowest AIC
predict.glmpath(gustopath, newx=full8$x, newy=full8$y, s=12)
```

### 13.4.3 Predictions after Shrinkage

Shrinkage leads to a less-extreme distribution of predictions in the GUSTO-I example. The linear predictor is shrunk towards the average compared with standard maximum likelihood, either with uniform shrinkage, penalized maximum likelihood estimation (PMLE), or the Lasso (Fig. 13.5).



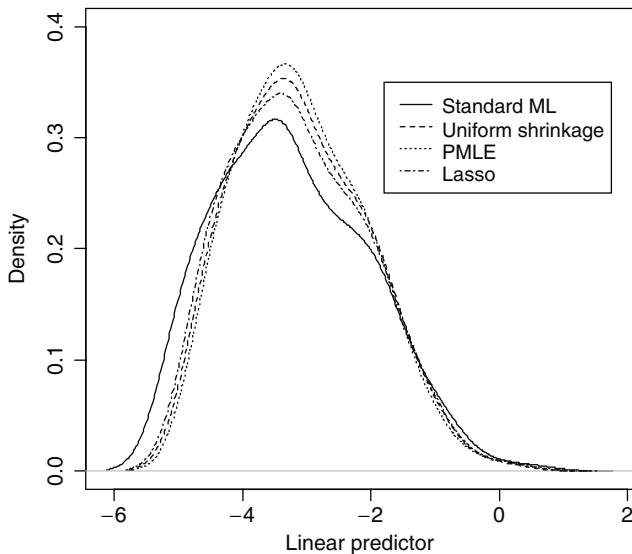
**Fig. 13.4** Coefficients and AIC for 8 and 17 predictor models according to the sum of the absolute values of the regression coefficients ( $|\beta|$ ) in sample4 from GUSTO-I ( $n=785$ , 52 deaths)

### 13.4.4 Model Performance after Shrinkage

We compared the performance of models constructed in small samples of the GUSTO-I data set in an independent test part (see Chap. 22 for design). Table 13.3 shows the discrimination and calibration with and without shrinkage. As expected, discrimination is not much affected by shrinkage. In contrast, the calibration slope is closer to 1 when shrinkage is applied. Shrinkage hence prevents that too extreme predictions are derived from the development data set.

## 13.5 Concluding Remarks

Shrinkage of regression coefficients is an important way to battle overfitting; too extreme predictions are prevented. Shrinkage is especially beneficial in small data sets, and/or situations with large numbers of candidate predictors. Using advanced



**Fig. 13.5** Distribution of the linear predictor in sample4 from GUSTO-I with standard and penalized maximum likelihood, uniform shrinkage, and the Lasso

**Table 13.3** Discrimination ( $c$  statistic) and calibration (calibration slope) of the 8 and 17 predictor models based on small and large subsamples (average,  $n=336$  and  $n=892$ , respectively), and based on the total training part ( $n=20,512$ ), as evaluated in the independent test part of GUSTO-I ( $n=20,318$ )

Training data	8 predictors		17 predictors	
	C statistic	Slope	C statistic	Slope
Total training ( $n=20,512$ , 1,423 deaths)	Standard ML	0.789	0.944	0.802
61 small subsamples ( $n=336$ , 23 deaths on average)	Standard ML	0.75	0.66	
	Uniform shrinkage	0.75	1.01	
	Penalized ML	0.76	0.93	
23 large subsamples ( $n=892$ , 62 deaths on average)	Lasso	0.75	0.83	
	Standard ML	0.78	0.86	0.78
	Uniform shrinkage	0.78	0.97	0.78
	Penalized ML	0.78	0.96	0.79
	Lasso	0.78	1.01	0.78

Mean values are shown for several estimation methods with a fixed selection of predictors

shrinkage procedures is readily possible with modern software, implemented in for example *R* (pentrace function in Design library for penalized estimation, glmpath for Lasso). Penalty factors are a general concept in smooth estimation of model parameters; they are also important in curve fitting (e.g. with splines) and generalized additive models.<sup>181</sup> The Lasso currently receives interest for analysis of genomic data.<sup>321</sup>

Shrinkage methods have been applied in a number of case studies. Moons et al. describe penalized maximum likelihood and illustrates the method with a nice case study.<sup>293</sup> Vach et al. compared the empirical behaviour of various shrinkage techniques.<sup>443</sup> The results from simulations in GUSTO-I were presented in more detail in other papers.<sup>408,409,410</sup>

## Questions

### 13.1 Shrinkage and model performance

Explain how shrinkage can influence (a) the predictions from a model, (b) calibration, and (c) discrimination.

### 13.2 Penalized maximum likelihood

- (a) Why might we label PML “shrinkage during estimation” (Table 13.1)
- (b) How is it possible that one penalty term leads to differential shrinkage in Table 13.2?
- (c) In a recent paper (Smits et al. 2007),<sup>391</sup> we can study the effect of PML on the various coefficients. Which coefficients are penalized most?

### 13.3 Shrinkage methods and stepwise selection (Sect. 13.3.3)

How can shrinkage and penalization be used when the model is developed with stepwise selection:

- (a) Uniform shrinkage with Van Houwelingen’s formula or bootstrapping?
- (b) Penalized maximum likelihood?

# Chapter 14

## Estimation with External Information

**Background** In this chapter we discuss methods that estimate regression coefficients based on the combination of findings from the sample under study with external information. We start with a simple “adaptation” method for univariate regression coefficients, which may be obtained from meta-analysis. This method was applied in a case study of operative mortality of abdominal aneurysm surgery. Next, we discuss some alternative approaches to estimate regression coefficients, including Bayesian estimation with explicit prior information.

### 14.1 Combining Literature and Individual Patient Data

We consider the common situation that several studies have already been published for a particular clinical prediction problem, in which the relation between patient characteristics and the outcome of interest is described. If the published papers describe comparable patient series, we may try to combine the available evidence quantitatively in a meta-analysis. The information in these papers is usually only sufficient to calculate a univariate regression coefficient for each of the patient characteristics.

Multivariable coefficients can directly be estimated if individual patient data are available from the published series, or if we know the correlation structure between predictors. This information is usually not available. Individual patient data may be especially hard to retrieve for papers published several years ago, and anyway requires a substantial research effort. Thus, typically the researcher may have access to individual patient data from one study (“own data set”) and univariate information from the literature (“publicly available”).

An “adaptation method” has been proposed to take advantage of the univariate literature data in the estimation of the multivariable regression coefficients in a prediction model.<sup>411</sup> The aim is better prediction of the outcome in individual patients. This adaptation method is closely related to an earlier proposal by Greenland for meta-analysis.<sup>151</sup> For example, when studying the relation between coffee consumption and acute myocardial infarction, one study may have corrected the regression coefficient for a confounder (for example alcohol consumption), while other studies have not. Greenland proposed to use the change from unadjusted to adjusted regression coefficient to adapt the unadjusted coefficients in the latter studies.

### 14.1.1 Adaptation Method 1

In our case of regression analysis on literature and individual patient data, the formula reads like

$$\beta_{m|I+L} = \beta_{u|L} + (\beta_{m|I} - \beta_{u|I}),$$

where  $\beta_{m|I+L}$  refers to the multivariable coefficient based on the combination of individual patient data and literature data (the “adapted coefficient”),  $\beta_{u|L}$  is the univariate coefficient from a meta-analysis of the literature, and  $\beta_{m|I} - \beta_{u|I}$  is the difference between multivariable and univariate coefficient in the individual patient data (the “adaptation factor”). Hence, we simply use the change from univariate to multivariable coefficient in our own data to adapt the meta-analysis coefficient.

For the variance of the adapted coefficient ( $\text{var}(\beta_{m|I+L})$ ), we may add the difference between variances of the multivariable and univariate coefficient to the variance of the univariate coefficient from the literature, ignoring all covariances:

$$\text{var}(\beta_{m|I+L}) = \text{var}(\beta_{u|L}) + \text{var}(\beta_{m|I}) - \text{var}(\beta_{u|I}).$$

### 14.1.2 Adaptation Method 2

A more general way to formulate the adaptation formula is as

$$\beta_{m|I+L} = \beta_{m|I} + c (\beta_{u|L} - \beta_{u|I}),$$

where  $c$  is a factor between 0 and 1. If  $c = 1$ , the same formula as proposed by Greenland arises. If  $c$  equals 0, the literature data is effectively discarded. The estimate of  $\beta_{m|I+L}$  is unbiased for any choice of  $c$ , if the expectation of  $\beta_{u|L} - \beta_{u|I} = 0$ , that is, the individual patient data form a random part from the studies included in the meta-analysis. It was found that we can derive a formula for  $c$  so as to minimize the variance of  $\beta_{m|I+L}$ :

$$C_{\text{opt}} = \rho(\beta_{m|I}, \beta_{u|I}) \frac{\text{SE}(\beta_{m|I}) \times \text{SE}(\beta_{u|I})}{\text{var}(\beta_{u|L}) + \text{var}(\beta_{u|I})},$$

where  $\rho(\beta_{m|I} - \beta_{u|I})$  refers to the correlation between multivariable and univariate coefficients in the individual patient data.

This variant of the adaptation method indicates that adaptation will be especially advantageous if the literature data set is larger (resulting in a smaller  $\text{var}(\beta_{u|L})$ ), or when the correlation  $\rho(\beta_{m|I} - \beta_{u|I})$  is larger. The latter correlation is expected to be large if the collinearity between covariates is small. The adaptation factor will then be close to 1, and method 1 may yield good results.

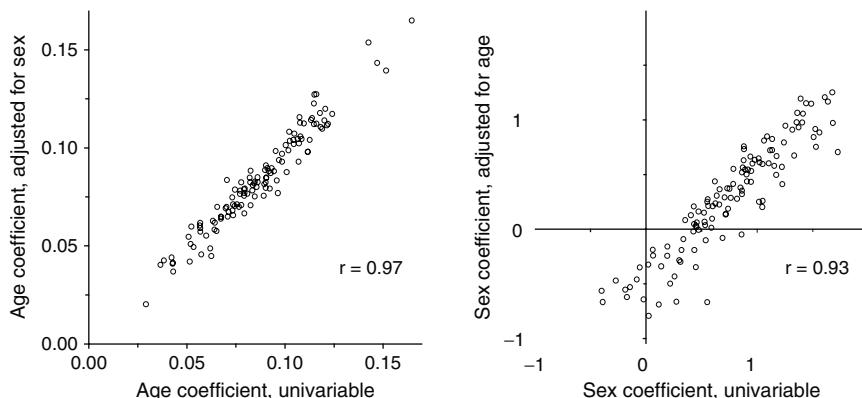
### 14.1.3 Estimation

Meta-analysis techniques may be used to estimate the univariate coefficients from the literature data. The literature data may include the individual patient data for maximal efficiency. The meta-analysis may assume fixed effects (for example, Mantel-Haenszel method, or conditional logistic regression), or random effects (for example, DerSimonian Laird method, or likelihood-based methods<sup>97</sup>). The calculations for method 1 use estimates that are readily available. For example, logistic regression analysis with standard maximum likelihood (ML) provides estimates of the univariate and multivariable coefficients in the individual patient data.

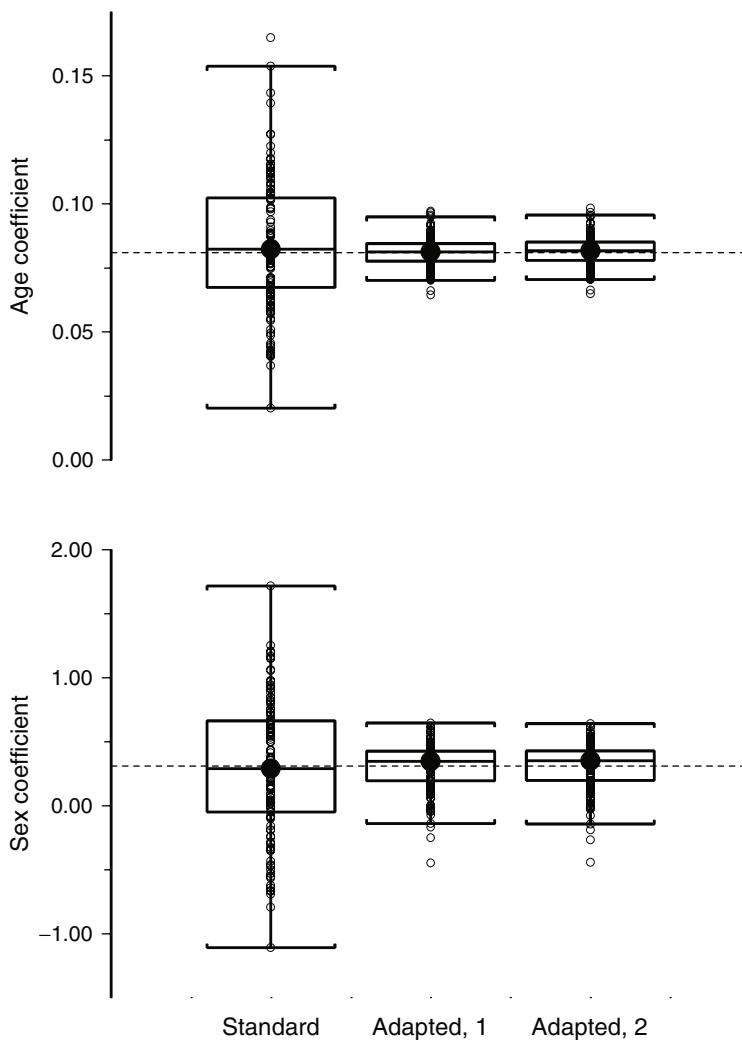
For the second method, the estimation of the optimal adaptation factor requires estimates of the variances of the regression coefficients, and an estimate of the correlation between univariate and multivariable coefficients. The latter correlation cannot easily be estimated with logistic regression methods. We therefore used bootstrap re-sampling to calculate the coefficients  $\beta_{m|I}$  and  $\beta_{u|I}$  repeatedly, and their correlation  $\rho$ .

### 14.1.4 Simulation Results

The adaptation method was tested in the GUSTO-I data.<sup>411</sup> First, we assessed the correlation between multivariable and univariate coefficients across 121 small subsamples. We observed a strong correlation for the combination of age and sex in a 2 predictor model (Fig. 14.1). Results were somewhat less favorable for predictors with stronger collinearity. For example, weight and height had a Pearson correlation coefficient of 0.54, and the correlation between their univariate and multivariable coefficients was 0.80 and 0.83 in a bivariate model respectively. Overall, the strong  $r(\beta_{m|I} - \beta_{u|I})$  supports the use of the adaptation method in medical data.



**Fig. 14.1** Correlations between univariate and multivariable regression coefficients in a 2 predictor model consisting of age and sex estimated in 121 small subsamples of the GUSTO-I data set<sup>411</sup>



**Fig. 14.2** Regression coefficients in the 2 predictor model consisting of age and sex. Box plots show the standard ML, and adapted estimates (methods 1 and 2) for 121 small subsamples; the line --- indicates the coefficient observed in the total GUSTO-I data set ( $n = 40,830$ )<sup>411</sup>

Next, we estimated the values of  $c_{\text{opt}}$ . Values were quite close to 1 ( $0.98 \pm 0.015$  and  $0.99 \pm 0.020$  for age and sex (mean  $\pm$  SD) in the 121 small subsamples). Hence, Greenland's method ( $c = 1$ ) and our method ( $c$  estimated with bootstrapping) resulted in very similar estimates of the adapted coefficients (Fig. 14.2). Both methods lead to much better estimates of the multivariable regression coefficients in the small subsamples. Specifically, a substantial reduction is noted in the variability

compared with the standard multivariable regression coefficients, i.e.  $\text{var}(\beta_{m|I_L}) \ll \text{var}(\beta_{m|I})$ . These very favorable results were obtained by using univariate results from approximately half of the GUSTO-I data ( $n = 20,000$ ). We also examined the influence of the size of the literature data. We applied the adaptation methods in the small subsamples, where univariate literature estimates were obtained from a neighboring, small subsample. This resulted effectively in a doubling of the sample size. This pattern was also reflected in the values of the adaptation factor from method 2; close to 1 with  $n = 20,000$  as literature data, around 0.50 with a neighbour subsample as literature data.

### **14.1.5 Performance of Adapted Model**

Finally, we compared the predictive performance of the adaptation method to the performance obtained with uniform shrinkage, penalized ML, or the Lasso in 23 large subsamples from GUSTO-I (Table 14.1). The discriminative ability improved slightly, but some problems were noted in calibration. Miscalibration was less than for the standard ML estimates, but some form of shrinkage should actually have been built into the adaptation method.

### **\*14.1.6 Improving Calibration**

To improve the calibration of the predictions resulting from applying the adaptation method, we considered two approaches. First, we shrunk the multivariable regression coefficients as estimated in the individual patient data. This approach was discarded

**Table 14.1** Discrimination ( $c$  statistic) and calibration (calibration slope) of the 8 and 17 predictor models based on large subsamples (average  $n=892$ , respectively), and based on the total training part ( $n=20,512$ ), as evaluated in the independent test part of GUSTO-I ( $n=20,318$ )

	<i>c</i> statistic		Calibration slope	
	8 predictors	17 predictors	8 predictors	17 predictors
Total training ( $n=20,512$ , 1,423 deaths)				
Standard ML	0.789	0.802	0.944	0.959
23 large subsamples ( $n=892$ , 52 deaths on average)				
Standard ML	0.78	0.78	0.86	0.76
Uniform shrinkage	0.78	0.78	0.97	0.95
Penalized ML	0.78	0.79	0.96	0.98
Lasso	0.78	0.78	1.01	0.93
Adapted 1	0.79	0.79	0.92	0.86
Adapted 2	0.79	0.79	0.92	0.86

Means are shown for two variants of the adaptation method and several other modern estimation methods (see Chap. 13)

because it led to better calibration (slope closer to 1), but a decrease in discriminative ability. The second approach was motivated by the observation that the miscalibration of the adapted estimates was approximately halfway that of shrunk estimates and the standard ML estimates. The proposed formula is

$$\beta_{m|l} = (1 + \text{shrinkage factor}) / 2[\beta_{m|l} + c(\beta_{u|l} - \beta_{u|l})]$$

where the shrinkage factor is the uniform shrinkage factor, either estimated with a heuristic formula, or by bootstrapping (see Chap. 13).

Evaluations of this correction with method 1 ( $c$  set to 1) or 2 ( $c$  estimated by bootstrapping) showed an improvement in calibration. Discriminative ability was identical to that without shrinkage, since the shrinkage did not affect the ordering of predictions.

## 14.2 Example: Mortality of Aneurysm Surgery

In our examples with GUSTO-I, no relevant differences were noted between adaptation methods 1 and 2. We applied adaptation methods 1 and 2 in the prediction of peri-operative mortality (in-hospital or within 30 days) after elective abdominal aortic aneurysm (AAA) surgery.<sup>421</sup> Individual patient data were available on a relatively small sample (246 patients, 18 deaths). Patients were operated on at the University Hospital Leiden between 1977 and 1988. Univariate literature data were available from 15 published series with 15,821 patients (1,153 deaths) in total. Predictors considered included age and sex, cardiac comorbidity (history of myocardial infarction (MI), congestive heart failure (CHF), and ischemia on the ECG), pulmonary comorbidity (COPD, emphysema or dyspnea), and renal comorbidity (elevated pre-operative creatinin level). These predictors were chosen since they were reported in at least two studies in the literature, and were also available in the Leiden data set.

### 14.2.1 Meta-Analysis

Univariate logistic regression coefficients were estimated both with fixed and random effects methods from the literature data. As expected, the estimates of the coefficients were very similar, but the SEs were somewhat larger with the random effect method (Table 14.2).

A number of practical issues merit discussion with respect to the meta-analysis of the literature data. First, definitions of predictors varied, especially for pulmonary and renal comorbidity. Despite these differences, it was considered reasonable to assume one single effect for each predictor across the studies (non-significant tests for heterogeneity of odds ratios, non-significant interaction terms between study and effect estimates in logistic regression).

**Table 14.2** Meta-analysis results for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	Fixed effect	Random effect
Age (per decade)	0.79 (0.06)	0.79 (0.11)
Female sex	0.36 (0.08)	0.36 (0.18)
History of MI	1.03 (0.27)	1.03 (0.32)
Congestive heart failure	1.59 (0.33)	1.59 (0.41)
ECG: Ischaemia	1.52 (0.31)	1.51 (0.38)
Impaired renal function	1.32 (0.25)	1.30 (0.26)
Impaired pulmonary function	0.89 (0.23)	0.85 (0.24)

**Table 14.3** Individual patient data results ( $n=246$ ) for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	Univariate	Standard ML	Shrunk	Penalized	$r(\beta_{\text{ml}}, \beta_{\text{ul}})$
Age (per decade)	0.98 (0.38)	0.58 (0.39)	0.48	0.34	0.91
Female sex	0.28 (0.79)	0.30 (0.86)	0.25	0.17	0.81
History of MI	1.50 (0.50)	0.74 (0.57)	0.61	0.57	0.88
Congestive heart failure	1.78 (0.55)	1.04 (0.59)	0.86	0.67	0.92
ECG: Ischaemia	1.72 (0.55)	0.99 (0.62)	0.83	0.63	0.87
Impaired renal function	1.24 (0.70)	1.12 (0.77)	0.93	0.74	0.85
Impaired pulmonary function	0.84 (0.53)	0.61 (0.59)	0.51	0.39	0.90

Second, the number of studies that described a predictor varied. The effect of age was reported in 15 studies, sex and renal function in 6, pulmonary function in 5, MI in 3, and CHF and ECG findings in only 2 studies. This somewhat limits the value of the adaptation method in this case study.

Third, the analysis of age as a continuous variable was hampered by the fact that mortalities were described in relatively large age intervals, for example, younger or older than 70 years. For logistic regression analysis, we estimated the mean ages in these age intervals using study-specific descriptions as far as available (mean and SE). We checked in a small simulation study that using the mean was better than using the median for age categories. The effect of age would have been estimated more accurately if smaller age intervals had been reported or more study characteristics had been published.

### 14.2.2 Individual Patient Data Analysis

In the individual patient data, multivariable logistic regression coefficients were usually smaller than the univariate coefficients, reflecting a predominantly positive correlation between predictors (Table 14.3). Correlations were strongest between the three cardiac comorbidity factors ( $r$ , 0.26, 0.32, and 0.45) and between these three factors and age ( $r>0.20$ ). We note that the number of predictors (7) was large

relative to the number of events (18 deaths). Bootstrapping estimated a shrinkage factor of 0.83 (200 replications, convergence in only 119), and penalized ML was performed with 14 as the penalty factor. The correlation  $\rho$  between univariate and multivariate coefficients was estimated between 0.81 and 0.91.

### 14.2.3 Adaptation Results

The literature and individual patient data were combined with the adaptation method, using the random effect estimates from the literature data. For adaptation method 1,  $c_{\text{opt}}$  was always set to 1 (Table 14.4; for method 2,  $c_{\text{opt}}$  was estimated between 0.63 and 0.86, results not shown). Compared with shrunk or penalized coefficients, the adapted estimates for sex and renal and pulmonary function were somewhat higher and lower for a history of MI.

For application in clinical practice, scores were created by rounding each adapted coefficient after multiplication by 10 and shrinkage of 90% ( $(1+\text{bootstrap shrinkage factor})/2 \approx 0.90$ ). The intercept was calculated with an offset variable in a logistic regression model. The offset was the linear combination of the scores (divided by ten) and the values of the covariables in the individual patient data. The intercept was estimated as -3.48.

The intercept was further adjusted for a presumably lower mortality in current surgical practice (5%) than that observed in the individual patient data (7.6%). This adjustment can be considered as a form of recalibration to contemporary circumstances. It was achieved by subtracting  $\ln(\text{odds}(5\%)/\text{odds}(7.6\%)) = -0.44$  from the previous intercept estimate:  $-3.48 - 0.44 = -3.92$ . This results in the following formula to estimate the risk of peri-operative mortality in current elective abdominal aortic aneurysm surgery:

$$p(\text{operativemortality}) = \frac{1}{[1 + \exp(-(\sum \text{score} / 10) - 3.92)]}.$$

The area under the ROC curve was 0.83 in the individual patient data with standard, shrunk or penalized estimation. But the optimism-corrected estimates were

**Table 14.4** Individual patient data results ( $n=246$ ) for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	$\beta_{\text{ml}} - \beta_{\text{ull}}$	$c$ method 1	Adapted 1	Score
Age (per decade)	-0.40	1	0.38 (0.14)	3
Female sex	+0.02	1	0.38 (0.40)	3
History of MI	-0.76	1	0.27 (0.41)	2
Congestive heart failure	-0.74	1	0.85 (0.47)	8
ECG: Ischaemia	-0.73	1	0.79 (0.48)	7
Impaired renal function	-0.12	1	1.18 (0.41)	11
Impaired pulmonary function	-0.23	1	0.62 (0.34)	6

Score: Rounded value of  $9 \times \text{"Adapted 1"}$

0.80 for standard or shrunk, estimation, and 0.81 for penalized estimation (bootstrapping with 200 replications). For the final model with adapted coefficients, we expect a performance at least as good as these methods, but this needs to be confirmed in further validation studies.

## 14.3 Alternative Approaches

Several alternative approaches are possible to adjust univariate results for use in a multivariable model. We discuss two approaches below: Using an overall calibration factor for the univariate literature coefficients and Bayesian methods.

### 14.3.1 Overall Calibration

One variant of naïve Bayes was already suggested in Chap. 4, i.e. use of a uniform, overall calibration factor for all univariate coefficients. In the case study of aortic aneurysm mortality, the calibration factor is 0.69 for a linear predictor based on the univariate coefficients from the literature multiplied with the predictor values in the individual patient data. The recalibrated coefficients are reasonably close to those estimated with our adaptation method. The overall calibration led to higher values for cardiac comorbidity factors (scores 7, 11, and 10 for MI, CHF, and Ischaemia vs. 2, 8, and 7 with the adaptation method, respectively). This is explained by the relatively strong correlations among these factors, while the overall calibration reflects an average correlation between all the seven predictors.

### 14.3.2 Bayesian Methods: Using Data Priors to Regression Modelling

Greenland has argued that a Bayesian perspective needs to be incorporated into basic biostatistical and epidemiological training.<sup>152</sup> In particular in small data sets with many predictors, Bayesian approaches may offer advantages over conventional frequentist methods. Estimation of regression coefficients is difficult for data sets with few or no subjects at crucial combinations of predictor values.

Bayesian estimation consists of setting prior values for the regression coefficients, which are combined with the estimates in the data to produce posterior estimates of the coefficients. When the prior values are all zero, the coefficients are pulled towards zero. This is similar to shrinkage, as discussed in Chap. 13. Setting a prior to zero may be reasonable for a variable with very doubtful value as a predictor. A negative or positive effect is then equally likely, making zero the best prior guess. We allow for the possibility that the effect is non-zero, but may consider

large values unlikely. The degree of shrinkage is then determined by the width of the prior distribution. The narrower the prior distribution, the more the prior shrinks the coefficient towards zero. The other factor determining shrinkage is how strongly the predictor is related to the outcome in the data under study; in an informative data set (many events, not a rare predictor), there will be limited shrinkage. The final estimate is an average of the prior expectation and the conventional estimate.

A more interesting role for Bayesian approaches in regression is in using informative priors. For example, we may hypothesize a priori that a predictor has an odds ratio of 2, with values smaller than 0.5 and larger than 8 being highly unlikely. Setting a reasonable informative prior is the most difficult task for Bayesian analysis. Expert judgment or literature review can be used. When using informative priors, the source of these priors should be well documented, and sufficient variability allowed in the prior distribution. Presentation of prior information can be presented as “informationally equivalent,” e.g. assuming knowledge of 100 patients with a certain outcome. This may be acceptable to some in the medical field, but will be met with scepticism by others, including traditional biostatisticians and applied clinical researchers.

#### \*14.3.3 Example: Predicting Neonatal Death

Greenland describes a case study of predicting neonatal-death risk in a cohort of 2,992 births with 17 deaths.<sup>152</sup> He estimates logistic regression models with 14 predictors, assuming small to large effects for most predictors. He finds that the predictive ability of the Bayesian model is better than a model based on standard ML. He also illustrates how Bayesian estimation can be achieved relatively easily with data augmentation: Records are added to a data set, reflecting predictive effects of predictors.<sup>153</sup> In the case of a multivariable model, the prior distributions refer to the multivariable effects of predictors, which may be more complicated to elicit from experts or from literature than univariate effects.

#### \*14.3.4 Example: Mortality of Aneurysm Surgery

In the prediction of peri-operative mortality of aortic aneurysms, we might try to use informative priors based on the literature. The meta-analysis however provides univariate effects, and we need to translate these to priors for multivariable effects. The difference between univariate and multivariable coefficients is directly related to the correlation between predictors. If we have some guesses for these correlations, this may give some hints on how the multivariable coefficients compare with the univariate coefficients. For example, with substantial correlations, we might halve all univariate coefficients; with no correlation, we keep the multivariate effect at the univariate estimate. Being on the conservative side with informative priors may be sensible to make Bayesian analysis more acceptable.

## 14.4 Concluding Remarks

The proposed adaptation methods emphasize the central role of subject knowledge in developing prediction models in small data sets. Literature data may guide the selection of predictors (Chap. 11), as well as improve the estimates of the regression coefficients (this chapter). Especially when the data set is relatively small, this strategy will result in more reliable regression models than using a strategy that considers a data set with individual patient data as the sole source of information.

A potential problem of meta-analyses is that publication bias may have led to overestimation of the regression coefficients. Also, performing a meta-analysis may not be realistic if definitions of risk factors vary substantially in the literature. Finally, the central assumption in the adaptation method is that the data set under study and the literature data are random subsamples from a common population, which implies that the correlations between predictors are similar in the individual patient data and in the literature data.

Bayesian methods provide another perspective on estimation of regression coefficients. If no effect is expected for a predictor, shrinkage of coefficients towards zero is achieved, quite similar to using uniform shrinkage or penalized ML. If other effects are assumed, coefficients will be pulled towards this prior value. As with any Bayesian method, the main criticism will be on the choice of prior distribution.

Many papers have been written about Bayesian approaches, but Bayesian methods have not yet made it to mainstream predictive modelling. A variant is empirical Bayes estimation, which will be discussed in Chaps. 20 and 21. Empirical Bayes methods have an important role in for example estimating centre effects, and provider profiling. With this variant, the prior distribution of centre effects is determined empirically from the data.

In some Bayesian applications, uninformative priors are used by default; these variants only use Bayesian calculations to achieve results that are difficult to calculate with frequentist methods, such as ML. These methods are becoming quite popular in medicine, e.g. using WinBUGS ([www.mrc-bsu.cam.ac.uk/bugs/](http://www.mrc-bsu.cam.ac.uk/bugs/)) with the Gibbs sampler as the core Bayesian method.<sup>134</sup>

## Questions

### 14.1 Key factors in adaptation method (Sect. 14.1 and 14.2)

We examine the key factors for the adaptation method, as illustrated in the aneurysm case study.

- (a) What would happen to the adapted coefficients when larger univariate coefficients were found in the literature?
- (b) What would happen to the adapted coefficients when the univariate coefficients were identical in the literature and in the individual patient data?
- (c) What would happen to the adapted coefficients when there was virtually no correlation between predictors?

### 14.2 Variance of adapted coefficients (Sect. 14.1.1)

In the simple variant, the variance of the adaption method is estimated as:

$$\text{var}(\beta_{m|I+L}) = \text{var}(\beta_{u|L}) + \text{var}(\beta_{m|I}) - \text{var}(\beta_{u|I})$$

When we have a literature data base ("L") of the same size as the individual patient data base ("I"), the variance decreased by a factor of 2 (SE decreases by  $1/\sqrt{2}$ , Sect. 14.1.4). What may be expected for the variance and SE of an adapted coefficient when we have a literature data base of 3 times the size of the individual patient data?

### 14.3 Adaptation method in aneurysm case study (Sect. 14.2)

For the aneurysm case study, the age effect is based on a very large sample size in the meta-analysis. The regression coefficient is 0.79 per 10 years; SE in random effect model, 0.14.

- (a) Verify that the adaptation factor  $\beta_{m|I} - \beta_{u|I}$  is -0.40.
- (b) Verify that the SE of the adapted coefficient becomes 0.14, while it was 0.39 in the original multivariable analysis (Table 14.3).

# Chapter 15

## Evaluation of Performance

**Background** When we develop or validate a prediction model, we want to quantify how good the predictions from the model are (“model performance”). Predictions are absolute risks, which go beyond assessments of relative risks, such as regression coefficients, odds ratios, or hazard ratios. We can distinguish apparent, internally validated, and externally validated model performance (Chap. 5). For all types of validation, we need performance criteria in line with the research questions, and different perspectives can be chosen. We first take the perspective that we want to quantify how close our predictions are to the actual outcome. Next, more specific questions can be asked about calibration and discrimination properties of the model, which are especially relevant for prediction of binary outcomes in individual patients. We will illustrate the use of performance measures in the testicular cancer case study, with model development in 544 patients, internal validation with bootstrapping, and external validation with 273 patients from another centre.

### 15.1 Overall Performance Measures

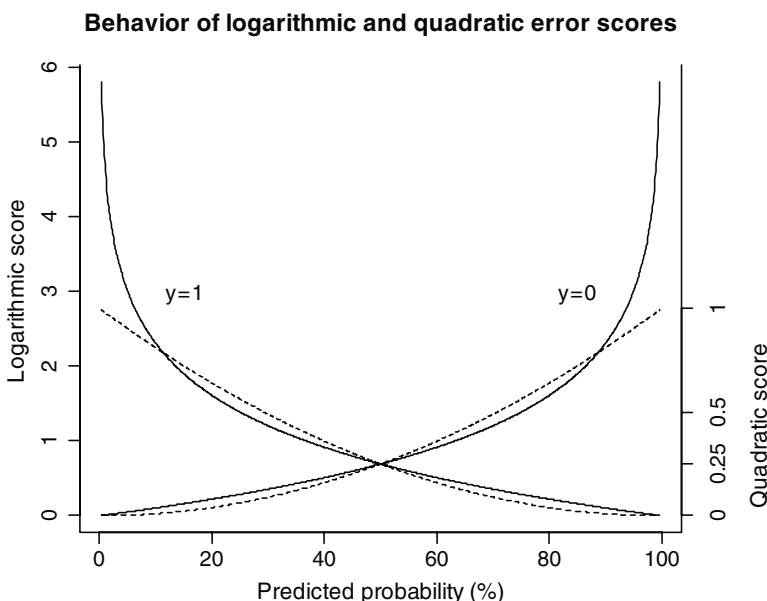
The distance between the predicted outcome and actual outcome is a central to quantify overall model performance from a statistical perspective.<sup>181</sup> The distance is  $Y - \hat{Y}$  for continuous outcomes. For binary outcomes,  $\hat{Y}$  is equal to the predicted probability  $p$ , and for survival outcomes it is the predicted time to an event. These distances between observed and predicted outcomes are related to the concept of “goodness-of-fit” of a model, with better models having smaller distances between predicted and observed outcome.

#### 15.1.1 Explained Variation: $R^2$

The amount of explained variation ( $R^2$ ) is an overall measure to quantify the amount of information in a model in a given data set.  $R^2$  is useful to guide various

model development steps for all types of predictive regression models, including linear and generalized linear models (e.g. logistic, Cox). With  $R^2$ , we can readily compare the impact of different encoding of predictors, different shapes of the relationship of continuous predictors to the outcome, different selections of predictors, and the impact of including interaction terms (see previous chapters).

$R^2$  is the most common performance measure for continuous outcomes. For generalized linear models, Nagelkerke's  $R^2$  can well be used.<sup>309</sup> As discussed in Chap. 4, this is a logarithmic scoring rule:  $(Y - 1) - (\log(1 - p)) + Y \times \log(p)$ . The logarithm of predictions  $p$  is compared with the actual outcome  $Y$ . For binary outcomes, the log likelihood for a patient with the outcome is  $\log(p)$ , without the outcome  $\log(1 - p)$ . When a very low prediction is made for a patient who actually had the outcome, this prediction has a severe score (Fig. 15.1). This may be a disadvantage for a prediction model that gives a prediction close to 0 or 1 while the outcome is discordant.



**Fig. 15.1** Logarithmic and quadratic error scores of a subject with ( $y = 1$ ) or without ( $y = 0$ ) the outcome in relation to predicted probability ( $p$ ). The logarithmic score was calculated as  $y \times \log(p) + (1 - y) \times (1 - p)$ , as in Nagelkerke's  $R^2$  (solid line). The quadratic score was calculated as  $(y - p)^2$ , as in the Brier score (dashed line). Lines were scaled such that they crossed at  $p = 50\%$ . We note that the logarithmic score severely penalizes false predictions close to 0 or 100%

### 15.1.2 *Brier Score*

An alternative for binary outcomes is to use a quadratic scoring rule, where the squared differences between actual outcomes  $y$  and predictions  $p$  are calculated. This calculation is done in the Brier score, which is simply defined as  $(Y - p)^2$ . We can also write this similar as the logarithmic score:  $Y \times (1 - p)^2 + (1 - Y) \times p^2$ , with  $Y$  the outcome and  $p$  the prediction for each subject. For a subject, the score can range from 0 (prediction and outcome equal) to 1 (discordant prediction); a prediction of 50% has a score of 0.25 both when the outcome is 0 or 1. The Brier score is less severe than Nagelkerke's  $R^2$  in penalizing false predictions close to 0% or 100% (Fig. 15.1). The Brier score for a model can range from 0% for a perfect model to 0.25 for a non-informative model with a 50% incidence of the outcome. When the incidence is lower, the maximum score for a model is lower, e.g. for 10%,  $0.1 \times (1 - 0.1)^2 + (1 - 0.1) \times 0.1^2 = 0.090$ . A disadvantage of the Brier score is hence that the interpretation depends on the incidence of the outcome.

Similar to Nagelkerke's approach to the LR statistic, we could scale Brier by its maximum score:  $\text{Brier}_{\text{scaled}} = 1 - \text{Brier} / \text{Brier}_{\text{max}}$ , where  $\text{Brier}_{\text{max}} = \text{mean}(p) \times (1 - \text{mean}(p))^2 + (1 - \text{mean}(p)) \times \text{mean}(p)^2$ , with  $\text{mean}(p)$  indicating the average probability of the outcome.  $\text{Brier}_{\text{scaled}}$  ranges between 0% and 100%.

### \*15.1.3 *Example: Performance of Testicular Cancer Prediction Model*

We consider a development sample containing 544 patients contributed by six study groups,<sup>417</sup> and a validation sample 273 patients treated at Indiana University Medical Centre.<sup>466</sup> We developed a logistic regression model with five predictors: teratoma elements in the primary tumor, pre-chemotherapy levels of AFP and HCG, post-chemotherapy mass size, and reduction in mass size.

Internal validation of performance was estimated with bootstrapping (200 replications). Bootstrap samples were created by drawing random samples with replacement from the development sample. The prediction model was fitted in each bootstrap sample and tested on the original sample.

The essential R code is:

```
# 5 predictors in data set n544; develop model
full <- lrm(NEC ~ TER+PREAFP+PREHCG+SQPOST+REDUC10, data=n544)
val.prob(logit=full$linear.predictor, y=full$y) # apparent
validate(full, B=200) #Internal validation with 200 bootstraps
# External validation; refit model for matrix x and
# comparison of coefs
```

**Table 15.1** Overall performance of testicular cancer prediction model

	Development	Internal validation	External validation
$R^2$	38.9%	37.6%	26.7%
Brier	0.174	0.178	0.161
Brier <sub>max</sub>	0.248	0.248	0.201
Brier <sub>scaled</sub>	29.8%	28.2%	20.0%

Development and internal validation with  $n=544$  patients, external validation in  $n=273$  patients. Internal validation with 200 bootstrap resamples using Harrell's validate function.  $\text{Brier}_{\text{scaled}} = 1 - \text{Brier} / \text{Brier}_{\text{max}}$

```
ext.full <- lrm(NEC~TER+PREAFP+PREHCG+SQPOST+REDUC10,
                  data=val, x=T, y=T)
lp <- ext.full$x %%
  full$coef[2:length(full$coef)] + full$coef[1]
val.prob(logit=lp, y=ext.full$y, riskdist="predicted") #external
```

Nagelkerke's  $R^2$  was 38.9% in the development sample, and slightly lower at internal validation (Table 15.1). At external validation, the  $R^2$  was estimated considerably lower, as 26.7%. Note that  $R^2$  is based on the difference between a Null model ("intercept only") and a model with recalibrated predictions (intercept + calibration slope $\times$ logit of predictions).<sup>174</sup> So, the  $R^2$  is estimated after recalibration of the predictions.

The Brier score was 0.174 and 0.178 at development and internal validation respectively. Remarkably, the Brier score was better at external validation (0.161). The external Brier score was simply calculated by comparing predictions with actual outcome, without recalibration as was done for  $R^2$ . The interpretation of the Brier score is easier with the scaled version, which compensates for the fact that the maximum Brier score was lower in the external validation set (necrosis in 76 of 273 (28%); Brier<sub>max</sub>, 0.20) than in the development set (necrosis in 245 of 544 (45%); Brier<sub>max</sub>, 0.25). The scaled Brier score was clearly lower at external validation than at internal validation (20% vs. 28%, Table 15.1).

#### \*15.1.4 Overall Performance Measures in Survival

Nagelkerke's  $R^2$  can readily be calculated for survival outcomes, based on the difference in  $-2 \log$  likelihood of a model without and a model with the linear predictor. Calculation of the Brier score is not directly possible because of censoring: Not all subjects are followed long enough for the outcome to occur. To address the censoring issue, we can define a weight function, which considers the conditional probability of being uncensored during time.<sup>146,375,374</sup> The assumption is that the censoring mechanism is independent of survival and the subject's history.

**Table 15.2** Classification of subjects according to a cutoff for the probability of an outcome (event or no event)

	Event	No event
Predicted probability $\geq$ cutoff	TP	FP
Predicted probability $<$ cutoff	FN	TN
$N_{\text{event}}$		$N_{\text{no event}}$

TP and FP: Numbers of true and false-positive classifications; FN and TN: Numbers of false and true-negative classifications, respectively.  $N_{\text{event}} = \text{TP} + \text{FN}$ ;  $N_{\text{no event}} = \text{FP} + \text{TN}$

We can hence calculate the Brier score at fixed time points. For example, we can compare predicted survival vs. observed survival at 1, 2, and 5 years of follow-up. Choosing many consecutive time-points leads to a time-dependent graph. This is useful to use a benchmark curve, based on the Brier score for the overall Kaplan-Meier estimator, which does not consider any predictive information. The survival estimates of the overall Kaplan-Meier curve only depend on time of follow-up, and are identical for all subjects alive at a certain point in time. An interesting example is provided by a case study on the disappointing contribution of microarray data to prediction of survival for patients with diffuse large-B-cell lymphoma.<sup>374</sup>

### \*15.1.5 Decomposition in Discrimination and Calibration

Overall statistical performance measures incorporate both calibration and discrimination aspects. For example, the Brier score can formally be decomposed into indicators of calibration and discrimination.<sup>303,38</sup> Discrimination relates to how well a prediction model can discriminate those with the outcome from those without the outcome. Calibration relates to the agreement between observed outcomes and predictions. Studying discriminative ability and calibration is often more meaningful than an overall measure such as  $R^2$  or Brier score when we want to appreciate the quality of model predictions for individuals. We therefore discuss these aspects further.

### 15.1.6 Summary Points

- $R^2$  is a common measure to express the amount of variability in outcomes that is explained by the prediction model
- The Brier score is another common performance measure for the distance between observed and predicted outcome, which can be decomposed in discrimination and calibration aspects

## 15.2 Discriminative Ability

Model predictions need to discriminate between those with and those without the outcome (Event vs. No event). Several measures can be used to indicate how good we classify patients in a binary prediction problem. The concordance ( $c$ ) statistic is the most commonly used performance measure to indicate the discriminative ability of generalized linear regression models. For a binary outcome  $c$  is identical to the area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of the sensitivity (true positive rate) against 1 – specificity (false-positive rate) for consecutive cutoffs for the probability of an outcome. We therefore consider sensitivity and specificity first.

### 15.2.1 Sensitivity and Specificity of Prediction Models

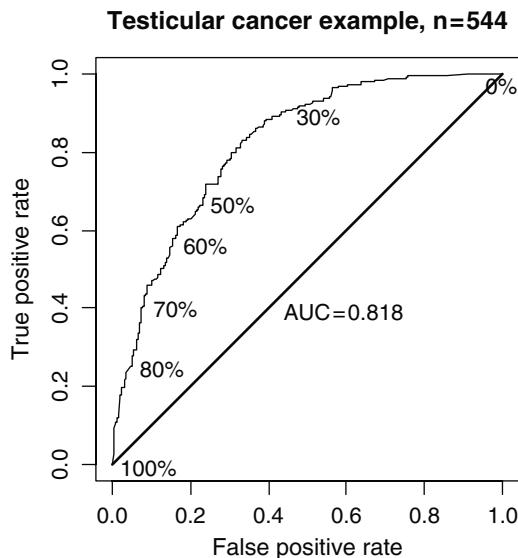
Sensitivity is defined as the fraction of true-positive (TP) classifications among the total number of patients with the outcome ( $TP/N_{\text{event}}$ ), and the specificity as the fraction of true-negative classifications among the total number of patients without the outcome ( $TN/N_{\text{no event}}$ , Table 15.2). To classify a patient as positive or negative, we need to apply a cutoff to the predicted probability. If the prediction is higher than the cutoff, the patient is classified as positive, otherwise as negative. It is common to use a cutoff of 50% for classification. This cutoff is often not defendable in a medical context, as we will discuss in detail in the next chapter (Chap. 16). We can examine sensitivity and specificity over the whole range of cutoffs from 0% to 100%. The results can be plotted in an ROC curve.<sup>172</sup>

### 15.2.2 Example: Sensitivity and Specificity of Testicular Cancer Prediction Model

If we classify patients as having necrosis when the probability of necrosis is over 50%, we have a sensitivity of 68% and a specificity of 77% (FP rate, 23%). With a higher cut-off, for example 70%, these numbers are 42% and 92%, respectively. This illustrates that a higher cutoff leads to better specificity, at the price of a lower sensitivity. This trade-off is visualized in an ROC curve (Fig. 15.2).

### 15.2.3 ROC Curve

A plot of an ROC curve has often been used in diagnostic research to quantify the diagnostic value of a test over its whole range of possible cutoffs for classifying patients as positive vs. negative. We can also make an ROC curve with consecutive cutoffs for the predicted probability of a binary outcome. We start with a cutoff of

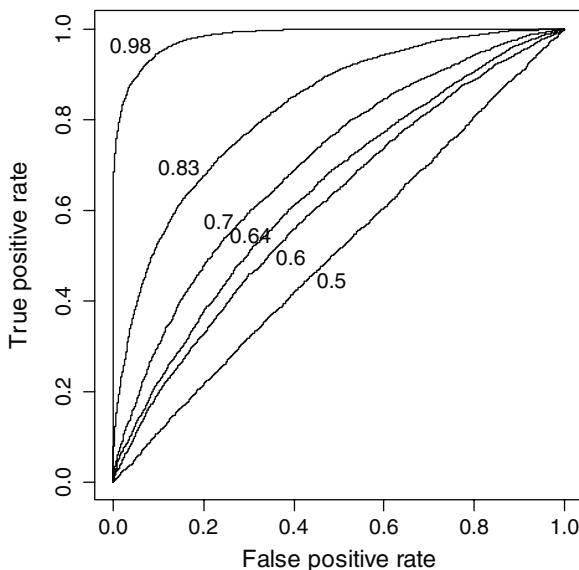


**Fig. 15.2** Receiver operating characteristic (ROC) curve for the testicular cancer model in the development data set of 544 patients. Using cutoffs for the predicted probability of necrosis (benign tissue) results in specific combinations of true-positive rate (sensitivity) and false-positive rate ( $1 - \text{specificity}$ ). The area under the curve is 0.818

0%, which implies that all subjects are classified as positive. The sensitivity is 100%, and the specificity 0% (upper-right point in Fig. 15.2). There are no false-negative classifications, and 100% false-positive classifications, since all subjects without the outcome are classified as positive. We then shift to a slightly higher cutoff, e.g. 1%, where sensitivity may still be 100%, but specificity above 0%. We follow all possible cutoffs till 100%, where all subjects are classified as negative. This is the lower-left point in Fig. 15.2. The sensitivity is then 0%, and specificity 100%. The curves are more to the upper left corner when the distributions of predictions are more separate between those with and without the outcome (Fig. 15.3).

We can draw a line between the 0%, 0% and 100%, 100% points, indicating a non-informative model. Note that the sum of TP and TN is 1 at every cutoff for such a model. This sum (also known as Youden's index) is larger than 1 for sensible prediction models.

The area under the curve can be interpreted as the probability that a patient with the outcome is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome.<sup>172</sup> An uninformative model, such as a coin flip, will hence have an area of 0.5. A perfect model has an area of 1. The interpretation hence is relatively straightforward, but assumes that we have a pair of patients, one with and one without the outcome. This is a rather artificial situation. Statistically, this conditioning on a pair of patients is attractive, since it makes the area independent of the incidence of the outcome, in contrast to  $R^2$  or the Brier score for example.



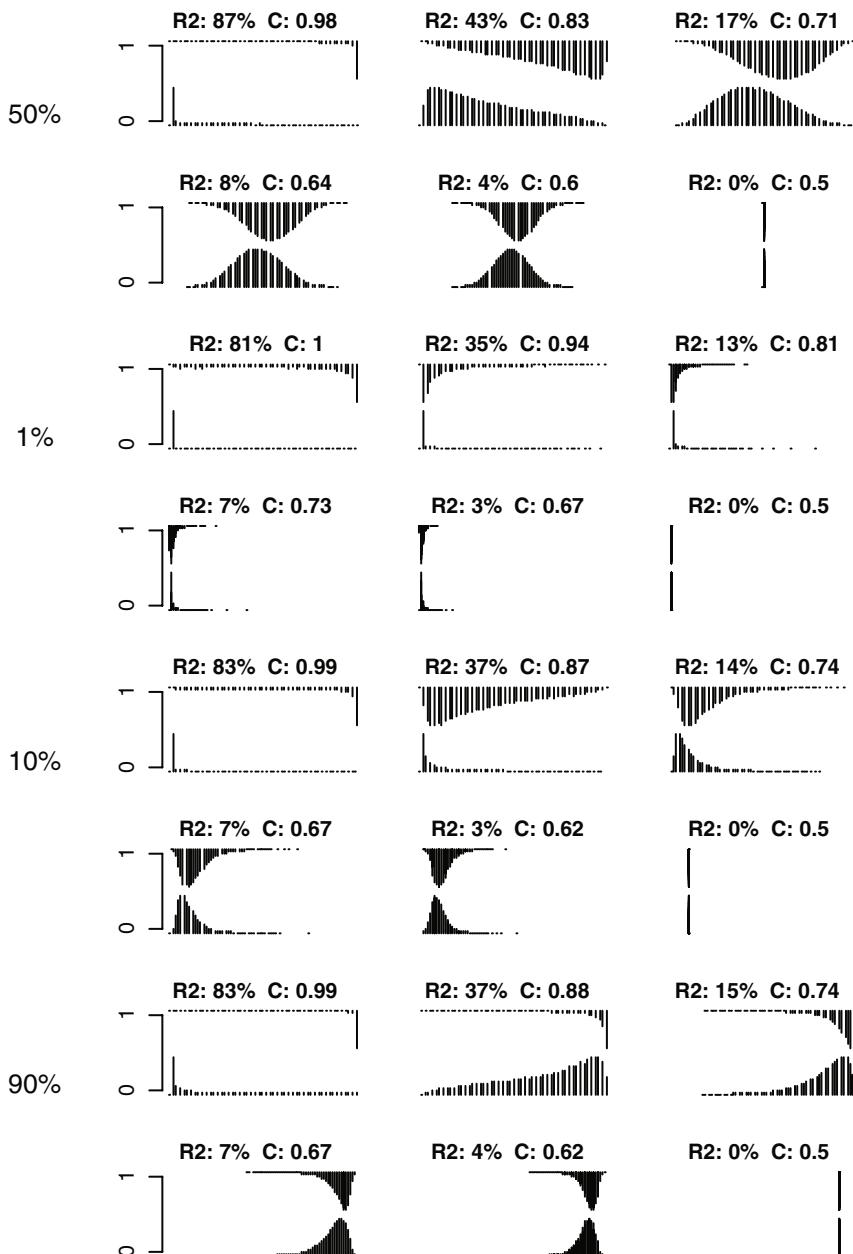
**Fig. 15.3** ROC plot for five hypothetical prediction models. Models were created with distributions as shown in Fig. 15.4 (see also Fig. 4.6). The  $c$  statistics were 0.5, 0.6, 0.64, 0.7, 0.83, and 0.98 at 50% incidence of the outcome

A generalization of the area under the ROC curve is provided by the concordance statistic ( $c$ ).<sup>175</sup> The  $c$  statistic is a rank order statistic for predictions against true outcomes, related to Somer's D statistic. As a rank order statistic, it is insensitive to errors in calibration such as differences in average outcome. For binary outcomes,  $c$  is identical to the area under the ROC curve.

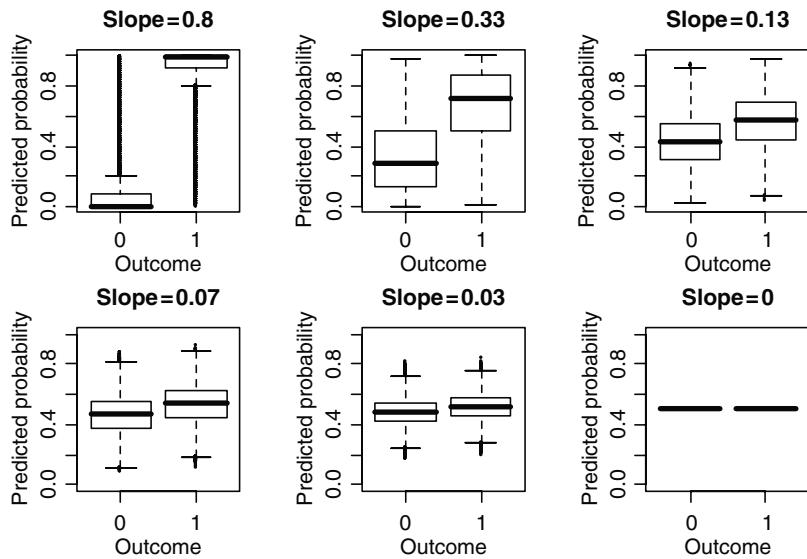
Confidence intervals for the area under ROC curve (or  $c$  statistic) can be calculated with various methods. Standard asymptotic methods may be problematic, especially when sensitivity or specificity are close to 0% or 100%.<sup>9</sup> Bootstrap resampling is a good choice for many situations. For example, differences in  $c$  between models fitted on the same data can be tested with standard formulas for the difference. But such formulas are only valid if the models were pre-specified. If one or both models were estimated on the same data, bootstrapping can be used for comparison of optimism-corrected estimates (see Chap. 17).

#### 15.2.4 $R^2$ vs. $c$

We compare the behavior of Nagelkerke's  $R^2$  and the  $c$  statistic in some simulations over a range of incidences of the outcome (1%, 10%, 50%, 90%, Fig. 15.4). At 50% incidence, a high  $c$  statistic such as 0.98 is associated with an  $R^2$  value of 87%. With lower incidence,  $R^2$  is somewhat lower.



**Fig. 15.4** Distribution of observed outcomes (0 or 1), in relation to predicted probabilities from hypothetical logistic models relating  $Y$  to a predictor  $X$ . The top six graphs relate to an incidence of 50%. The next sets of  $3 \times 6$  graphs relate to incidences of 1%, 10%, and 90% respectively. For each hypothetical model, Nagelkerke's  $R^2$  and  $c$  statistic are listed. If  $c=0.5$  (and  $R^2=0\%$ ), predictions are at the incidence of the outcome for all subjects, with or without the outcome, indicated with a single spike. If  $c$  is close to 1 ( $R^2$  close to 100%), predictions are close to 0% for those without the outcome, and close to 100% for those with the outcome. Note that  $R^2$  and  $c$  statistics differ somewhat between 10% and 90% incidence, because of random noise in the simulation procedure



**Fig. 15.5** Box plots for predictions from six hypothetical prediction models with different discriminative ability (see Fig 15.4). The discrimination slopes are calculated as the difference in means of predictions for those with and those without the outcome (mean incidence, 50%)

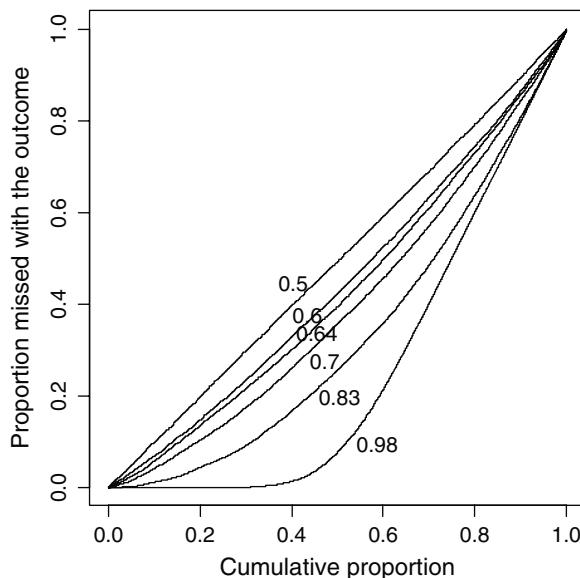
### \*15.2.5 Box Plots and Discrimination Slope

The discrimination slope has been proposed as a simple measure for how well subjects with and without the outcome are separated. It is easily calculated as the absolute difference in average predictions for those with and without the outcome.

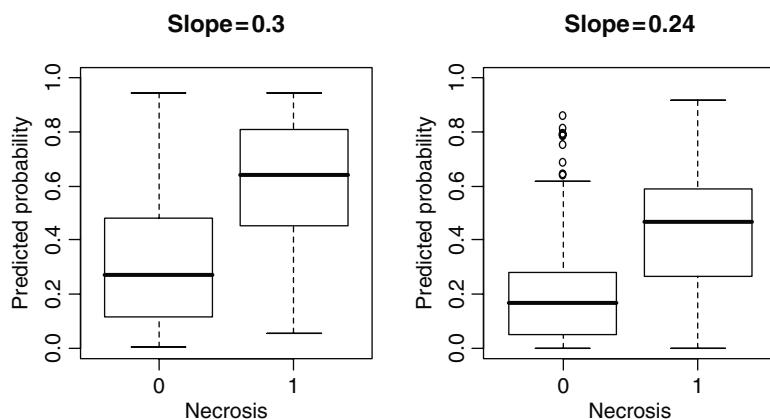
Visualization is readily possible with a box plot (Figs. 15.5 and 15.7). The box plot may be a simple and intuitive way to communicate the extent of risk differentiation achieved by the model. The same information can be shown by histograms, which will show less overlap between those with and those without the outcome for a better discriminating model (Fig. 15.4). Similar to Fig. 15.4, the incidence of the outcome determines the visual expression that a box plot makes, and the magnitude of the discrimination slope. With low incidence, the slope is somewhat lower, for the same  $c$  statistic.

### \*15.2.6 Lorenz Curve

An alternative way to judge discriminative ability is the Lorenz curve (Fig. 15.6). The Lorenz curve has been used in economics to characterize the distribution of wealth in a population.<sup>267</sup> This curve has been used to plot the cumulative distribution of wealth against the cumulative distribution of the population, ranked on the basis of individual wealth.



**Fig. 15.6** Lorenz curve showing proportion missed with the outcome vs. the cumulative proportion of patients according to rank order of predictions, for an outcome incidence of 50%. We note that a near perfect model ( $c=0.98$ ) follows a horizontal line and then rises steeply to 100% false-negative rate from the point of 50% cumulative proportion.



**Fig. 15.7** Box plot showing predictions by actual outcome (necrosis) for testicular cancer patients ( $n=544$  and 273, respectively)

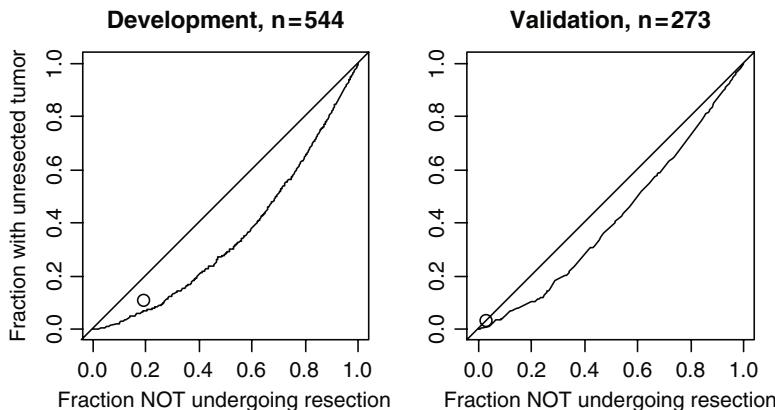
**Table 15.3** Summary of some measures for discriminative ability of a prediction model for binary outcomes

Measure	Calculation	Visualization	Pros	Cons
Concordance statistic	Rank order statistic	ROC curve	Insensitive to outcome incidence; interpretable for pairs of patients with and without the outcome	Interpretation artificial
Discrimination slope	Difference in mean of predictions between outcomes	Box plot	Easy interpretation, nice visualization	Depends on the incidence of the outcome
Lorenz curve	Shows concentration of outcomes missed by cumulative proportion of negative classifications	Concentration curve	Shows balance between finding true positive subjects vs. total classified as positive	Depends on the incidence of the outcome

For prediction models we can plot the cumulative proportion of the population on the  $x$  axis, ranked by predicted probability. On the  $y$  axis, we plot the cumulative proportion of subjects with the outcome. For example, we can show the proportion of subjects developing cancer against the cumulative proportion of the population ranked by cancer risk.<sup>31</sup> In terms of ROC curves, we plot the cumulative rate of false-negative classifications against the total of negative predictions. With incidences of the outcome around 50%, the ROC and Lorenz curves look very similar, except that the Lorenz curve is flipped vertically and horizontally. In case of a non-informative model, a straight line arises, since every rate of the population classified as negative corresponds to the same rate classified as negative among those with the outcome. A good model has a curve under this straight line, with a relatively large proportion of the population classified as negative having only a small part of the outcomes (low false-negative rate). On the upper end of the  $x$  axis, a small part of the population should contain many subjects with the outcome. In the ideal case, a cutoff is used that classifies the fraction as positive, equal to the prevalence, and all these have the outcome. Indeed, we note that a  $c$  statistic of 0.98 leads to a nearly horizontal line till the 50% cumulative proportion point on the  $x$  axis, and increases more or less linearly to 100% after that.

The Gini index is often calculated as a summary measure for the Lorenz curve. The Gini index is the ratio between the area ( $A$ ) between the Lorenz curve of the prediction model and the line for a non-informative model and the area under the line for an non-informative model (0.5). Hence,  $G = 2A$ .

Other summaries are related to quantiles of the cumulative distribution. For example, we can consider the number of missed outcomes when 25% of the population is classified as negative. If we want to be sure not to miss the outcome, usually only



**Fig. 15.8** Lorenz curves for prediction of necrosis vs. residual tumor. Patients classified as necrosis would not undergo surgical resection (x axis). With increasing fractions not undergoing resection, the fraction with unresected tumor increases (“missed tumor”). With 75% undergoing resection, 56% of the tumors are resected, leaving 44% unresected

few can be classified as negative, unless a model is used with very good discriminative ability. At the upper end of the range, we can consider how many outcomes are concentrated in the upper quartile (above 75 percentile). We will illustrate these percentiles for the testicular cancer prediction case study (Fig. 15.8).

An advantage of the Lorenz concentration curve is that the trade-off is clearly visualized between how many subjects can be classified as negative without missing many with the outcome. A disadvantage is that the appearance of the Lorenz curve depends strongly on the incidence of the outcome; with low incidence, the graph looks impressive, and with high incidence, the graph looks rather poor. As an example, consider a screening setting with 1% of subjects having the disease of interest. Only few cases with disease are missed at 25% classified negative when we use a model with a  $c$  statistic of 0.83. The top 25% then easily contains most cases. With a more frequent outcome, more cases are missed at the point of 25% classified negative, and fewer of the cases are in the top 75 percentile.

### 15.2.7 Discrimination in Survival Data

For survival data, Harrell’s overall  $c$  statistic indicates the proportion of all pairs of subjects who can be ordered such that the subject with the higher predicted survival is the one who survived longer.<sup>175</sup> Ordering is possible if both subjects have an observed survival time, or when one has the outcome and a shorter survival time than the censored survival time of the other subject. Ordering is not possible if both

subjects are censored, or if one has the outcome with a survival time longer than the censored survival time of the other subject. Some alternative definitions of  $c$  have been proposed, which lead to time-dependent performance curves.<sup>183</sup>

In oncology, prognostic groups are often created after constructing a prognostic model. A common procedure is to base these groups on quartiles of predicted survival; the lower 25% should have the worst survival and the highest 25% the best survival. This approach can well illustrate the discriminative ability of a model. An example is shown in Chap. 23 (Fig. 23.8).

### 15.2.8 Example: Discrimination of Testicular Cancer Prediction Model

We continue the example of predicting a benign histology in testicular cancer patients after chemotherapy. The  $c$  statistic was 0.818 at model development, with small optimism according to bootstrap validation (decrease by 0.006 to 0.812). At external validation, the  $c$  statistic was 0.785, with a relatively wide 95% confidence interval of 0.73 to 0.84 (Table 15.4).

The discrimination slope was 0.30 at model development, with small optimism according to bootstrap validation (decrease to 0.29). At external validation, the slope was much smaller (0.24). Part of this decrease is attributable to the lower average prevalence of necrosis (76 of 273, 28%, vs. 245 of 544, 45%). This lower prevalence is also evident from the box plots (Fig. 15.7).

The Lorenz curves were created with  $x$  axis as the cumulative fraction classified as necrosis, i.e. not having tumor, and hence classified as not undergoing surgical resection (Fig. 15.8). The  $y$  axis was the fraction of missed tumors, i.e. tumor masses left unresected. The point of 25% classified as necrosis corresponds to using a cutoff of 68% for the probability of necrosis; only patients with

**Table 15.4** Discriminative ability of testicular cancer prediction model

	Development (n=544, 245 necrosis)	Internal validation	External validation (n=273, 76 necrosis)
$c$ statistic	0.818	0.812	0.785
[95% CI]	[0.783–0.852]	[0.777–0.847] <sup>a</sup>	[0.726–0.844]
Discrimination slope	0.301	0.294	0.237
[95% CI]	[0.235–0.367] <sup>b</sup>	[0.228–0.360] <sup>a</sup>	[0.178–0.296] <sup>b</sup>
Lorenz curve p25, tumors missed	9%	–	13%
Lorenz curve p75, tumors missed	58%	–	65%

Development and internal validation with  $n=544$  patients, external validation in  $n=273$  patients.

Internal validation with 200 bootstrap resamples using Harrell's validate function

<sup>a</sup>Assuming the same SE applies as estimated for model development

<sup>b</sup>Based on bootstrap resampling

a probability over 68% are not resected. We miss 9% of the tumors with that cut-off. Hence, sparing surgery in 25% leads to missing 9% of the tumors. The point of 75% classified as necrosis corresponds to using a low cutoff (21%), and missing 58% of the tumors. Hence 42% of the tumors are concentrated in the upper quartile of the distribution.

At external validation, the curve looks worse, which is related to a lower discriminative ability and to a lower average prevalence of necrosis (28% vs. 45%). The 25% and 75% cumulative fractions correspond to cutoffs of 40% and 8% for the probability of necrosis, and lead to 13% and 65% missed tumors, respectively.

As a reference, we consider the current widely used policy of resection if the residual mass size exceeds 10 mm.<sup>418</sup> This policy uses only one of the five predictors in the model (post-chemotherapy mass size), and hence has less discriminative ability (the point is closer to the 45° line in Fig. 15.8). In the development sample, 107 of the 544 patients (20%) had residual masses  $\leq 10$  mm, but among them 30 with tumor (fraction tumor missed, 30 of 299, 10%). In the validation sample, only 9 of the 273 patients (3.3%) had residual masses  $\leq 10$  mm, but among them, 6 with tumor (fraction tumor missed, 6 of 197, 3%). Hence, the reference policy did not perform well in the validation sample.

### \*15.2.9 Verification Bias and Discriminative Ability

In the testicular cancer validation sample, only nine patients had very small residual masses. This reflects the policy for resection in the specific centre, where patients with such very small masses were not considered candidates for resection.<sup>466</sup> This leads to verification bias; we do not know the histology of these masses, since they were not resected, and cannot evaluate predictions for these patients. We know that the estimation of regression coefficients is not biased by this selection, if we include the selection criterion (residual mass size) in the prediction model. Hence model predictions are valid even with verification bias.<sup>497</sup> But performance measures such as sensitivity and specificity suffer from this verification bias.<sup>30</sup> The  $c$  statistic may not be affected too much because verification bias makes that we merely shift on the ROC curve to a different combination of sensitivity and specificity.

### \*15.2.10 R Code

The boxplot is created simply with the `boxplot` command, based on a “full model,” including five predictors in the development data:

```
lp <- full$linear.predictors
boxplot(plogis(lp ~ full$y)) # Fig 15.7
```

The discrimination slope is the difference between the mean predicted probabilities by outcome:

```
mean(plogis(lp[full$y==1])) - mean(plogis(lp[full$y==0]))
```

Lorenz curves are created with the ROCR package:

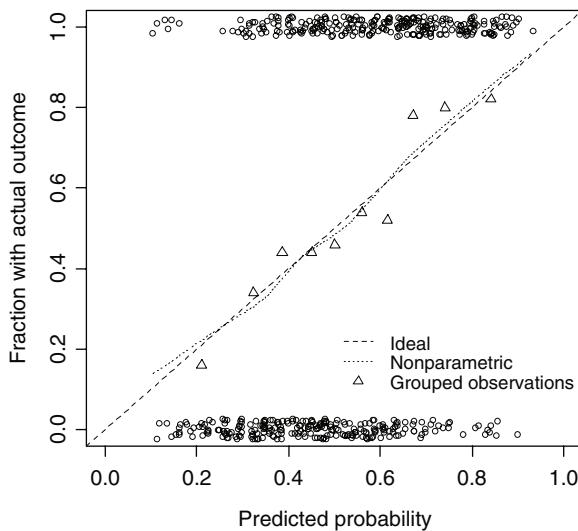
```
library(ROCR)
# Make ROC object with predicted probability for outcome
pred.full <- prediction(plogis(lp), full$y)
# Lorenz curve data and plot
perf1      <- performance(pred.full, "fpr", "rpp")
plot(perf1, xlab="NOT undergoing resection",
      ylab="with unresected tumor")
abline(a=0, b=1)      # Fig 15.8
```

## 15.3 Calibration

Another important property of a prediction model is calibration, i.e. the agreement between observed outcomes and predictions. For example, if we predict 70% probability of benign tissue for a testicular cancer patient, the observed frequency of benign tissue should be 70 out of 100 such patients.

### 15.3.1 Calibration Plot

A calibration plot has predictions on the  $x$  axis, and the outcome on the  $y$  axis. A line of identity helps for orientation: Perfect predictions should be on the  $45^\circ$  line. For linear regression, the calibration plot results in a simple scatter plot. For binary outcomes, the plot contains only 0 and 1 values for the  $y$  axis. Probabilities are not observed directly. However, smoothing techniques can be used to estimate the observed probabilities of the outcome ( $p(y=1)$ ) in relation to the predicted probabilities. The observed 0/1 outcomes are replaced by values between 0 and 1 by combining outcome values of subjects with similar predicted probabilities, e.g. using the loess algorithm.<sup>174</sup> We can also plot results for subjects grouped by similar probabilities (quantiles), and thus compare the mean predicted probability to the mean observed outcome. For example, we can plot observed outcome by decile of predictions (Fig. 15.9). This makes the plot a graphical illustration of the Hosmer-Lemeshow goodness-of-fit test (see Sect. 15.3.8 and 15.3.10). A better discriminating model has more spread between such deciles than a poorly discriminating model. The choice of quantiles is important for the visual impression of calibration; if small groups are plotted, the variability will be large.



**Fig. 15.9** Calibration plot of actual outcome vs. predictions for a hypothetical model with  $c$  statistic 0.7,  $n=500$ . The distributions of actual 0 and 1 values are shown at the *bottom* and at the *top* of the graph; the loess smoother is close to the ideal 45° line; actual outcomes by deciles of risk are shown by triangles (each triangle,  $n=50$ )

### 15.3.2 Calibration in Survival

In a survival context, the calibration of a model is usually studied at fixed time points. For these time points, we can consider grouped patients, with sufficient numbers per group to allow for calculation of survival rates with the Kaplan-Meier method. This observed survival is compared with the mean predicted survival from the prognostic model. Harrell suggests to use at least 50 subjects per group, depending on the hazard of the outcome.<sup>174</sup> It would be interesting to plot a smoothed curve as for binary outcomes, but this is not easy.

### 15.3.3 Calibration-in-the-Large

A calibration plot can easily be made for the data set used to develop a model. This indicates the apparent calibration. In model development, the average of predictions is the average of the outcomes:  $\text{mean}(Y) = \text{mean}(\hat{Y})$ . For example,  $\text{mean}(\text{observed BP}) = \text{mean}(\text{predicted BP})$  in linear regression, and  $\text{mean}(\text{observed 30-day mortality}) = \text{mean}(\text{predicted 30-day mortality})$ . This correspondence is guaranteed by the intercept in a (generalized) linear model. This correspondence of average outcomes remains at internal validation with bootstrapping. When we apply the model to external data, this correspondence may be less. The difference between  $\text{mean}(\hat{Y})$  and  $\text{mean}(Y_{\text{new}})$  is referred to as “calibration-in-the-large.”

### 15.3.4 Calibration Slope

Another important calibration measure is related to the average strength of the predictor effects. For linear regression, we can write  $Y_{\text{new}} = a + b_{\text{overall}} \hat{Y}$ , and for generalized linear models  $f(Y_{\text{new}}) = a + b_{\text{overall}}$  linear predictor, where the linear predictor is the combination of regression coefficients from the model and the predictor values in the new data. A link function  $f$  is used for  $Y_{\text{new}}$ , e.g. logodds (or logit) in logistic regression. The  $b_{\text{overall}}$  is named the calibration slope.<sup>86</sup> Ideally, the calibration slope  $b_{\text{overall}} = 1$ . With apparent validation,  $b_{\text{overall}} = 1$  because this yields the best fit on the data under study with either least squares or maximum likelihood methods. At internal validation, the calibration slope reflects the amount of shrinkage that is required for a model ( $b_{\text{overall}} < 1$ ).<sup>81</sup> It indicates how much we need to reduce the effects of predictors on average to make the model well calibrated for new patients from the underlying population. The calibration slope can hence be used as a shrinkage factor to adjust a model for future use (Chap. 14). At external validation, the calibration slope reflects the combined effect of two issues: overfitting on the development data and true differences in effects of predictors.

### 15.3.5 Estimation of Calibration-in-the-Large and Calibration Slope

For continuous outcomes, calibration-in-the-large can be assessed easily by comparing the mean ( $\hat{Y}$ ) and mean( $Y_{\text{new}}$ ), and testing the differences  $Y_{\text{new}} - \hat{Y}$ , e.g. with a one-sample  $t$ -test. This test indicates the statistical significance of the mean under- or overestimation of the observed outcome: mean( $Y_{\text{new}} - \hat{Y}$ ). In a linear regression model, we can estimate an intercept  $a$  in the model with as outcome the residual  $Y_{\text{new}} - \hat{Y}$ :  $Y_{\text{new}} - \hat{Y} = a$ . The recalibration model is simply  $Y_{\text{new}} = a + b_{\text{overall}} \hat{Y}$ . The deviation of the calibration slope from 1 can be tested in linear regression by a model that studies the residuals:  $Y_{\text{new}} - \hat{Y} = a + b_{\text{overall}} \hat{Y}$ . The significance of  $b_{\text{overall}}$  is then determined as usual in regression, and indicates on average stronger or weaker effects of the predictors in a model.

For binary outcomes, calibration-in-the-large again refers to the difference between mean  $\hat{Y}$  and mean( $Y_{\text{new}}$ ). A simple comparison can directly be made, with an odds ratio indicating the average under- or overestimation of the outcome:

$$\text{OR} = \text{odds}(\text{mean}(\hat{Y})) / \text{odds}(\text{mean}(Y_{\text{new}})) = \\ [\text{mean}(\hat{Y}) / (1 - \text{mean}(\hat{Y}))] / [\text{mean}(Y_{\text{new}}) / (1 - \text{mean}(Y_{\text{new}}))].$$

For statistical testing of the difference we need to be more careful. In logistic regression, the relationship between the outcome  $y$  and the linear predictor is non-linear (i.e. logistic). We have to compare  $\text{logit}(Y_{\text{new}} = 1)$  to  $\text{logit}(\hat{Y})$ , where  $\text{mean}(\text{logit}(Y_{\text{new}} = 1)) - \text{logit}(\hat{Y})$  is not equal to  $\text{mean}(\text{logit}(Y_{\text{new}} = 1)) - \text{mean}(\text{logit}(\hat{Y}))$ .

In a model, we could write  
 $\text{logit}(Y_{\text{new}} = 1) - \text{logit}(\hat{Y}) = a;$

$$\text{or } \text{logit}(Y_{\text{new}} = 1) = a + \text{logit}(\hat{Y}) = a + \text{offset (linear predictor)}.$$

The intercept  $a$  then reflects the difference in logodds between predictions and observed outcome, adjusted for the linear predictor. The offset makes that predictions are taken literally, as in linear regression. Values of the offset variable are subtracted from the actual outcomes  $Y_{\text{new}}$  (as in Poisson regression). Equivalently we can think of a regression coefficient for the offset variable that is fixed at unity. The statistical significance of intercept  $a$  can be tested with standard regression tests, such as the Wald test or the likelihood ratio (LR) test.

The calibration slope can be estimated from the recalibration model

$$\text{logit}(Y_{\text{new}} = 1) = a + b_{\text{overall}} \times \text{logit}(\hat{Y}) = a + b_{\text{overall}} \times \text{linear predictor}.$$

The deviation of the calibration slope from 1 (“miscalibration”) can be tested by a model that includes an offset variable:

$$\text{logit}(Y_{\text{new}} = 1) = a + b_{\text{miscalibration}} \times \text{linear predictor} + \text{offset (linear predictor)}.$$

The slope coefficient  $b_{\text{miscalibration}}$  reflects the deviations from the ideal slope of 1, and can be tested with Wald or LR statistics.

Calibration-in-the-large cannot be detected with a refitted Cox regression model, since the baseline hazard  $h_0$  is usually left free in fitting such a model. For a survival outcome, the calibration slope can be assessed as:

$$\text{log(hazard}(y_{\text{new}} = 1)) = h_0 + b_{\text{overall}} \times \text{linear predictor}.$$

The model for deviation from a slope of 1 is:

$$\text{log(hazard}(y_{\text{new}} = 1)) = h_0 + b_{\text{miscalibration}} \times \text{linear predictor} + \text{offset (linear predictor)}.$$

Testing of coefficient  $b_{\text{miscalibration}}$  is as usual, i.e. with a Wald test or LR test.

With a parametric survival model, we can specify parameters that reflect differences in average survival, after adjustment for predictor effects. Van Houwelingen hereto transformed the baseline hazard from a Cox model to a Weibull model.<sup>456</sup> The Weibull model has two parameters to describe the baseline hazard parametrically (Chap. 4). These two parameters can be refitted for external validation data, together with the linear predictor, to estimate a recalibrated model.

### \*15.3.6 Other Calibration Measures

Various other measures are available for calibration. An intuitively appealing measure of calibration is the absolute difference between smoothed observed outcomes

**Table 15.5** Calibration tests for prediction model  $y \sim a + b_{\text{overall}} \hat{Y}$ 

	$H_0$	$H_1$	$df$
Calibration-in-the-large	$a=0 \mid b_{\text{overall}} = 1$	$a <> 0 \mid b_{\text{overall}} = 1$	1
Calibration slope	$b_{\text{overall}} = 1$	$b_{\text{overall}} <> 1$	1
Recalibration	$a = 0$ and $b_{\text{overall}} = 1$	$a <> 0$ or $b_{\text{overall}} <> 1$	2

$H_0$  and  $H_1$  indicate the Null and alternative hypothesis respectively

and predicted probabilities (Harrell's  $E$  statistic).<sup>174</sup> This measure is related to the calibration plot, and depends on the way the 0/1 outcomes are smoothed. The difference between smoothed observed outcomes and predicted probabilities can also be judged visually in a calibration plot such as Fig. 15.9.

### 15.3.7 Calibration Tests

Statistical tests can be performed with various null hypotheses for calibration, phrased in the formulation of the recalibration model  $y \sim a + b_{\text{overall}} \hat{Y}$  (Table 15.5). Tests for calibration-in-the-large and calibration slope have one  $df$ ; the calibration test has two  $df$ . The test for calibration-in-the-large requires that the predictions are taken literally ( $b_{\text{overall}} = 1$ ). In generalized linear models, this can be achieved with an offset variable. The calibration slope can easily be estimated in the recalibration model. The recalibration test has several advantages (Table 15.6). It can pick-up common patterns of miscalibration, i.e. systematic differences between the new data and the model development data, and overfitting of the effects of predictors. Moreover the test parameters  $a$  and  $b_{\text{overall}}$  are well interpretable, provided that  $a \mid b_{\text{overall}} = 1$  is reported (rather than  $a$  with  $b_{\text{overall}}$  left free). The slope  $b_{\text{overall}}$  can directly be taken from the re-calibration model (where  $a$  is left free).

Statistical testing for calibration has a number of drawbacks. First, the null hypothesis is of good calibration. Hence, if we test calibration in a small study, we have low power and will not reject the null hypothesis unless miscalibration is very severe. On the other hand, even a model with very good, but not perfect, calibration will fail if the sample size is sufficiently large.

### 15.3.8 Goodness-of-Fit Tests

Calibration is related to goodness-of-fit, which relates to the ability of a model to fit a given set of data. Typically, there is no single goodness-of-fit test that has good power against all kinds of lack of fit of a prediction model. Examples of lack of fit are missed non-linearities, interactions, or an inappropriate link function between the linear predictor and the outcome. Goodness-of-fit can be tested with a  $\chi^2$  statistic.

For binary outcomes, the Hosmer-Lemeshow (H-L) goodness-of-fit test is often used.<sup>199</sup> Usually, patients are grouped by decile of predicted probability. The sum

**Table 15.6** Summary of some measures for calibration of a prediction model for binary outcomes

Performance aspect	Calculation	Visualization	Pros	Cons
Calibration-in-the-large	Compare mean( $y$ ) vs. mean( $\hat{y}$ )	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration slope	Regression slope of linear predictor	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration test	Joint test of calibration-in-the-large and calibration slope	Calibration graph	Efficient test of two key issues in calibration	Insensitive to more subtle miscalibration
Harrell's E Statistic	Absolute difference between smoothed $y$ vs. line of identity	Calibration graph	Conceptually easy, summarizes miscalibration over whole curve	Depends on smoothing algorithm
Hosmer-Lemeshow test	Compare observed vs. predicted in grouped patients	Calibration graph or table	Conceptually easy	Interpretation difficult; low power in small samples
Goeman-Le Cessie test	Consider correlation between residuals	-	Overall statistical test; supplementary to calibration graph	Very general
Subgroup calibration	Compare observed vs. predicted in subgroups	Table	Conceptually easy	Not sensitive to various miscalibration patterns

of predicted probabilities is the number of expected outcomes; this expected number is compared with the observed number in the ten groups with a  $\chi^2$  test. In model development, this  $\chi^2$  test has eight degrees of freedom; at external validation the degrees of freedom is 9. There are many drawbacks to the H-L test.<sup>198,174</sup> First, there are some technical issues: Should we always use deciles of predictions, or make the quantiles dependent on the sample size? Can we group by risk-interval, e.g. 0–10%, 11–20%, etc (“interval grouping”)? Second, the test has poor power to detect miscalibration in the common form of systematic differences between outcomes in the new data and the model development data, or to detect overfitting of the effects of predictors. Some proposed that the H-L test should only be used in model development, in addition to more specific tests on model assumptions, such as tests for linearity (adding non-linear transformations) and additivity (adding interaction terms). Reported H-L tests are usually non-significant if they reflect apparent validation on the data that were also used to construct the model. Such non-significant results may contribute to the face validity of a model as perceived by some readers, but have no scientific meaning.

Alternative goodness-of-fit tests have been proposed with better statistical properties, such as the Goeman-Le Cessie goodness-of-fit test.<sup>250,141</sup> It assesses the alternative hypothesis that any nonlinearities or interaction effects have been missed in a logistic regression model. Such neglected effects can be detected by looking for patterns in the residuals: Observations close to each other in covariate space, which deviate from the model in the same direction. The approach is to smooth the regression residuals and to test whether these smoothed residuals have more variance than expected under the null hypothesis, which occurs when residuals that are close together in the covariate space are correlated. The test statistic is a sum of squared smoothed residuals.

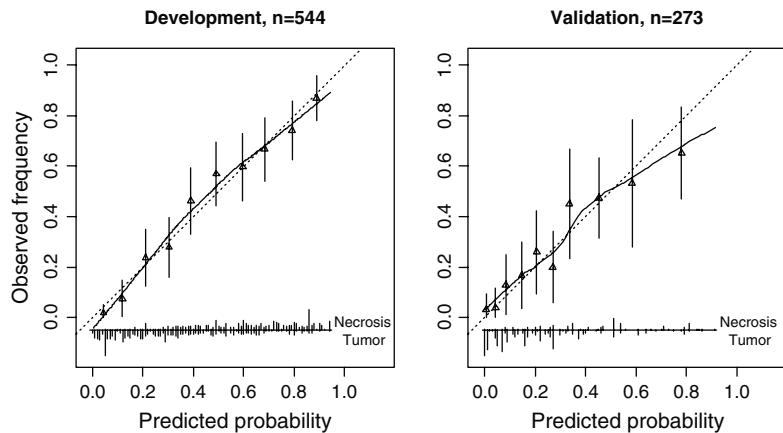
Another approach to goodness-of-fit is to study observed vs. expected outcomes in subgroups of patients. For example, we can assess the difference between observed vs. expected outcomes in males and females, or other subgroups of patients. If the effect of the subgroup is not well modelled, e.g. an interaction was missed, this might be reflected in this assessment. There are however more direct ways of assessing the influence of subgroup characteristics, as was discussed in Chap. 13 on model specification. So, this check for calibration is also more for face validity of the model and for convincing potential users than a serious check of calibration. Measures for assessment of calibration are compared in Table 15.6.

### ***15.3.9 Calibration of Survival Predictions***

For survival outcomes, formal tests similar to the H-L test are possible by comparison of observed K-M percentages with average predictions across groups of patients. Furthermore, we can study the distribution of Cox-Snell residuals, in a plot of the cumulative hazard vs. the residuals, which should form a straight line.<sup>174</sup>

### ***\*15.3.10 Example: Calibration in Testicular Cancer Prediction Model***

For the prediction model of residual mass histology, we plot the actual outcome vs. predicted for the development sample and the validation sample (Fig. 15.10). We include the distribution of predicted risks, such that discrimination can also be judged. The results by decile of predicted risk are shown in Table 15.7, to clarify the calculation of the Hosmer-Lemeshow statistic. Other tests for miscalibration included the overall test for calibration-in-the-large and calibration slope, and the Goeman–Le Cessie test, which were non-significant for model development and external validation (Table 15.8).



**Fig. 15.10** Validity of predictions of necrosis in the development sample ( $n=544$ ) and in the validation sample ( $n=273$ ). The distribution of predicted probabilities is shown at the bottom of the graphs, separately for those with necrosis and those with residual tumor. The triangles indicate the observed frequencies by deciles of predicted probability

**Table 15.7** Hosmer-Lemeshow test for calibration of the testicular cancer prediction model

Decile	$P(\%)$	Development <sup>a</sup>			Validation <sup>b</sup>			
		$N$	Predicted	Observed	$P(\%)$	$N$	Predicted	Observed
1	<7.3	56	2.4	1	<1.8	31	0.2	1
2	7.3–16.5	53	6.3	4	1.8–7.3	25	1.1	1
3	16.6–26.5	55	11.6	13	7.4–11.1	31	2.6	4
4	26.6–34.7	54	16.4	15	11.2–17.5	30	4.4	5
5	34.8–43.6	54	21.0	25	17.6–24.3	27	5.6	7
6	43.7–54.0	58	28.5	33	24.4–31.0	30	8.1	6
7	54.1–63.5	52	31.0	31	31.1–37.2	20	6.7	9
8	63.6–73.8	54	36.9	36	37.3–54.6	38	17.2	18
9	73.9–85.0	54	42.8	40	54.7–64.7	15	8.8	8
10	>85.0	54	48.0	47	>64.7	26	20.3	17
		544	245	245		273	74.9	76

<sup>a</sup> $\chi^2=5.9$ ,  $df=8$ ,  $p=0.66$  <sup>b</sup> $\chi^2=9.2$ ,  $df=9$ ,  $p=0.42$

**Table 15.8** Calibration of testicular cancer prediction model

	Development	Internal validation	External validation
Calibration-in-the-large	0	0	-0.03
Calibration slope	1	0.97 <sup>a</sup>	0.74
Calibration tests			
Overall miscalibration	$p=1$	-	$p=0.13$
Hosmer-Lemeshow	$p=0.66$	-	$p=0.42$
Goeman – Le Cessie <sup>b</sup>	$p=0.63$	-	$p=0.94$

Development and internal validation with  $n=544$  patients, external validation in  $n=273$  patients.

Internal validation with 200 bootstrap resamples using Harrell's validate function

<sup>a</sup>Equivalent to the uniform shrinkage factor as discussed in Chap. 14

<sup>b</sup>Test statistics of squared smoothed residuals calculated with R program from Jelle Goeman, available from website

### **15.3.11 Calibration and Discrimination**

The calibration plot can be extended into a “validation plot” as a central tool to visualize model performance. Calibration is shown by observed outcomes being close to prediction, while discrimination aspects can be indicated with the distribution of the predicted probabilities. The distribution can be shown by a histogram or density distribution. We can also make separate histograms for those with and without the outcome for further insights (see e.g. Fig. 15.10). It also helps to see the separation according to quantiles of predicted probabilities. For example, when deciles are used, these will be relatively far apart for a good discriminating model.

Calibration-in-the-large is a phenomenon that is fully independent of discrimination. For example, we can change the incidence of the outcome in a case-control study, but the discrimination will be unaffected. The calibration slope however has a direct relationship with discrimination. If the calibration slope is below unity, the discrimination is lower. Hence, overfitted models will show both poor calibration and poor discrimination when validated in new patients (Chap. 19).

Perfect calibration is possible with poor discrimination, for example when the range of predicted probabilities is small, such as between 9 and 11% for an average incidence of the outcome of 10%. At external validation, such a small range in predictions may arise from a narrow selection of patients (homogeneous case-mix). A drop in discriminative ability compared with the development setting can hence be explained by overfitting (calibration also poor), or a more homogeneous in case-mix (independent of calibration, see Chap. 19). On the other hand, a well discriminating model can have poor calibration, which can be corrected with various updating methods (Chap. 20).

### **\*15.3.12 R Code**

The Hosmer-Lemeshow test is implemented in a simple function `hl.ext` at the book’s website. The user can specify the number of groups (ten by default) and degrees of freedom (groups – 2 for model development, groups – 1 for model validation).

Calibration plots are made by an extension of Harrell’s `val.prob` function, called `val.prob.ci`. This function also provides assessment of calibration-in-the-large, calibration slope, and the calibration test *p*-value. Goeman provided R code for the functions `mlogit` (for binary or multinomial logistic regression), `smoothU` (for calculation of smoothed residuals), and `testfit` (for the Goeman-Le Cessie goodness-of-fit test).

## **15.4 Concluding Remarks**

In this chapter we have discussed a number of performance measures for prediction models; many more can be used, as systematically discussed in work by Hilden, Bjerregaard, and Habbema in the 1970s.<sup>161,162,163,191,192</sup> Many performance measures are related to each other; e.g. the *c* statistic is related to the Mann-Whitney U statistic,

which is calculated as a rank order test for the difference between predictions by outcome. The  $c$  statistic is also linearly related to Somer's  $D$  statistic ( $c=D/2 + 0.5$ ).

From a simple statistical perspective we want a small distance between observed outcome  $Y$  and predicted outcome  $\hat{Y}$ . Explained variation ( $R^2$ ) can then be used to indicate performance, and indicates the predictability of the outcome: How much do we know already about the phenomena that lead to the outcome.<sup>372</sup> Diagnostic prediction models would hence be expected to have higher  $R^2$  than prognostic models with long-term outcome. Indeed, prognostic models usually have  $R^2$  around 0.20. This indicates that substantial uncertainty remains at the individual level; we can only provide probabilities, and no certainty on the individual outcome.<sup>13,112</sup>

We have focused on measures that are in wide use in medical journals nowadays, including the concordance statistic (' $c$ ', or area under the ROC curve) for discrimination, and various tests for calibration and goodness-of-fit. The  $c$  statistic has been criticized by some, and should not be the only criterion in assessment of model performance. Especially,  $c$  may be rather insensitive to inclusion of additional predictors in prediction models, such as novel biomarkers.<sup>79,330</sup> But our theoretical examples and case study show that the  $c$  statistic is a key measure; it is closely related to other performance measures such as  $R^2$  and Brier score.

In principle we might focus our modelling strategy on optimizing performance measures such as the  $c$  statistic. Indeed, estimation algorithms have been described that maximize the  $c$  statistic rather than the log likelihood.<sup>332</sup>

Compared with current practice, calibration should receive more attention when evaluating prediction models. The recalibration test and its components (calibration-in-the-large and calibration slope) should be used routinely in performance assessment in external validation of prediction models.

### 15.4.1 *Bibliographic Notes*

The framework of a recalibration model was already proposed by Cox,<sup>86</sup> and has been supported by many other researchers for evaluation of model performance.<sup>81,174,290,291,458</sup> Nice illustrations of diagnostic test evaluation with ROC curves are available at:

<http://www.anaeasthetist.com/mnm/stats/roc/>

Nice illustrations of Lorenz curves and the Gini index are at:

[http://en.wikipedia.org/wiki/Gini\\_coefficient](http://en.wikipedia.org/wiki/Gini_coefficient)

## Questions

### 15.1 Overall performance measures

Overall performance measures for logistic regression models include Brier score and  $R^2$  type of measures, such as Nagelkerke's  $R^2$ .

- (a) What values can Brier scores and  $R^2$  take?
- (b) What types of scoring rules are Brier and  $R^2$ ?
- (c) What are disadvantages of Brier and  $R^2$ ?

### 15.2 Lorenz curve and incidence (Fig. 15.6)

In a Lorenz curve, the visual impression of a model with a  $c$  statistic of 0.80 depends on the incidence of the outcome.

- (a) What happens when a Lorenz curve is made for situation with 1% incidence?
- (b) And what for 99% incidence?

### 15.3 Interpretation of validation graph (Fig. 15.10)

Validity of predictions can well be judged graphically. How do you judge

- (a) calibration-in-the-large?
- (b) calibration slope?
- (c) discrimination?

### 15.4 Relationship between calibration, discrimination, and overall performance.

Explain the differences and the relation between calibration, discrimination, and overall performance measures.

# Chapter 16

## Clinical Usefulness

**Background** In addition to performance measures such as discrimination and calibration, we may want to know whether a prediction model is clinically useful: Is the model beneficial in clinical practice to guide diagnostic work-up, or decision making on therapy. For such decisions, we need a cutoff for the predicted probability (“decision threshold,” or “classification cutoff,” see Chap. 2). Patients with predictions above the cutoff are classified as positive; those under the cutoff as negative. We will use the term *clinical usefulness* for a model’s ability to make such classifications better than a default policy without the prediction model.

We consider performance measures for classification from a decision-analytic perspective, and discuss their relationships with performance measures as discussed in the previous chapter. Finally, we discuss study designs for measuring the actual impact of decision rules in clinical practice. We will illustrate the use of clinical usefulness measures in the testicular cancer case study, with model development in 544 patients and external validation with 273 patients from another centre.

### 16.1 Clinical Usefulness

In the previous chapter we saw that the distance between the predicted outcome and actual outcome ( $Y - \hat{Y}$ ) is central to quantify overall performance for regression models.<sup>181</sup> For classification, we replace  $\hat{Y}$  by a binary classifier, such that the distance becomes 0 for a correct classification and 1 for an incorrect classification. This is known as the error rate, which is simply the sum of false classifications divided by  $n$ , the number of subjects in the sample.

A critical issue is the choice of cutoff to classify subjects as positive or negative. Traditionally, the cutoff is set to 50%. This implies that false-positive and false-negative classifications are equally important. This is seldom the case in medicine. Often missing a patient with the outcome is more important than incorrect classification of a patient without the outcome; false-negative errors are more important than false-positive errors. We will consider informal and formal approaches to determining the optimal cutoff for a specific medical problem. The optimal cutoff is defined by the decision context, not by statistical criteria. Once a cutoff is chosen,

clinical usefulness measures can be defined. These consider the relative weight of false-positive and false-negative classifications, e.g. in a weighted error rate. A further approach is to study model performance over the whole range of possible cut-offs, as is done with the receiver operating characteristic (ROC) curve, but now for a “decision curve.”<sup>469</sup>

### ***16.1.1 Intuitive Approach to the Cutoff***

We consider two situations: treatment for bacterial meningitis and abandoning treatment for patients with traumatic brain injury. Bacterial meningitis is a severe infectious disease, with usually good outcome when treated early with antibiotics, but poor outcome when not treated in time. Several prediction models have been developed to predict the diagnosis “bacterial meningitis” among patients presenting at the emergency ward. If the probabilities from such models are used for decision making, we should use a rather low cutoff, such as not to miss bacterial meningitis cases.

Several prediction models have been developed for patients with traumatic brain injury. If presenting characteristics are dismal (e.g. high age, severe trauma, poor remaining brain function), the risk of a poor long-term outcome is high. Some researchers have tried to define patients who should not be treated because of very high risk of poor outcome. The cutoff was set close to 100%, since we only want to refrain from treatment in case of near certainty of a poor outcome.

### ***16.1.2 Decision-Analytic Approach to the Cutoff***

The cutoff for treatment against no treatment can formally be defined with a decision-analytic approach (Table 16.1). The loss (or “costs” in a broad sense) can include patient outcomes (mortality, morbidity, quality of life) as well as economic costs (including diagnostic work-up, therapeutic interventions, admission costs, costs of follow-up, etc).

We define two groups of subjects: those with the event if not treated, and those without the event if not treated. In the first group the costs relate to undertreatment

**Table 16.1** Costs of classification of subjects according to a decision threshold (“cutoff”)

	Event	No event
Treatment: Risk $\geq$ cutoff	cTP	cFP
No treatment: Risk $<$ cutoff	cFN	cTN

cTP and cFP: Costs of true and false-positive classification;  
 cFN and cTN: Costs of false and true-negative classifications, respectively

(false-negative classifications); in the second group overtreatment (false-positive classifications). The costs of false-negative classifications are referred to as cFN in Table 16.1. These should be compared with the costs of true-positive classifications (cTP); the difference cFN–cTP is the benefit of treatment for those who would have the event without treatment.

In the second group, relevant costs are for those without the event if not treated, who are treated. The costs of these false-positive classifications (cFP) should be compared with the costs of true-negative classifications (cTN); the difference cFP–cTN is the harm of overtreatment for those who would not have the event anyway.

Specifying a mathematical loss function leads to a simple definition for the optimal decision threshold: The odds of the cutoff corresponds to the relative weight of harm vs. benefit.<sup>132</sup>

$$\text{Odds(cutoff)} = \frac{(cFP - cTN)}{(cFN - cTP)},$$

where cTP and cFP refer to costs of true and false-positive classifications, and cFN and cTN to costs of false and true-negative classifications, respectively.

As discussed, benefit occurs for those with the event when not treated (cFN – cTP), and harm for those without the event when not treated (cFP – cTN). So, we note that only the differences between treated and non-treated situations are relevant to decision-making. Harm relates to the unnecessary treatment of those without the outcome; benefit to the correct treatment of those with the outcome. Patients with predicted risks (odds) above the threshold should be treated, and those below the threshold not treated.

### 16.1.3 Error Rate and Accuracy

If benefit and harm are weighted equally, the odds of the threshold is 1:1, or a threshold probability of 50%. This cutoff is by default considered in the calculation of the error rate, which is defined as (FN+FP)/N (Table 16.2). The complement is the accuracy rate: (TN+TP)/N. Often FN classifications are more important than FP classifications, which makes the accuracy rate not a sensible indicator of clinical usefulness. Other disadvantages include that the accuracy rate by definition is high for a frequent or infrequent outcome. For example, if the average mortality is 7% after an acute MI, the accuracy is 93% when we classify all patients as survivors.

**Table 16.2** Classification of subjects according to a decision-threshold

	Event	No event
Treatment: Risk $\geq$ cutoff	TP	FP
No treatment: Risk $<$ cutoff	FN	TN

TP and FP: Numbers of true and false-positive classifications; FN and TN: numbers of false and true-negative classifications, respectively

### ***16.1.4 Accuracy Measures for Clinical Usefulness***

The accuracy rate is usually calculated at the simplistic cutoff of 50%, but can also be calculated at clinically defendable thresholds. The accuracy then indicates the proportion of the population that receives the optimal treatment if the predictions from the model are followed. Similarly, sensitivity and specificity can be calculated at the optimal decision threshold: Sensitivity = TP / (TP+FN); Specificity = TN / (FP+TN).

The harm to benefit ratio that underpins the choice of the cutoff can also be used to calculate a weighted accuracy, and its complement, the weighted error rate. We can express the TN classifications in units of the TP classifications, such that the weighted accuracy is calculated as  $(TP + w TN) / (N_{\text{event}} + w N_{\text{no event}})$ . Similarly, the weighted error rate can be calculated as  $(FN + w FP) / (N_{\text{event}} + w N_{\text{no event}})$ . These rates can also be calculated for a default policy, which would be followed without using the prediction model. The default policy could be treat all, or treat none. The improvement that is obtained by making decisions based on predictions from the model is the difference between the weighted accuracy obtained with the model vs. the weighted accuracy of the default policy.

### ***16.1.5 Decision Curves***

In practice, it is often difficult to define the optimal threshold precisely. Difficulties may lie at the population level, i.e. that we do not have sufficient data to quantify harms and benefits for the typical threshold in a decision problem. Moreover, the relative weight of harms and benefits may differ from patient to patient, necessitating individual thresholds.

An impression of the order of magnitude of the typical threshold can usually be obtained from clinical experts. We could consider lower and upper values for the threshold, with a grey area in between. This approach was for example followed in classifying patients with possibly indolent prostate cancer, where those with probabilities < 30% were advised to undergo surgery, and those with probabilities > 60% to undergo active surveillance, and a grey area in between.<sup>424</sup>

It is also possible to consider the whole range of decision thresholds, ranging from 0% to 100%. This approach was worked out by Vickers and Elkin.<sup>469</sup> They constructed a “decision curve,” which considers a threshold over the range 0–1. The method starts as we did before, by noting that the threshold is directly related to the harm to benefit ratio. Next they create a plot which shows the net benefit (NB) of treating patients according to the prediction model. The formula for NB goes back to work published long ago<sup>328</sup>:

$$\text{Net benefit} = \text{NB} = \frac{(TP - wFP)}{N},$$

where TP is the number of true-positive classifications, FP the number of false-positive classifications, and  $w$  is a weight equal to the odds of the threshold

$(p_t/(1 - p_t)$ , or the ratio of harm to benefit. For example, a threshold of 10% means that the FP classifications are valued at 1/9th of a TP classification.

The NB of a prediction model should be compared with default policies of “treat none,” or “treat all.” Treat none means that the NB is zero (since TP and FP are zero). The NB for “treat all” depends on the threshold and the incidence of the outcome. The NB of a well-calibrated prediction model is at the maximum when the threshold is at the incidence of the outcome. At this point, the policy “treat all” has an NB of zero, as well as the policy “treat none.”

If the prediction model required efforts such as obtaining data from extra medical tests that were invasive, burdensome, or costly, a different version of the NB formula can be used:

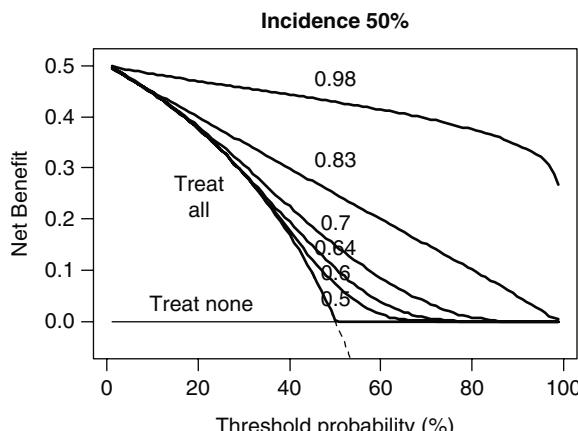
$$\text{NB} = (\text{TP} - w\text{FP}) / N - \text{test harm}$$

where test harm is expressed per patient in units of the TP result.<sup>469</sup>

The interpretation of the NB is in units of the true positives; how many more patients are correctly treated (TP decisions) at the same rate of not treating those who do not need treatment (TN decisions). For the interpretation of a decision curve we need to identify a range of plausible threshold probabilities for treatment, and then see whether the model has benefit at all values within this range. If so, the model can clearly be recommended for clinical use.

### \*16.1.6 Examples of NB in Decision Curves

We present decision curves for prediction models with increasing  $c$  statistics in Fig. 16.1, based on the distributions as shown in Fig. 15.4. With a  $c$  statistic of 0.5, the NB



**Fig. 16.1** Decision curves for prediction models with increasing  $c$  statistics, based on the distributions as shown in Fig. 15.4. We note that the net benefit strongly depends on the  $c$  statistic, but that a near perfect model ( $c=0.98$ ) is always clinically useful. “Treat all” is associated with a negative NB for thresholds over 50%

**Table 16.3** Maximum net benefit of various prediction models at decision thresholds equal to incidences of the outcome

Prediction model	Incidence			
	1%	10%	50%	90%
$c=0.6$	0.0028	0.019	0.073	0.150
$c=0.64$	0.0032	0.026	0.104	0.218
$c=0.7$	0.0048	0.036	0.149	0.324
$c=0.83$	0.0071	0.059	0.250	0.511
$c=0.98$	0.0096	0.092	0.429	0.817

The default policies of treat all or treat none have a net benefit of zero at these points. Net benefit is expressed in units of true-positive decisions (TP)

is identical to the strategy of treat all; no gain is obtained from using the decision model. With a near perfect model ( $c=0.98$ ), the NB appears to be substantial over the whole range of thresholds from 0% to 100%. With 50% incidence, the maximum NB is 0.5 at a threshold of 0%, i.e. treat all (50% correct, 50% incorrect).

The maximum gain from using a prediction model is when the threshold is equal to the incidence of the outcome. The default policies of “treat all” or “treat none” both have an NB of zero at these points. Table 16.3 shows that the maximum gain is less than 1% at an incidence of 1%; but increases to over 0.8 for a near perfect prediction model at incidence 90%. NB refers here to identifying the cases. If we reverse the coding, we consider identifying the non-cases, and we can keep the incidence between 0 and 50%.

### 16.1.7 Example: Clinical Usefulness of Prediction Model for Testicular Cancer

In the testicular cancer example, the residual mass histology is classified as benign (necrosis) vs. malignant (residual tumor). Malignant histology should surgically be resected, but resection of benign tissue is harmful (risks of surgical complications, hospital admission, costs). A decision analysis suggested a threshold of 70% for the probability of benign histology.<sup>422</sup> This implies a ratio of 7:3 for missing malignant histology vs. unnecessary surgery of benign histology, or equivalently, a ratio of 3:7 for unnecessary vs. necessary surgery.

At model development, the sensitivity and specificity were 92% and 42%, respectively, at a cutoff of 70% (Table 16.4). At external validation, only 23 patients had predictions over 70%. The specificity was lower (21%), but the sensitivity higher (96%) than at model development. The accuracy rate was  $(102+275) / 544 = 69\%$  at development, and remarkably, slightly better at validation:  $(16+190) / 273 = 75\%$ . The error rates are the complements of the accuracy rate (31% and 25%).

The weighted accuracy rate is expressed in necessary resections of tumor (TP), where an unnecessary resection of necrosis (FP) is weighted at 3/7 of a missed

**Table 16.4** Classification table for the development ( $n=544$ ) and validation ( $n=273$ ) sets of testicular cancer patients at a cutoff for the probability of necrosis of 70% (or tumor)=30%

	Development ( $n=544$ )		Validation ( $n=273$ )	
	Necrosis	Tumor	Necrosis	Tumor
Prediction $\geq 70\%$	102	24	16	7
Prediction $<70\%$	143	275	60	190
	245	299	76	197
Spec=42%	Sens=92%		Spec=21%	Sens=96%
Accuracy=69%, wAcc=74%			Accuracy=75%, wAcc=86%	
$NB_{model}=0.393$	$NB_{treat\ all}=0.357$		$NB_{model}=0.602$	$NB_{treat\ all}=0.602$
Increase in NB=0.036			Increase in NB=0	

Sensitivity, specificity, accuracy and weighted accuracy are calculated for both data sets. Sensitivity calculated for patients with tumor (the more severe outcome), and specificity for patients with necrosis (the less severe outcome).

tumor. Resection of all masses leads to 299 tumor resections, but also to 245 unnecessary resections. This is a better choice than resection in none, since the average probability of benign tissue was 45%, which is below the threshold of 70%. This default policy has a weighted accuracy rate of  $299/(299+3/7 \times 245) = 74\%$ . Using the model with a cutoff of 70% would lead to 275 necessary resections plus 102 correct omissions of resection in patients with necrosis. The weighted accuracy rate is  $(275+3/7 \times 102) / (299 + 3/7 \times 245) = 79\%$ . So, missing  $299 - 275 = 24$  patients with tumor (FN decisions) is more than compensated by the increase in correct omission of resection from 0 to 102 (TN decisions). The  $NB = (TP - w FP) / N = (275 - 3/7 \times 143) / 544 = 0.393$ , in contrast to  $(299 - 3/7 \times 245) / 544 = 0.357$  for resection in all.

In the validation sample, resection of all masses would lead to 197 tumor resections, but also 76 unnecessary resections, for a weighted accuracy rate of  $197/(3/7 \times 76 + 197) = 86\%$ . Using the model with a cutoff of 70% would lead to 190 necessary resections plus 16 correct omissions of resection in patients with necrosis. The weighted accuracy rate is also 86%  $([3/7 \times 16 + 190] / [3/7 \times 76 + 197])$ . The  $NB = (TP - w FP) / N = (190 - 3/7 \times 60) / 273 = 0.602$ , in contrast to  $(197 - 3/7 \times 76) / 273 = 0.602$  for resection in all. Hence, the model is not clinically useful in the validation setting. Put simply, sparing resection in 16 patients with necrosis does not compensate missing tumor in 7, when we weigh tumor as 7/3 of necrosis (Table 16.4).

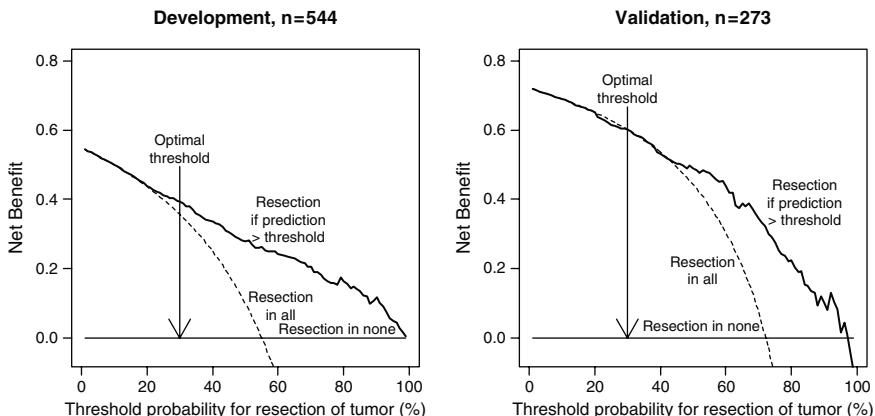
### 16.1.8 Decision Curves for Testicular Cancer Example

Thus far we assumed a constant utility function, i.e. a weight of 7/3, for the decision to perform surgery on a residual mass in a testicular cancer patient. The corresponding threshold of 30% for the probability of tumor is an average and may vary for individual patients based on their personal weighing of surgical risks against

increased chances of long-term survival.<sup>469</sup> Instead of a single relative weight, we may consider a range of weights in a decision curve. This curve shows that the model is clinically useful for thresholds over 20% in the development sample, equivalent to thresholds for the probability of necrosis below 80% (Fig. 16.2). In the validation sample, the range is 55–95%, confirming that the model is not clinically useful in the validation setting when we assume a decision threshold of 30%.<sup>426</sup> The lines for resection in all and resection in none cross at the frequencies of tumor (55% and 72% respectively). At these points the model has maximum gain in NB.

### \*16.1.9 Verification Bias and Clinical Usefulness

As mentioned in the previous chapter, the policy for resection was generally that residual masses should be detected on CT scan. This usually means that the radiologist considers a lymph node as enlarged, i.e. >10 mm. Patients with masses  $\leq 10$  mm are generally not considered candidates for resection, and these patients are hence not included in our evaluations of clinical usefulness.<sup>466</sup> Such verification bias affects performance measures such as sensitivity and specificity, and also the clinical usefulness measures. The conclusion from Fig. 16.2 is that the prediction model has limited to no clinical usefulness for the patients in these samples; hence we cannot easily reduce resections in those who underwent resection under current policies. There is however a large group of patients who are currently not considered for resection.



**Fig. 16.2** Decision curves of predictions of tumor in patients with testicular cancer in the development sample ( $n=544$ ) and in the validation sample ( $n=273$ ). We note that the model is not clinically useful in the validation setting. The lines for resection in all and resection in none cross at the frequencies of tumor (55% and 72% respectively)

Among these patients, with small or “normal” residual masses, some will harbor residual tumor cells. These patients likely would benefit from resection, while they are currently not considered candidates in most centres. An exploratory analysis in 241 patients from a MRC/EORTC trial suggested that 84 (31%) of these might be candidates for resection, using the decision threshold of 30% risk of tumor.<sup>467</sup>

### \*16.1.10 R Code

Classification tables can readily be calculated with the `table` command, with accuracy and weighted variants. Andrew Vickers provided a function `dca` for *R* and Stata, which enables drawing decision curves (<http://www.mskcc.org/mskcc/html/74366.cfm>). For the development data, the commands are as follows:

```
# tumor as outcome; predictions were for necrosis,
# hence "1 - y," and "1 - prob" in dca function from Vickers
dca.dev <- dca (yvar=1-full$y, xmatrix=1-plogis(lp), prob="Y")
# plot 3 lines: net benefit using model; treat all; treat none
plot(x=dca.dev$threshold, y=dca.dev[,1], ...)
lines(x=dca.dev$threshold, y=dca.dev[,2:3]) # Fig 16.2
```

## 16.2 Discrimination, Calibration, and Clinical Usefulness

From a statistical perspective, some have argued that discrimination is the primary criterion of a prediction model, since miscalibration can relatively easily be corrected when we apply a model in a new setting. However, when we apply a model, we do not know about any miscalibration yet. Hence, clinical usefulness of a model in clinical practice depends at least on the combination of discrimination and calibration. Of course some discriminative ability is important for any clinical usefulness, as is clear from the decision curves in Fig. 16.1. Some refer to  $c$  statistics over 0.7 as “acceptable” or “modest,” over 0.8 as “good,” and over 0.9 as “excellent.” This is, however, very problematic: It is not possible to indicate a minimum value for the  $c$  statistic to make a model clinically useful. In addition to not considering calibration aspects, the consequences of decisions are not considered in the  $c$  statistic. Once the ratio of harms to benefits is used to define a clinically relevant threshold, the distribution of predictions around this threshold has a major influence on clinical usefulness. The NB of a model with well-calibrated predictions is maximal if the decision threshold has the same value as the incidence of the outcome. Approximately half of the predictions are then above and the other half below the threshold. If all predictions are above or below the threshold, the model is not clinically useful, even with a “good”  $c$  statistic of 0.8 for example.

As an example, we review the performance of the testicular cancer prediction model according to various criteria (Table 16.5). We note that overall performance,

**Table 16.5** Summary of performance measures for prediction model in testicular cancer

Aspect	Measure	Development <sup>a</sup>	Validation
Overall performance	$R^2$	38%	27%
	Brier <sub>scaled</sub>	28%	20%
Discrimination	c statistic	0.81	0.79
	Discrimination slope	0.29	0.24
	Lorenz curve p25	9%	13%
	p75	58%	65%
Calibration	Calibration-in-the-large	–	-0.03
	Calibration slope	0.97	0.74
	Test for miscalibration	p=1	p=0.13
Clinical usefulness at cutoff	Sensitivity	92%	96%
	Specificity	42%	21%
p(tumor)=30%	Accuracy	69%	75%
	Weighted accuracy	74%	86%
	Net benefit – resection in all	0.39 – 0.36 = 0.03	0.60 – 0.60 = 0

<sup>a</sup>Internally validated measures if available

discrimination, and calibration look quite satisfactory, although predictive effects were slightly less strong than anticipated at external validation (calibration slope, 0.74). The external validation data set was relatively small, hence providing limited power for tests of miscalibration. The clinical usefulness measures show a less-fortunate pattern. Sensitivity is quite high, but specificity low. Hence, we can spare only few patients with necrosis a resection. Indeed, there was no clinical usefulness at the cutoff of 30% in the external data set. Hence, good calibration and discrimination are necessary but not sufficient for clinical usefulness.

### 16.2.1 Aim of the Prediction Model and Performance Measures

Performance measures have different relevance in relation to the aim of the prediction model (see also Chap. 2). As discussed earlier, clinical usefulness requires considering a decision threshold, which is determined by the relative weight of harms and benefits of a treatment. Clinical usefulness then depends on the combination of calibration, discrimination, and the distribution of predictions around the decision threshold.

One application of a model is in targeting preventive activities to certain “high risk” groups for efficient use of sparse resources. Discrimination is then the primary requirement; the main issue is to reasonably order subjects according to risk. If sparse resources are not an issue, the targeting should be based on harm to benefit considerations, making clinical usefulness the most relevant aspect of performance.

If the aim is to inform or make decisions in clinical practice, calibration is an essential requirement. Miscalibration implies that we provide biased information,

which can lead to worse decision making than with a default policy that ignores the model predictions (a loss in NB, Chap. 19). Of course discriminative ability is also required, but limited discrimination with a well-calibrated model will lead to a limited, but never a negative, NB. Miscalibration can lead to a negative NB.

Prediction models may have several roles in research. In RCTs, inclusion criteria can be according to a model, e.g. to select high-risk groups for investigation of a new treatment. For such an application, calibration is essential. Vickers et al. have described a method to determine eligibility for an RCT based on NB considerations, including the expected effect of a treatment.<sup>470</sup> For covariate adjustment in an RCT, discrimination is most important. If no strong predictors are known, covariate adjustment has no benefit over unadjusted analysis of the treatment effect. Calibration is not an issue when covariate effects are included in a model to estimate adjusted treatment effects. When a prediction model is used for confounder adjustment or case-mix adjustment, calibration is also automatically corrected for. Confounder adjustment can be achieved with various approaches, including traditional regression analyses, including the exposure and confounders, and propensity score adjustment. Discrimination of a model with confounders can range from low to high values, which does not make the adjustment less or more valid. With very high discrimination, we may even suspect that we adjust for a predictor that is too close to the outcome that we want to analyze; hence very high discrimination is suspicious. Similarly, a high  $c$  statistic of a propensity score does not imply validity; it merely means that we can predict who gets treatment and who does not. Most relevant is that all relevant confounders are included in the adjustment, i.e. covariates that are associated with treatment decisions and with outcome. The latter requires subject knowledge rather than statistical criteria.

### 16.2.2 Summary Points

- Discriminative ability is the primary requirement of a prediction model if we want to identify a high-risk group, or perform covariate adjustment of a randomized controlled trial.
- For informing patients and medical decision making, calibration is the primary requirement, which determines clinical usefulness together with discrimination and the distribution of prediction around the decision threshold.

## 16.3 From Prediction Models to Decision Rules

Prediction models provide diagnostic or prognostic probabilities. They may assist clinical decision making without telling to clinicians what to do precisely. One motivation for providing probabilities only is that decision thresholds may differ from patient to patient. Some argue, however, that prediction models will more

likely have an impact on clinical practice when clear actions are defined in relation to the predictions. They favor presentation as a decision rule rather than as a prediction model.

Decision rules may be most useful when decision making is complex, when the clinical stakes are high, or when there are opportunities to achieve cost savings without compromising patient care.<sup>282</sup> But few rules have undergone formal analysis to determine whether they improve outcomes when used in clinical practice (“impact analysis”). The medical impact of most published prediction or decision rules is unknown.

For application as a decision rule, prediction models may require simplification to provide clear advise on actions with high and low predictions. A decision threshold has to be defined, either chosen informally or by formal decision analysis, based on the relative weights of false-negative and false-positive decisions. In some diagnostic rules, we may not want to miss any patient with the outcome of interest (e.g. Ottawa Ankle rules,<sup>147</sup> CT head rules<sup>391</sup>). This implies that we aim for a sensitivity of 100%, and hope for reasonable specificity. We accept false-positive classifications, since the 100% sensitivity implies an infinite cost of false-negative classifications.

Reilly and Evans have proposed a set of criteria for assessing the impact of prediction models as decision rules (Table 16.6).<sup>344</sup> These progressive evidentiary standards emphasize that a prediction model rises to the level of a decision rule only if clinicians use its predictions to help make decisions for patients.

The first level of evidence is at the development of a prediction model. Reilly and Evans emphasize the model selection aspects (“identification of predictors”) and blinded assessment of outcomes. We have seen that overfitting and measures to prevent overoptimistic expectations of model performance are especially important.

Levels 2 and 3 are related to model validation, which indeed is essential before application of a model can be recommended. Validation in multiple settings is required to gain confidence in the applicability of a model for a new setting.

Levels 4 and 5 consider impact analysis, where a prediction model is used as a decision rule. We assess whether the rule improves physicians’ decisions (quality or cost-effectiveness of patient care).

### ***16.3.1 Performance of Decision Rules***

Sensitivity and specificity are often used as performance criteria for a decision rule. As discussed before, these criteria may also be used for validation of a prediction model at certain cutoffs. Reilly and Evans note that decision rules generally improve physicians’ specificity more than sensitivity; physicians ascribe greater value to true-positive decisions (provide care to patients who need it) than to true-negative decisions (withhold care from patients who do not need it). This is equivalent to weighing FN more than FP classifications, or a ratio of harm to benefit less

**Table 16.6** Developing and evaluating clinical prediction models and decision rules (based on Reilly and Evans<sup>344</sup>)

Level of evidence	Definitions and standards of evaluation	Clinical implications
Level 1		
Derivation of prediction model	Identification of predictors for multivariable model; blinded assessment of outcomes	Needs validation and further evaluation before using in actual patient care
Level 2		
Narrow validation of prediction model	Assessment of predictive ability when tested prospectively in one setting; blinded assessment of outcomes	Needs validation in varied settings; may use predictions cautiously in patients similar to sample studied
Level 3		
Broad validation of prediction model	Assessment of predictive ability in varied settings with wide spectrum of patients and physicians	Needs impact analysis; may use predictions with confidence in their accuracy
Level 4		
Narrow impact analysis of prediction model used as decision rule	Prospective demonstration in one setting that use of decision rule improves physicians' decisions (quality or cost-effectiveness of patient care)	May use cautiously to inform decisions in settings similar to that studied
Level 5		
Broad impact analysis of prediction model used as decision rule	Prospective demonstration in varied settings that use of decision rule improves physicians' decisions for wide spectrum of patients	May use in varied settings with confidence that its use will benefit patient care quality or effectiveness

than 1. This implies a threshold for treatment below 50%. An important issue is that the sensitivity and specificity of a decision rule in clinical practice is not only influenced by the quality of the prediction model, but also by the adherence of clinicians to the rule. Validation of a prediction model may indicate the efficacy of a rule (the maximum that can be attained with 100% adherence), but impact analysis will indicate the effectiveness in practice.

Clinicians may choose to overrule the decision rule, which may improve sensitivity and/or specificity. But overruling may also dilute the effects of the rule.<sup>56</sup> There may be various barriers to the clinical use of decision rules. Barriers include issues in attitude such as skepticism of guidelines (in general and with respect to the specific rule), questions on the clinical sensibility of the rule, too high confidence in clinical judgment, fear of medicolegal risks, concern that important factors are not addressed by the decision rule, concern on patient safety, and practical issues such as availability of the rule at the time of decision making, and ease of use.

An impact analysis should ideally be designed as an RCT. Randomization by centre is an obvious approach to organizational changes such as using a decision rule in practice. But there is a risk for contaminating intervention and control groups and the logistic and economic challenges of multi-centre studies are formidable. Some previous evidence of impact is required, which may come from a single centre. Such evaluation measures the actual effects of using the rule in clinical practice, which is critical information when planning multi-centre (level 5) studies. The Ottawa Ankle rules provide an excellent case study for model development, validation, and impact assessment.<sup>427,334</sup>

### \****16.3.2 Treatment Benefit in Prognostic Subgroups***

Prediction models may indicate subgroups of patients with a poor prognosis, often suggesting that these patients may need more aggressive treatment. Note that this assumes a curative intend; e.g. in oncology, palliative treatments are generally considered when cure is not possible, and more aggressive treatment may do more harm than good. On the other hand, good prognosis groups may be defined, where less-intensive treatment may be sufficient. This distinction is for example made by the International Germ Cell Classification (IGCC).<sup>5</sup> In this clinical area, several RCTs have been performed that follow the IGCC classification. More aggressive treatment was studied in “poor risk” patients (e.g. high dose instead of standard dose chemotherapy), and less intensive therapy in “good prognosis” patients (e.g. three instead of four cycles of cisplatin-based chemotherapy).

### \****16.3.3 Evaluation of Classification Systems***

In the near future, new classification systems will come up, which include genetic profiles or other novel biomarkers. Systematic studies are required to validate these new systems, and provide evidence on any treatment benefits in subgroups as indicated by such new classification systems. For example, a clinical trial has started, which will use a genetic profile to test which early-stage breast cancer patients need chemotherapy.<sup>483</sup> For efficient design, the trial can focus on the patients whose risk classification is discordant between the genetic profile and the traditional classification with clinical-pathological information only. For example, women who are determined to be at high risk for relapse of breast cancer by both the genetic profile and traditional clinical pathological criteria may be treated with chemotherapy, as is current standard practice. Those who are low risk by both criteria will not receive chemotherapy. However, the women who are determined to be at high risk for distant relapse by one criterion and low risk by the other will be randomly assigned to one of two arms.<sup>439</sup> Such a study is important for the evaluation of treatment benefit,

but also validates the prognostic model by comparing outcomes between various prognostic groups under standard treatment. The design of trials of markers in oncology is discussed in more detail elsewhere.<sup>364</sup>

A limitation of this design is that differences between groups may be small, leading to large sample sizes to be studied. Moreover, the prognostic value of a classification can usually well be determined in observational data, e.g. in a prospective validation study. For example, the prediction model for the residual histology in testicular cancer was validated in three validation studies.<sup>467</sup> We did consider setting up an RCT in discordant pairs of patients: Those with an indication for surgery according to either the model or current policy, but no indication according to the other. The benefits of surgery were however considered to be clear once the histology was known. This is similar to other diagnostic studies, where knowledge of the reference standard is considered sufficient to estimate further treatment benefit. In contrast, the definition of “indolent prostate cancer” causes a lot of debate among urologists, with uncertainty as to whether such “indolent cancers” can be safely selected for active surveillance rather than surgery (see discussion on papers presenting nomograms for “indolent cancer”<sup>424,227</sup>).

## 16.4 Concluding Remarks

In this chapter we have discussed measures for clinical usefulness of prediction models. From a statistical perspective we may simply calculate error rates, with implicit equal weighing of false-positive and false-negative classifications. We noted that the  $c$  statistic was not sufficient to indicate clinical usefulness, although a low  $c$  statistic made it unlikely that a model was clinically useful. Good calibration was required, and the distribution of predictions had to be on both sides of the decision threshold. Usefulness is high for a perfectly calibrated model, with a substantial  $c$  statistic, in a clinical problem where the decision threshold is equal to the incidence of the outcome. A further discussion follows in Chap. 19.

Note that the determination of the decision threshold is fully independent from developing and validating the decision model. It should ideally be based on a formal weighting of harms and benefit of a treatment, compared with the alternative of no treatment and treatment for all. Clinical usefulness is hence problem dependent, and not in the hands of the modeler. Final impact of a prediction model as a decision rule is one further step in the evaluation.

Compared with current practice, calibration should receive more attention when evaluating prediction models. The recalibration test and its components (calibration-in-the-large and calibration slope) should be used routinely in performance assessment in external data. Also, measures of clinical usefulness should be considered more often. Decision curves are promising tools by providing simple graphs to summarize a model’s quality for the full range of possible decision thresholds.

### ***16.4.1 Bibliographic notes***

For comparison of performance of alternative prediction models, Cook recently proposed to start with a comparison of overall model performance (e.g.  $R^2$ ), followed by calibration and discrimination. She suggests to consider reclassification of individuals across categories, and outcome for these reclassified individuals.<sup>78</sup> Similarly, a “predictiveness curve” has been proposed to assess the usefulness of a prediction model for a population. For a diagnostic problem, it shows the predicted probabilities (y-axis) vs. the cumulative distribution (x-axis, as in Lorenz curves) [<http://www.bepress.com/uwbiostat/paper282/>]. Pencina et al. recently proposed some statistical summary measures, which initiated substantial discussion.<sup>329,330</sup> These proposals all lack a thorough decision-analytic motivation, in contrast to for example, decision curves.

## Questions

### 16.1 Calculation of net benefit (sect. 16.1.5)

Net benefit is defined as:  $NB = (TP - w \cdot FP) / N$ , where  $TP$  means a true-positive classification,  $w \cdot FP$  weighted false-positive classification, and  $N$  the sample size.

- (a) What is the NB if we classify all subjects as positive, in a setting of 50% incidence of the outcome, and a relative weight of FP classifications as 1:1?
- (b) And what if the relative weight of FP classifications is 1:2?
- (c) Recalculate the sensitivity, specificity, accuracy, and NB for the 273 validation patients in Table 16.4.

### 16.2 Decision curves (Fig. 16.1)

- (a) Why is the “treat all” strategy in Fig 16.1 associated with a negative NB for thresholds over 50%?
- (b) What will happen to the decision curves when a lower incidence than 50% is considered? Or a higher incidence?

### 16.3 Verification bias (Sect. 16.1.9)

What is verification bias? How does it effect clinical usefulness, e.g. in the right panel of Fig 16.2?

### 16.4 Usefulness for decision making vs. research purposes

When would you consider a model clinically useful? And useful for research?

### 16.5 Errors in an Editorial

Consider the Editorial in JNCI on discrimination, calibration, and interpretation of risks.<sup>112</sup>

- (a) What is wrong in Fig. 2?
- (b) What is the decision threshold for this problem? What is the basis for this threshold?
- (c) How clinically useful is the Gail model with this threshold, according to weighted accuracy and NB?

# Chapter 17

## Validation of Prediction Models

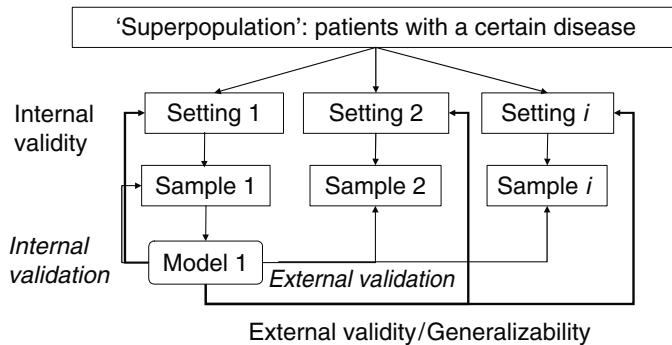
**Background** The purpose of a predictive model is to provide valid outcome predictions for new patients. Essentially, the data set to develop a model is not of interest other than to learn for the future. Validation hence is an important aspect of the process of predictive modelling. An important distinction is between internal and external validation. We discuss internal and external validation techniques in this chapter, with illustrations in case studies.

### 17.1 Internal vs. External Validation, and Validity

A general framework for validation and validity concepts is shown in Fig. 17.1. We develop a model within a representative sample of patients from an underlying population. This underlying population has specific characteristics, e.g. a specific hospital with a certain profile of how patients come to this hospital. By necessity, the sample is historic in nature, although we generally will aim for recent data, which are representative of current practice. At least we should determine the internal validity (or “reproducibility”) of our predictive model for this underlying population. We do so by testing the model in our development sample (“internal validation”). Internal validation is the process of determining internal validity. Internal validation assesses validity for the setting where the development data originated from.

Another aspect is the external validity (or “generalizability” / “transportability”) of the prediction model to populations that are “plausibly related.”<sup>222</sup> Generalizability is a desired property from both a scientific and practical perspective. Scientifically speaking hypotheses and theories are stronger when their generalizability is larger. Practically, we hope to be able to validly apply a prediction model to our specific setting.

The definition of “plausibly related” populations is not self-evident, and requires subject knowledge and expert judgment on epidemiological study design aspects. We consider “plausibly related” as that populations can be thought of as parts of a “superpopulation” (Fig. 17.1). We could also state that we consider populations that would be reasonable to apply the previously developed model to. Populations will



**Fig. 17.1** A conceptual framework on internal vs. external validation, and validity. We consider a superpopulation, consisting of several subpopulations (referred to as “settings”). We develop a model in sample 1 from setting 1. Internal validation is the process of determining internal validity for setting 1. External validation is the process of determining generalizability to settings 2 to *i*

be slightly different, e.g. treated at different hospitals or in different time frames. Various aspects may differ between these populations, e.g. the selection of patients (e.g. referral centre vs. more standard setting), and definitions of predictors and outcome. For example, a superpopulation could be formed by “patients with an acute MI,” with the GUSTO-I data representing one population, defined by the inclusion criteria for this trial, the participating centres, and the time of accrual.

We learn about external validity by testing the model in other samples (sample 2 to *i* in Fig. 17.1, “external validation”). These samples are fully independent from the development data and originate from different but plausibly related settings. The more often the model is externally validated and the more diverse these settings, the more confidence we gain in the generalizability of the model. This is similar to the approach to assessing any scientific hypothesis.<sup>222</sup>

## 17.2 Internal Validation Techniques

Several techniques are available to assess internal validity. Some of the most common techniques in medical research are discussed here (Table 17.1).

### 17.2.1 Apparent Validation

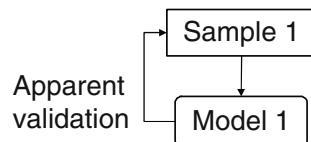
With apparent validation, model performance is assessed directly in the sample where it was derived from (Fig. 17.2). Naturally this leads to an optimistic estimate of performance (biased assessment), since model parameters were optimized for

**Table 17.1** Overview of characteristics of some techniques for internal validation

Method	Development	Validation
Apparent	Original 100%	Original 100%
Split-sample	50–67% of original	Independent 50–33%
Cross-validation <sup>a</sup>		
Classical	2 × 50% – 10 × 90% of original	Independent 2 × 50% – 10 × 10%
Jack-knife	$N \times (N - 1)$ of original	Independent $N \times 1$ patient
Bootstrap	Bootstrap sample of size $N$	Original 100%

<sup>a</sup> More stable cross-validation results are obtained by repeating the cross-validation many times, e.g. 50 times (“multi-fold cross-validation”)

**Fig. 17.2** Apparent validation refers to assessing model performance in the sample where the model was derived from



the sample. However, we use 100% of the available data to develop the model, and 100% of the data to test the model. Hence, the procedure gives optimistic but stable estimates of performance.

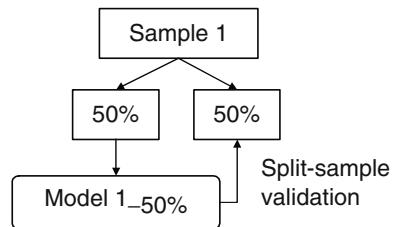
### 17.2.2 Split-Sample Validation

With split-sample validation, the sample is randomly divided into two groups. This very classical approach is inspired by the design of an external validation study. However, the split in derivation and test set is at random. In one group the model is created (e.g. 50% of the data) and in the other the model performance is evaluated (e.g. the other 50% of the data, Fig. 17.3). Typical splits are as 50%:50% or 2/3:1/3.

Several aspects need attention when a split-sample validation is performed. If samples are split fully at random, substantial imbalances may occur with respect to distribution of predictors and the outcome. For example, if we perform split-sample validation with a small subsample from GUSTO-I ( $n=429$ ), the average incidence of 30-day mortality is 5.6% (24/429), but it may easily be 4% in a 50% random part and 7% in another part. Similarly, the distribution of predictors may vary. For predictors with skewed distributions the consequences may be even worse. For example, a random development sample may not contain any patient with shock, which occurred in only 1.6% (7/429). A practical possibility is to stratify the random sampling by outcome and relevant predictors.

The drawbacks of split-sample methods are numerous.<sup>174,292,374</sup> One major objection is related to variance. Only part of the data is used for model development, leading to less-stable model results compared with development with all development

**Fig. 17.3** Split-sample validation refers to assessing model performance in a random part of the sample, with model development in the other part



data. Also, the validation part is relatively small, leading to unreliable assessment of model performance. Further, the investigator may be unlucky in the split; the model may show a very poor performance in the random validation part. It is not more than human that the investigator is tempted to repeat the splitting process until more favorable results are found. Another objection is related to bias. We obtain an assessment of the performance when a part of the data is used, while we want to know the performance of a model based on the full sample.

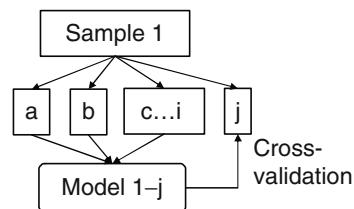
In sum, split-sample validation is a classical but inefficient approach to model validation. It dates from the time before efficient but computer-intensive methods were available, such as bootstrapping.<sup>108</sup> Simulation studies have shown that rather large sample sizes are required to make split-sample validation reasonable.<sup>413</sup> But with large samples, the apparent validity is already a good indicator of model performance. Hence, we may conclude that split-sample validation is a method that works when we do not need it. It should be replaced in medical research by more efficient internal validation techniques, and by attempts of external validation.

### 17.2.3 Cross-Validation

Cross-validation is an extension of split-sample validation, aiming for more stability (Fig. 17.4). A prediction model is again tested on a random part that was left out from the sample. The model is developed in the remaining part of the sample. But this process is repeated for consecutive fractions of patients. For example, the data set may be split in deciles (containing 1/10 of the patients), with model development in nine of the ten and testing in one of the ten, which is repeated ten times (“ten-fold cross-validation”). In this way, all patients have served once to test the model. The performance is commonly estimated as the average of all assessments.<sup>174</sup>

Compared with split-sample validation, cross-validation can use a larger part of the sample for model development (e.g. 90%). This is an advantage. However, the whole cross-validation procedure may need to be repeated several times to obtain truly stable results, for example 50 times ten-fold cross-validation. The most extreme cross-validation is to leave out each patient once, which is equivalent to the jack-knife procedure.<sup>108</sup> With large numbers of patients, this procedure is not very efficient.

**Fig. 17.4** Cross-validation refers to assessing model performance consecutively in a random part of the sample, with model development in the other parts. With ten-fold cross-validation, deciles of the sample serve as validation parts



A problem is that cross-validation may not properly reflect all sources of model uncertainty, such as caused by automated variable selection methods. We provide an example at the book's website, where we consider the stability of a backward stepwise selection procedure in the large subsample from GUSTO-I (sample4,  $n=785$ , 52 deaths). A ten-fold cross-validation procedure suggests a quite stable selection of “important predictors”: SHO, A65, HIG, and HRT. In contrast, bootstrapping shows a much wider variability. The underestimation of variability is easily recognized for jack-knife cross-validation, where the development sample is identical to the full sample except for one patient. Hence, largely the same predictors will generally be selected in each jack-knife sample as in the full sample. Such model uncertainty can better be reflected with bootstrap validation.

### 17.2.4 Bootstrap Validation

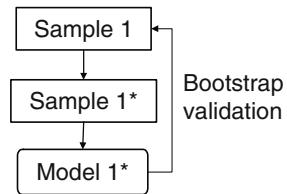
As discussed in Chap. 5, bootstrapping reflects the process of sampling from the underlying population (Fig. 17.5). Bootstrap samples are drawn with replacement from the original sample, reflecting the drawing of samples from an underlying population. Bootstrap samples are of the same size as the original sample.<sup>108</sup> In the context of model validation, 100–200 bootstraps may often be sufficient to obtain stable estimates, but in one simulation study we reached a plateau only after 500 bootstrap repetitions.<sup>401</sup> With current computer power bootstrap validation is a feasible technique for most prediction problems.

For bootstrap validation a prediction model is developed in each bootstrap sample. This model is evaluated both in the bootstrap sample and in the original sample. The first reflects apparent validation, the second test validation in new subjects. The difference in performance indicates the optimism. This optimism is subtracted from the apparent performance of the original model in the original sample.<sup>174,108,409,413</sup> The bootstrap was illustrated for estimation of optimism in Chap. 5.

Advantages of bootstrap validation are various. The optimism-corrected performance estimate is rather stable, since samples of size  $N$  are used to develop the model as well as to test the model. This is similar to apparent validation, and an advantage over split-sample and cross-validation methods. Compared with apparent validation, some uncertainty is added by having to estimate the optimism. When sufficient bootstraps are taken, this additional uncertainty is however negligible.

Moreover, simulations have shown that bootstrap validation can appropriately reflect all sources of model uncertainty, especially variable selection.<sup>401</sup> The bootstrap

**Fig. 17.5** Bootstrap validation refers to assessing model performance in the original sample for a model ( $\text{Model } 1^*$ ) that was developed in a bootstrap sample ( $\text{Sample } 1^*$ ), drawn with replacement from the original sample



also seems to work reasonably in high-dimensional settings of genetic markers, where the number of potential predictors is larger than the number of patients (“ $p>n$  problems”), although some modifications may be considered.<sup>374</sup> Disadvantages of bootstrap validation, and other resampling methods such as cross-validation, include that only automated modelling strategies can be used, such as fitting a full model without selection, or following an automated stepwise selection approach. In many analyses, intermediate steps are made, such as collapsing categories of variables, truncation of outliers or omission of influential observations, assessing linearity visually in a plot, testing some interaction terms, studying both univariate and multivariable  $p$  values, or assessing proportionality of hazards for a Cox regression model. It may be difficult to repeat all these steps in a bootstrap procedure.

In such situations, it may be reasonable to at least validate the full model containing all predictors to obtain a first impression of the optimism. For example, when we consider 30 candidate predictors, and build a final model with predictors that have multivariable  $p<0.20$  in a backward stepwise selection procedure, but after univariate screening with e.g.  $p<0.50$ , the optimism can be estimated by validating the full 30 predictor model. Another reasonable approximation for the optimism in this example may be to simply perform backward stepwise selection with  $p<0.20$ , ignoring the univariate screening. We would definitely be cheating if we validated the finally selected model and ignored all selection steps. In one study we found an optimism estimate of 0.07 for the  $c$  statistic when we replayed all modeling steps (based on univariate and multivariable  $p$  values) in contrast to 0.01 when we considered the final model as pre-defined.<sup>401</sup>

### 17.3 External Validation Studies

External validation of models is essential to support general applicability of a prediction model. Where internal validation techniques are all characterized by random splitting of development and test samples, external validation considers patients that differ in some respect from the development patients (Fig. 17.1). External validation studies may address aspects of historic (or temporal), geographical (or spatial), methodological, and spectrum transportability.<sup>222</sup> Historic transportability refers to performance when a model is tested in different historical periods. Especially relevant is validity in more recently treated patients. Geographic

transportability refers to testing in patients from other places, e.g. other hospitals or other regions, see e.g. a recent study in stroke patients.<sup>240</sup> Methodological transportability refers to testing with data collected by using alternative methods, e.g. when comorbidity data are collected from claims data rather than from patients' charts. Spectrum transportability refers to testing in patients who are, on average, more (or less) advanced in their disease process, or who have a somewhat different disease.<sup>222</sup> Spectrum transportability is relevant when models are developed in secondary care and validated in primary care, or models developed in randomized trials are validated in a broader, less-selected sample.

In addition to these aspects, we may consider whether external validation was performed by the same investigators who developed the model, or by investigators not involved at the development stage. If model performance is found adequate by fully independent investigators, in their specific setting, this is more convincing than when this result was found by investigators who also developed the model.

A simple distinction in types of external validation studies is shown in Table 17.2. We distinguish temporal validation (validation in more recent patients), geographic validation (validation in other places), and fully independent validation (by other investigators at other sites). Mixed forms of these types can occur in practice. For example, we validated a testicular cancer prediction model in 172 patients: 100 more recently treated patients from hospitals that participated in the model development phase and 72 from a hospital not included among the development centres.<sup>412</sup>

### **17.3.1 Temporal Validation**

With temporal validation, we typically validate a model in more recently treated patients. A straightforward approach is to split the development data into two parts: one part containing early treated patients to develop the model and another part containing the most recently treated patients to assess the performance.

Also, we may aim for a prospective application of the model in a specifically collected cohort. An example is from a study in patients suspected of Lynch syndrome (see Chap. 10).

**Table 17.2** Summary of types of external validation studies (based on Justice et al.<sup>222</sup>)

Method	Characteristics
Temporal validation	Prospective testing, more recent patients
Geographic validation	Multi-site testing
Fully independent validation	Other investigators at another site

### \*17.3.2 Example: Development and Validation of a Model for Lynch Syndrome

We aimed to predict the prevalence of Lynch-syndrome related genetic defects (*MLH1* or *MSH2* mutations) based in proband and relative characteristics (“family history”). Predictors included type of cancer diagnosis, age, and number of affected relatives. We developed a model with 898 patients who were tested at Myriad Genetics between 2000 and 2003. This model was tested in a validation sample containing 1,016 patients who were tested between 2003 and 2004 (Table 17.3).

In the validation sample, the outcome definition was slightly different, since not only mutations but also deletions of genes were assessed. This led to a slightly higher prevalence of mutations (15% at validation versus 14% at development),

**Table 17.3** Multivariable analysis of Lynch syndrome prediction model

Predictors	Development OR [95% CI]	Validation OR [95% CI]	Combined OR [95% CI]
<b>Proband</b>			
CRC 1	2.2 [1.9 – 2.5]	7.0 [6.0 – 8.1]	3.8 [3.6 – 4.1]
CRC 2+	8.2 [5.6 – 12]	37 [25 – 55]	16 [14 – 20]
Adenoma	1.8 [1.5 – 2.2]	1.5 [1.2 – 1.7]	1.5 [1.4 – 1.6]
Endometrial cancer	2.5 [2.1 – 3.1]	7.1 [6.1 – 8.2]	4.2 [3.9 – 4.6]
Other HNPCC cancer	2.1 [1.7 – 2.5]	1.4 [1.1 – 1.8]	1.8 [1.6 – 2.0]
<b>Family history</b>			
CRC in 1st/2nd degree <sup>a</sup>	2.3 [2.1 – 2.5]	3.0 [2.8 – 3.3]	2.6 [2.5 – 2.7]
CRC 2 in 1st degree	3.1 [2.6 – 3.6]	4.2 [3.6 – 4.8]	3.6 [3.4 – 3.8]
Endometrial cancer 1st/2nd degree <sup>a</sup>	2.7 [2.4 – 3.2]	2.7 [2.3 – 3.1]	2.6 [2.4 – 2.8]
Endometrial cancer 2 in 1st degree	6.5 [1.8 – 24]	26 [6.0 – 113]	12 [6.3 – 23]
Other HNPCC cancer	1.5 [1.4 – 1.7]	1.4 [1.2 – 1.6]	1.5 [1.4 – 1.6]
<b>Age at diagnosis</b>			
CRC <sup>b</sup>	1.5 [1.4 – 1.6]	1.4 [1.2 – 1.5]	1.4 [1.3 – 1.5]
Endometrial cancer <sup>c</sup>	1.3 [1.2 – 1.4]	1.4 [1.3 – 1.5]	1.3 [1.2 – 1.4]
<b>Model performance</b>			
<i>c</i> statistic	0.79 [0.76–0.83] <sup>d</sup>	0.80 [0.76–0.84] <sup>e</sup>	0.80 [0.77–0.83] <sup>d</sup>
Mean observed vs. predicted	14% vs. 14%	15% vs. 13% <sup>e</sup>	15% vs. 15%
Calibration slope	0.85 <sup>d</sup>	1.26 [1.03–1.49] <sup>e</sup>	0.94 <sup>d</sup>

Odds ratios of predictors are shown for the development ( $n=898$ ) and validation ( $n=1,016$ ) patients, as well as in the combined data set ( $n=1,914$ ) used for estimation of the final prediction model. Model performance includes assessment of discrimination and calibration

<sup>a</sup> Family history coded as first-degree + 0.5 second-degree relatives, with first-degree and second-degree relatives coded as 0, 1, 2+

<sup>b</sup> Age effect for colorectal cancer and/or adenoma in probands, and colorectal cancer in first- and second-degree relatives

<sup>c</sup> Age effect for endometrial cancer in probands, in first degree, and in second-degree relatives

<sup>d</sup> Internal validation by bootstrapping for *c* statistic and calibration slope

<sup>e</sup> External validation for *c* statistic, mean observed and predicted probabilities, and calibration slope

while the case-mix remained similar (mean predicted probability for validation sample, 13%). This difference in prevalence of the outcome could easily be adjusted by using a slightly higher intercept in the logistic regression model (+0.25, indicating 25% higher odds). The effects of the predictors were similar in the development and validation samples. Also, the discriminative ability remained at a similar level as at development with  $c$  statistic around 0.80.

The good performance at external validation may not be too surprising given that definitions of predictors were exactly the same. For the final model, both data sets were combined, such that 1,914 patients were analyzed, leading to smaller confidence intervals for the effects of the predictors and the  $c$  statistic.

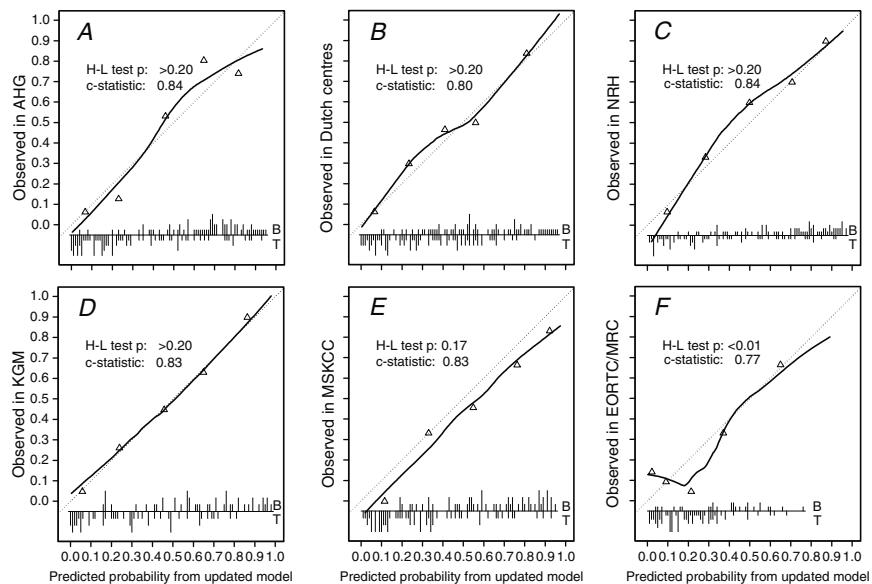
### 17.3.3 Geographic Validation

With geographic validation, we evaluate a predictive model according to site. Geographic validation can be seen as a variant of cross-validation. It could be labelled “leave-one-centre-out cross-validation.” Importantly, standard cross-validation makes splits in the data at random; with geographic validation the splits are not at random. Some examples are shown in Table 17.4. Geographic validation is often possible with collaborative studies, and more meaningful than a standard cross-validation.

A drawback of such geographical validations is that validation samples may get quite small, leading to unreliable results. Results may easily be overinterpreted, for example as showing that “the model was not valid for hospital X.” For example, in the testicular cancer case study, we found a systematic difference in calibration for patients treated in one centre (Fig. 17.6).<sup>467</sup> In fact, we perform multiple, small, subgroup analyses. Emphasis should be on general consistency (if this is observed) rather than on differences that will always occur with small numbers. On the other hand, remarkable findings for a specific setting may indicate a need for further validation before applying the model in this setting, and trigger further research.

**Table 17.4** Examples of studies with external validation according to site (“leave-one-centre-out cross-validation”)

Model	Development	Validation	Site
Testicular cancer	$6 \times 5$ groups	$6 \times 1$ group	A group consisted of a single hospital or a previously published patient series
<i>Chlamydia trachomatis</i>	$4 \times 3$ regions	$4 \times 1$ region	Municipality health centres organizing regional case finding
DRASTIC study	$5 \times 4$ hospitals	$5 \times 1$ hospital	Large hospitals participating in an RCT and a category “other”



**Fig. 17.6** Results of external validation by centre for the testicular cancer prediction model. We note  $c$  statistics around 0.8 for all sites, and non-significant miscalibration according to the Hosmer-Lemeshow test (H-L test), except in graph(F). B, benign tissue; T, tumor<sup>467</sup>

### 17.3.4 Fully Independent Validation

Finally we mention external validation by independent investigators (“fully independent validation”). Other investigators may use slightly different definitions of predictors, outcome, and study patients that were differently selected compared with the development setting. An example of that is a prostate cancer model developed for clinically seen patients and validated in patients selected by a systematic screening program (European Study on Prostate Cancer, ERSPC).<sup>424</sup> Here, case-mix seemed similar, but a severe underestimation of relatively innocent (“indolent”) cancer probability was found (Table 17.5). This phenomenon was addressed by a new logistic model intercept, while keeping the regression coefficients close to their original values.

Similarly it was found that a prediction model for the selection of patients undergoing in vitro fertilization for single embryo transfer needed an adjustment when a model developed at one hospital was applied in another centre. Again, a systematic difference remained even after adjustment for well-known and important predictors.<sup>206</sup> This difference in average outcome (“calibration-in-the-large”) is an important motivation for recalibration of model predictions as a simple but important updating technique (see Chap. 19).

Some examples of fully independent validation studies with their main conclusions are listed in Table 17.6. It seems that fully external validation studies often

**Table 17.5** Prediction accuracy of three previous nomograms for indolent prostate cancer developed by Kattan et al.<sup>227</sup> for 247 ERSPC patients<sup>424</sup>

Performance parameter	Kattan et al.	Nomogram		
		Base	Medium	Full
Area under the ROC curve	Kattan et al.	0.64	0.74	0.79
	ERSPC	0.61 [0.54–0.68]	0.72 [0.66–0.78]	0.76 [0.70–0.82]
Calibration-in-the-large	Predicted	24%	22%	15%
	Observed	49% [43–55%]	49% [43–55%]	49% [43–55%]
Calibration slope	Predicted	1	1	1
	Observed	0.78 [0.32–1.24]	0.87 [0.55–1.19]	1.07 [0.74–1.40]

Base: Serum PSA + clinical stage + biopsy Gleason grade 1 and 2

Medium: Base + US volume + %positive cores

Full: Base + US volume + mm cancerous tissue + mm non-cancerous tissue

**Table 17.6** Examples of studies with fully independent external validation

Model	Development	Validation	Conclusions
Prostate cancer	Two hospitals <sup>227</sup>	Screening setting (ERSPC) <sup>424</sup>	Intercept problem
Aneurysm mortality	One hospital + meta-analysis <sup>421</sup>	UK small aneurysm trial <sup>54</sup> and another hospital <sup>231</sup>	Missing predictors; poor/moderate performance
Renal artery stenosis	RCT <sup>243</sup>	One French hospital <sup>278</sup>	“Reasonably valid”

provide more unfavourable results than a temporal or geographical external validation. This is also illustrated by other examples of fully independent validation, showing generally poor results, in a review by Altman and Royston.<sup>13</sup>

### 17.3.5 Reasons for Poor Validation

Unfavourable results at validation may often be explained by inadequate model development. Sample size may have been relatively small, or patients were selected from a single centre. This was for example noted in a review of over 25 models in traumatic brain injury.<sup>333</sup> Also, statistical analysis may often have been suboptimal, e.g. with stepwise selection in relatively small samples with many potential predictors, and no shrinkage of regression coefficients to compensate overfitting.

Other explanations include true differences between development and validation settings, especially in coding of predictors and outcome. The problem of transportability of models that incorporate laboratory test results was already recognized in the 1980s for a prediction rule for jaundice, where units of measurement were not consistent.<sup>379</sup> Indeed, the validation of a model that was previously developed by others is often more difficult than anticipated. If a nomogram is presented with some

non-linear terms, it is not so easy to derive a formula to calculate outcome predictions for new patients. So, it is quite likely that errors are made at such external validation studies. Units of measurement and the intercept value require special attention. Contacting the authors may help to prevent mistakes.

Moreover, variables required for a model may not be available at validation. A constant value can be filled in (e.g. the mean or median), but obviously this limits the external performance of a model. For example, a Dutch model for abdominal aneurysm mortality was validated in the UK small aneurysm study, while two of the seven predictors were not available.<sup>54</sup> In a validation study with patients from Rotterdam, all predictors except one were available and a better external performance was found.<sup>231</sup>

## 17.4 Concluding Remarks

We considered several approaches to internal and external validation. For internal validation, bootstrapping appears most attractive, provided that we can replay all modelling steps. This may sometimes be difficult, e.g. when decisions on coding of predictors, and selection of predictors are made in the modelling process. Several variants of bootstrapping are under study, which may be more efficient than the procedure described here. Also, the optimism may in fact be larger than estimated by bootstrapping when the ratio of predictors considered to the sample size is very unfavourable, such as in genetic marker research.<sup>292,220,221</sup>

Any internal validation technique should be seen as validating the modelling process rather than a specific model.<sup>181</sup> For example, when a split-sample validation is followed, e.g. to convince physicians who are skeptical of modern developments, the final model should still be derived from the full sample. It would be a waste of precious information if the final model were only based on a random part of the sample. Differences in regression coefficients will generally be small, since the split was at random, and the data have overlap, but the estimates of the full sample will be more stable. If a stepwise selection procedure was followed in the random sample, it should be repeated in the full sample. This may result in a different model specification, but this is preferable to sticking to results from only part of the available data.

The same reasoning holds for cross-validation and bootstrap validation. Especially with bootstrap validation we may well illustrate the instability of stepwise selection procedures (see Chap. 11). The final model may only be selected in a few of the bootstrap samples. This model uncertainty has to be taken into account in the optimism estimate for the final model.

If external validation has been performed, we may similarly define the final model from the combined data set. This was for example done in the Lynch syndrome case study (Table 17.3).<sup>25</sup> This combination of data implies that the two samples represent the same population, which is not necessarily the case. If relevant differences are found, a setting-specific intercept or setting-specific interaction terms may be included (see Chaps. 19–21).

## Questions

### 17.1 Stability of internal validation techniques (Table 17.1)

- (a) Split-sample validation is notoriously unstable. In contrast, apparent validation and bootstrap validation share stability in the estimation of model performance. Do you agree?
- (b) Cross-validation eventually uses 100% of the sample for validation; why might multi-fold cross-validation help?

### 17.2 Interpretation of external validation (Fig. 17.6)

Fig. 17.6 can be interpreted in different ways. One perspective is to emphasize the similarity in performance between settings. Alternatively, we might focus on graph E and F, which show a systematic miscalibration. What would be your view on the performance of this centre? Consider a fixed effect and random effect perspective (see also Chap. 20).

### 17.3 Problems with internal validation<sup>50</sup>

Interpret the published results on “internal validation” in Table 2 of an Ann Int Med paper (<http://www.annals.org/cgi/reprint/143/4/265.pdf>).

- (a) What do you think went wrong?
- (b) What do you think of the interpretation provided in the text?
- (c) What do you think about the “corrected Table 2,” published as an erratum?  
<http://www.annals.org/cgi/reprint/144/8/620.pdf>

# Chapter 18

## Presentation Formats

**Background** The presentation of a prediction model deserves careful attention. Epidemiologic regression analyses commonly concentrate on estimation of relative effects, and present tables with odds ratios or hazard ratios, and their confidence intervals. Such tables are usually not sufficient to calculate absolute risks, which requires a model intercept (for continuous or binary outcomes) or a baseline hazard (for survival outcomes). We need to separate presentations that generate predictions (“clinical prediction models”) from presentations that generate advice for a decision (“clinical decision rules”). Various presentation formats are possible for prediction models and for decision rules, some of which will be discussed in this chapter. We illustrate the creation of some formats at a technical level for the testicular cancer case study.

### 18.1 Prediction vs. Decision Rules

A clinical prediction model provides an estimate of absolute risk, based on the combination of several patient characteristics. For a good prediction model, the prediction for an individual patient can span a wide range, from relatively low to relatively high. The interpretation of the prediction and any actions following from it are left to the treating physician and/or the patient. We can also present a decision rule, where a specific course of action is suggested depending on the combination of patient characteristics. Decision rules are hence not synonymous with prediction models.<sup>344</sup> Decision rules require more subject matter input, e.g. from clinical experts, especially on the choice of a cutoff point for predictions (see Chap. 16).

Some have argued that presentation as a decision rule leads more easily to a wide application of a model. Examples include decision rules for traumatic injuries to the ankle or foot, knee, cervical spine, and head (“Ottawa rules”). The developers of the rules suggest substantial impact, and conclude that emergency physicians should adopt these clinical decision rules to standardize care and reduce costs.<sup>344</sup> Decision rules may also be a natural extension of heuristic rules that humans tend to use.<sup>135</sup>

We discuss several options for presentation of prediction models and clinical decision rules (Table 18.1). Formats differ on aspects such as the medium by which

**Table 18.1** Some examples of presentation formats for clinical prediction models and clinical decision rules

Rule	Characteristics	Pros	Cons	Example
<i>Prediction models</i>				
Regression formula	Simple, follows directly from analysis	Can be implemented in computerized format	Leaves work to the user; difficult to calculate confidence intervals	Predicted response dose, formula in abstract <sup>29</sup>
Spreadsheet	Includes exact calculations, exact 95% confidence intervals	Standard software, familiar to many	Needs user to open specific file	Survival after surgery for lung cancer <sup>36</sup>
Prognosis program	Includes exact calculations, exact 95% confidence intervals	Easy to download and install	Needs user to get acquainted with specific software	Oncologiq ( <a href="http://oncologiq.nl/">http://oncologiq.nl/</a> )
Nomogram	Includes quite exact calculations, approximate 95% confidence intervals	Quite exact predictions	Difficult to understand at first sight	Prostate cancer recurrence <sup>226</sup>
Score chart	Includes approximate calculations, approximate 95% confidence intervals	Easy to understand	Approximate predictions	DRASTIC prediction rule for renal artery stenosis <sup>243</sup>
Table	Includes averaged calculations, approximate 95% confidence intervals can be added	Very easy to understand and use	Predictions by predictor combination; continuous predictors have to be categorized	Framingham risk equation to identify candidates of statin therapy <sup>487</sup>
Specific formats	Based on specific interest of audience	Should appeal specifically to target audience	Less easy to understand for non-target audience	Relevance of pre- and post-chemotherapy mass size in testicular cancer prediction example <sup>49</sup>
<i>Decision rules</i>				
Regression tree	Simple, large groups	Very easy to understand and use	Unstable if based on limited data	Goldman diagnostic index for acute MI <sup>142</sup>
Score chart rule	Score based on highly rounded coefficients	Rule simple to understand	Continuous predictors have to be categorized	CT rule for minor head injury <sup>391</sup>
Survival by group	Simple, large groups	Very easy to understand and use	Inaccurate predictions	IGCC classification for testicular cancer <sup>5</sup>
Meta-model tree	Simple, large groups	Very easy to understand and use	Stable but cut-offs based on distribution of risk rather than decision-analytic considerations	Testicular cancer group with >70% benign tissue <sup>418</sup>

they are presented (on paper, or electronically), the level of detail in the predictions (rough indications of risk, or exact probabilities), presence of indicators of uncertainty (e.g. 95% confidence intervals around predictions), and user-friendliness (simple to complex formats).

## 18.2 Clinical Prediction Models

Clinical prediction models provide probabilities of diagnostic or prognostic outcomes. We discuss detailed presentations with a regression formula, a nomogram, or a score chart (Table 18.1).

### 18.2.1 Regression Formula

Clinical prediction models can be presented in various formats. The simplest form is to present the final regression formula. An example is the regression formula presented in the abstract of a study in anovulatory infertile women (Box 18.1).<sup>209</sup>

Calculation of predictions with a regression formula incorporates two steps. The first step is to calculate the linear predictor. The linear predictor is central to regression models such as linear, logistic, polytomous, Cox, or parametric survival models. In the linear predictor, we multiply regression coefficients with predictor values. Definitions and encoding of the predictors have to be clear to the user. For binary predictors, a 0/1 coding is convenient, which makes that patients without a characteristic have a score of zero. For categorical predictors, dummy variables are usually constructed. The reference category for these dummy variables can be based on frequency (e.g., the most common category is the reference), or on clinical considerations. For continuous variables, the units have to be clear. For example units for concentrations are important (by weight, e.g. mg/dl, or molecular count, e.g. mmol/l). Also, continuous predictors are sometimes centred to the mean value, which should then be subtracted from the original value when using the regression formula.

The second step is to translate the linear predictor values to units on the outcome scale. For a logistic model, we use the logistic transformation to estimate probabilities

**Box 18.1** Regression formula for prediction of the individual follicle-stimulating hormone threshold<sup>209</sup>

FSH response dose = 4 body mass index (in kg/m<sup>2</sup>) + 32 clomiphene citrate resistance (yes = 1 or no = 0) + 7 initial free insulin-like growth factor-I (in ng/mL) + 6 initial serum FSH level (in IU/L) – 51

of the outcome ( $p(Y=1)$ ). For survival, we can estimate survival probabilities, e.g. at 1,2, or 5-year, median survival, or other quantiles of survival. With a Cox model, we need the baseline hazard function to estimate these survival probabilities  $S(t) = h_0(t) \times \exp(\text{linear predictor})$ , where  $h_0(t)$  indicates the baseline hazard function for time  $t$ . Parametric survival models have an intercept similar to other regression models. Predictions from such parametric models are straightforward to calculate, and are more stable at the end of follow-up (Chap. 4).

Heuristic shrinkage can be incorporated in the translation from linear predictor to predictions. One way is to standardize predictor values, such that the average of the linear predictor is zero.<sup>459</sup> We can then multiply the linear predictor with the shrinkage factor. The average of the predictions will then remain reasonably correct. However, a systematic error will arise when the range of predictions is wide, or the shrinkage severe, because of the non-linearity in the translation from linear predictor to prediction. As an alternative, we can shrink regression coefficients and re-estimate the intercept for proper calibration-in-the-large.

Regression formulas can serve as the basis for computerized implementation, in PDAs, mobile phones, hospital information systems or electronic patient records, webpages, or spreadsheets (see <http://www.nomograms.org>). One example is a spreadsheet to show survival after surgery for lung cancer, where a model is presented, including seven predictors. The predicted survival curve was calculated according to the individual predictor values, with an approximate 95% confidence interval.<sup>36</sup>

Similarly, specific programs can be developed for presentation of the prediction model. An example is the OncologIQ program (<http://oncologiq.nl/>). The program provides individualized survival predictions for cancer patients in graphical and tabular format, similar to what can be done with a spreadsheet. The program also provides documentation on the prediction model. Finally, Web-based calculators become more and more common. A wide collection is provided commercially at <http://www.infopoems.com/>.

### \*18.2.2 *Confidence Intervals for Predictions*

Uncertainty around predictions for linear regression models can be presented with *confidence intervals* and *prediction intervals*. Confidence intervals indicate the uncertainty around the average and become very small with very large sample size. For example, a growth curve predicting length by age will have a very tight confidence interval when based on thousands of adolescents. Prediction intervals for individual subjects will however remain of substantial size because of the variability in the population.

For predicted probabilities, the fact that a probability is estimated reflects the inherent uncertainty of the prediction process. Confidence intervals around predicted probabilities can become quite small with large sample size, but the prediction for an individual remains a probability. With regression analysis, predictions can

approach, but never reach, 0 or 1. Uncertainty in survival can be indicated around probabilities at time points in follow-up. We can also indicate uncertainty around survival duration, e.g. median survival surrounded by 2.5%, 5%, 10%, 25%, 75%, 90%, 95%, 97.5% quantiles. The latter quantiles will always cover a substantial width, even with infinite sample size, similar to the prediction intervals in linear regression.

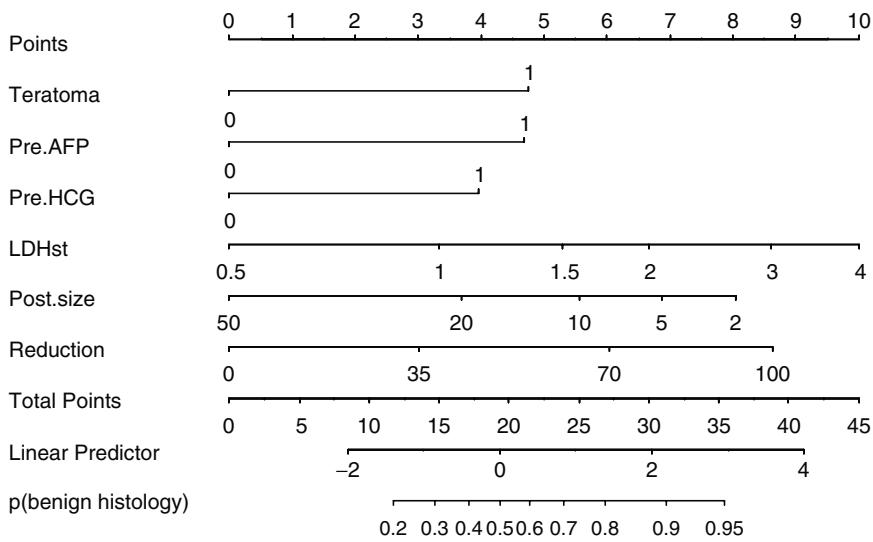
Confidence intervals are only a valid indication of the uncertainty of the prediction model if there is no systematic bias in the predictions. The total uncertainty in a prediction is the sum of systematic and random errors. Miscalibration of predictions is an example of a systematic error, which may be due to various differences between the development setting and the setting where the model is applied, e.g. encoding of predictors, missed predictors with different distributions, and truly differential effects. Hence we must be cautious in the interpretation of predictions when the confidence interval is small because of a large sample size. On the other hand, a model derived from a small data set will show large confidence intervals, which is a useful warning against overinterpretation of predictions. Further, one might argue that the values of the predictions remain of primary interest for decision making, and uncertainty is less relevant. If we cannot make a better estimate than the one provided by the model, following that estimate is the best we can do, even when the estimate is uncertain.

Technically, confidence intervals are calculated with the standard error of a prediction. The standard error is calculated from the covariance matrix of the regression model. If shrinkage was applied, it may be reasonable to still use the covariance matrix of the original, unshrunken model. With a penalized model we can use the covariance matrix of the penalized model, which will result in slightly smaller standard errors of predictions than the original model.

Every combination of predictor values leads to a different standard error of the prediction. The same linear predictor value can have a different standard error, since different combinations of predictor values may lead to the same sum. This is handled easily in a regression formula and in a spreadsheet, but more complicated in other presentation formats such as nomograms and score charts. In the latter formats, using the mean or median standard error can be considered to indicate uncertainty for a given linear predictor value.

### 18.2.3 *Nomograms*

Nomograms are graphical presentations of a prediction model, and have a long history in the precomputer era, with a more recent role as presentation format of a clinical prediction model.<sup>269</sup> Again, two steps are discerned. The calculation of the linear predictor is essentially as for a regression formula. The nomogram has a reference line for reading scoring points (e.g. 0–100 or 0–10, Fig. 18.1). The user manually totals the points and the total corresponds linearly to the linear predictor. The second step is the transformation of the linear predictor to predictions, which



**Fig. 18.1** Nomogram for benign histology based on six predictors in a penalized logistic regression model. Teratoma, teratoma elements in the primary tumor; Pre.AFP, pre-chemotherapy AFP normal / elevated; Pre.HCG, pre-chemotherapy HCG normal / elevated; LDHst, standardized pre-chemotherapy LDH (LDH divided by upper limit of normal values; 1 means values equal to upper normal); Post.size, post-chemotherapy mass size in mm. Reduction, % reduction in size during chemotherapy, e.g. 50–10mm =80%

can be read at the bottom of the nomogram. Predictions can be in the form of a probability, median survival, or other quantities. Harrell's nomogram function is a valuable tool to develop these presentations.<sup>174</sup>

Nomograms have especially been promoted for urological tumors, such as prostate cancer, by Kattan et al.<sup>351,72</sup> Advantages are several. The relative importance of the predictors can be judged by the length of the lines with nomogram scores, provided that the predictor values on the lines are based on the distribution of the predictor in the data under study. Hence, the reader obtains a visual impression of the relevance of a predictor in the model, relative to the other predictors. Furthermore, interaction terms can be handled well. Separate lines are constructed, such that always only one axis has to be read to obtain a score corresponding to a predictor value. Complex models, e.g. survival models with time-dependent covariates, can also be presented as nomograms.<sup>228</sup> The translation of the total points to the probability or survival scale is relatively easy. Scales can be extended with approximate confidence intervals (with the median standard error per decile of predicted risk), or additional scales for the outcome, e.g. 25 and 75 survival percentiles.

Disadvantages of nomogram presentations include the relative complexity at first sight, the inaccuracy of readings when many predictors are included, and the inaccuracy of translation to the final outcome. It is not clear yet whether clinicians prefer nomograms to other formats such as score charts.

### \*18.2.3.1 Instructions for nomogram

*Instruction to physicians using the model in their care:* Determine the patient's value for each predictor, and draw a straight line upwards to the points axis to determine how many points towards benign histology the patient receives. Sum the points received for each predictor and locate this sum on the total points axis. Draw a straight line down to find the patient's predicted probability of benign histology.

*Instruction to patient:* "Mr. X, if we had 100 men exactly like you, we would expect that the chemotherapy was fully successful in approximately <predicted probability from nomogram × 100>, as reflected in fully benign disease at surgical resection of your abdominal lymph nodes." (text based on Kattan et al. <sup>227</sup>)

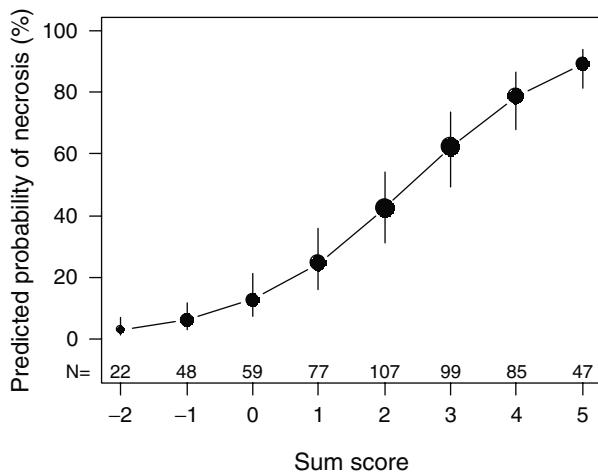
## 18.2.4 Score Chart

Score charts are another simple presentation format for clinical prediction models. The first step is to round regression coefficients to scores. A simple approach is to multiply coefficients by ten, and round them. However, we can often find lower numbers for multiplication that still allow for a sufficiently refined prediction. Some define scores by dividing through the smallest coefficient of a binary predictor, which has then by definition a score of 1. The other predictors get rounded scores. This procedure is suboptimal, since it capitalizes on the estimate of one coefficient. This leads to unnecessary extra uncertainty in the rounded coefficients. It is preferable to search for a common denominator across all coefficients. A score chart for the testicular cancer model is shown in Table 18.2, with corresponding probabilities in Fig. 18.2.

**Table 18.2** Score chart for estimation of the probability of benign tissue after chemotherapy for metastatic testicular cancer with continuous predictors

Characteristic	Scores						Sum score
Primary tumor							
Teratoma elements		1					
Prechemotherapy tumor markers							
AFP elevated		1					
HCG elevated		1					
LDH times normal							
Values	0.6	1	1.6	2.5	4	6	
Scores <sup>a</sup>	-0.5	0	0.5	1	1.5	2	
Postchemotherapy size (mm)							
Values	<5	10	20	40	70		
Scores <sup>a</sup>	+0.5	0	-0.5	-1	-1.5		
Reduction in size							
Values	Increase	0%	50%	100%			
Scores <sup>a</sup>	-1	0	1	2			
Total score (add all)							...
Probability of benign tissue and 95% CI from Fig. 18.2							...% [...% – ...%]

<sup>a</sup>Intermediate scores can be estimated with linear interpolation



**Fig. 18.2** Probability of necrosis (benign tissue) in relation to the sum score from Table 18.2. Number of patients with each score are indicated at the bottom, and reflected in the size of the dots. 95% confidence intervals are shown around the predicted probabilities

### 18.2.5 Tables with Predictions

Predictions can sometimes be presented in table format, but this may require some simplifications of the model. Especially, we need to categorize continuous predictors, which implies a loss of information (see for an example from the Framingham Heart Study<sup>487</sup>: <http://www.nhlbi.nih.gov/about/framingham/riskabs.htm>). Also, the Adult Treatment Panel III presents a number of tools for detailed calculations on the Web (<http://hp2010.nhlbihin.net/atpiii/calculator.asp?usertype=prof>). An interactive risk assessment tool is presented to estimate 10-year risk for “hard” coronary heart disease outcomes (myocardial infarction and coronary death), and calculators are downloadable for use on a Palm OS or as a spreadsheet.

A simple table has been successful in providing indications for statin treatment. This table defines absolute 10-year risks of cardiovascular events by smoking status, hypertension, diabetes, cholesterol to HDL-cholesterol ratio, and sex. Moreover, colors were added corresponding to treatment advice: treat with a statin, do not treat with a statin, or treat in the presence of a family history of cardiovascular disease. This presentation was followed in several Dutch guidelines for prevention of cardiovascular disease.

A tabular presentation was considered as a simple way to present the testicular cancer model.<sup>416</sup> The advantage is that decision guidelines can easily be coupled to the predictions. In this case, a clear treatment advice was linked to predictions over 90% (follow-up) and prediction below 60% (resection, Table 18.3). In between is a gray area, where treatment decision making is not straightforward and may depend on various factors, such as feasibility of close follow-up, experience of surgeon, and the

**Table 18.3** Probability of benign tissue in relation to the sum of five favourable characteristics and mass size for the testicular cancer case study

Residual mass size (mm)	Sum of favourable characteristics <sup>a</sup>					
	0	1	2	3	4	5
0–9				$p > 60\%$	$p > 70\%$	$p > 80\%$
10–19			Resection			
20–29				$p > 60\%$		
30–49					$p > 70\%$	$p > 80\%$
$\geq 50$ or increased mass	$p \leq 60\%$					

<sup>a</sup>Sum of five characteristics: primary tumor teratoma negative; pre-chemotherapy AFP normal; Pre-chemotherapy HCG normal; Pre-chemotherapy LDH elevated; reduction in mass size  $\geq 70\%$

technical difficulty of the resection. All patients with a large ( $\geq 50$  mm) or increased mass should undergo resection, as well as all with less than two favorable characteristics. This tabular format however implies a severe loss of discriminative ability ( $c$  decreases from 0.839 to 0.773).

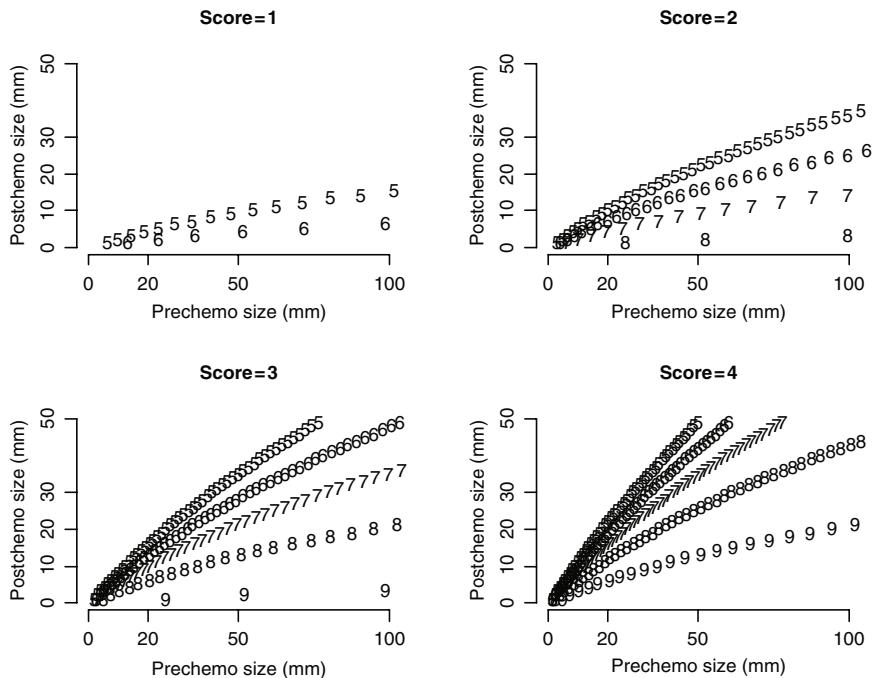
### \*18.2.6 Specific Formats

Specific formats may appeal to certain audiences. For example, radiologists are important in the monitoring of treatment of cancer. They usually compare images obtained during or after treatment with images made before treatment. Hence, a presentation of prediction model might focus on the information in such images. This was attempted for the relevance of pre- and post-chemotherapy mass size in the testicular cancer prediction example (Fig. 18.3).<sup>419</sup> We created iso-probability curves for combinations of pre- and post-chemotherapy mass size, based on the underlying logistic regression function. The graph shows that the post-chemotherapy size was more relevant than the pre-chemotherapy size; probabilities increase sharply with smaller post-chemotherapy size. This is caused by a direct effect of post-chemotherapy size, in combination with a strong effect of reduction in size (larger reductions with smaller post-chemotherapy sizes).

## 18.3 Case Study: Clinical Prediction Model for Testicular Cancer Model

### 18.3.1 Regression Formula from Logistic Model

In the testicular cancer case study we concentrate on the prediction of a benign histology (“necrosis”) after chemotherapy for metastatic disease. A logistic regression



**Fig. 18.3** Predictions for benign histology based on pre-chemotherapy and post-chemotherapy mass size, and by score of four prognostic characteristics (no teratoma elements in primary tumor, normal AFP, normal HCG, or elevated LDH). Lines are labelled with 5 for 50%, 6 for 60%, 7 for 70%, 8 for 80%, and 9 for 90% probability of benign tissue. Patients with a score of zero always had predicted probabilities below 50%. For example, a post-chemotherapy size of 20 mm after a prechemotherapy size of 100 mm results in a probability around 90% when the score is 4, but a probability around 65% when the score is 2

**Table 18.4** Regression coefficients in logistic regression models for post-chemotherapy histology in testicular cancer, with uniform shrinkage ( $s = 0.95$ ), penalized ML estimation (penalty factor 4), and the scores for a score chart (multiplication by 10, or 10/8 to achieve simpler scores)

Predictor	Coef <sub>orig</sub>	Coef <sub>shrunken</sub>	Coef <sub>pen</sub>	10* Coef <sub>pen</sub>	10/8* coef
Teratoma	0.909	0.872	0.873	9	1
Pre.AFP	0.903	0.865	0.860	9	1
Pre.HCG	0.783	0.750	0.729	7	1
Log(LDHst)	0.985	0.944	0.884	9	1
Sqrt(Post.size)	-0.292	-0.279	-0.261	-3	0
Reduction (%)	0.016	0.015	0.016	0	0

model with six predictors was fitted. Bootstrapping suggests a shrinkage factor of  $s$  of  $\sim 0.95$ . Further, a penalty factor of 4 was used in a penalized maximum likelihood procedure (Table 18.4).

The formula with shrunk coefficients is:

$$\text{lp}_{\text{shrunk}} = -0.97 + 0.84 \times \text{Teratoma} + 0.83 \times \text{Pre.AFP} + 0.72 \times \text{Pre.HCG} + 0.91 \times \ln(\text{LDHst}) - 0.27 \times \text{sqrt}(\text{Post.size}) + 0.014 \times \text{Reduction},$$

where Teratoma = 1 if teratoma elements were present in the primary tumor, 0 otherwise; Pre.AFP = 1 if pre-chemotherapy AFP was elevated, 0 if normal; Pre.HCG = 1 if pre-chemotherapy HCG was elevated, 0 if normal;  $\ln(\text{LDHst})$  refers to the natural logarithm of the pre-chemotherapy LDH value, standardized to the upper limit of local normal limits;  $\text{sqrt}(\text{Post.size})$  refers to the square root of the post-chemotherapy size in mm; Reduction refers to the reduction in size during chemotherapy in %.

The formula with penalized coefficients is:

$$\text{lp}_{\text{penalized}} = -1.1 + 0.87 \times \text{Teratoma} + 0.86 \times \text{Pre.AFP} + 0.73 \times \text{Pre.HCG} + 0.88 \times \ln(\text{LDHst}) - 0.26 \times \text{sqrt}(\text{Post.size}) + 0.016 \times \text{Reduction}$$

The formula to calculate predicted probabilities is simply:

$$P = \frac{1}{(1 + \exp(-\text{lp}))}$$

If we want to calculate confidence intervals, we need to consider the covariance matrix, which looks like:

	Intercept	Teratoma	Pre.AFP	Pre.HCG	LDHst	Post.size	Reduction
Intercept	0.3700	-3.1e-02	-3.0e-02	-1.4e-02	0.04200	-0.04300	-2.7e-03
Teratoma	-0.0310	4.6e-02	5.7e-03	-2.4e-03	-0.00150	0.00100	5.6e-05
Pre.AFP	-0.0300	5.7e-03	5.4e-02	-9.2e-03	0.00320	0.00130	8.6e-05
Pre.HCG	-0.0140	-2.4e-03	-9.2e-03	5.3e-02	0.01100	-0.00160	8.0e-06
LDHst	0.0420	-1.5e-03	3.2e-03	1.1e-02	0.04400	-0.00970	-4.1e-04
Post.size	-0.0430	1.0e-03	1.3e-03	-1.6e-03	-0.00970	0.00660	3.1e-04
Reduction	-0.0027	5.6e-05	8.6e-05	8.0e-06	-0.00041	0.00031	2.7e-05

The square root of the diagonal indicates the variance of the regression coefficients. The off-diagonal numbers are used for the calculation of variance of specific combinations of predictor values:  $\text{SE}_{\text{prediction}} = \text{transpose}(X) \times \text{covariance matrix} \times X$ . A detailed example is provided for the EuroSCORE, which predicts cardiac operative risks.<sup>286</sup> The predictions for the testicular cancer histology are presented with 95% confidence in an Excel spreadsheet, which is freely available at the Web.

### 18.3.2 Nomogram

A nomogram can easily be constructed with Harrell's Design library:

```
nomogram(full.pen, fun=plogis, lp=T, lp.at=c(-2,0,2,4),
LDHst=c(.5,1,1.5,2,3,4), post.size=c(50,20,10,5,2),
Reduction=c(0, 35, 70, 100),
fun.at=c(seq(.1, .9, by=.1), 0.95), funlabel="p(benign histology)",
vnames="lab", maxscale=10)
```

We used a maximum of 10 points in Fig. 18.1, accepting some loss in accuracy in summing the points corresponding to each predictor value. The total points correspond linearly to the linear predictor, which correspond to  $p(\text{benign histology})$  through the logistic transformation.

### \*18.3.3 Score Chart

A score chart for the testicular cancer prediction model was shown in Table 18.2. We consider the following steps with some technical details:

1. multiply and round regression coefficients of binary predictors and dummy variables of categorical predictors
2. search scores for continuous predictors, continuous or categorized
3. estimate the multiplication factor for the scores
4. estimate the intercept and present as score chart

#### 18.3.3.1 Rounding coefficients

The first step is to multiply regression coefficients and round them to scores. A simple approach is to multiply coefficients by ten. But we can also search for smaller rounded scores. For example, the coefficients of the binary predictors Teratoma, Pre.AFP, and Pre.HCG were quite similar (~0.8 for the penalized coefficients, Table 18.4). We multiply by 10/8 to give these three predictors each a score of 1. In general, we can often find lower numbers for multiplication that still allow for a refined prediction.

The R command for a search of scaling factor for the penalized coefficients was:

```
for(i in seq(1, 10, by=0.5)) {
cat("Multiply by:", 10/i, "i=", i, "Coefs:",
round(full.pen$coefficients[-1] * 10 / i), "\n") }
```

#### 18.3.3.2 Scores for continuous predictors

Three continuous predictors are considered in the prediction model, with a log transformation (LDHst), square root transformation (Post.size), and linear (Reduction in size during treatment). These continuous predictors need to be rescaled in such a way that a one point change in score corresponds approximately to a 10/8 increase in logodds. We

first treat these predictors as continuous variables, and later consider categorization as a further simplification. We go through several steps:

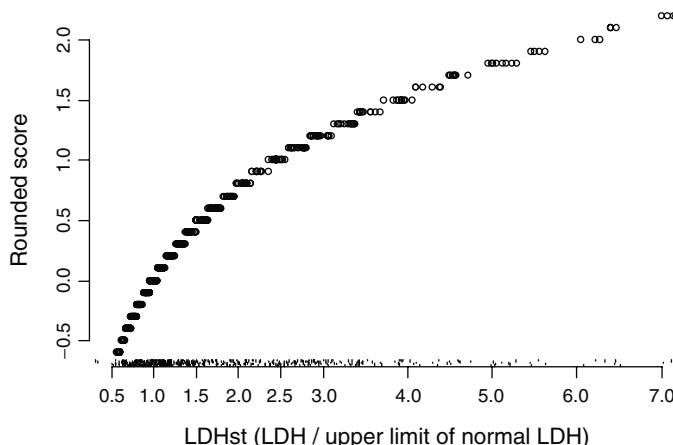
- Score 0: convenient? With a predictor such as age, it is often strange to have a score of 0 at 0 years. We may need to change the reference point to a sensible value, which we subtract from the original value.
- Score points: We aim to find values of the predictors where the scores are +1, +2, etc. points, depending on the distribution of predictor values and the predictor effects.
- Intermediate scores: We have to think about intermediate values, which have e.g. a score of +0.5 points. Other values can then be scored by linear interpolation.

We consider these three steps for the three continuous predictors in the testicular cancer histology prediction model (LDH, postchemotherapy size, and reduction in size).

## LDH

A score of zero is obtained for  $\log(1)$ , i.e. when LDH values are equal to the upper limit of normal. This is convenient as a reference.

The log transformed variable happens to have a penalized coefficient of 0.88, which can be rounded directly to one point when multiplied by  $10/8$  ( $0.88 \times 10/8 = 1.1$ ). Hence, when the natural log = 1, the score is one point. The LDHst value then is  $\exp(1) = 2.7$ . So an LDH value 2.7 times the upper limit of normal has a score of 1. The score of 1.1 was actually somewhat larger than 1. Hence, we can set one point at 2.5 times normal, and a score of zero at 1 times normal levels. A score of 2 points is achieved at  $2.5 \times 2.5$  is  $\sim 6$  times normal levels.



**Fig. 18.4** Relation between LDHst and rounded values of the score in the testicular cancer model. We can read the scores for increasing values of LDHst. The density of the data is plotted at the bottom of the graph

A more general approach is to study the relation between the score corresponding to LDH values. We calculate the score (rounded at one decimal) for the LDHst values (Fig. 18.4). We note that intermediate levels of LDH are common; the median value was 1.36 times normal. We note that a score of +0.5 is found at 1.6 times normal LDH, and we show this value in the score chart; a score of +1.5 is found at 4 times normal LDH. Scores to be shown for LDHst are 0.6, 1, 1.6, 2.5, 4, and 6 times upper limit of normal, with scores of -0.5, 0, 0.5, 1, 1.5, and 2 respectively.

### Post-size and Reduction in Size

Similarly, points were chosen for post-chemotherapy sizes (<5 mm, +0.5; 10mm, 0; 20 mm, -0.5; 40 mm, -1, and 70 mm, -1.5 point), and for reduction in size (<0%: -1, 0%, 0; 50%, 1; 100%, 2 points). Details are presented at the Web.

We check the regression coefficients for each rescaled predictor with a logistic regression model:

```
lrm(NEC ~ Teratoma+Pre.AFP+Pre.HCG+LDHr+SQPOSTr+REDUC5, data=n544)
```

The coefficients are 0.91, 0.91, 0.77, 0.89, 0.89, and 0.80. Hence the rescaling worked to obtain effects around 0.8, which was the typical value of the coefficients of the three dichotomous predictors.

#### 18.3.3.3 Multiplication factors

After finding a suitable set of weights we need to find the multiplication factor for the scores. In the testicular cancer model we round after multiplication with the factor 10/8 in a logistic model. To compensate for this multiplication, the actual multiplier was 0.86. This factor can be multiplied with the shrinkage factor (in this case, 0.95) to obtain shrunk predictions (shrunken.beta=0.81).

The calculation is as follows, where we omit the intercept from the set of coefficients:

```
score.fit <- lrm(NEC~Teratoma+Pre.AFP+Pre.HCG+LDHr+SQPOSTr+
                  REDUC5, data=n544, x=T, y=T)
rounded.lp <- score.fit$x %*% rep(1,6) # All scores a weight of 1
# multiplier makes the rescaled factors for logistic formula
multiplier <- lrm.fit(y=score.fit$y, x=rounded.lp)$coef[2]
shrunken.beta<- 0.95 * multiplier # shrinkage * 0.86
shrunken.beta # shrunk multiplier for better predictions 0.81
```

#### 18.3.3.4 Final steps

We estimate the intercept corresponding to the scores, using the rounded coefficients multiplied with the shrunken.beta coefficient as an offset variable:

```
lrm.fit (y=score.fit$y,offset=shrunken.beta * rounded.lp)
# The formula becomes lp = -1.94 + 0.81×score.
```

We check the deterioration in discriminative performance. The  $c$  statistic of the original model was 0.839; uniform shrinkage does not affect this value. We find that the  $c$  statistic with rounded scores and using continuous predictors is only 0.001 lower (0.838). This is in line with evaluations of alternative versions of the EuroSCORE, where a simple version had a  $c$  statistic that was 0.002 lower than a full logistic version.<sup>287</sup>

The final score chart can be constructed in several ways. Especially, the presentation of values for continuous predictors is possible with scores horizontally or vertically.

#### \*18.3.4 Coding with Categorization

Categorization of the three continuous predictors can also be considered in the testicular cancer example. It should consider the distribution and the predictive effects of the predictors. Strong predictors, with a wide range of predictor values, should have more categories than weaker predictors.

For LDHst, we could simply categorize the predictor as normal vs. abnormal, as was done for AFP and HCG. For post-chemotherapy size, we could use three categories: <20mm, 20–49mm, and  $\geq 50$ mm, with 2, 1, and 0 as scores respectively. For reduction in size, we could create three categories, with increase, 0–49% reduction, and  $\geq 50\%$  reduction in size having -1, 0, and +1 points respectively.

These categorizations can also be checked in a regression analysis, where the coefficient for the categorized predictors should have values around 0.8. Indeed, this is the case when we fit the following model with categorized predictors, where coefficients were 0.92, 0.86, 0.66, 0.91, 0.86, and 0.79:

```
lrm(NEC ~ Teratoma+Pre.AFP+Pre.HCG+PRELDH+POST2+REDUCr, data=n544)
```

Of note, using categorized versions of the continuous predictors led to a substantial drop in  $c$  statistic (from 0.839 to 0.808). Hence, categorizing simplifies presentation at the cost of performance. A score chart with the categorized version is available at the Web.

#### 18.3.5 Summary Points

- Several presentation formats are possible for the testicular cancer model that predicts benign tissue after chemotherapy
- Userfriendliness may vary, but empirical evidence on what formats clinicians prefer is limited
- The discriminative ability may suffer from very simple presentations (Table 18.5)

**Table 18.5** Discriminatory ability of different formats of presentation of the testicular cancer prediction model

Format	Table / figure	c statistic
Logistic formula / nomogram / graphical	Table 18.4 / Fig. 18.1 / Fig. 18.3	0.839
Rounded scores	Table 18.2 / Fig. 18.2	0.838
Categorized scores	Tables at Web site	0.808
Tabular	Table 18.3	0.773

## 18.4 Clinical Decision Rules

### 18.4.1 Regression Tree

Some modelling techniques, such as regression and classification trees, more naturally lead to decision rules. A regression tree classifies patients according to a (usually limited) number of characteristics. It is therefore often possible for clinical experts to link treatment recommendations to the various groups that are defined by the tree. Once these treatments have been defined, the tree can often be reformatted for easier application. This was for example done for the Goldman diagnostic index for acute MI. Based on a tree analysis of 482 patients suspected of acute MI, a decision protocol was constructed in the format of a simple flow chart considering nine clinical factors.<sup>142</sup> As discussed before, deriving a stable, reliable tree requires relatively large amounts of data: Trees are data-hungry. But tree presentations are generally considered to be very easy to understand.

### 18.4.2 Score Chart Rule

Scores can be based on severely rounded coefficients, e.g. counting each predictor as one point. This may be reasonable when the actual regression coefficients are similar in magnitude. When coefficients vary in magnitude, an alternative is to define minor and major risk factors. Such major rounding of coefficients leads to less-accurate predictions than the original rule. Especially, calibration may suffer.<sup>287</sup> The advantage of severe rounding is that it is possible to remember such decision rules by heart, in contrast to more refined prediction models.

A simple rule is that exceeding a certain score is an indication for a certain action. An example is the difficult issue of which patients should have a CT scan after minor head injury (defined as having sustained blunt injury to the head, with normal or minimally altered level of consciousness upon presentation. In one recent study, a detailed prediction model was developed, from which a simple decision rule was derived. Major and minor risk factors were defined based on rounding of the regression coefficients from a logistic model,

and categorization of continuous predictors (such as age: <40, zero score; 40–59, minor; >=60, major). The decision rule was CT scan in case of presence of at least one major or two minor risk factors (out of a list of ten major and eight minor risk factors).<sup>391</sup> With this rule, the sensitivity was 100% for neurosurgical interventions, which was considered a clinically very important outcome that should not be missed. Internal validation showed that we should not expect 100% sensitivity in new patients. The average sensitivity from a bootstrap validation procedure was 96%, with 100% sensitivity in 56% of the samples. On the other hand, many CT scans would still be made in those without an important outcome (specificity 25%, or a false-positive CT scan rate of 75%). Implementation of the decision rule was expected to reduce the number of CTs by ~25%. Hence most patients with minor head injury should have a CT scan if we want to exclude serious injury.

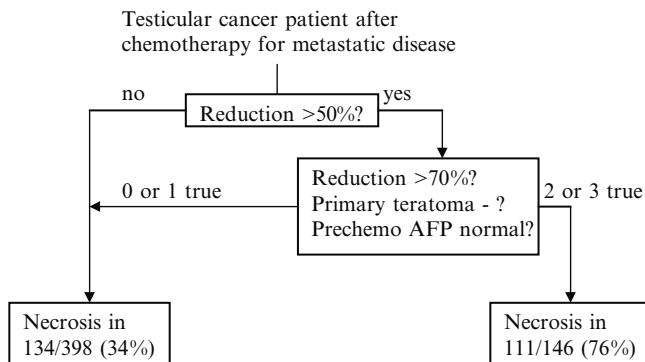
### ***18.4.3 Survival Groups***

Results from survival analyses are often presented by grouped predictions, e.g. quartiles. Such groupings can be linked to treatment recommendations. Survival can also be shown in relation to specific combinations of risk factors, similar to a regression tree. This approach was e.g. followed for the IGCC classification.<sup>5</sup> Five predictors were considered: two were coded as dichotomous predictors (poor vs. good), and three tumor markers were coded as low, intermediate, and high according to their level. A good prognosis group contained patients without intermediate or poor risk characteristics. An intermediate group contained patients with intermediate levels of tumor markers, but no poor risk characteristics. A poor prognosis group contained patients with at least one high tumor marker or a poor risk factor. The numbers of patients were 50%, 35%, and 15%, with 5-year survival of 92%, 80%, and 50%, respectively. The choice of risk group definitions was motivated by the idea to study more aggressive new therapy in the poor risk group (e.g. stem-cell therapy), and less aggressive therapy in the good prognosis group (e.g. three instead of four cycles of chemotherapy).

Such risk classifications present predictions for combined groups of patients, which is expected to lead to stable predictions at a group level. But the definition of the risk groups is often motivated by the distribution of risk rather than by decision-analytic considerations.

### ***18.4.4 Meta-Model***

Another option is to develop a meta-model, which describes an underlying, more complex model with a certain level of accuracy. A meta-model is a model that predicts the predictions from an underlying model. It aims to capture the general



**Fig. 18.5** Decision rule for patients with testicular cancer <sup>418</sup>

patterns and inherits any shrinkage of coefficients from the underlying more complex model. For a decision rule, we may categorize the predictions from the underlying model at a relevant cut-off, e.g. as needing treatment vs. no treatment. Subsequently, we can predict membership of either category. The meta-model can be presented in various forms, for example as a tree. A tree is an attractive format for this step because of its intuitive communication (see e.g. Fig. 18.5,<sup>418</sup> details at the Web site).

## 18.5 Concluding Remarks

The presentation format is an important issue in predictive modelling and deserves more attention than is usually done in current practice. The overview in this chapter is probably not complete, but intends to give inspiration for presentation of prediction models and decision rules. The format should match with the intended audience; some clinical areas have a more quantitative orientation than do others for example. Also, some formats have become more or less standard in certain areas (e.g. nomograms for prostate cancer).<sup>72</sup> Graphical presentations may sometimes be considered, e.g. to show predictions in relation to a single continuous predictor and one or two categorical predictors. There is no convincing evidence on the preference of certain presentation formats over others for optimal communication of individualized predictions.<sup>433</sup>

We can imagine that the ongoing automatization, e.g. with electronic patient records, will enable the direct and easy availability of predictions from detailed and rather complex prediction models. Hence, computerized presentations may have the future, both for prediction models and decision rules.

## Questions

### 18.1 Testicular cancer presentation formats

Calculate predicted probabilities for a man with a post-chemotherapy mass of 12 mm, which was 24 mm before chemotherapy, who had no teratoma elements in his primary tumor, elevated AFP, normal HCG, and 3 times normal LDH levels, using

- (a) the nomogram (Fig. 18.1)
- (b) the score chart (Table 18.2 with Fig. 18.2)
- (c) the simplified table (Table 18.3)
- (d) the graphs for radiologists (Fig. 18.3)
- (e) the penalized regression formula (Sect. 18.3.1)
- (f) the classification tree (Fig. 18.5)

### 18.2 Continuous predictors in a score chart (Sect. 18.3.3)

- (a) What specific challenges are posed by continuous predictors in a score chart?
- (b) What is the disadvantage of categorizing scores for a score chart (see Table 18.5)?

### 18.3 Odds ratios or regression coefficients for scores<sup>297</sup>

Several investigators have used odds ratios to derive scores for logistic regression models, which are added in a sum score.

- (a) Why is this incorrect?
- (b) What kind of deviations will occur if some odds ratios are small and some very large?

### 18.4 Prediction models and decision rules

- (a) What is the difference between a prediction model and a decision rule?
- (b) How can we derive a decision rule from a prediction model?

## **Part III**

# **Generalizability of Prediction Models**

Generalizability refers to the external validity of predictions from a model. The quality of predictions can be quantified by various performance measures, e.g. related to calibration, discrimination, and clinical usefulness. These measures are determined by validity of regression coefficients and the specific case-mix of the external setting.

For generalizability, internal validity is a minimum prerequisite. To achieve internal validity, we need to follow the seven steps outlined in Part II. In part III, we first consider differences between populations that may affect generalizability (Chap. 19). Next, we consider approaches to updating of a prediction model for a specific setting (Chap. 20). Finally, we consider the situation that a prediction model is applied in multiple settings. Detection of differences between settings may then actually be the purpose of the analysis, for example the comparison of quality of hospitals in a league table (“provider profiling,” Chap. 21).

# Chapter 19

## Patterns of External Validity

**Background** Generalizability depends on the quality of the prediction model as developed for the development setting (internal validity), and on characteristics of the population where the model is applied (validity of regression coefficients and distribution of predictor values). The general framework of validity of predictions was discussed in Chap. 17 (see in particular Fig. 17.1). Here, we first consider a number of typical situations that we may encounter when a prediction model is applied in an external setting. Theoretical relationships are illustrated with a large sample simulation and findings in some case studies. Approximate power calculations are given for tests of invalidity of a prediction model.

### 19.1 Determinants of External Validity

We concentrate on the external validity of predictions for a binary outcome  $Y$ . We consider a number of differences between populations that determine this external validity, related to case-mix and regression coefficients  $\beta$  (Table 19.1).

#### 19.1.1 Case-Mix

With case-mix we refer to the distribution of predictors  $X$  that are included in the regression model  $Y \sim X$ , as well as the distribution of predictors that are not included in the model, either observable or unobservable. Predictors not included in the model are referred to as “missed predictors,” despite the fact that some may in fact be observable. Since the linear predictor (lp) is a linear function of the predictors  $X$ , we will for simplicity consider one predictor “ $x$ ” in the model  $Y \sim x$ . Here,  $x$  represents a linear combination of  $X$ . Similarly, the missed predictors  $Z$  are represented as one variable “ $z$ ” in the regression model  $Y \sim x + z$ .

**Table 19.1** Differences between populations that determine external validity

Scenario	Characteristic	Differences	Example
Case-mix	Distribution of observed predictors ("X")	Different selection, e.g. more-severe patients are selected; or inclusion criteria smaller/wider	Validation in referral centre; validation in/outside RCT
	Distribution of missed predictors ("Z")	Different selection based on predictors not considered in the model	Validation in different setting
	Distribution of outcomes ("Y")	Retrospective sampling of cases and controls	Case-control design
Coefficients	Coefficients $\beta$ smaller than expected	Overfitted model is validated	Validation of model from small development sample
	Coefficients $\beta$ different	Truly different population	Validation in different setting

### 19.1.2 *Differences in Case-Mix*

A different case-mix may be encountered because the setting differs compared with the development situation; e.g. model development in secondary care and validation in a primary or tertiary care setting. Or a model was developed in patients participating in a randomized controlled trial (RCT) and is applied in a less selected population. Such situations make that the distribution of observed predictors  $X$  is different between development and validation setting. The distribution of missed predictors  $Z$  may also differ when we apply a model in a different setting; per definition, such differences cannot be excluded a priori. Missed predictors  $Z$  may be fully independent of  $X$ , or be correlated. Finally, the design of a study may cause differences in incidence of the outcome  $Y$ , and may hence influence case-mix indirectly. For example, a case-control design can be followed, where the ratio of cases to controls is different than in the population.

### 19.1.3 *Differences in Regression Coefficients*

Regression coefficients  $\beta$  can be different between settings because of true differences between populations. Various reasons can be thought of, including definitions of predictors, the definition of the outcome, and a different selection of patients.

In practice, the coefficients  $\beta$  are not known for the development setting, but only estimated from a finite sample size. The same holds for a validation sample from a validation setting. This makes it impossible that the same regression coefficients are found when a regression model is re-estimated in a validation sample. Even if the underlying true coefficients are identical, some of the re-estimated coefficients will be larger and some smaller than in the development sample.

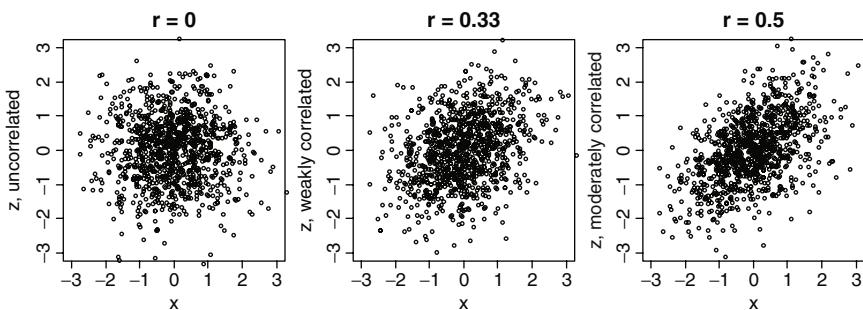
Another problem is that regression coefficients may on average have been estimated too large because of overfitting in the development data set. Such overfitting is most likely for models developed in small data sets with a relatively large number of (candidate) predictors (see e.g. Chap. 5, 11, and 13). Shrinkage of coefficients at model development should have prevented overestimation of coefficients for predictive purposes, but this is not the case for many currently developed models.

## 19.2 Impact on Calibration, Discrimination, and Clinical Usefulness

In the following we will consider various scenarios for differences between populations (Table 19.1). We will study the impact of these differences on calibration, discrimination, and clinical usefulness of prediction models for binary outcomes. We simulate an outcome  $y$ , which depends on  $x$  and a missed predictor  $z$  (both with standard normal distribution). In the development population, we estimate a logistic regression model with an intercept  $\alpha_0$  and coefficient  $\beta_1$  for  $x$ , while in fact the outcome  $y$  is determined by  $x$  and  $z$ . The missed predictor  $z$  and  $x$  are uncorrelated, weakly correlated, or moderately correlated (Pearson correlation coefficients  $r$ , 0, 0.33, 0.5, Table 19.2 and Fig. 19.1). Findings are summarized in Table 19.3.

**Table 19.2** Design of simulations with predictor  $x$  and missed predictor  $z$ , for a logistic regression model  $Y \sim x + z$  (adjusted analysis) and  $Y \sim x$  (unadjusted analysis)

Correlation $x - z$	Adjusted coefficients	Unadjusted coefficient
Pearson $r=0, r^2=0\%$	$2.05*x + 1.5*z$	$1.5*x$
Pearson $r=0.33, r^2=11\%$	$1.5*x + 1.5*z$	$1.5*x$
Pearson $r=0.5, r^2=25\%$	$1.18*x + 1.5*z$	$1.5*x$



**Fig. 19.1** Correlation between  $x$  (represented in the linear predictor) and  $z$  (a missed predictor), with or without correlation. Illustration for  $n=1,000$ ;  $n=500,000$  in further simulations

**Table 19.3** Patterns of invalidity for a prediction model for binary outcomes in relation to differences between development and validation populations

Scenario	Characteristics	Differences	$a b=1$	$b$	c stat	NB
Development setting	$y = 1.5*x$ ( $x \sim N(0,1)$ )	–	0	1	0.81	0.055
Case-mix in validation setting	Distribution of observed predictors (“ $x$ ”)	More-severe patients Less-severe patients More heterogeneous Less heterogeneous	0 0 0 0	1 1 1 1	0.77 0.77 0.90 0.75	0.006 0.104 0.104 0.030
	Distribution of missed predictors (“ $z$ ”)	More-severe patients <sup>a</sup> Less-severe patients <sup>a</sup> More heterogeneous <sup>a</sup> Less heterogeneous <sup>a</sup>	0.70 -0.70 0 0	1 1 1 1	0.81 0.81 0.83 0.81	0.001 0.109 0.062 0.053
	Distribution of outcomes (“ $y$ ”)	2 times more cases 2 times less cases	log(2) -log(2)	1 1	0.81 0.81	NA NA
Coefficients in validation setting	Coefficients $\beta$ smaller than expected	Slope 0.8 Slope 0.6	0 0	0.8 0.6	0.77 0.72	0.037 0.014
	Coefficients $\beta$ different	X effects * 0.5 or 1.5 X effects * 0.25 or 1.5	0 0	0.84 0.68	0.78 0.74	0.040 0.023

<sup>a</sup>For correlation  $x - z$  of 0.33; detailed results in Figs. 19.5 and 19.6;  $a|b=1$ , intercept given that calibration slope is 1, indicating “calibration-in-the-large”;  $b$ , calibration slope;  $c$  stat,  $c$  statistic to indicate discriminative ability; NB, net benefit; NA, not applicable

### 19.2.1 Simulation Set-Up

We create a validation population where we apply the developed model. Various differences are simulated for the validation population compared with the development population. We first consider populations ( $n=500,000$ ) and later samples of smaller size to illustrate sampling variability and statistical power. We consider a scenario inspired by the testicular cancer case study, with average incidence of tumor close to 50%, and a decision threshold for the probability of tumor of 30% (Chaps. 15 and 16). We consider a good discriminating model, with a  $c$  statistic of 0.81. This  $c$  statistic is achieved with a logistic regression model with a single predictor  $x$ , with  $x$  normally distributed and regression coefficient  $\beta$ , 1.5. We can hence define the linear predictor “lp” as  $lp=1.5*x$ .

We generate the outcome  $y$  with inclusion of the missed predictor  $z$  (uncorrelated or correlated). In the underlying model the lp is a function of  $x$  and  $z$ . With uncorrelated  $x - z$ , we define the lp as  $lp=2.05*x + 1.5*z$ . The adjusted regression coefficient for  $x$  is 2.05 rather than 1.5. This may be surprising, but is related to the “stratification” or “conditioning” effect in non-linear models such as logistic regression and Cox regression models. In such models, adjusted effects are more extreme than unadjusted effects when a covariate is considered that is related to the outcome, but uncorrelated to other covariates. This is well known in the analysis of randomized clinical trials (see Chaps. 2 and 22).<sup>133,182,348,403</sup> In unadjusted analysis, the coefficient for  $x$  is 1.5 (Table 19.2).

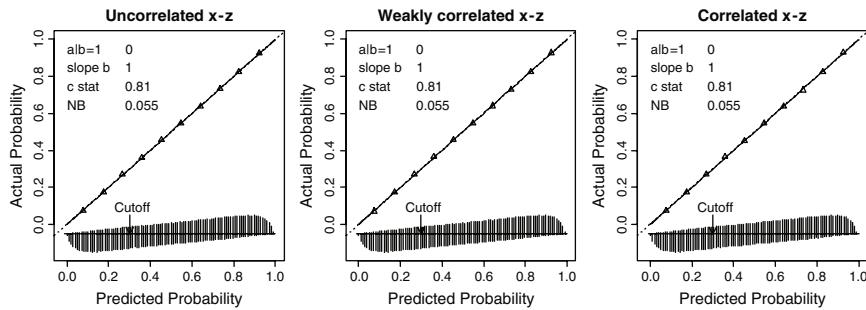
For moderately correlated  $x - z$  data ( $r=0.5$ ), we define the lp as  $lp = 1.18*x + 1.5*z$ . Now the adjusted regression coefficient is 1.18 rather than 1.5, which is caused by the positive correlation between  $x$  and  $z$ . The confounding effect of this correlation was stronger than the stratification effect. In unadjusted analysis, the coefficient for  $x$  is again 1.5. An intermediate situation was identified by trial and error, where the correlation between  $x$  and  $z$  was 0.33, such that the negative effect of confounding and positive effect of stratification on  $z$  are exactly balanced in the adjusted analysis. The model is  $lp = 1.5*x + 1.5*z$ .

In both development and validation settings, we study predictions only in relation to  $x$ , since  $z$  is a missed predictor. The observed relation is  $lp = 1.5*x$  at development, with a  $c$  statistic of 0.81. At validation we hope to see  $y = 0 + 1*lp$  in a logistic regression model (see Chap. 15 and 20 for more background on this calibration model).<sup>86</sup> This relation between  $y$  and  $lp$  may be influenced by changes in the distribution of the  $x$ ,  $z$ , or  $y$ , or differences in the regression coefficients that determine the  $lp$  (see Table 19.1).

### **19.2.2 Performance Measures**

In the following, we concentrate on a limited number of indicators of calibration, discrimination, and clinical usefulness, although many more performance measures can be considered for validation of predictions for binary outcomes (see Chaps. 15 and 16). For calibration we consider calibration-in-the-large (intercept given that slope  $b$  is set to 1,  $a|b=1$ ) and the calibration slope ( $b$ ). Both are determined in logistic regression models:  $y \sim lp$ . The linear predictor  $lp$  is entered as an offset variable ( $a|b=1$ ), or as the only predictor (to estimate slope  $b$ ) in a logistic regression model estimated in the validation data. The  $c$  statistic is used to indicate discriminative ability (Chap. 15). For clinical usefulness, we calculate the net benefit (NB), with the formula  $NB = (TP - wFP) / N$ , where TP is the number of true-positive classifications, FP the number of false-positive classifications, and  $w$  is a weight equal to the odds of the threshold (cutoff/(1 – cutoff)), or the ratio of harm to benefit (see Chap. 16). We compare the NB of the model with a cutoff at 30% with the strategy with the next best NB (“treat all” or “treat none”). With an incidence of the outcome at 50% and a threshold of 30%, “treat all” has the next best NB: for every 100 patients, 50 true-positive classifications are made, and 50 false-positive classifications (which are weighted as 3/7). The NB of treat all hence is  $50 - 3/7 * 50 = 28.6 / 100$  patients. A clinically useful model should have an NB higher than this reference value.

When the considered model is applied in the development population, the calibration is perfect ( $a|b=1 = 0$ ; slope  $b = 1$ ) and discrimination good ( $c = 0.81$ , Fig. 19.2). The increase in NB by 0.055 means that 5.5 more true-positive classifications are obtained per 100 patients, at the same number of false-positive classifications (see Chap. 16). The model performance is identical whether uncorrelated or correlated missed predictors are present.



**Fig. 19.2** Calibration, discrimination, and clinical usefulness when the prediction model is applied in a population with identical distribution of predictors  $x$  and missed predictor  $z$  (from left to right:  $r=0, 0.33, 0.5$ ).  $a|b=1$ ; intercept given slope  $b$  is 1; slope  $b$ , calibration slope in a model  $y \sim lp$ ;  $c$  stat,  $c$  statistic indicating discriminative ability; NB, net benefit compared with “treat all.” The value of 0.055 means that 5.5 more true positive decisions are taken per 100 patients, at the same number of false-positive decisions (see Chap. 16). Triangles represent deciles of patients grouped by similar predicted probability. The distribution of patients is indicated with spikes at the bottom of the graph, separately for those with and without the outcome

## 19.3 Distribution of Predictors

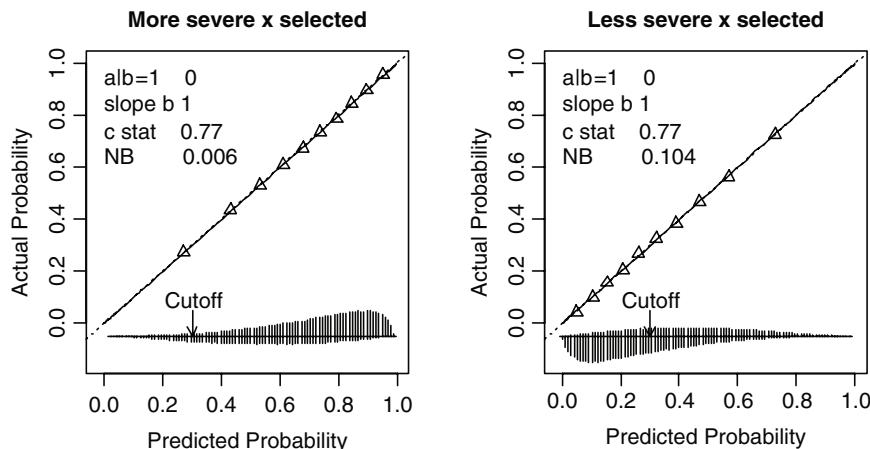
We consider various selection mechanisms based on observed predictors  $X$  and missed predictors  $Z$ . Such selection is an example of missing not at random (MNAR, Chap. 7). We already know that regression coefficients of a predictor  $X$  remain unbiased with an MNAR mechanism. Hence, calibration is expected to remain unaffected. Of interest is any influence on discrimination and clinical usefulness.

### 19.3.1 More- or Less-Severe Case-Mix According to $X$

Subjects may be more likely to be included in the validation setting because they have higher  $X$  values (“more suspect cases”). For example, we may assume that only the higher  $X$  values are represented (correlation with missingness, 0.62;  $R^2$ , 39%). This leads to a more-severe case-mix at validation.

```
n <- 500000
x <- rnorm(n)                                     #standard normal x
xM <- ifelse(rnorm(n=n, sd=.8)<x, x, NA)       #50% missing, r=0.62
```

In our particular simulation, the more-severe case-mix is associated with somewhat less spread in predictions, and hence a lower  $c$  statistic (0.77 instead of 0.81, Fig. 19.3 left panel). Moreover, only few patients have predictions below the postulated threshold of 30%, reducing the NB to 0.006 instead of 0.055. The prediction model would be judged of very limited use in this validation setting. If the missingness was even more selective ( $r>0.75$ ), the NB would become zero, meaning



**Fig. 19.3** Influence of selection of more- or less-severe cases according to observed predictor values (“ $x$ ”). 50% of the subjects were selected, with higher or lower likelihood of selection with higher  $x$  values. Validation with a less-severe case-mix makes the prediction model clinically more useful (right panel)

that “treat all” would be as good a policy as using the model. In contrast, a less-severe case-mix led to a higher NB (NB, 0.104; Fig. 19.3, right panel). These patterns were identical with uncorrelated or correlated  $z$ .

### \*19.3.2 Example: Interpretation of Testicular Cancer Validation

These findings are important for the interpretation of the external validity of the testicular cancer example presented in Chaps. 15 and 16. When applied in more-severe patients treated at a tertiary referral centre (Indiana University Medical Center), we noted a decrease in clinical usefulness of the prediction model. But we have to realize that not all testicular cancers undergo surgical resection; there is “verification bias.”<sup>30</sup> Typically a selection is made towards those with a suspicion of residual tumor (e.g. larger residual masses). If all testicular cancer patients were considered, the model would also indicate resection in some patients who are not candidates for resection under current policies. Clinical usefulness would hence be judged higher.

### 19.3.3 More or Less Heterogeneous Case-Mix According to $X$

Another situation is that a more heterogeneous setting is considered, which is fully represented by the  $X$  values. For example, inclusion criteria may be wider in surveys of patients with traumatic brain injury (TBI) compared with randomized controlled trials (RCTs). RCTs typically have a list of inclusion and exclusion criteria. If these

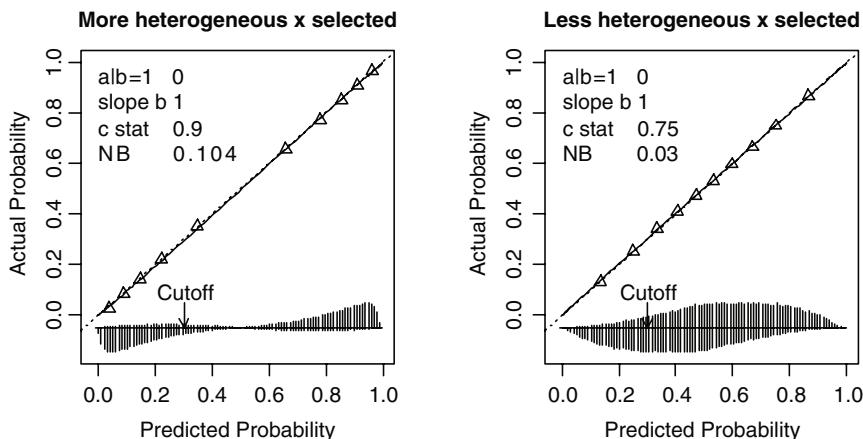
criteria apply to predictors that are all considered in the prediction model, the distribution of  $X$  values will be more heterogeneous in surveys. Note that the selection on  $X$  values may lead to extrapolation of model predictions in the validation data beyond observed  $X$  values in the development data.

The heterogeneity in case-mix translates into a higher discriminative ability; we can distinguish more patients with very low or very high prediction ( $c$  statistic, 0.90 instead of 0.81, Fig. 19.4, left panel). More patients have predictions below the postulated threshold of 30%, doubling the NB (0.104 instead of 0.055). The prediction model would be judged quite useful in this more heterogeneous validation setting. The reverse is found for validation in a setting with less heterogeneity (lower  $c$  statistic, 0.75; lower NB, 0.03, Fig. 19.4, right panel). These patterns were identical with uncorrelated or correlated  $z$ .

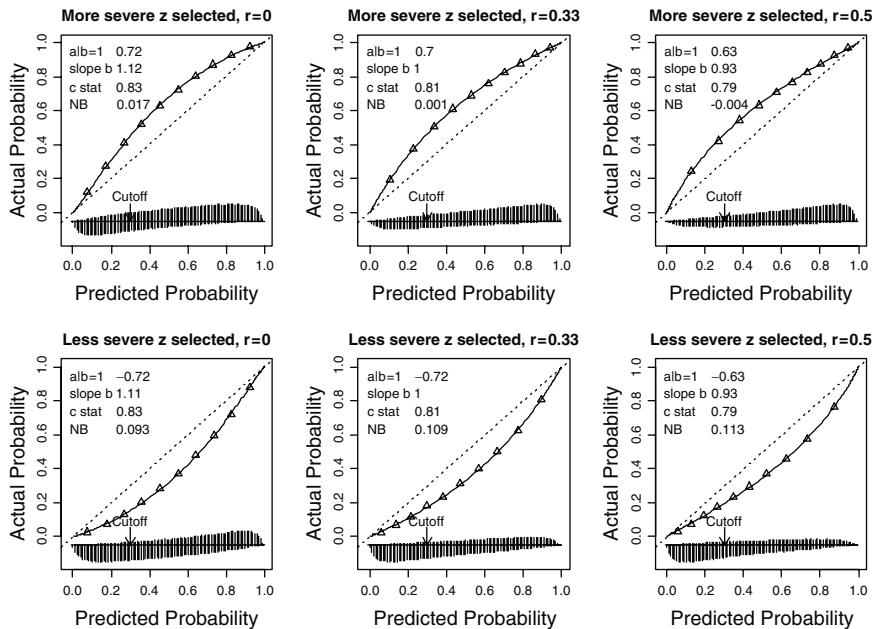
### 19.3.4 More- or Less-Severe Case-Mix According to Z

Similar to distributions of observed predictors, distributions of missed predictors  $Z$  may also differ between development and validation settings. We will see that the correlation between observed predictors  $X$  and missed predictors  $Z$  is especially relevant for calibration.

The first situation is that a prediction model is applied in a setting of more- or less-severe cases, according to predictors that are not (or not fully) captured in the prediction model. A more-severe case-mix mainly causes a systematic miscalibration of predictions (Fig. 19.5, top row). The calibration-in-the-large ( $a|b=1$ ) values are around 0.7, which reflects that approximately twice as many cases are found



**Fig. 19.4** Influence of selection of more or less heterogeneous cases according to observed predictor values for “ $x$ .” Approximately 35% of the subjects were selected, with higher or lower likelihood of selection with more extreme  $x$  values. Validation with a more heterogeneous case-mix makes the prediction model more discriminatory and clinically more useful (*left panel*)

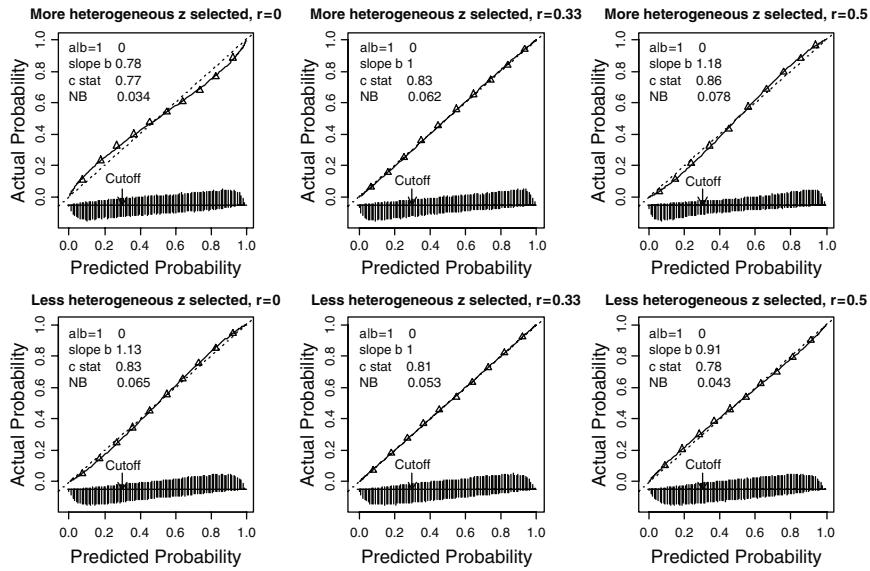


**Fig. 19.5** Influence of selection of more- or less-severe cases according to a missed predictor (“z,”  $x - z$  correlation, 0, 0.33, or 0.5). 50% of the subjects were selected, with higher or lower likelihood of selection with higher  $z$  values

than predicted (odds ratio,  $\exp(0.7) = 2.0$ ). The calibration slope is around 1. Without correlation between  $x$  and  $z$  ( $r=0$ , Fig. 19.5, upper left panel), the slope is 1.12, which is explained by the reduced stratification effect of  $z$  in the regression model. In the development setting, the stratification effect was such that the adjusted coefficient was 2.05 for an unadjusted coefficient of 1.5 for  $x$ ; with less stratification, the unadjusted coefficient is  $1.12 \times 1.5 = 1.68$ . With moderate correlation ( $r=0.5$ , Fig. 19.5, upper right panel), the confounding effect is weaker, leading to an unadjusted coefficient of  $0.93 \times 1.5 = 1.4$  for  $x$ .

The discrimination follows the same pattern as the calibration slope, with values around the original estimate of 0.81. The poor calibration causes the model to have at most small clinical usefulness. The NB of the model may even become negative ( $-0.004$  in Fig. 19.5, upper right panel). This means that worse decisions are made with the model than the reference strategy of “treat all.” This can be understood by realizing that the model assigns patients with a prediction under 30% to “no treatment,” while predictions are systematically miscalibrated. Hence, many among those with a prediction under 30% have actual probabilities over 30% and should have been classified for “treat.” On balance, the loss of inappropriately withholding treatment from those with actual probabilities over 30% was larger than the gain of reducing false-positive classifications (100% with a “treat all” strategy).

The reverse pattern is noted when selection is on less-severe patients according to some missed predictor (Fig. 19.5, second row). Calibration-in-the-large is the



**Fig. 19.6** Influence of selection of more or less heterogeneous cases according to a missed predictor (“ $z$ ,”  $x - z$  correlation, 0, 0.33 or 0.5). Approximately 35% of the subjects were selected, with higher or lower likelihood of selection with more extreme  $z$  values

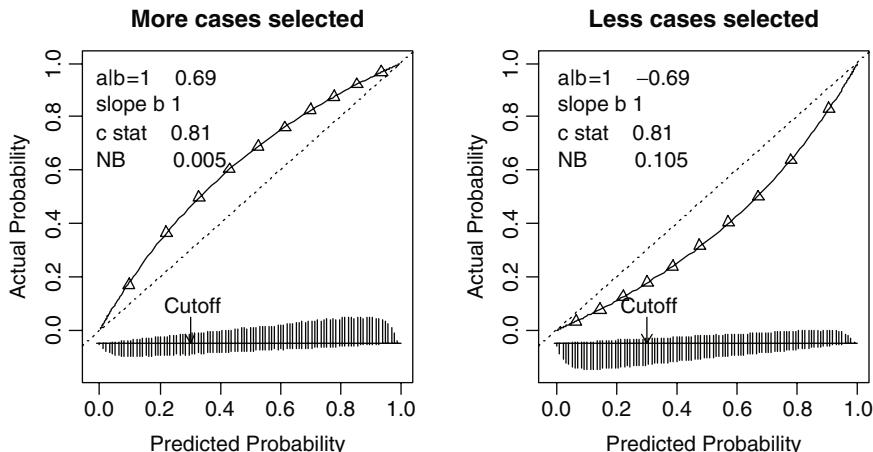
main problem. But interestingly, the clinical usefulness is now increased, despite this miscalibration.

### 19.3.5 More or Less Heterogeneous Case-Mix According to Z

Similar to observed predictors, we can imagine that missed predictors may have a more or less heterogeneous distribution in a validation setting. Such distributional changes affect the calibration slope, but not calibration-in-the-large (Fig. 19.6). The specific patterns can again be explained by the magnitude of stratification and confounding effects. Discrimination and clinical usefulness were better with higher calibration slopes.

## 19.4 Distribution of Observed Outcomes Y

A case-control design allows for separate sampling of cases ( $y=1$ ) and controls ( $y=0$ ). Cases and controls should come from the same underlying populations as would be considered in a cohort study (Chap. 3). In our examples, the ratio of cases and controls was 1:1 (50% incidence of the outcome  $Y$ ). The effect of manipulating the outcome incidence is reflected in calibration-in-the-large. With a ratio of 2 cases to 1 control, the odds ratio of the intercept is 2. Indeed, the coefficient is 0.69,



**Fig. 19.7** Influence of a case–control design on the model intercept; calibration slope and discriminability remain unaffected. The ratio of cases to controls was set to 2:1 (left) and 1:2 (right panel)

or  $\log(2)$  (Fig. 19.7, left panel). Conversely, a ratio of 1 case to 2 controls leads to an intercept of  $-0.69$ . With a proper case–control design, the effects of predictors remain identical (calibration slope = 1), as well as the  $c$  statistic (0.81). Calculation of clinical usefulness is only sensible after correction of the intercept, which can be seen as translating a case–control design back to clinical practice.

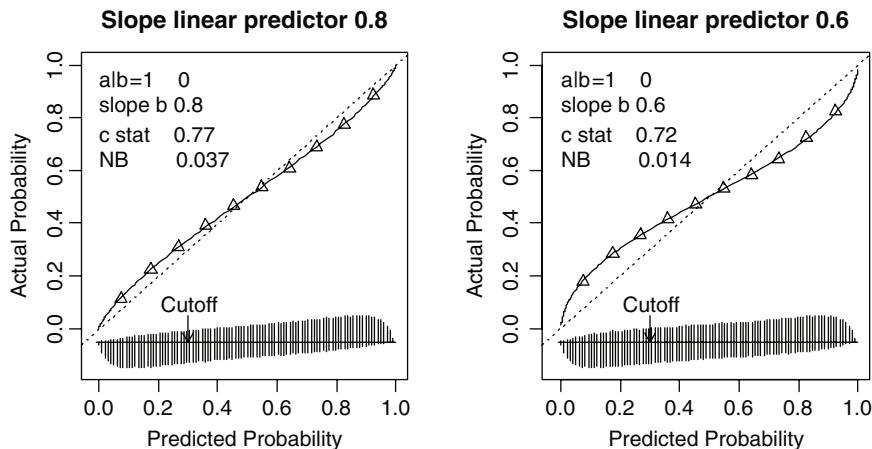
In a traditional case–control design, the number of controls is unknown. This makes it impossible to correctly adjust the intercept. In a *nested* case–control design, we sample the cases and controls from a defined underlying cohort. The number of controls is known in such a design, which makes it straightforward to adjust the intercept, for example by weighting the controls by the inverse of their sampling ratio.

## 19.5 Coefficients $\beta$

### 19.5.1 Coefficient of Linear Predictor $< 1$

Overfitting is a major problem of predictive modelling (Chaps. 4–18). At external validation, we may often find less predictive effect of the linear predictor  $l_p$ . This reduced effect might have been detected already at internal validation, and might have led to incorporation of a shrinkage factor to compensate for overfitting. True differences in predictive effects may also play a role, for example caused by definition and selection issues.

A typical shrinkage factor found at internal validation is 0.8; more-severe overfitting might lead to a shrinkage factor of 0.6. At external validation, we find that such patterns of overfitting lead to a reduction in discriminative ability ( $c$ , 0.77 or 0.72 instead of 0.81) and a reduction in clinical usefulness (NB, 0.037 or 0.014 instead of 0.055, Fig. 19.8).



**Fig. 19.8** Influence of overfitting in model development. The slope of the linear predictor is 0.8 or 0.6, with lower discriminative ability ( $c=0.77$  or  $0.72$ ), and lower clinical usefulness (net benefit,  $0.037$  or  $0.014$ )

### 19.5.2 Coefficients Different

In addition to being overestimated on average, regression coefficients may truly differ between development setting and validation setting. Various causes can be imagined, all related to the validation population not being “plausibly related” anymore to the development population.<sup>222</sup> Terrin et al. considered various scenarios of different effects of predictors in a validation setting. In simulation studies, they simulated weaker effects of predictors, motivated by clinical scenarios, and found reductions in  $c$  statistic from 0.75 to 0.72.<sup>431</sup>

In Chap. 5, we used an arbitrary example of differences in predictor effects, with half of the predictors having 0.5 and half having 1.5 times the effect of the development setting. We use this example here for illustration, and a more extreme situation, with half of the predictors having a very small effect at validation (0.25).

### \*19.5.3 R Code

The programming code may help to understand how simulations were performed. First 10  $x$  variables were created, with decreasing standard deviation:

```

n      <- 500000
x1    <- rnorm(n, sd=1)
x2    <- rnorm(n, sd=.9)
...
x10   <- rnorm(n, sd=.1)
```

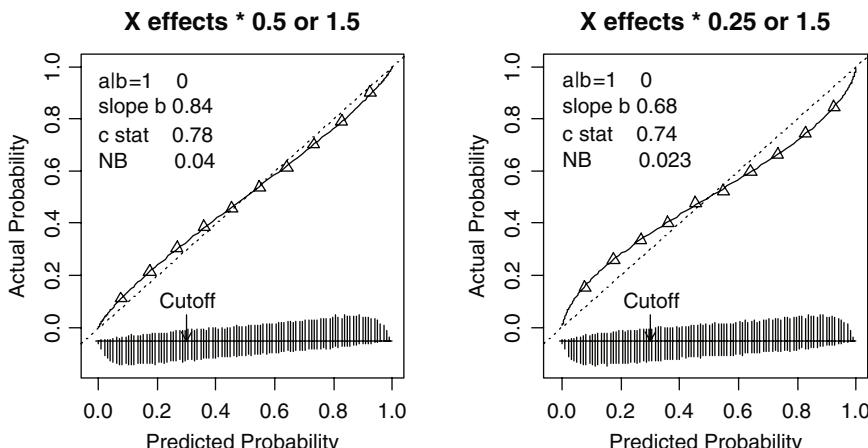
For the development setting, we assume that each  $x$  has a coefficient of 1; but in the two validation settings these weights are different.

```
#development data
xsum <- x1+x2+x3+x4+x5+x6+x7+x8+x9+x10
#validation data: 2 scenarios
xval <- .50*x1+1.5*x2+.50*x3+1.5*x4+...+.50*x9+1.5*x10
xval2 <- .25*x1+1.5*x2+.25*x3+1.5*x4+...+.25*x9+1.5*x10
```

Logistic regression models were constructed with the `xsum`, `xval`, and `xval2` variables. When the latter<sup>2</sup> variables are multiplied by 0.76, the  $c$  statistics are 0.81 for both models. We validate predictions from the model with `xsum` as predictor in settings where  $0.76 \cdot xval$  or  $0.76 \cdot xval2$  is the true linear predictor determining outcome.

#### 19.5.4 Influence of Different Coefficients

Calibration-in-the-large may remain unaffected when predictive effects are different (Fig. 19.9). However, the calibration slope was smaller than 1. When effects in the validation setting remained close to the effects at development, the slope was 0.84, and discrimination was slightly decreased (0.78 instead of 0.81). When differences in coefficients were more substantial (Fig. 19.9, right panel), the calibration slope was 0.68, the  $c$  statistic 0.74, and clinical usefulness smaller (0.023 instead of 0.055). Hence, differences between effects in the development setting vs. the validation setting may seriously deteriorate model performance.



**Fig. 19.9** Influence of differences in regression coefficients between development and validation setting. Regression coefficients were 0.5 or 1.5 times as large in the *left panel*, and 0.25 or 1.5 times as large in the *right panel*. In the right panel, miscalibration was severe (slope, 0.68), discriminative ability and clinical usefulness modest ( $c$  statistic, 0.74; net benefit, 0.023)

### \*19.5.5 Other Scenarios of Invalidity

Thus far, we considered one element at a time for differences between development and validation populations. All simulation results depend on the specific parameters chosen; with more extreme parameters, differences will be larger. We can consider other scenarios that are also plausible in medical research, where we combine differences in distribution of  $x$ ,  $z$ , and regression coefficients (Table 19.4). Detailed results are provided at the book's Web site.

### 19.5.6 Summary of Patterns of Invalidity

- Calibration

In the development setting, the calibration was perfect, the  $c$  statistic 0.81 and the NB of applying the model 0.055. Calibration remained perfect when the validation setting consisted of more- or less-severe patients according to predictor values, or more or less heterogeneous patients according to observed or missed predictor values. Calibration can be systematically disturbed by a more- or less-severe distribution of missed predictor values ( $z$ , e.g. intercept +0.70 or -0.70). A similar disturbance can be caused by a case-control design; however, the case:control ratio is under the influence of the researcher, while the distribution of a missed predictor usually is not. Calibration can also be affected by overfitting at model development (e.g. slope, 0.8 or 0.6), or truly differential predictive effects (coefficients of individual predictors 0.25 / 0.5 / 1.5 times as large).

- Discrimination

Discriminative ability is related to the calibration slope, with a lower  $c$  statistic associated with a lower calibration slope. Another reason for a lower  $c$  statistic is a less heterogeneous case-mix (e.g., slope = 1, but  $c = 0.75$  instead of 0.81, Fig. 19.4, right panel). A high  $c$  statistic such as 0.90 was found for a more heterogeneous

**Table 19.4** Combinations of differences between development and validation populations and their impact on validity of a prediction model for binary outcomes

Scenario	$x$	$z$	Coefficients	$a b=1$	$b$	$c$ stat	NB
Change of setting	–	More severe	$X$ effects * 0.5 or 1.5	0.67	0.87	0.78	-0.008
	–	Less severe		-0.67	0.87	0.78	0.098
RCT vs. survey	More heterogeneous	More severe	$X$ effects * 0.5 or 1.5	0.64	1.04	0.88	0.027
	Less heterogeneous	More severe		0.69	0.59	0.65	-0.036
	More heterogeneous	Less severe		-0.68	1.03	0.88	0.167
	Less heterogeneous	Less severe		-0.68	0.59	0.65	0.037

setting (Fig. 19.4, left panel). More heterogeneity in missed predictors had only small effects (Fig. 19.6). These examples illustrate that discrimination is determined by validity of estimated regression coefficients  $\beta$  and case-mix. Poor discrimination can hence result from both aspects (i.e. poor calibration and/or relatively homogeneous case-mix).

- Clinical usefulness

Tables 19.3 and 19.4 highlight the importance of calibration for clinical usefulness. A systematic miscalibration, e.g. caused by a more-severe case-mix according to a missed predictor  $z$ , may lead to a model without clinical usefulness. With incorrect calibration, we can even make systematically wrong decisions. This is not the case if predictions are well calibrated. Discrimination and calibration slope are linked, with a low calibration slope or low discrimination both associated with a low clinical usefulness.

Perfect calibration and good discrimination do not guarantee clinical usefulness. Discrimination is important; better discrimination may lead to better decision making. If a model has no discriminative ability, it cannot be clinically useful. Discrimination is hence a *necessary but not sufficient* condition for clinical usefulness.

When applying the model in more- or less-severe patients, the  $c$  statistic was 0.77 for both settings, but clinical usefulness was 0.006 for a more-severe setting and 0.104 for a less-severe setting. These findings are in line with the lack of clinical usefulness of the testicular cancer case study in Chap. 16, where we noted that few patients had a prediction above the threshold of 70% for the probability of benign tissue at external validation.

Case-mix is also very relevant. The case-mix in observed predictors ( $X$ ) affects clinical usefulness through the distribution of predictions around the decision threshold, while leaving calibration largely intact. The case-mix in missed predictors ( $Z$ ) may predominantly affect clinical usefulness through poor calibration-in-the-large.

## 19.6 Reference Values for Performance

The distribution of predictors  $X$  can be taken into account in the calculation of “reference values” for model performance. Reference values indicate a model’s performance under the condition that the model predictions are valid in the validation sample. For a regression model this means that the regression coefficients for predictors  $X$  and the model intercept are fully correct for the validation setting. Such reference values may be very useful to obtain insight in what is happening at validation: Are there differences in case-mix or differences in regression coefficients compared with the development setting?

### 19.6.1 Calculation of Reference Values

For calibration, obvious reference values are 0 for calibration-in-the-large, and 1 for the calibration slope. For discrimination, we noted that the  $c$  statistic can vary

with case-mix. Similarly, overall performance measures such as  $R^2$  and Brier score depend on the case-mix in observed predictors  $X$ .<sup>202</sup>

A practical approach is to simulate the outcome  $Y$  for the observed case-mix in  $X$ , given that the prediction model is correct. This is simply obtained by first calculating the predictions for each subject in the validation data, and subsequently randomly assigning an outcome  $Y$  based on this prediction. With at least 100 repetitions for each patient, a stable estimate of the reference values is obtained. We illustrate the calculation below for 1,000 repetitions per patient in a logistic regression model.

### \*19.6.2 R Code

```
# fit in development data
fit      <- lrm(y~x1 + x2, data=dev.data)
# linear predictor for validation data
lp       <- predict(fit, newdata=val.data)
# External validation
val.prob(logit=lp, y=val.data$y, ...)
# Start simulation of outcomes
n        <- nrow(val.data)
nsamples <- 1000 # for stable results
perf.m   <- matrix(nrow=nsamples*n, ncol=2)
perf.m[,1] <- rep(lp, nsamples) # repeat lp nsamples times
# Generate y for validation data
perf.m[,2] <- ifelse(runif(length(perf.m[,1])) <=
                     plogis(perf.m[,1]), 1, 0)
# Determine reference values
val.prob(logit=perf.m[,1], y = perf.m[,2], ... )
```

### 19.6.3 Performance with Refitting

Another type of reference value is the performance obtained by refitting the model in the validation data. The regression coefficients are then optimal for the validation data, and hence provide an upper bound for the performance, which would be obtained if the coefficients from the development setting were exactly equal to those in the validation setting. However, this upper bound does not only depend on case-mix, but also on the effects of predictors in the validation setting. It is hence not simple to compare performance between development and validation settings: Differences may be attributable to both case-mix and/or coefficients.

**Table 19.5** Examples of reference values for performance of two prediction models, developed in one setting and applied in another setting

Example	Measure	Apparent	Internally validated	Externally validated	Reference	Refitted
Testicular cancer	<i>c</i> stat	0.818	0.812	0.785	0.824	0.819
	$R^2$	38.9%	37.4%	26.7%	37.0%	34.2%
	Brier	0.174	0.178	0.161	0.144	0.147
Traumatic brain injury	<i>c</i> stat	0.767	0.765	0.816	0.804	0.819
	$R^2$	27.9%	27.3%	37.1%	35.3%	38.0%
	Brier	0.186	0.188	0.180	0.181	0.168

#### \*19.6.4 Examples: Testicular Cancer and TBI

We apply the calculation of reference values to the testicular cancer and traumatic brain injury (TBI) case studies (Table 19.5). The apparent performance is calculated for  $n=544$  testicular cancer patients ( $n=245$  (45%) with benign histology) and  $n=2,036$  TBI patients ( $n=798$  (39%) with unfavorable 6-month outcome). The 544 testicular cancer patients are mostly from secondary care centres,<sup>417</sup> while validation was done in 273 patients from a tertiary care centre (Indiana). A benign outcome was less frequent among these patients ( $n=76/273$ , 28%).<sup>466</sup> The 2,036 TBI patients were from the Tirilazad randomized controlled trials,<sup>203</sup> with validation in three largely unselected series (UK 4 centre study, European Brain Injury Consortium survey, Traumatic Coma Databank,  $n=2,090$ ).<sup>271</sup> These patients more often had an unfavourable outcome at 6 months ( $n=1,249/2,090$ , 60%) compared with the development sample.

In the testicular cancer case study, the apparent *c* statistic was 0.818, with 0.006 optimism according to a bootstrap procedure. At external validation, the *c* statistic was 0.785, while 0.824 was expected based on the case-mix of the predictor variables (“reference,” Table 19.5). When the model was refitted, the performance was slightly lower than this reference value (*c* statistic, 0.819 vs. 0.824). A similar pattern was noted for the  $R^2$  and Brier statistics. We might test the statistical significance of these differences in performance, but concentrate here on the point estimates.

In the TBI case study, the apparent *c* statistic was 0.767, with negligible optimism. Surprisingly, the *c* statistic was higher at external validation (0.816), while 0.804 was expected based on the case-mix of the predictor variables. When the model was refitted, the performance was also higher than the reference (*c* 0.819 vs. 0.804). A similar pattern was noted for the  $R^2$  and Brier statistics.

The interpretation of Table 19.5 is as follows:

1. Internal validation corrects for the statistical problem of overfitting in the development setting; case-mix is unchanged

2. External validation tests the model in a sample from a new setting, where both case-mix and coefficients may be different than in the development sample
3. The reference performance corrects for the new case-mix according to predictor values in the validation sample, while keeping the coefficients at the values from the development setting (that would ideally have been corrected with shrinkage to correct for overfitting)
4. The refitted performance corrects for the new case-mix and estimates optimal regression coefficients in the validation sample

The poorer external performance of the testicular cancer model is not explained by case-mix, at least not in the distribution of observed predictor values, since the reference performance was very similar to that in the development sample. The poorer external validity should hence be attributed to differences in regression coefficients between the settings. The refitted performance was similar to the reference performance, indicating that the predictors had similar predictiveness in both settings when refitted.

The better external performance of the TBI model is partly explained by case-mix, since the reference performance was higher than in the development sample. The surprisingly good external validity should further be attributed to differences in regression coefficients between the settings; predictive effects were overall stronger in the validation setting (calibration slope, 1.08), in line with the even better refitted performance (refitted  $c$ , 0.819, Table 19.5).

## 19.7 Estimation of Performance

Thus far, we examined theoretical patterns of invalidity with very large simulated samples, which can be considered as populations. The testicular cancer and TBI case studies considered more limited sample sizes for model development and validation; differences in model performance might at least partly be attributed to chance. Performance parameters such as model intercept ( $a|b=1$ ), calibration slope ( $b$ ),  $c$  statistic, and measures of clinical usefulness are subject to sampling error in real life.

### 19.7.1 Uncertainty in Validation of Performance

We illustrate the empirical behaviour of measures for calibration and discrimination of logistic regression models. The prediction model is the same as before, with a linear predictor defined by ten normally distributed  $x$  variables, each with a regression coefficient of 0.76. The model has a  $c$  statistic of 0.812. We consider small to large sample sizes for model development ( $N_{\text{dev}} = 100\text{--}10,000$ ) and for model validation ( $N_{\text{val}} = 100\text{--}10,000$ ), with outcome incidence 50% or 10%. Simulations are first performed under the Null hypothesis, i.e. that both samples originate from the same underlying population (Table 19.6). Case-mix and regression coefficients

**Table 19.6** Estimation of calibration and discrimination of logistic regression models in small to large sample sizes for model development and for model validation

Scenario	Events/ $N_{\text{dev}}$	Events/ $N_{\text{val}}$	$a b=1$	slope $b$	$c$ statistic
<i>Incidence 50%</i>					
Large sizes	5,000/10,000	5,000/10,000	$0 \pm 0.03$	$1.00 \pm 0.03$	$0.81 \pm 0.004$
Small development samples	50/100	5,000/10,000	$0 \pm 0.28$	$0.64 \pm 0.15$	$0.77 \pm 0.017$
	100/200		$0 \pm 0.17$	$0.82 \pm 0.13$	$0.79 \pm 0.010$
	250/500		$0 \pm 0.12$	$0.92 \pm 0.09$	$0.80 \pm 0.006$
	500/1000		$0 \pm 0.08$	$0.95 \pm 0.07$	$0.81 \pm 0.005$
	1,000/2,000		$0 \pm 0.06$	$0.97 \pm 0.05$	$0.81 \pm 0.004$
Small validation samples	5,000/10,000	50/100	$0 \pm 0.24$	$1.06 \pm 0.24$	$0.82 \pm 0.043$
		100/200	$0 \pm 0.16$	$1.03 \pm 0.17$	$0.81 \pm 0.030$
		250/500	$0 \pm 0.11$	$1.01 \pm 0.10$	$0.80 \pm 0.018$
		500/1,000	$0 \pm 0.08$	$1.00 \pm 0.07$	$0.81 \pm 0.014$
		1,000/2,000	$0 \pm 0.06$	$1.00 \pm 0.05$	$0.81 \pm 0.009$
Small development samples, half size validation	50/100	25/50	$0 \pm 0.52$	$0.71 \pm 0.31$	$0.77 \pm 0.070$
	100/200	50/100	$0 \pm 0.34$	$0.83 \pm 0.25$	$0.79 \pm 0.048$
	250/500	100/200	$0 \pm 0.20$	$0.95 \pm 0.18$	$0.80 \pm 0.030$
	500/1,000	250/500	$0 \pm 0.13$	$0.98 \pm 0.11$	$0.81 \pm 0.018$
	1,000/2,000	500/1,000	$0 \pm 0.10$	$0.99 \pm 0.09$	$0.81 \pm 0.014$
Small development samples and equal size validation samples	50/100	50/100	$0 \pm 0.44$	$0.66 \pm 0.23$	$0.77 \pm 0.051$
	75/150	75/150	$0 \pm 0.32$	$0.77 \pm 0.22$	$0.78 \pm 0.039$
	100/200	100/200	$0 \pm 0.27$	$0.82 \pm 0.19$	$0.79 \pm 0.033$
	175/350	175/350	$0 \pm 0.19$	$0.89 \pm 0.15$	$0.80 \pm 0.023$
	250/500	250/500	$0 \pm 0.15$	$0.93 \pm 0.13$	$0.80 \pm 0.019$
	500/1,000	500/1,000	$0 \pm 0.11$	$0.97 \pm 0.09$	$0.81 \pm 0.014$
	1,000/2,000	1,000/2,000	$0 \pm 0.08$	$0.99 \pm 0.07$	$0.81 \pm 0.010$
<i>Incidence 10%</i>					
Large sizes	1,000/10,000	1,000/10,000	$0 \pm 0.05$	$1.00 \pm 0.05$	$0.83 \pm 0.007$
Selected combinations of development and validation sample sizes	50/500	50/500	$0 \pm 0.25$	$0.85 \pm 0.18$	$0.81 \pm 0.033$
		100/1,000	$0 \pm 0.23$	$0.85 \pm 0.15$	$0.81 \pm 0.021$
		200/2,000	$0 \pm 0.19$	$0.86 \pm 0.14$	$0.81 \pm 0.018$
		1,000/10,000	$0 \pm 0.18$	$0.86 \pm 0.14$	$0.81 \pm 0.010$
	100/1,000	50/500	$0 \pm 0.22$	$0.93 \pm 0.17$	$0.82 \pm 0.032$
		100/1,000	$0 \pm 0.18$	$0.93 \pm 0.13$	$0.82 \pm 0.021$
		200/2,000	$0 \pm 0.15$	$0.93 \pm 0.11$	$0.82 \pm 0.015$
		1,000/10,000	$0 \pm 0.13$	$0.93 \pm 0.11$	$0.82 \pm 0.008$
	200/2,000	50/500	$0 \pm 0.19$	$0.95 \pm 0.15$	$0.82 \pm 0.031$
		100/1,000	$0 \pm 0.18$	$0.96 \pm 0.13$	$0.82 \pm 0.022$
		200/2,000	$0 \pm 0.11$	$0.97 \pm 0.10$	$0.83 \pm 0.017$
		1,000/10,000	$0 \pm 0.09$	$0.96 \pm 0.07$	$0.82 \pm 0.007$
	1,000/10,000	50/500	$0 \pm 0.17$	$1.01 \pm 0.15$	$0.83 \pm 0.030$
		100/1,000	$0 \pm 0.13$	$0.99 \pm 0.10$	$0.83 \pm 0.021$
		200/2,000	$0 \pm 0.09$	$1.00 \pm 0.07$	$0.83 \pm 0.015$

Numbers are mean  $\pm$  standard error, as observed in simulations (100–1,000 repetitions for sufficiently stable results)

were hence identical in both settings, and estimates may only vary because of finite sample sizes at development and/or validation.

With 50% incidence of the outcome in very large development and validation sizes ( $N_{\text{dev}} = 10,000$  and  $N_{\text{val}} = 10,000$ ), the standard errors (SEs) are small: The SE around the calibration-in-the-large and calibration slope  $b$  is 0.03, around the  $c$  statistic 0.004. With 10% incidence,  $N_{\text{dev}} = 10,000$  and  $N_{\text{val}} = 10,000$ , the SEs are larger, corresponding to the lower number of events (1,000 instead of 5,000).

We find that the calibration-in-the-large is 0 on average in all scenarios; the SE depends on the size of the development sample and the size of the validation sample. With only 100 subjects for model development, the SE is 0.28 if validation is in 10,000 subjects; if validation is in 50 or 100 subjects, the SE is much larger ( $\pm 0.52$  and  $\pm 0.44$  respectively). A quite low SE ( $\pm 0.06$ ) is found with  $n=2,000$  for model development and 10,000 for model validation, or with a reversal of this design (development  $n=10,000$ , validation  $n=2,000$ ).

The calibration slope is below 1 when small samples are used for model development (e.g. slope  $b=0.65$  with  $N_{\text{dev}} = 100$  and  $N_{\text{val}} = 10,000$ , reflecting clear overfitting and a need for shrinkage of coefficients). In contrast, small validation samples lead to an upward bias for the slope (e.g. slope  $b=1.08$  with  $N_{\text{dev}} = 10,000$  and  $N_{\text{val}} = 100$ ). The SE is somewhat larger with small validation samples than with small development samples (e.g.  $N_{\text{dev}} = 100$ : SE  $\pm 0.15$ ;  $N_{\text{val}} = 100$ : SE  $\pm 0.25$ ).

The discriminative ability ( $c$  statistic) was 0.81 in the population, but smaller with small development samples (e.g.  $c=0.77$  with  $N_{\text{dev}} = 100$ ,  $N_{\text{val}} = 10,000$ ). Again small validation samples led to an upward bias (e.g.  $c=0.82$  with  $N_{\text{dev}} = 10,000$  and  $N_{\text{val}} = 100$ ). The SE was markedly higher with small validation samples (e.g.  $N_{\text{dev}} = 100$ : SE  $\pm 0.017$ ;  $N_{\text{val}} = 100$ : SE  $\pm 0.043$ ). Apparently, small development samples lead to poor discriminating models, which can reliably be quantified with large validation samples, but small validation samples lead anyway to uncertain estimates of discrimination.

### \*19.7.2 Estimating Standard Errors in Validation Studies

In Table 19.6, we calculate SEs empirically by studying the distribution of coefficients over samples. We can also use the asymptotic SE for the performance measures. The SE of calibration-in-the-large and calibration slope can be obtained from the variance estimates in logistic regression models. The SE of the  $c$  statistic can be calculated with standard formulas for rank order statistics.<sup>172</sup> We found that the asymptotic SEs agreed rather well with the empirical estimates.

### 19.7.3 Summary Points

- Variability is substantial with small development samples, but especially with small validation samples

- The effective sample size is largely determined by the number of events rather than the total sample size
- SEs can be estimated with asymptotic formulas or from simulations (“empirically”)

## 19.8 Design of External Validation Studies

The variability in performance has implications for the design and power of validation studies (see references for validation of linear regression models).<sup>392,336</sup> We have seen in Chap. 17 that the bootstrap is generally preferable for internal validation purposes. Despite its inefficiency, some researchers may like a split-sample approach to convince their readership. This design was discouraged in Chap. 17. A common ratio in such a design is 2/3 of the sample for model development and 1/3 for validation. According to Table 19.6, a lower variability of performance is obtained with a half–half split-sample design; but this design has more optimism in calibration slope and discrimination. A 2:1 ratio may be a reasonable balance between optimizing bias and variability.

For external validation we may well choose a temporal validation design.<sup>222</sup> But we then face the same question on how to choose the size of the development data set vs. the size of the more recent validation set. With spatial validation, e.g. “leave-one-centre-out” cross-validation, the validation sets may be much smaller than the development set. The results in Table 19.6 show that this makes the performance quite uncertain in each validation part per se.

Another situation is that a model was published, and we simply wish to externally validate this model for our setting. We set up a fully independent external validation study, and wonder about a reasonable sample size, accepting the developed model as reasonable to test. This design requires some estimates of power to detect relevant differences in performance.

### 19.8.1 Power of External Validation Studies

Power calculations depend on various quantities: Statistical Type I and Type II error; the variability in the quantity we want to test, and the “clinically relevant” difference we do not want to miss. Type I error is conventionally set at 5%, and type II error at 20% (power 80%). The variability of performance measures is shown in Table 19.6. Note that these are empirically derived SEs for one specific logistic regression model (with ten normally distributed predictors). In practice, we may only know the asymptotic (i.e. estimated) SE of some measures such as the model intercept. Clinically relevant differences may be context-dependent. For logistic regression models we might consider a systematic over or underestimation by 1.5 times the odds of the outcome (intercept + or –  $\ln(1.5)$ ), a calibration slope less than

0.8 (difference 0.2 with ideal slope of 1), and a decrease in  $c$  statistic by more than 0.05 (given the same case-mix).

Some specific issues come up in power calculations for validation studies. The first is whether we should perform one-sample or two-sample tests. If we consider the prediction model as a system generating predictions, a one-sample test is reasonable to test whether the validation performance deviates from hypothesized values. For calibration, these values are obvious: 0 for calibration-in-the-large and 1 for calibration slope. For the  $c$  statistic, we may consider the reference value given the case-mix in the validation setting (see Sect. 19.11). For the  $c$  statistic we might also consider a two-sample test, including uncertainty in the estimate from the development setting. A further issue is whether we should perform one-sided or two-sided tests. Calibration-in-the-large asks for a two-sided test, since the incidence in the validation setting may be higher or lower than predicted. But for calibration slope we could test for  $\text{slope} < 1$ , rather than  $\text{slope} \neq 1$ . Similarly, only a decrease in discrimination is an interesting alternative hypothesis.

Finally, one might argue that we should consider assessment of validity as a non-inferiority design. This implies that we change the Null hypothesis to stating that the model is invalid, and test whether the model performance is within reasonable limits from the expected value. The reasonable limits may be context dependent, similar to defining “clinically relevant” differences in traditional sample-size calculations.

### \*19.8.2 Required Sample Sizes for Validation Studies

We first approximate the power given the SE under the null hypothesis, i.e. the model was actually valid in both development and validation setting. We consider SEs for model development with a large sample size in Table 19.6, such that the predominant source of variability is the validation sample size. For simplicity we use one-sample tests for all measures. For calibration-in-the-large, we use a two-sided test; for calibration slope, a one-sided test ( $\text{slope} < 1$ ); for the  $c$  statistic, a one-sided test ( $c < c_{\text{reference}}$ ). The critical values<sup>1</sup> for power calculations are determined by Type I and Type II error, which we set at 5% (one-sided or two-sided) and 20% (one-sided). The critical value is  $1.96+0.84=2.80$  for two-sided tests, and  $1.64+0.84=2.49$  for one-sided tests. We multiply these critical values with the SE to obtain the minimum differences that can be detected with 80% power (Table 19.7).

As expected, small validation sizes only have 80% power to detect substantial invalidity. For example, if we validate a model in a sample with 50 events and 50 non-events, we only have enough power to detect a calibration-in-the-large problem with twice too high, or twice too low predictions (odds ratio, 1.96); a dramatically poor calibration slope (less than 0.4), and a decrease in  $c$  statistic over 0.1 (Table 19.7). To detect a more modest calibration-in-the-large problem, such as 1.5 times too low or too high predictions, we would need at least 100 events and 100 non-events

---

<sup>1</sup>Critical value: the value that a test statistic must exceed for the null hypothesis to be rejected.

**Table 19.7** Required sample size for 80% power when validating a logistic regression model in a setting with 50% or 10% incidence of the outcome

Scenario	Events/ $N_{\text{val}}$	$a b=1 \leftrightarrow 1^a$	slope $b < 1^b$	$c_{\text{validation}} < c_{\text{reference}}^c$
Incidence 50%	50/100	$\pm 0.67$ , OR=1.96	<0.40	<-0.107
	100/200	$\pm 0.45$ , OR=1.57	<0.58	<-0.077
	250/500	$\pm 0.31$ , OR=1.36	<0.75	<-0.045
	500/1,000	$\pm 0.22$ , OR=1.25	<0.83	<-0.035
	1,000/2,000	$\pm 0.17$ , OR=1.18	<0.88	<-0.022
Incidence 10%	50/500	$\pm 0.45$ , OR=1.61	<0.63	<-0.075
	100/1000	$\pm 0.34$ , OR=1.44	<0.75	<-0.052
	200/2000	$\pm 0.25$ , OR=1.29	<0.83	<-0.037

OR, Odds ratio

<sup>a</sup> Asymptotic SE and minimum OR that can be detected with 80% power<sup>b</sup> Minimum slope that can be detected with 80% power<sup>c</sup> Minimum differences in  $c$  statistic that can be detected with 80% power**Table 19.8** Power for slope  $< 1$  (true value, 0.84) and  $c$  statistic decrease (true decrease from, 0.821 to 0.778, -0.043)

Scenario	Events/ $N_{\text{val}}$	slope $b$ 0.84	$c$ statistic -0.043
Incidence 50%	50/100	15%	11%
	100/200	25%	24%
	250/500	50%	57%
	500/1,000	78%	87%
	1,000/2,000	97%	99%

(total sample size  $> 200$ ). This sample size would also have 80% power for a slope less than 0.58, and a decrease in  $c$  by 0.077. With more non-events (incidence of outcome, 10%), the picture is slightly better in terms of number of events required, but the total sample size should be at least 1,000 (100 events) for reasonable power.

In a secondary analysis, we simulate power in the case that the prediction model is invalid. We create a model with coefficients 0.76 for ten normally distributed predictors  $x_1$  to  $x_{10}$ , and validate in a setting where the coefficients are 0.5 or 1.5 times as large (see Fig. 19.9). In the validation setting, calibration-in-the-large is fine (average, 0), but the slope is 0.84 instead of 1, and the  $c$  statistic is 0.778 instead of 0.821 in the development setting (decrease, 0.043). From Table 19.7, we expect that the power for detecting that the slope is lower than 0.84 is slightly below 80% with 500 events; indeed we find 78% power with this sample size (Table 19.8). For a decrease in  $c$  statistic by -0.043, we expect that more than 250 events and 250 non-events are required; indeed the power is 57% with these numbers, and 87% with 500 events.

### 19.8.3 Summary Points

- The variability of external validation assessments depends on the size of the development sample and the size of the validation sample

- For statistical testing, we may accept the prediction model as given, and hence perform one-sample tests in the validation data
- For such tests to have reasonable power, we need at least 100 events and at least 100 non-events in external validation studies, but preferably more (>250 events). With lower numbers the uncertainty in performance measures is large.

## 19.9 Concluding Remarks

The performance of a prediction model in a new setting (“generalizability” or “transportability”) essentially depends on two aspects: the validity of the regression coefficients, and the case-mix in the validation setting. The validity of regression coefficients can be assessed by comparing regression coefficients between settings. Indeed we note that many validation studies report on the coefficients in their sample and compare these to the previous estimates. With relatively small development and validation samples it would be highly coincidental if coefficients agreed well. Even if the two samples came from exactly the same underlying population, chance processes will cause the coefficients in both samples to differ from each other to some extent, with some coefficients larger and some smaller than expected from the development sample.

Differences in case-mix between development and validation setting are usually considered informally, by comparing patient characteristic in a kind of “Table 1.” One usually makes only informal comparisons to the case-mix in the development sample. Some statistical measures have previously been proposed for a more formal assessment of comparability, such as the “ $M$  statistic” to compare trauma populations.<sup>53</sup> With this approach, survival probabilities of trauma patients are grouped, for example as 0–25%, 26–50%, 51–75%, 76–90%, 91–95%, and 96–100%. The fraction of patients in these groups at validation is compared with the fraction at model development. The smaller of the two fractions is summed over all groups. This creates a number ranging from 0–1.  $M$  values close to 1 indicate a perfect match with the development case-mix, while 0 indicates a total discrepancy between the two samples. An arbitrary cutoff point of 0.88 has been suggested, and studies with  $M$  values below 0.88 should be “interpreted cautiously.”<sup>53</sup>

We followed a more systematic approach to study the influence of differences in case-mix. Differences in predictor distributions (“ $X$ ”) do not affect calibration, and only discrimination aspects, as long as the model is correctly specified for the range of  $X$  values examined. If non-linearities and/or interactions had been missed at model development, we can imagine that shifting to another predictor distribution may impact on calibration as well. Furthermore, we may assume that a very different distribution in  $X$  implies that differences in missed predictors (“ $Z$ ”) are also likely. Differences in missed predictors between settings may severely invalidate a prediction model, both with respect to calibration (especially calibration-in-the-large) and discrimination. When predictions are systematically miscalibrated, we can make systematically wrong decisions based on the model. This may lead to a negative NB of using the model, com-

pared with a default policy without using the model. It is therefore important to perform external validation studies.<sup>40</sup>

We also noted that the distribution of predictors can formally be taken into account in the calculation of reference values for model performance, given that the model is valid in the validation sample. This may be very useful to obtain insight in what is happening at validation: differences in case-mix or differences in regression coefficients.

Finally, we studied design issues of validation studies for predictive regression models. If a temporal split is made, a 2:1 ratio may be reasonable. This limits overfitting at development, and still gives reasonable power at validation. A validation data set should contain at least 100 events and 100 non-events for reasonable power.<sup>401,465</sup> For the detection of smaller but still quite relevant invalidity, higher sample sizes are advisable, e.g. 250 events and 250 non-events or 100 events and 900 non-events.<sup>450,493</sup>

## Questions

### 19.1 Differences between populations (Table 19.1)

Consider a hypothetical model that is developed with logistic regression analysis in a sample of 100 patients in a clinical setting. The model is validated in a screening setting. What differences would you expect with respect to

- (a) case-mix
- (b) regression coefficients

### 19.2 Validity of a model

What would happen to the calibration and discrimination of a prediction model if

- (a) units of measurement were wrong, e.g. mg/dl vs. mmol/L?
- (b) a different measurement device was used, with random deviations compared with the measurements in the development setting
- (c) a more heterogeneous case-mix was present in the validation setting
- (d) a treatment that was very effective for all patients was used
- (e) a treatment that was very effective for one subgroup was used

### 19.3 Influence of case-mix on clinical usefulness

A less-severe case-mix led to a higher net benefit than a more-severe case-mix (NB 0.104 vs. 0.006, Fig. 19.3). How do you explain this finding?

### 19.4 Disturbance of calibration (Sect. 19.8)

We found that calibration is not disturbed when the validation setting consists of (a) more-or less-severe patients according to predictor values, or (b) more or less heterogeneous patients according to observed or missed predictor values.

- (a) What disturbs calibration-in-the-large?
- (b) What disturbs the calibration slope?

**19.5 Discrimination and clinical usefulness**

Why is discrimination a *necessary but not sufficient* condition for clinical usefulness?

**19.6 Reference values for performance (Sect. 19.6)**

Reference values indicate a model's performance under the condition that the model predictions are valid in the validation sample. How is it possible that the reference value for performance can be better than the performance estimate in the development setting?

**19.7 Power of validation studies (Table 19.7)**

Suppose we wish to detect a possible deterioration in calibration-in-the-large of an odds ratio of 1.5, and a calibration slope  $< 0.8$ . What sample size would you recommend?

**19.8 Study design: epidemiologic and statistical aspects**

Suppose we can do a single centre study with 1,000 patients, where 200 (20%) will have the event of interest. Alternatively, we can do a multi-centre study with three centres, each contributing 300 patients. Among the 900, we expect 180 (20%) patients with the event of interest.

Which design would you prefer? Explain why, weighing epidemiological considerations (such as generalizability) and statistical considerations (such as standard error).

# Chapter 20

## Updating for a New Setting

**Background** A prediction model ideally provides valid predictions of outcome for individual patients at another setting than where the model was developed, e.g. differing in time and place. The validity of predictions can be assessed by comparing observed outcomes and predictions when empirical data from this setting are available. Various patterns of invalidity may however be observed as we have seen in the previous chapter. Detection of calibration-in-the-large problems should have top priority since miscalibration can cause systematically wrong decision making with the model (negative net benefit). Obviously, we may subsequently aim to update the model to improve predictions for future patients from the new setting. We discuss several approaches for updating a previously developed model. The risk is that simply re-estimating all regression coefficients in a model might replace reliable but slightly biased estimates by unbiased but very unreliable ones, particularly if the validation data set is relatively small.

We start with considering updating methods that focus on re-calibration (re-estimation of the intercept and/or updating of the slope of the linear predictor). Next, we turn to more structural model revisions (re-estimation of some or all regression coefficients, model extension with more predictors). For illustration we consider case studies with updating of a previously developed logistic regression model, a regression tree, and a previously developed Cox regression model. We conclude that parsimonious updating methods may often be preferable to more extensive model revisions, which should only be attempted with relatively large validation samples, in combination with shrinkage of differences between the updated model and the previously developed model.

### 20.1 Updating the Intercept

The external validity (or generalizability) of model predictions is important when a previously developed model is applied in another setting, such as another medical centre, and/or in a more recent time period. When empirical data are available, we can assess the external validity according to measures such as calibration and discrimination. Also, we may consider updating a previously

developed model, such that the prediction model is adjusted to local and/or contemporary circumstances.

The first issue to consider is calibration-in-the-large. The mean observed outcome should be equal to the mean of the predicted outcomes; for a survival outcome, the number of observed deaths should agree with the predicted number. Calibration-in-the-large is controlled by the model intercept for continuous and dichotomous outcomes and by the baseline hazard function in a survival model. Several approaches can be followed to adjust the intercept for a new setting.

### ***20.1.1 Simple Updating Methods***

A simple approach is to consider the mean observed outcome in the new setting, and compare this to the mean of the development setting. The difference is used to update the intercept. This is a naïve Bayesian approach, based on a univariate comparison of outcomes incidences in the development and validation setting. This approach has been shown to work reasonably well in a number of case studies, suggesting that differences in mean outcome are often largely attributable to factors outside the model.<sup>299,67</sup>

Similarly, it is possible to present a prediction model with the explicit option to use a setting-specific intercept. An example is the score chart for operative mortality for elective aortic aneurysm surgery (Chap. 14).<sup>421</sup> Another example is a model to guide the indication for a CT scan in patients with minor head injury.<sup>391</sup> The model was developed in a setting with 243 of 3,181 (7.6%) patients presenting with intracranial traumatic lesions. The model was presented with a range from 2.5% to 15% for the “prior probability” of an intracranial traumatic lesion. Such a simple adjustment is directly applicable if the case-mix between development and validation samples is fully comparable with respect to the predictors in the model. A variant is to use the mean outcome and the mean of predictor values in the calculation of the required update of the intercept.<sup>264</sup> The intercept adjustment reflects differences between settings in other aspects than captured by the predictors.

A special case is infectious disease prediction, where seasonal patterns are important and epidemics occur. These background incidences have impact on the intercept of prediction rules for infectious diseases.<sup>346,123</sup>

### ***20.1.2 Bayesian Updating***

Another approach, similar to shrinkage (see Chap. 13), is to use Bayesian estimation methods for model updating. We assume that the development and validation samples come from an underlying superpopulation with some heterogeneity between settings. We use the estimated heterogeneity to obtain Bayesian estimates

of updated coefficients. An updated intercept  $\alpha_{\text{updated}}$  can be obtained with a standard formula:<sup>300</sup>

$$\alpha_{\text{updated}} = \mu + \tau^2 / (\tau^2 + \sigma_{\text{estimated}}^2) * (\alpha_{\text{estimated}} - \mu)$$

where  $\mu$  is the overall mean estimate,  $\tau^2$  is the variance between development and validation settings (“heterogeneity”), and  $\alpha_{\text{estimated}}$  and  $\sigma_{\text{estimated}}^2$  are the estimated intercept and its variance in the validation sample. A relatively large sampling uncertainty (large  $\sigma_{\text{estimated}}^2$ ) implies substantial shrinkage for  $\alpha_{\text{estimated}}$  towards the overall mean  $\mu$ . In contrast, large heterogeneity (large  $\tau^2$ ) implies that  $\alpha_{\text{estimated}}$  is not much shrunk towards the overall mean  $\mu$ . The extreme is that  $\tau^2$  is infinity, i.e. each  $\alpha_{\text{estimated}}$  is used as estimate for  $\alpha_{\text{updated}}$ . Every setting is considered as unique and may have any intercept. The latter is implausible, and argues for some form of Bayesian analysis. The problem of such a Bayesian analysis is however that we need to specify a value for  $\tau^2$ ;  $\mu$  is readily obtained from the previous model.

A full Bayesian approach is to elicit  $\tau^2$  from experts; they may for example state that it is unlikely that the incidence of the outcome (adjusted for the prediction model) is more than 4 times lower or higher than the original incidence.<sup>152</sup> Interpreting these limits as 95% credibility intervals means that  $\tau \approx \log(4)/2 = 0.69$  and  $\tau^2 = 0.48$ . Stating that the limits are 2 times lower to 2 times higher incidence implies  $\tau \approx \log(2)/2 = 0.35$  and  $\tau^2 = 0.12$ , leading to more shrinkage.

The Empirical Bayes approach is to estimate  $\tau^2$  from the distribution of intercepts in different validation samples. This approach will be followed in the following chapter. In addition to the intercept, we can in principle consider other model parameters in a Bayesian framework, for example the calibration slope or individual regression coefficients.

## 20.2 Approaches to More-Extensive Updating

In addition to calibration-in-the-large, further aspects of calibration need to be considered. These may conveniently be studied in the context of a general calibration model, where the linear predictor based on the previously developed model is the only covariate.<sup>86</sup> This model has only two free parameters: intercept  $\alpha$  and calibration slope  $\beta_{\text{overall}}$ . A simple updating method might focus on re-calibration, i.e. that the updated model has a new intercept  $\alpha$  and new regression coefficients based on multiplication of the original coefficients with  $\beta_{\text{overall}}$ . This re-calibration approach has been followed for updating of a previously developed model in the context of risk-adjustment<sup>95,212</sup> and prediction.<sup>456,291</sup> We may also consider more extensive updating methods (“model revision”), such as re-estimation of regression coefficients of some or all predictor variables, and considering more covariables for inclusion of the model (“model extension,” following terminology proposed by Van Houwelingen).<sup>456</sup>

### \*20.2.1 A comparison of Eight Updating Methods

We consider eight updating methods for predictions of binary outcomes (Table 20.1). For illustration we assume that a previously developed logistic regression model is available with eight predictors, but that eight more are of interest as potential predictors for the validation setting. The described methods generalize to updating of any previously developed prediction model. The methods are ordered according to the number of parameters that are estimated for updating of the original model.<sup>402</sup>

- *No updating*

The first method is not to allow for any updating, that is to keep all regression coefficients fixed at their original value, including the intercept. The linear predictor  $lp$  for method 1 ( $lp_1$ ) is calculated as

$$lp_1 = \alpha_{\text{orig}} + \beta_{\text{orig}} * X_{1..8},$$

where  $\alpha_{\text{orig}}$  indicates the intercept from the original study;  $\beta_{\text{orig}}$  the regression coefficients from the original study; and  $X_{1..8}$  the eight predictors in the new (validation) sample. This method provides a reference upon which improvement should be obtained with any updating method.

**Table 20.1** Updating methods considered for a previously developed logistic regression model with eight predictors, in a validation sample where eight more potential predictors are available

No.	Label	Notation	Predictors considered	Parameters tested	Parameters estimated
<i>No updating</i>					
1	Apply original prediction model	–	8	0	0
<i>Re-calibration</i>					
2	Update intercept	$\alpha$	8	0	1
3	Re-calibration of intercept and slope	$\alpha + \text{calibration slope}$ $\beta_{\text{overall}}$	8	0	2
<i>Model revision</i>					
4	Re-calibration + selective re-estimation	$\alpha + \beta_{\text{overall}} + \gamma_{1..8 p \leq 0.05}$	8	8	2–9
5	Re-estimation	$\alpha + \beta_{1..8}$	8	0	9
<i>Model extension</i>					
6	Re-calibration + selective re-estimation + selective extension	$\alpha + \beta_{\text{overall}} + \gamma_{1..8 p \leq 0.05} + \beta_{9..16 p \leq 0.05}$	16	16	2–17
7	Re-estimation + selective extension	$\alpha + \beta_{1..8} + \beta_{9..16 p \leq 0.05}$	16	8	9–17
8	Re-estimation + extension	$\alpha + \beta_{1..16}$	16	0	17

- *Re-calibration*

The second and third methods are simple re-calibration methods. Updating of the intercept  $\alpha$  intends to correct “calibration-in-the-large,” i.e. to make the average predicted probability equal to the observed overall event rate:

$$lp_2 = \alpha_{\text{new}} + lp_1.$$

Hereto we may fit a logistic regression model in the validation sample with the intercept  $\alpha$  as the only free parameter and the linear predictor  $lp_1$  as an offset variable (i.e. the slope is fixed at unity).

In method 3, we update both the intercept  $\alpha$  and the overall calibration slope  $\beta_{\text{overall}}$  by fitting a logistic regression model in the validation sample with  $lp_1$  as the only covariate:

$$lp_3 = \alpha_{\text{new}} + \beta_{\text{overall}} \times lp_1.$$

This method has also been labelled “logistic calibration.”<sup>176</sup>

- *Model revision*

Methods 4 and 5 re-estimate more parameters in the model, referred to as “model revision.” With method 4, we first perform method 3, and then test whether predictors have an effect that is clearly different in the validation sample. We hereto perform likelihood ratio tests of model extensions in a forward stepwise manner, considering the predictor with the strongest difference first. We may extend the revised model until all differences in predictive effects have  $p > 0.05$  for each predictor (or another  $p$  value or use AIC). As a maximum, seven predictors could be selected, since  $\beta_{\text{overall}}$  was always included in the model. The number of estimated parameters could hence vary between two and nine. The linear predictor becomes:

$$lp_4 = \alpha_{\text{new}} + \beta_{\text{overall}} lp_1 + \gamma_{1..8|p \leq 0.05} \times X_{1..8|p \leq 0.05},$$

where a maximum of 7 of the 8 predictors is selected, and  $\gamma_i$  indicates the deviation from the re-calibrated coefficient:  $\gamma_i = \beta_i - \beta_{\text{overall}} lp_1$ . We estimate  $\gamma_i$  with a logistic regression model in the validation sample with the re-calibrated linear predictor  $lp_3$  as an offset variable (i.e. the slope is fixed at unity).

With method 5 we fit the 8 predictor model in the validation data:

$$lp_5 = \alpha_{\text{new}} + \beta_{\text{new}} \times X_{1..8},$$

where  $\alpha_{\text{new}}$  and  $\beta_{\text{new}}$  indicate the intercept and eight regression coefficients for the validation sample. Note that method 4 falls in between method 3 and 5: If selection of  $\gamma_i$  is extremely stringent ( $p$  value of 0), method 4 is equal to method 3 (no individual coefficients re-estimated), and if selection is extremely liberal ( $p$  value of 1), method 4 is equal to method 5 (all individual coefficients re-estimated). We therefore label method 4 re-calibration + selective re-estimation.

- *Model extension*

Methods 6–8 consider additional predictors, and are labelled “model extension” methods. Method 6 is a variant of method 4: We re-calibrate the original model

with an intercept  $\alpha$  and the overall calibration slope  $\beta_{\text{overall}}$  and test 16 predictors for statistically significant effects. The linear predictor becomes:

$$lp_6 = \alpha_{\text{new}} + \beta_{\text{overall}} lp_1 + \gamma_{1..8|p \leq 0.05} \times X_{1..8|p \leq 0.05} + \gamma_{9..16|p \leq 0.05} \times X_{9..16|p \leq 0.0}$$

where at most 15 of the 16 predictors are selected.

Method 7 is a variant of method 5, where we re-estimate the original model and selectively extend the model with more predictors  $X_{9..16}$  that have statistically significant predictive effects in the validation sample:

$$lp_7 = \alpha_{\text{new}} + \beta_{\text{new}} \times X_{1..8} + \gamma_{9..16|p \leq 0.05} \times X_{9..16|p \leq 0.0}$$

With method 8 we fit a model with 16 predictors, i.e. eight from the original model and eight additional predictors:

$$lp_8 = \alpha_{\text{new}} + \beta_{\text{new}} \times X_{1..16}$$

## 20.3 Case Study: Validation and Updating in GUSTO-I

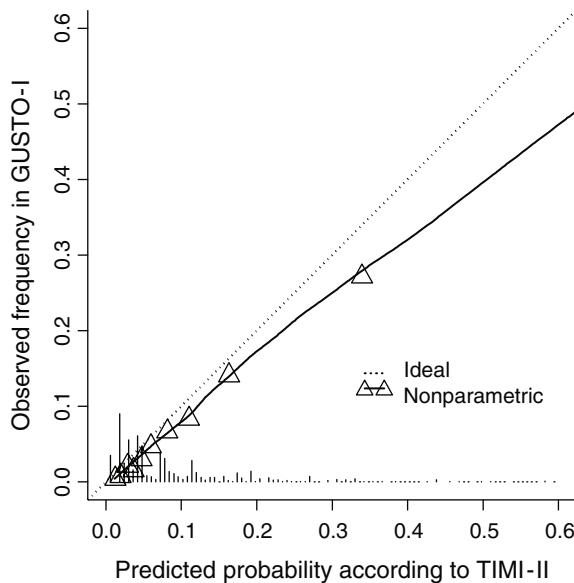
For illustration of updating methods we consider a prediction model for patients with acute MI that was developed with logistic regression analysis in the TIMI-II trial.<sup>302</sup> This trial included 3,339 patients treated in 50 US centres between 1986 and 1988.<sup>1</sup> The model was developed with backward stepwise selection methods and some continuous predictors were dichotomized. Although these approaches may be considered suboptimal for model development, we may still consider the “TIMI-II model” relevant for generating predictions in GUSTO-I.

The TIMI-II model included eight dichotomous predictors: shock, age > 65 years, high risk (anterior infarct location or previous MI), diabetes, hypotension (systolic blood pressure < 100 mmHg), tachycardia (pulse > 80), relief of chest pain > 1 h, and female gender. The outcome was 42-day mortality, in contrast to 30-day mortality in GUSTO-I.

### 20.3.1 Validity of TIMI-II Model for GUSTO-I

We construct a calibration plot for a first impression of validity of the TIMI-II model for the GUSTO-I patients (Fig. 20.1). We note that the observed mortality is systematically lower than that predicted. This may be attributed to the slight difference in outcome definition (30-day mortality in GUSTO-I vs. 42-day mortality in TIMI-II) and improvements in care for acute MI patients.

The validity is further assessed by comparing the regression coefficients between TIMI-II and GUSTO-I (Table 20.2). We note that the coefficients are reasonably



**Fig. 20.1** Calibration plot of the TIMI-II model (developed in  $n=3,339$ ) to predict 30-day mortality after acute myocardial infarction in GUSTO-I ( $n=40,830$ ). The *solid line* represents a non-parametric smooth curve for the relation between observed frequency and predicted probability. Perfect calibration is represented by the *dotted line* through the origin. *Triangles* are based on deciles of patients with similar predicted probabilities. The distribution of predicted probabilities is shown above the  $x$  axis (*vertical lines*). We note that the predicted risks are systematically too high; e.g. the highest decile has a mean predicted probability of 35% while the observed frequency is 27%<sup>402</sup>

**Table 20.2** Logistic regression coefficients  $\pm$  standard error in the TIMI-II data and in parts of the GUSTO-I data<sup>402</sup>

Predictors	TIMI-II ( $n=3,339$ )	GUSTO-I total ( $n=40,830$ )	GUSTO-I US patients ( $n=23,034$ )	GUSTO-I W region ( $n=2,188$ )	GUSTO-I sample 5 ( $n=429$ )
Shock	$1.79 \pm 0.29$	$1.60 \pm 0.08$	$1.56 \pm 0.11$	$2.39 \pm 0.41$	$2.96 \pm 0.92$
Age>65	$0.99 \pm 0.18$	$1.43 \pm 0.05$	$1.34 \pm 0.06$	$1.64 \pm 0.22$	$1.37 \pm 0.49$
High risk	$0.92 \pm 0.26$	$0.71 \pm 0.04$	$0.70 \pm 0.06$	$0.85 \pm 0.21$	$0.76 \pm 0.50$
Diabetes	$0.74 \pm 0.19$	$0.28 \pm 0.05$	$0.31 \pm 0.07$	$0.07 \pm 0.25$	$-0.11 \pm 0.64$
Hypotension	$0.69 \pm 0.27$	$1.19 \pm 0.06$	$1.19 \pm 0.07$	$1.22 \pm 0.25$	$1.39 \pm 0.57$
Tachycardia	$0.59 \pm 0.16$	$0.62 \pm 0.04$	$0.61 \pm 0.06$	$0.65 \pm 0.20$	$0.88 \pm 0.49$
Time to relief	$0.53 \pm 0.20$	$0.50 \pm 0.05$	$0.51 \pm 0.06$	$0.26 \pm 0.21$	$0.68 \pm 0.54$
Sex	$0.47 \pm 0.19$	$0.43 \pm 0.04$	$0.47 \pm 0.06$	$0.62 \pm 0.20$	$-0.04 \pm 0.51$
Intercept	$-4.47 \pm 0.35$	$-4.82 \pm 0.06$	$-4.84 \pm 0.09$	$-5.09 \pm 0.30$	$-5.19 \pm 0.72$

similar, although the coefficients of age>65 and hypotension are somewhat larger in GUSTO-I, and those of shock, high risk, and diabetes smaller.

We further study the estimated coefficients in smaller parts of the GUSTO-I data set. A total of 23,034 patients are included from the United States. Within the United

States, 2,188 patients are from the West region, including 429 patients in sample 5. The logistic regression coefficient of diabetes is close to zero in the West region and negative in sample 5. The effect of sex has vanished in the smallest sample.

### 20.3.2 Updating the TIMI-II Model for GUSTO-I

We illustrate the application of the updating methods 2, 3, and 4 in Table 20.3. Corresponding to the observed miscalibration in Fig. 20.1, the intercepts are negative (around  $-0.3$ ) when method 2 is applied, with somewhat more extreme estimates in the smaller validation sets. The corresponding odds ratios are between  $0.63$  in sample 5 ( $OR = e^{-0.47}$ ,  $p=0.03$ ) and  $0.76$  in the total GUSTO-I data set ( $OR = e^{-0.28}$ ,  $p<0.001$ ), indicating that the predicted probabilities are 1.3 to 1.6 times too high. The calibration slopes are close to 1 (method 3).

Method 4 updates the original model as in method 3 plus estimation of coefficients that are clearly different from overall re-calibrated values. We find that the differences in effects of age $>65$ , high risk, diabetes, hypotension, and tachycardia are statistically significant in the total GUSTO-I data set. No statistically significant deviations are observed in the smallest sample, obviating a clear need for re-estimation of individual coefficients (Table 20.3).

**Table 20.3** Illustration of updating of the TIMI-II model in parts of the GUSTO-I data according to calibration methods (method 2 and 3) and model revision with statistically significant different coefficients (method 4)

	GUSTO-I total (n=40,830)	GUSTO-I US patients (n=23,034)	GUSTO-I region 1 (n=2,188)	GUSTO-I sample 5 (n=429)
<i>Re-calibration</i>				
Method 2				
$\alpha$ : Intercept	$-0.28 \pm 0.02$	$-0.34 \pm 0.03$	$-0.36 \pm 0.09$	$-0.47 \pm 0.22$
Method 3				
$\alpha$ : Intercept	$-0.28 \pm 0.03$	$-0.39 \pm 0.05$	$-0.10 \pm 0.16$	$-0.26 \pm 0.47$
$\beta_{\text{overall}}$ : Slope	$0.99 \pm 0.02$	$0.98 \pm 0.03$	$1.13 \pm 0.09$	$1.11 \pm 0.22$
<i>Model revision</i>				
Method 4 <sup>a</sup>				
$\alpha$ : Intercept	$-0.76 \pm 0.15$	$-0.62 \pm 0.17$	$-0.25 \pm 0.36$	$-0.26 \pm 0.47$
$\beta_{\text{overall}}$ : Slope	$0.91 \pm 0.04$	$0.94 \pm 0.04$	$1.14 \pm 0.12$	$1.11 \pm 0.22$
$\gamma_1$ : Shock	+0	+0	+0	+0
$\gamma_2$ : Age $>65$	$+0.53 \pm 0.06$	$+0.42 \pm 0.07$	$+0.49 \pm 0.24$	+0
$\gamma_3$ : High risk	$-0.12 \pm 0.06$	$-0.17 \pm 0.07$	+0	+0
$\gamma_4$ : Diabetes	$-0.39 \pm 0.06$	$-0.38 \pm 0.08$	$-0.79 \pm 0.27$	+0
$\gamma_5$ : Hypotension	$+0.56 \pm 0.07$	$+0.52 \pm 0.08$	+0	+0
$\gamma_6$ : Tachycardia	$+0.09 \pm 0.05$	+0	+0	+0
$\gamma_7$ : Time to relief	+0	+0	+0	+0
$\gamma_8$ : Sex	+0	+0	+0	+0

<sup>a</sup>The updated regression coefficients  $\beta_i$  can be calculated as  $\beta_{\text{overall}} \times \beta_{i, \text{timi}} + \beta_i$

The results of method 5 (re-estimating all model coefficients), are already shown in Table 20.2. For updating methods 6–8, eight additional predictors are considered. These are height, weight, hypertension, smoking, hypercholesterolaemia, previous angina, family history, and ST elevation in >4 leads. These eight additional predictors are to some extent correlated to the eight TIMI-II predictors. In a 16-predictor model, the eight additional predictors are each statistically significant ( $p < 0.01$ ) in the full GUSTO-I data set ( $n = 40,830$ ) and the US part ( $n = 23,034$ ), but their predictive effects are smaller than those of the eight predictors from the TIMI-II model. In the West region, only weight and ST elevation have statistically significant effects, while none of the additional predictors have statistically significant effects in the smallest sample.

### 20.3.3 Performance of Updated Models

We hope that updating improves the performance of the prediction model. The calibration problem as noted in Fig. 20.1 is solved when the intercept is updated (all methods except method 1). The  $c$  index of the TIMI-II model was around 0.78 with methods 1–3 (Table 20.4). Updating of some (method 4) or all (method 5) of the coefficients led to a somewhat higher apparent discriminative ability ( $c$  around

**Table 20.4** Number of parameters estimated and apparent performance of updated versions of the TIMI-II model in parts of the GUSTO-I data

	Method	GUSTO-I total ( $n = 40,830$ )	GUSTO-I US patients ( $n = 23,034$ )	GUSTO-I W region ( $n = 2,188$ )	GUSTO-I sample 5 ( $n = 429$ )
Parameters estimated	1	0	0	0	0
	2	1	1	1	1
	3	2	2	2	2
	4	7	6	4	2
	5	9	9	9	9
	6	17	13	5	2
	7	17	17	11	9
	8	17	17	17	17
Discrimination ( $c$ statistic)	1	0.782	0.780	0.795	0.776
	2	0.782	0.780	0.795	0.776
	3	0.782	0.780	0.795	0.776
	4	0.793	0.791	0.810	0.776
	5	0.793	0.790	0.819	0.793
	6	0.802	0.800	0.819	0.776
	7	0.802	0.800	0.828	0.793
	8	0.802	0.800	0.830	0.852

Results are shown for methods 1–8 as defined in Table 20.1

0.80 in the larger samples). The extension of the TIMI-II model with more predictors increased the apparent discriminative ability further, although the increase was small in the total data set (from 0.79 to 0.80).

Since the apparent performance may be a severely optimistic estimate of performance in new patients, we studied the internal validity of the updated prediction models as identified with method 3, 5, and 8 for the smallest sample ( $n=429$ ). Models were developed in 200 bootstrap samples (drawn with replacement from the validation sample) and tested in the validation sample to estimate the optimism in apparent performance measures. The optimism was smallest for the 2 parameter model (method 3), and largest with the 17 parameter model (method 8), where discrimination was expected to decrease from 0.852 to 0.770. The highest internal validity was found for method 3, with optimism-corrected  $c$  0.772. This suggests that a model with updating of fewer parameters may perform better in independent data than a more extensively updated model. This issue is systematically studied in Sect. 20.4.

### \*20.3.4 R Code for Updating Methods

We start with defining 2 models in the GUSTO-I sample (sample 5,  $n=429$ ):

```
full8 <- lrm(DAY30~SHO+A65+HIG+DIA+HYP+HRT+TTR+SEX,
               data=gusto5, x=T, y=T)
full   <- lrm(DAY30~SHO+A65+HIG+DIA+HYP+HRT+TTR+SEX+
               HEI+WEI+HTN+SMK+LIP+PAN+FAM+ST4, data=gusto5, x=T, y=T)
```

The eight coefficients in TIMI-II are in the same order as the full8 model:

```
timi8.par <- c(-4.465, 1.79, 0.99, 0.92, 0.74, 0.69, 0.59,
               0.53, 0.47)
```

For method 1, we calculate the linear predictor:

```
lp1 <- full8$x %*% timi8.par[-1] + timi8.par[1]
```

For methods 2 and 3, we update the intercept or re-calibrate the model:

```
lp2 <- lrm.fit(y=full8$y, offset=lp1$linear.predictor)
lp3 <- lrm.fit(y=full8$y, x=lp1$linear.predictor)
```

For method 4, we test for deviations of effects, while always updating the intercept and slope:

```
for (i in 1:8)
  {fit4 <- lrm.fit(y=full8$y, x=cbind(full8$x[,i], lp1))
  ...} # some printing of results of fit4
```

For methods 5 and 8, we simply refit the model

```
lp5 <- full8$linear.predictor
lp8 <- full$linear.predictor
```

For methods 6 and 7 we again examine contributions of predictors beyond the effect of lp1. For example, method 7 works like:

```
for (i in 9:16)
  {fit7 <- lrm.fit(y=full$y, x=cbind(full$x[,i], full8$x))
  ...} # some printing of results of fit7
```

## 20.4 Shrinkage and Updating

Traditionally, regression coefficients can be shrunk towards zero (see Chap. 13). For model updating, we consider shrinkage of regression coefficients of revised models towards their re-calibrated values.<sup>456,402</sup> This implies that some regression coefficients are pulled to higher values rather than towards zero.

In traditional model development, a simple heuristic shrinkage factor can be defined as  $(\text{model } \chi^2 - df) / \text{model } \chi^2$  (see Chap. 13).<sup>81</sup> Here model  $\chi^2$  refers to the difference in  $-2 \log$  likelihood between a model with and without predictors, and  $df$  refers to the degrees of freedom used by the predictors. We can use the same formula in the context of model revision (methods 4 and 5) and model extension (methods 6–8, Table 20.1). The model  $\chi^2$  then refers to the difference in  $-2 \log$  likelihood between a model with re-estimated predictors and the re-calibrated model, and  $df$  correspond to the difference in degrees of freedom of these models. Regression coefficients can be pulled towards their re-calibrated values as obtained with method 3. A theoretical motivation for this shrinkage approach was developed by Van Houwelingen and is presented at the Web.

### 20.4.1 Example: Shrinkage towards Re-calibrated Values in GUSTO-I

We apply shrinkage towards re-calibrated values as obtained with method 3 for the TIMI-II model, when applied in GUSTO-I. Re-estimated coefficients for the first eight predictors are pulled towards  $\beta_{\text{overall}} \times \beta_{i,\text{TIMI}}$  with methods 4 and 5. The coefficients of the additional eight predictors considered in methods 6–8 are shrunken towards zero, since these predictors were not included in the TIMI-II model. The intercept of the shrunken model was re-estimated to ensure that the sum of predicted probabilities equaled the sum of observed outcomes (in our case, deaths). When stepwise regression is applied to select predictors for the model, the degrees of freedom of the candidate predictors should be considered in the formula.<sup>176,459</sup>

As an alternative, we may shrink coefficients towards the original TIMI coefficients. This is also straightforward with penalized maximum likelihood for model re-estimation. Hereto we use the original model predictions as an offset variable in the re-estimated logistic regression model.

For illustration, we consider updating of the TIMI-II model for the West region in GUSTO-I ( $n=2,188$ , Table 20.5). Re-estimated coefficients were somewhat different from the re-calibrated coefficients, with larger effects for shock, age, and hypotension, and smaller effects for diabetes and time to relief. The re-calibrated model had a model  $\chi^2$  of 170, which increased by 24 to 94 for the re-estimated model. The traditional shrinkage factor is  $(\text{model } \chi^2 - df) / \text{model } \chi^2 = (194 - 8)/194 = 0.96$ . This factor is used to shrink coefficients towards zero. The

**Table 20.5** Logistic regression coefficients in updated models for the West region of GUSTO-I ( $n=2,188$ ). Shrinkage and penalization were applied towards zero or towards (re-calibrated) values of coefficients from the TIMI-II model

Predictor	Re-estimated	Re-calibrated	TIMI	Shrunken towards			Penalized towards	
				zero	re-cal.	TIMI	zero	TIMI
Shock	2.40	2.02	1.79	2.30	2.29	2.21	2.37	2.38
Age>65	1.64	1.12	0.99	1.57	1.49	1.44	1.53	1.60
High risk	0.85	1.04	0.92	0.81	0.90	0.87	0.80	0.85
Diabetes	0.07	0.84	0.74	0.07	0.29	0.27	0.07	0.10
Hypotension	1.22	0.78	0.69	1.17	1.09	1.06	1.16	1.19
Tachycardia	0.65	0.67	0.59	0.62	0.65	0.63	0.61	0.64
Time to relief	0.26	0.60	0.53	0.25	0.36	0.34	0.25	0.28
Female Sex	0.62	0.53	0.47	0.60	0.60	0.58	0.61	0.62

re-calibration shrinkage factor is  $(24 - 7)/24 = 0.71$ . This factor is used to shrink coefficients towards the re-calibrated values. We can also examine the improvement of re-estimated coefficients over using the original TIMI coefficients; this appears to be a model  $\chi^2$  of 27. With  $df = 8$ , the shrinkage factor towards TIMI coefficients becomes (model  $\chi^2 - df$ ) / model  $\chi^2 = (27 - 8)/27 = 0.70$ .

The coefficients are surprisingly similar when shrunken to zero or pulled to re-calibrated values. The largest discrepancy is for diabetes, where the re-estimated coefficient was close to zero (0.07), while the re-calibrated value was much higher (0.84, Table 20.5). Shrinkage towards zero leaves the coefficient at 0.07, but pulling towards the re-calibrated value of 0.84 leads to a value of 0.29. Pulling towards TIMI-II coefficients leads to slightly smaller coefficients. Shrinkage towards zero is in the spirit of Bayesian analysis with an uninformative prior (coefficients are assumed to be zero); pulling towards (re-calibrated) coefficients assumes that the TIMI-II model is relevant for the new setting (coefficients are assumed to be close to the TIMI-II values).

For comparison, we examine results from penalized maximum likelihood procedures. In the re-estimated 8 predictor model, the optimal penalty factor is 6. The same value is found when the TIMI coefficients are used as an offset variable in the logistic regression model. The resulting penalized coefficients in the standard formulation of the penalized model are quite similar to the “shrunken to zero” coefficients. When penalized towards TIMI-II values, all coefficients are slightly larger, and closer to the re-estimated coefficient values.

#### \*20.4.2 R code for Shrinkage and Penalization in Updating

We start with re-estimating the 8 predictor model in the West region

```
full18 <- lrm(DAY30~SHOCK+A65+HIGH+DIA+HYP+HRT+TTR+SEX,
  data=West, x=T, y=T)
```

The original TIMI coefficients are in linear predictor 1

```
timi8.par <- c(- 4.465, 1.79, 0.99, 0.92, 0.74, 0.69, 0.59,
               0.53, 0.47)
lp1      <- full18$x %*% timi8.par[-1] + timi8.par[1]
```

Coefficients with traditional heuristic shrinkage are calculated as  $(\chi^2 - df)/\chi^2$

```
s.orig   <- (full18$stats[3]-full18$stats[4]) / full18$stats [3]
full18.coef.s.orig <- s.orig * full18$coef[-1]
```

Shrinkage towards re-calibrated values is calculated as

```
full3     <- lrm.fit(y=full18$y, x=lp1) # re-calibration model
model.chi2 <- deviance(full3)[2] - deviance(full18)[2]
s.recal   <- (model.chi2 - (full18$stats[4] - full3$stats[4])) /
               model.chi2
full18.coef.s.recal <- full3$coef[2]*timi8.par[-1]
                  + s.recal* (full18$coef[-1] - full3$coef[2]*timi8.par[-1])
```

Shrinkage towards TIMI-II values is calculated as

```
full18.off    <- update(full18, offset=lp1) # offset model
s.off        <- (full18.off$stats[3]-full18.off$stats[4]) / full18.
off$stats[3]
full18.coef.s.off <- s.off * full18.off$coef[-1] + timi8.par[-1]
```

Standard penalized maximum likelihood estimation is as

```
p          <- pentrace(full18, 0:20, maxit=50)
full.pen   <- update(full18, penalty=p$penalty)
```

Penalization towards TIMI-II values is calculated as

```
p.off      <- pentrace(full18.off, 0:20, maxit=50)
full.off.pen <- update(full18.off, penalty=p.off$penalty)
full.off.pen.coef <- full.off.pen$coef + timi8.par
```

## 20.5 Sample Size and Updating Strategy

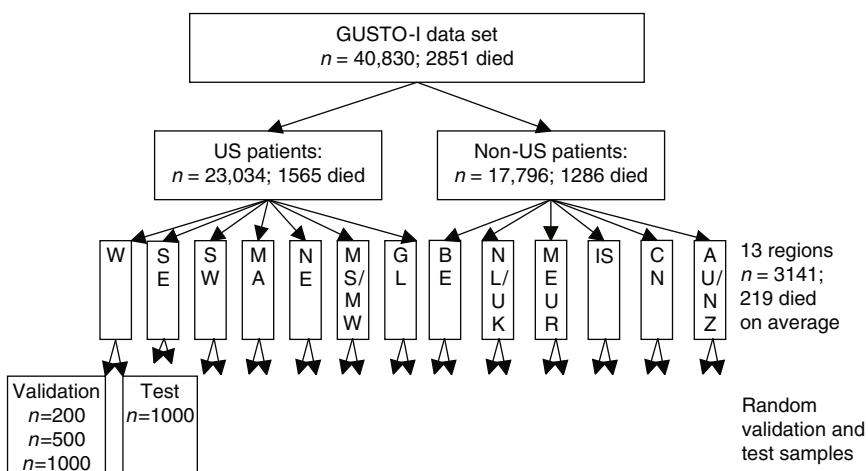
The choice of updating method depends on various factors. The first requirement is that it is reasonable to apply the previously developed model in the new setting from a clinical point of view. The model should not evidently be overfitted, include predictors with plausible effects, and have been developed with adequate statistical methods given the sample size. The relevance of the model should be supported by reasonable validity in the sample from the new setting, i.e. some correlation should be present between predictions and outcomes. If this is not the case, we should not consider updating methods, but may essentially consider the situation as developing a new model.<sup>458</sup> Also, the size of the development sample may have been too small to consider updating seriously. Possibly we can then start our updated model with the selection as considered in the previous model, but directly re-estimate coefficients (method 5, Table 20.1).

In the situation of a large development sample, we may have good confidence in the previously estimated regression coefficients. If we only have a small validation sample size, we should be modest in updating the model and re-calibration may be sufficient (methods 2 and 3, Table 20.1). In contrast, if we have a large validation sample, more rigorous updating is probably reasonable.

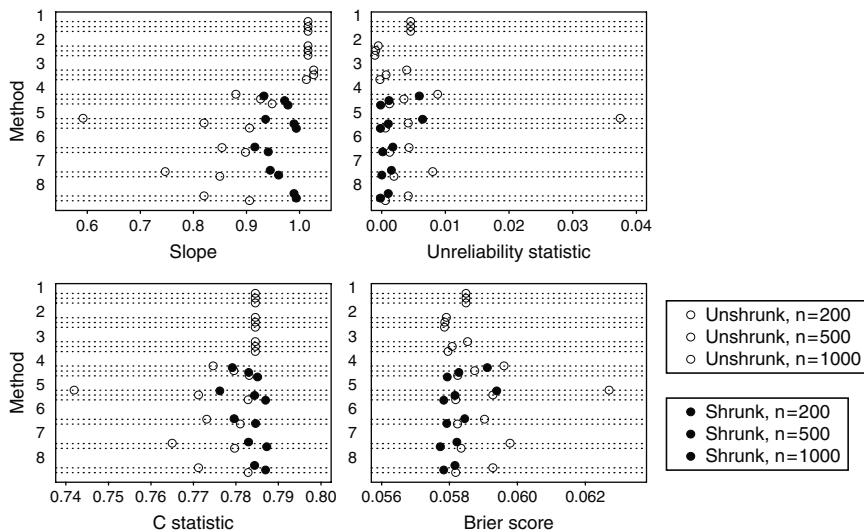
### \*20.5.1 *Simulations of Sample Size, Shrinkage, and Updating Strategy*

Simulation studies were performed in GUSTO-I to increase our insights in the relationship between sample size and updating strategy (Fig. 20.2). Validation sample size ranged from  $n=200$  to 1,000 in Fig. 20.3, and from  $n=1,000$  to 10,000 in Fig. 20.4. We note that a modest improvement in discriminative ability may be achieved by model re-estimation and model revision (methods 4–8), if validation sample sizes are relatively large and shrinkage is used. But with a relatively small validation sample we should only attempt to improve calibration, i.e. with updating of the model intercept (method 2) and calibration slope (method 3). Shrinkage is essential to prevent overfitting in updated models from small validation samples (Figs. 20.3 and 20.5).

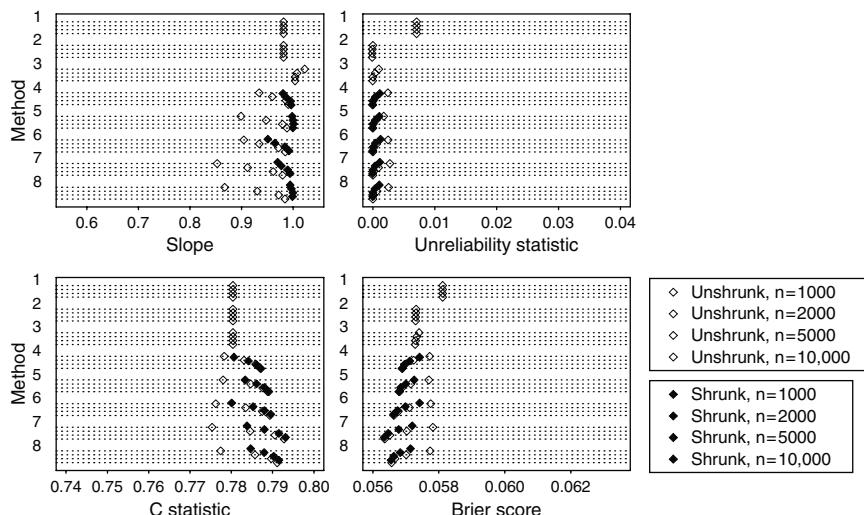
More extensive updating is beneficial if the previous model was based on a relatively small sample ( $n=500$  instead of  $n=3,339$ ), while a relatively large validation sample was available (Fig. 20.5). See Web for more details, and a paper in *Stat Med*.<sup>402</sup>



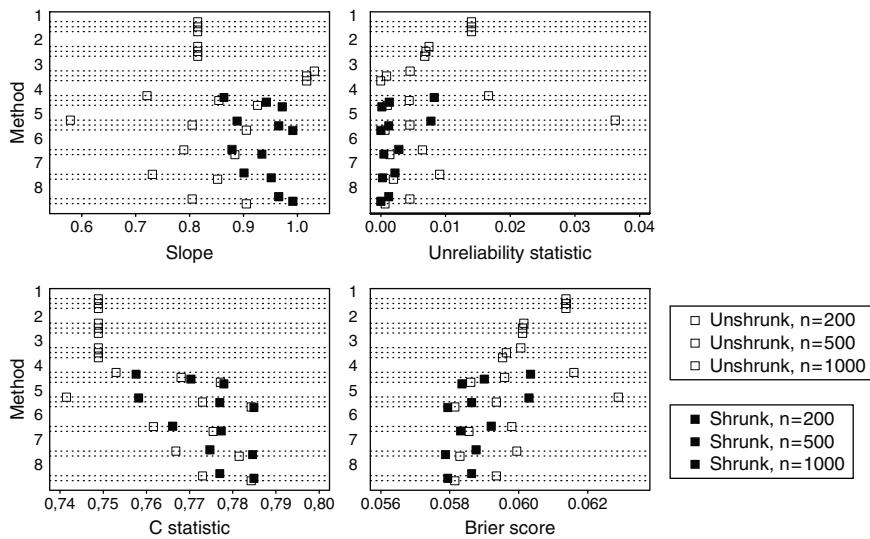
**Fig. 20.2** Schematic presentation of the sampling design of the simulation study in GUSTO-I. The GUSTO-I data was split in 13 regions. The seven US regions were West (W), South-East (SE), South-West (SW), Massachusetts (MA), New England (NE), Mid-South/Mid-West (MS/MW), and Great Lakes (GL). The six non-US regions were Belgium (BE), the Netherlands/United Kingdom (NL/UK), middle Europe—including France, Spain, Germany, Poland—(MEUR), Israel (IS), Canada (CN), and Australia/New Zealand (AU/NZ). Updating methods 1–8 were applied in random samples from each region with sizes of 200, 500, or 1,000 patients. Updated models were tested in independent test samples with 1,000 patients from the same region as where the validation sample originated from.



**Fig. 20.3** Dotcharts showing the average results for eight updating methods (numbers 1–8, Table 20.1) with or without application of shrinkage in the updating of regression coefficients. For methods 1–5, validation sample sizes were 200, 500, or 1,000 (three rows). For methods 6–8, validation sample sizes were 500 or 1,000 (two rows). Validation samples were drawn from 13 regions within the GUSTO-I study (Fig. 20.2). Slope, calibration slope, unreliability statistic,  $\chi^2$  test for calibration intercept and slope. Performance was determined in independent test samples with  $n=1,000$



**Fig. 20.4** Dotcharts showing the results of simulation studies in the US patients from the GUSTO-I study. Average results are shown for eight updating methods (numbers 1–8), with or without application of shrinkage in the updating of regression coefficients. Validation sample sizes were 1,000, 2,000, 5,000, or 10,000 (four rows for each method), with test sample sizes of  $n=10,000$



**Fig. 20.5** Dotcharts showing the results of simulation studies with smaller development samples ( $n=500$  instead of 3,339 for the original TIMI-II model as shown in Fig. 20.1). Average results are shown for eight updating methods (numbers 1–8), with or without application of shrinkage in the updating of regression coefficients. For methods 1–5, validation samples contained 200, 500, or 1,000 patients (three rows). For methods 6–8, sample sizes were 500 or 1,000 (two rows)

## 20.6 Validation and Updating of Tree Models

Prediction models developed with CART methods, or recursive partitioning, are attractively presented as trees (see Chap. 4). Usually, predicted outcomes are presented for each branch. Validation can then be performed in different ways.

A radical validation approach is to try to re-develop a new tree in a validation sample, and compare the structure. For example, Van Dijk re-developed a tree for survival of testicular cancer patients.<sup>452</sup> They found that the statistically optimal tree was very different from the tree as developed in a relatively small sample of German patients ( $n=332$ ).<sup>239</sup> This approach to validation is similar to developing a model with stepwise methods in a validation sample, if stepwise methods were applied in a development sample. As discussed in Chap. 11 it is highly unlikely that this model building results in the same selection of predictors. Such model re-development gives insight in the instability of the modelling procedure, but does not directly answer the question to what extent the outcomes in the validation data are adequately predicted by the old model.

Another validation strategy could be to accept the tree structure, but to re-estimate the predictions of the outcome. For a binary outcome, these estimates are simply the observed frequencies of the outcome in the branches. This is analogous to updating method 5 (model re-estimation while accepting the model structure). Some revision of the tree structure might be inspired by these findings; e.g. when

two branches lead to similar outcomes at validation, the split might be omitted for future predictions.

A more parsimonious strategy is to use a re-calibration model, similar to method 3 (Table 20.1). For a binary outcome we model the outcome  $y$  as a function of a new intercept  $\alpha$  and calibration slope  $\beta_{\text{overall}}$ :

$$y \sim \alpha + \beta_{\text{overall}} * \hat{y},$$

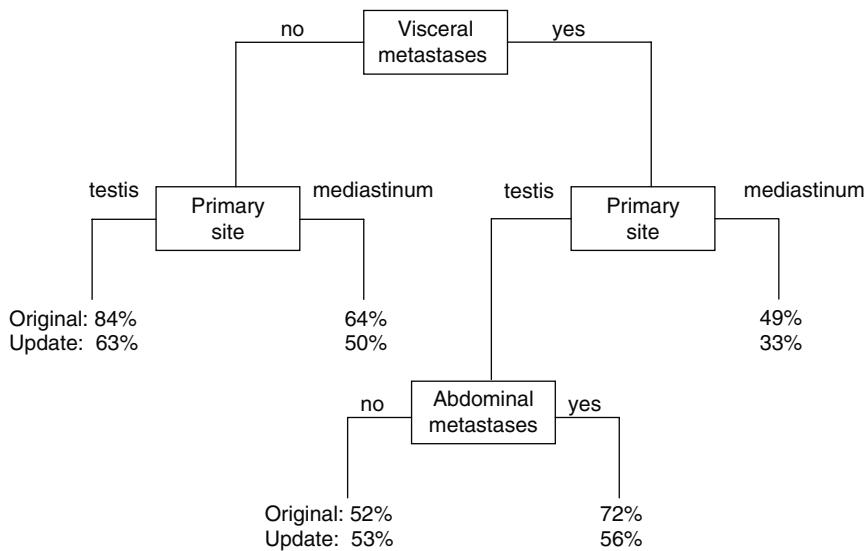
where  $y$  is the outcome,  $\alpha$  the updated intercept,  $\beta_{\text{overall}}$  the calibration slope, and  $\hat{y}$  the predicted outcome. If the outcome is binary, we need to transform the  $\hat{y}$  to e.g.  $\log(\text{odds}(\hat{y}))$ ; for survival outcomes we could use the  $\log(\text{cumulative hazard})$  of the Kaplan-Meier estimates at certain time points during follow-up:  $\log(-\log(S(t|\text{branch})))$ , with  $S(t|\text{branch})$  indicating survival at time  $t$  in a branch of the tree. This approach preserves the relative effects of the tree, but updates the predictions to obtain calibration-in-the-large (updating of intercept), and compensates for any overfitting that may have occurred at model development ( $\beta_{\text{overall}} < 1$ ).

### \*20.6.1 Example: Tree Modelling in Testicular Cancer

Patients with metastatic non-seminomatous germ-cell tumors nowadays have a long-term cure rate over 80%, due to highly effective chemotherapy. Because of the high overall cure rate, interest has shifted to reducing treatment-related toxicity for patients with a relatively good prognosis. On the other hand, patients with a relatively poor prognosis should be considered for more aggressive treatment, such as dose intensification and high-dose chemotherapy with stem-cell support. The International Germ Cell Cancer Collaborative Group (IGCCCG) developed the International Germ Cell Consensus (IGCC) classification to distinguish patients according to prognosis.<sup>5</sup> A poor prognosis group was defined by the presence of any “poor risk factor.” These were: mediastinal primary site, (non-pulmonary) visceral metastases,  $\alpha$ -fetoprotein (AFP) poor ( $>10,000$  ng/ml), human chorionic gonadotrophin (HCG) poor ( $>10,000$  ng/ml), and lactate dehydrogenase (LDH) poor ( $>10$  times the upper limit of normal).

Tree modelling was used to find subsets within 332 poor prognosis patients.<sup>239</sup> The risk factors visceral metastases, primary site, and abdominal mass were used. This resulted in a tree with five poor prognosis subgroups (Fig. 20.6). The subgroups differed in 2-year survival, ranging from 49% to 84%. Some subgroups however had only a small number of patients, and their identification might be the result of pure chance. Such subgroups may not be present when new data are considered. Furthermore, survival estimates of small groups are often unreliable. This was illustrated by the group of patients with visceral metastases and primary site testis, in which patients with an abdominal mass had a higher 2-year survival (72%; 95% confidence interval (CI), 64–80%) than patients without (52%; 95% CI, 27–77%).

We validated the 2-year survival probabilities in the poor prognosis patients in the IGCCCG database ( $n=456$ ).<sup>452</sup> We found that the survival probabilities were



**Fig. 20.6** Regression tree as developed in 332 poor prognosis patients with non-seminomatous germ-cell cancer. The 2-year survival is shown for the development sample (“original”) and for the validation sample (“update”)<sup>452</sup>

generally lower in the IGCCCG patients compared with the development setting (Fig. 20.6). A total of 125 patients was expected to have died by 2 years, while the observed number was 199 (i.e., 1.6 times more deaths). We assessed the calibration slope in the model:

$$\log(-\log(S_{\text{IGCCCG}}(t \mid \text{branch})) = \alpha + \beta_{\text{overall}} * \log(-\log(S_{\text{development}}(t \mid \text{branch}))),$$

where  $S(t|\text{branch})$  refers to the observed Kaplan-Meier survival probabilities for tree branch, and  $\beta_{\text{overall}}$  is the calibration slope. We found  $\alpha = -0.19$  and  $\beta_{\text{overall}} = 0.46$ . The predictive effects in the IGCCCG data were hence much less than at model development ( $\beta_{\text{overall}} << 1$ ), consistent with the hypothesis that the original tree was overfitted. This same pattern was noticed from a comparison of the discriminative ability. The  $c$  statistic was 0.63 at model development, and only 0.56 at validation.

## 20.7 Validation and Updating of Survival Models

Predictions of survival models involve a time dimension, e.g. for the fraction of patients surviving 1, 2, or 5 years after start of follow-up. The most common prognostic model in medicine is the Cox proportional hazards model, which can combine multiple prognostic factors to predict survival at different time points:

$$S(t | X) = S_0(t)^{\exp(\beta X)},$$

where  $S(t|X)$  denotes the probability of being alive at time  $t$  for a patient with predictors  $X$ ;  $S_0(t)$  denotes the baseline survival function for time  $t$  (usually for the average of predictor values), and  $\beta X$  indicates the linear predictor (multiplication of regression coefficients  $\beta$  with predictor values  $X$ ). We can also write the survival function based on the baseline cumulative hazard  $H_0(t)$  as  $S(t|X) = \exp(-H_0(t) * \beta X)$ . The baseline cumulative hazard  $H_0(t) = -\log(S_0(t))$ .

Hence, making predictions with the Cox model for individual patients requires that we know the baseline survival (or baseline cumulative hazard) function as well as the regression coefficients  $\beta$ .

- The full baseline survival function is usually not specified in publications, but sometimes survival at clinically relevant time points is provided (e.g. 1-, 2-, and 5-year survival). Also Kaplan-Meier curves can provide the baseline survival function graphically.
- The regression coefficients  $\beta$  are often provided in a table as hazard ratios ( $\exp(\beta)$ ). This makes it possible to calculate a detailed linear predictor for new patients. But often a simplified version of the model is presented as a “prognostic index,” e.g. based on a sum score, or a count of the number of adverse prognostic factors.

### **20.7.1 Case Study: Validation of a Simple Index for Non-Hodgkin’s Lymphoma**

A Cox regression model for overall survival for aggressive non-Hodgkin’s lymphoma was developed by an international group of investigators.<sup>3</sup> Five pre-treatment clinical characteristics were considered: age, Karnovsky score, Ann Arbor stage, extra nodal sites, and LDH scores. The five predictors are dichotomized for use in the “international prognostic index” (IPI). The IPI score counts the number of unfavorable predictors. The more extreme categories 0 and 1, and 4 and 5 are combined, resulting in groups with IPI 1–4. The 2-year survival probabilities ranged from 34 to 84%, and the 5-year probabilities from 26% to 73% (Table 20.6).

The validity of the IPI was studied by Hermans from a pragmatic perspective,<sup>187</sup> and later by Van Houwelingen from a more methodological perspective.<sup>456</sup> The validation sample was a Dutch cohort from a population-based registry of non-Hodgkin’s lymphoma patients. Hermans constructed Kaplan-Meier curves for each of the four IPI groups, which showed a clear separation. However, the observed survival probabilities were lower than expected (Table 20.6). This discrepancy was attributed to the selection of patients: From clinical trials at model development, and from a population-based registry at model validation. The validation cohort was less selected, e.g. with respect to age.

**Table 20.6** Validity of the original and updated IPI for a Dutch cohort of 426 non Hodgkin's lymphoma patients

IPI	2-year survival (%)			5-year survival (%)		
	Original	K-M	Re-calibrated	Original	K-M	Re-calibrated
1 ( <i>n</i> =148)	84	78	78	73	61	58
2 ( <i>n</i> =110)	66	54	55	51	35	31
3 ( <i>n</i> =85)	54	39	41	43	15	23
4 ( <i>n</i> =83)	34	24	21	26	10	9

Updating was with Kaplan-Meier curves (Sect. 20.7.2) for the four IPI groups and a re-calibration procedure (Sect. 20.7.3, for groups by time points)<sup>456</sup>

The Kaplan-Meier curves answer the qualitative question on whether the discriminative ability of the original model was retained in an external setting. More quantitative questions relate to calibration: Is there a systematic difference between predicted and observed survival for all IPI groups, and what is the predictive strength of the IPI in the validation setting? These questions can well be studied in the re-calibration framework.<sup>456</sup>

## 20.7.2 *Updating the Prognostic Index*

The observed Kaplan-Meier probabilities can be considered as updated estimates of survival for future Dutch non-Hodgkin's lymphoma patients (Table 20.6). However, this update only considers the grouping of the IPI, and discards any further prognostic information from the development sample on survival during follow-up. The Kaplan-Meier curves are non-parametric, and allow for non-proportional hazards of the IPI risk groups. Identical results can be obtained from a Cox regression model in the validation sample with the four IPI groups as strata.

Re-calibration of the IPI probabilities is an alternative approach, which may be especially valuable in relatively small validation samples. The Dutch cohort of 426 patients may be considered sufficiently large for the Kaplan-Meier approach, but the standard error (SE) around the survival estimates in Table 20.6 is around 0.05. This means that 95% CIs are  $\pm 10\%$  around the Kaplan-Meier survival probabilities in Table 20.6. With a smaller size the updated survival probabilities would have been even more uncertain.

## 20.7.3 *Re-calibration for Groups by Time Points*

Simple re-calibration is possible for the two time points (2 and 5 years), comparing the predicted survival with the observed survival for groups of patients in a calibration model on the log hazard scale:

$$\log(-\log(S(t | g))) = \alpha + \beta * \log(-\log(S_{\text{model}}(t | g))),$$

where  $S(t|g)$  refers to the observed Kaplan-Meier survival probabilities for the groups  $g$ , and  $S_{\text{model}}(t|g)$  to the predicted survival probabilities for these groups. Setting  $\beta$  to 1 means that we accept the hazard ratios for the four IPI groups as estimated in the development data set. This is analogous to method 2 for logistic regression models (Tables 20.1 and 20.7).

With  $\beta = 1$ , Van Houwelingen reports that  $\alpha = 0.37$  at 2 years, and  $\alpha = 0.56$  at 5 years.<sup>456</sup> Hence, we make somewhat different corrections on the log hazard scale for the two time points. The re-calibrated survival probabilities are shown in Table 20.6, calculated with the formula

$$S_{\text{cal}}(t | g) = \exp(-\exp(\alpha + \log(-\log(S_{\text{model}}(t | g)))).$$

#### 20.7.4 Re-calibration with a Cox Regression Model

A further validity assessment is to study the calibration slope  $\beta_{\text{overall}}$  in a Cox regression model:

$$S_{\text{cal}}(t | \beta X) = S_{0,\text{new}}(t)^{\exp(\beta_{\text{overall}} * \beta X)},$$

where  $S_{\text{cal}}(t | \beta X)$  refers to the re-calibrated survival,  $S_{0,\text{new}}(t)$  to the re-calibrated baseline survival function, and  $\beta_{\text{overall}}$  to the calibration slope for the linear predictor  $\beta X$ . A Cox regression with the linear predictor  $\beta X$  as the single covariate assumes proportional effects of the IPI during follow-up. The baseline hazard function is updated, and a calibration slope is identified to calibrate the linear predictor to the new setting. This approach is more or less analogous to method 3 for logistic regression models (Tables 20.1 and 20.7).

Such re-calibration requires that we know the linear predictor for the four IPI classes. The original regression coefficients for the four IPI classes were not published, but we can try to calculate the coefficients from the published 2-year and 5-year survival probabilities.<sup>456</sup> Hereto we rewrite the Cox survival formula  $S(t|X) = S_0(t)^{\exp(\beta X)}$  as  $\log(-\log(S(t|X))) = \log(-\log(S_0(t))) + \beta X$ . The Weibull model can be used for the baseline survival function  $S_0(t)$ , which specifies that  $\log(-\log(S_0(t))) = \beta_0 + \beta_1 \log(t)$ . The Weibull is attractive since it specifies the baseline survival with only two parameters, but other parametric models can also be used. The Weibull model reads like  $\beta_0 + \beta_1 \times \log(t_j) + \beta_i X_i$  with  $j = 2, 5$  and  $i = 1, 2, 3, 4$  for the four IPI groups. Since we do not have access to the original IPI data, Van Houwelingen uses a simple linear regression model to fit the parameters. The IPI-Weibull model becomes:

$$\log(-\log(S_0(t))) = -0.319 + 0.439 * \log(t) + \beta X,$$

**Table 20.7** Updating approaches for the IPI in non-Hodgkin's lymphoma

Method	Approach	Proportionality assumption and baseline hazard	$\beta_{\text{IPI}}$
-	Kaplan-Meier	Non-proportional, free	Free
2	Kaplan-Meier re-calibration	Non-proportional, free	Original
3a	Cox, re-calibrate IPI	Proportional, free	Re-calibrated
3b	Weibull, re-calibrate IPI	Proportional, re-calibrated	Re-calibrated

with  $\beta X = -1.638; -0.824; -0.514$ ; and 0 for IPI=1, 2, 3, and 4, respectively. The resulting survival curves are plotted in Fig. 20.7. A reasonable fit is found for the observed 2- and 5-year estimates.

When this PI is used in a Cox regression model, the coefficient becomes 1.03 (SE = 0.10). This indicates a very similar predictive effect of the IPI in the validation sample compared with the development sample.<sup>456</sup>

### \*20.7.5 Parametric Re-calibration

Instead of re-calibration with a Cox regression model, Van Houwelingen also describes how we can use an exponential model or a Weibull model.<sup>456</sup>

The baseline cumulative hazard function of the IPI-Weibull model is defined as:

$$H_{0, \text{model}}(t) = -\ln(S_{0, \text{model}}(t)) = \exp(-0.319 + 0.439 * \log(t)) = 0.727 * t^{0.439}.$$

We use this transformation for the time  $T$  in the validation sample:

$$T^* = 0.727 \times T^{0.439}.$$

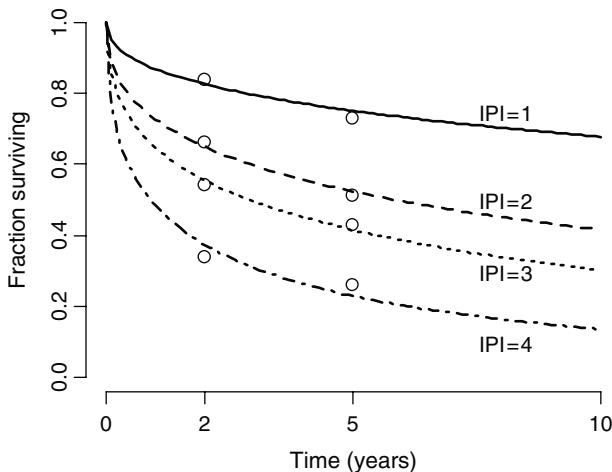
This transformed time  $T^*$  follows an exponential distribution if the IPI - Weibull model is valid.<sup>458</sup>

An assessment of calibration-in-the-large is possible in various ways. We can use an exponential survival model with  $\log(T^*) = \alpha + \text{PI}$ , where  $\alpha$  refers to a constant that controls the level of the log(hazard), adjusted for the IPI effects in the prognostic index (PI, based on the IPI coefficients as defined in Sect. 20.7.4). A simple alternative is to directly compare the number of observed deaths to the number predicted (correcting for censoring). Van Houwelingen reports that  $\alpha = 0.436$  (SE, 0.06).<sup>456</sup> This is equivalent to a hazard ratio of  $\exp(0.436) = 1.55$ . Hence, the overall survival was 1.55 times worse in the validation sample than in the development sample, adjusted for IPI score.

Re-calibration can also be assessed with an exponential model with the PI as the single predictor:

$$\log(T^*) = \alpha + \beta * \text{PI}.$$

Van Houwelingen however prefers re-calibration assessment with a Weibull model, which allows for a different shape of the baseline hazard in the validation sample:



**Fig. 20.7** Survival according to the International Prognostic Index (IPI) for non-Hodgkin's lymphoma patients. The reported 2- and 5-year survival probabilities are shown (O), with the Weibull approximation of survival with lines from 0 to 10 years of follow-up (“IPI-Weibull model”)<sup>456</sup>

$$\log(T^*) = \alpha + \beta * \text{PI} + \gamma * e.$$

Here  $\alpha$  refers to a constant that controls the level of the log(hazard),  $\beta$  refers to the effect of the prognostic index PI, based on the IPI coefficients, and  $\gamma$  controls the shape of the hazard function, and  $e$  indicates the exponential distribution. If  $\gamma = 1$ , the shape of the baseline hazard function is maintained in the validation data. If  $\beta = 0$ ,  $\beta = -1$ , and  $\gamma = 1$ , the calibration is perfect. The ratio of  $-\beta$  and  $\gamma$  is the usual calibration slope.

Re-calibration of the IPI-Weibull model results in estimates of  $\alpha = -0.24$  (SE, 0.06),  $\beta = -0.68$  (SE, 0.07), and  $\gamma = 0.65$  (SE, 0.03).<sup>456</sup> Since  $\gamma$  is clearly different from 1, the shape of the hazard function is different in the validation data than estimated with the IPI-Weibull model. We hence cannot simply adjust the baseline hazard from the IPI-Weibull model by a constant factor, such as an hazard ratio of 1.55. This is consistent with the finding of different values for  $\alpha$  when re-calibration is done by time points (Sect. 20.7.3).

In the proportional hazards interpretation of the Weibull model the calibration slope for the linear predictor is  $-\beta / \gamma = 0.68/0.65 = 1.04$ . This is in line with the Cox re-calibration model (see Sect. 20.7.4), which also indicated that the predictive effect of the IPI was remarkably similar in the validation setting.

The Weibull re-calibration procedure updates the IPI survival predictions by estimating three parameters. The parameters  $\alpha$  and  $\gamma$  are used to update the baseline survival;  $\beta$  is used to re-calibrate the IPI effect. It is a parsimonious (and hence attractive) alternative to re-calibration with a Cox model.

More extensive model updating is also possible. For example we can reweight the five components of the IPI (method 4 in Table 20.1). This is actually similar to for example reweighting components of a comorbidity summary score as discussed in Chap. 9. Van Houwelingen reports that the predictor age >60 had a statistically significant stronger effect than that assumed in the IPI.<sup>456</sup> We can also update a model with inclusion of non-proportionality of the hazards between prognostic groups. For further details see papers by Van Houwelingen.<sup>456,458</sup>

### 20.7.6 Summary Points

- The IPI could be updated in at least four ways with different freedom for the baseline hazard, with or without a proportionality assumption on the effect of the IPI, and with different assumptions on the validity of the previously estimated regression coefficients (Table 20.7).
- Kaplan-Meier estimates can be generated for the four IPI groups separately, which implies re-estimation of the baseline hazard, and new, separate effects for the relative effects implied by IPI.
- Kaplan-Meier estimates can also be used in a re-calibration procedure per time point, preserving the original IPI effects.
- A Cox regression model can be used to re-estimate the baseline hazard, while recalibrating the IPI effects
- A Weibull model can be used for a more parametric re-calibration of the baseline hazard and the relative effects implied by IPI.

## 20.8 Continuous Updating

So far, we assumed that a validation sample with a fixed size was available for model updating. The updating strategy then depends, among other considerations, on the size of this validation sample, and on the size of the development sample. We can also imagine a more dynamic situation, where a previously developed model is applied in a new setting, with accumulation of patient numbers over time. The prediction model should gradually adapt to the new setting. It is reasonable to start with parsimonious updating methods, such as re-calibration, and gradually move to model revision and model extension following the framework set out in Table 20.1.

As a continuous updating strategy for logistic regression models, we start with accepting the original model to generate predictions. After a relatively limited number of patients has been enrolled, we can consider updating of the model intercept to correct calibration-in-the-large (method 2, Table 20.1). This updating attempts to correct for any systematic differences between the development and validation setting. This correction of calibration-in-the-large should have top priority

since miscalibration can cause systematically wrong decision making with the model (Chap. 19). Next, we can consider model re-calibration, i.e. update the intercept and slope of the linear predictor. Updating of the slope is important to correct for overfitting that may have occurred in the development sample.

When more patients are enrolled we may turn to re-estimation of regression coefficients (method 5), with shrinkage of updated coefficients towards the re-calibrated values. An intermediate approach would be to test each predictor for a deviation of its re-calibrated effect (method 4). Finally, when a substantial number of patients is enrolled, we may consider model extension with more predictors (method 8), or intermediate methods that involve testing of the effects of additional predictors (methods 6 and 7, Table 20.1).

### 20.8.1 A Continuous Updating Strategy

The question is when to move on to more extensive updating in the dynamic situation with gradually increasing numbers of patients. We should not use more extensive updating methods too early, since updated predictions may then be unbiased but quite imprecise, and lead to poorer model performance instead of better performance for the new setting. Statistically, we can try to set a minimum number of patients before thinking about updating the intercept. In Table 19.6, we reported (SE) for the intercept for different sample sizes in the situation that the prediction model was fully valid for the validation setting. With 50 events among 100 or 500 subjects, the SEs were 0.24 and 0.17; with 100 events 0.16 and 0.13 respectively. So, if we would always update the intercept, considerable variability would be introduced. On the other hand we should not update too late, i.e. in the situation that the model is only partly valid and updating is in fact indicated. A compromise is to consider statistical testing of the difference in intercept. Testing is technically already possible after a few events that have occurred. An important issue is the  $p$  value to consider for updating; we may use  $p < 0.05$  as a default selection rule, but we should feel free to use higher  $p$  values.

A similar discussion holds for the calibration slope. In Table 19.6 we found that the SE was between 0.10 and 0.15 for 50 and 100 events respectively. Again, we may test for a deviation from the ideal value of 1, and requiring  $p < 0.05$  before updating the slope. If the model is developed in a small sample, a slope below 1 is likely in the validation data, arguing for a higher  $p$  value and/or a one-sided test for the alternative hypothesis “slope  $< 1$ .”

Re-estimation of coefficients and model extension with new predictors should not be considered too early, since our simulations indicated that the predictive performance of an updated model can be worse than the original model (e.g. a lower discriminative ability).<sup>402</sup> We propose to perform an overall test for the improvement in performance of a model with re-estimated coefficients compared with a re-calibrated model. Hereto we compare the  $-2 \log$  likelihood ( $-2\text{LL}$ ) of the re-calibrated and re-estimated models for a likelihood ratio test. The difference in

$-2\text{LL}$  follows a  $\chi^2$  distribution. Similarly, predictions from an extended model are assessed in comparison with a re-estimated model. We may require  $p < 0.05$  in overall tests before updating of coefficients is considered. The updating should include shrinkage. We note that the shrinkage factor is zero unless the  $\chi^2$  is larger than the  $df$  used in model estimation ( $s = (\text{model } \chi^2 - df) / \text{model } \chi^2$ ). This also sets an effective limit to the  $p$  values for testing; e.g. with 8  $df$ , the  $\chi^2$  has to be larger than 8, which is equivalent to  $p < 0.43$ .

### \*20.8.2 Example: Continuous Updating in GUSTO-I

For illustration we consider continuous updating of the TIMI-II model in the West region of GUSTO-I. Tests for model improvement are considered for increasingly complex models (Table 20.8). Updating of the intercept uses 1  $df$ ; the re-calibration model has 2  $df$ ; the re-estimated model estimates eight regression coefficients and a new intercept (9  $df$ ), while the extended model estimates 16 regression coefficients and a new intercept (17  $df$ ). The differences in  $-2\text{LL}$  are tested with the difference in  $df$  between models.

In the full sample of  $n = 2,188$  patients, the  $-2\text{LL}$  of the original TIMI model was 862. With a new intercept ( $-0.36$ , see Table 20.3), the  $-2\text{LL}$  improves from 862 to 846 ( $\chi^2, 16, 1 df, p < 0.001$ ). A small further improvement is obtained by using a calibration slope of 1.13 (Table 20.3). The  $-2\text{LL}$  improves from 846 to 844 ( $\chi^2, 2, 1 df, p < 0.15$ ). Re-estimation of the eight predictors as defined in the TIMI-II model leads to a  $-2\text{LL}$  of 819, at the price of estimating seven parameters more ( $\chi^2, 24.6, 7 df, p < 0.001$ ). Model extension with eight more predictors leads to a  $-2\text{LL}$  of 803 for the 16 predictor model. This extension is a statistically significant improvement over the re-estimated 8 predictor model ( $\chi^2, 15.8, 8 df, p = 0.045$ ). In sum, an important updating aspect in this example is to re-calibrate the model intercept; re-calibration of the linear predictor is not necessary; some further improvement can be obtained with model revision and model extension.

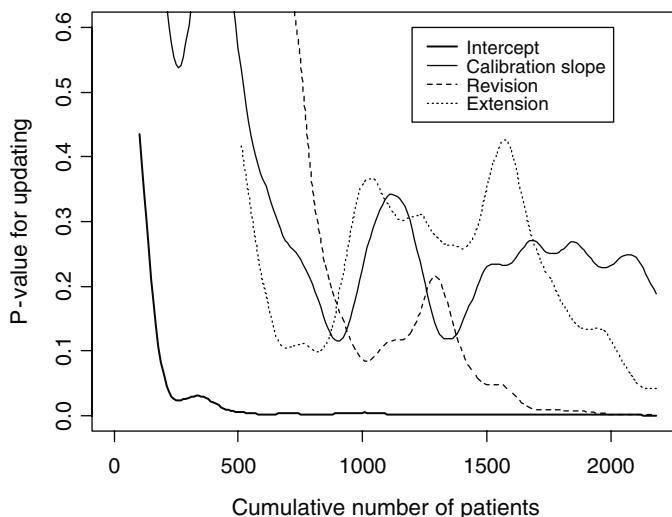
We now turn to the dynamic situation of increasing sample size. Sample size refers in this context to the number of patients with predictors and the outcome known. In the GUSTO-I example, the outcome is 30-day mortality, which is

**Table 20.8** Tests for model improvement in a dynamic updating strategy for the TIMI-II model in GUSTO-I

Method	Label	Parameter	$df_{\text{model}}$	$df_{\text{model improvement}}$
2	Update intercept	Intercept	1	1 (vs. TIMI-II model)
3	Re-calibration	Intercept and slope	2	1 (vs. updated intercept)
5	Model revision	Re-estimate coefficients	9	7 (vs. re-calibrated model)
8	Model extension	Re-estimation + extension	17	8 (vs. re-estimated model)

hence known to the analyst without much delay. If a more long-term outcome is specified, e.g. 1-year survival, the delay is obviously longer before updating analyses can start. We arbitrarily start testing for a difference in intercept from a sample size after including 100 patients in the validation sample, which implies approximately 7 events with an incidence of 30-day mortality of 7% in GUSTO-I. Inclusion is supposed to increase with calendar time. We note that the  $p$  value for a different intercept is still high at  $n=100$  ( $p=0.64$  in this example, Fig. 20.8). The  $p$  value decreases rapidly, to  $p<0.05$  at  $n=170$  in this particular example. The calibration slope is not statistically different from 1 in the full sample of  $n=2,188$ ; in the dynamic situation, the  $p$  value was over 0.50 for  $n<500$ . From  $n=500$ , we also start testing for model revision and model extension. The “model extension” method approaches statistical significance around  $n=650$ , while model revision does so only at  $n>1,500$  (Fig. 20.8). Around  $n=650$ , we might only re-calibrate the effects of the first eight predictors, and extend the model with shrunken effects for eight more predictors. Model extension is not statistically significant between  $n=700$  and 2,000; with shrinkage of regression coefficients the difference would anyway be small between predictions with or without the eight additional predictors. From  $n>1,500$  we would re-estimate the effects of the first eight predictors.

This example illustrates how continuous updating can be applied. Somewhat higher  $p$  values might also be used for testing of updating parameters. We may then rely on the shrinkage methods to prevent “over-updating,” just as shrinkage prevents overfitting in standard model development.



**Fig. 20.8** Continuous updating with accumulating numbers of patients. The  $p$  value for validity of the intercept is significant from  $n>170$ ; the  $p$  value for the calibration slope does not reach significance even at  $n=2,188$ ; model revision and model extension are statistically significant from  $n>1,500$  and 2,000

## 20.9 Concluding Remarks

Re-calibration methods are attractive because of their stability, which is related to the fact that few parameters are estimated.<sup>86,456</sup> The disadvantage of simple re-calibration methods is a potential for bias in the individual regression coefficients. In contrast, model revision may lead to a lower bias but higher variance in the updated model, since more parameters are estimated.<sup>402</sup>

From a clinical perspective, the question needs to be answered whether a previously developed model is reasonable to apply in a new setting. This requires subject knowledge rather than statistical methods. Three examples are in Table 20.9. From a statistical perspective, the sample sizes of both the validation data set and the development data set are crucial in the choice of an updating method. Our simulations in Figs. 20.3–20.5 supported the idea that substantial sample sizes are required before an improvement in discriminative ability is achieved by updating of regression coefficients.<sup>402</sup>

Shrinkage methods in model updating may not only improve calibration, but also discrimination. This is in contrast to traditional model development, where shrinkage does merely improve calibration and has no substantial impact on discrimination.

A specific situation of model updating is that we consider a new predictor, which was not part of a previously developed model. For example, a new biomarker may be promising, with prognostic value shown in a meta-analysis. If we know the correlation of this biomarker with traditional predictors, we may try to update the regression coefficients in a multivariable model with both the traditional predictors and the biomarker. An illustration is available for coronary heart disease.<sup>200</sup>

**Table 20.9** Examples of re-calibration of previously developed prediction models

Patients	Outcome	Development	Validation	Updating
Children with growth hormone deficiency <sup>93</sup>	Growth	Kabi Pharmacia International Growth Study database (n=593)	Dutch Growth Foundation (n=136)	$\hat{Y}_c = \hat{Y}_o + (2.15 - 0.19 \times \hat{Y}_o)$ , where $\hat{Y}_c$ and $\hat{Y}_o$ are the calibrated and original predictions
Men undergoing prostatectomy for prostate cancer <sup>424</sup>	Indolent cancer	Clinical series (n=409)	European Randomized Study on Screening for Prostate Cancer (n=247)	Re-calibration of intercept and rounding of coefficients for score chart
Patients undergoing surgery <sup>215</sup>	Severe post-operative pain	Clinical series Amsterdam	Clinical series Utrecht (n=752 + 283)	Five updating methods performed similarly

## Questions

### 20.1 Simple updating of model intercept

Suppose a model predicts an average operative mortality for elective aortic aneurysm surgery of 8%, but we observe 10 deaths out of 200 (5%) in another series from another hospital.

- (a) What would be the most naïve update of the model intercept?
- (b) What problems should be considered in such a naïve update?

### 20.2 Model updating framework (Table 20.1)

Which updating methods can be seen as nested models, i.e. that a next updating method is an extension of a previous, simpler, method?

### 20.3 Updating strategies (Table 20.1)

What updating strategy makes sense when major improvements in care have taken place

- (a) for all patients
- (b) for a subgroup of patients

### 20.4 Shrinkage and re-calibration (Table 20.5)

We note that the shrunken coefficients for female sex are very similar, whatever method is applied (0.60, 0.60, and 0.58 for shrinkage towards zero, re-calibrated coefficients, or TIMI coefficients respectively). How is this possible?

### 20.5 Performance of updated models (Table 20.4 and Figs. 20.3–20.5)

We note that the  $c$  statistic for method 8 (Re-estimation + extension, 16 predictors) seems to perform best in all parts of GUSTO-I. Performance seems especially good in the smallest sample (sample5,  $n=429$ ,  $c=0.852$ ).

- (a) How do you explain this high apparent  $c$  statistic?
- (b) How is it possible that re-estimation can lead to a poorer performing model at validation in independent patients (Fig. 20.3)?
- (c) Does consideration of eight more predictors in methods 6–8 lead to better models compared with method 5 in Figs. 20.3–20.5?

### 20.6 Continuous updating (Fig. 20.8)

In Fig. 20.8, we note that the  $p$  value for updating of the intercept decreases quickly to small, statistically significant, values. How do you explain this pattern?

### 20.7 Validation and updating of a Framingham model

Consider the paper by D'Agostino on validity of the Framingham risk function to other populations.<sup>90</sup> What is the essential strategy for validation and updating of predictions?

# Chapter 21

## Updating for Multiple Settings

**Background** Updating of a prediction model can be considered for a single new setting, but also for a range of settings, such as multiple hospitals. We can consider such settings as parts of an underlying superpopulation, making them to some extent related. We first quantify the distribution of differences between settings, and subsequently update the model to setting-specific values considering this distribution. This approach is well possible with random effects models or Empirical Bayes estimation. We illustrate the approach for logistic regression models.

We may specifically be interested in differences between centres in the context of quality assessment. We illustrate modern methods for estimation of differences and rank ordering between centres for patients with stroke (“provider profiling”).

### 21.1 Differences Between Settings

#### 21.1.1 Testing for Calibration-in-the Large

We first concentrate on systematic differences between settings in outcome, reflected in calibration-in-the-large. We consider the situation of differences between hospitals in logistic regression models, and subsequently turn to survival models. For logistic regression models, we can simply include “hospital” as a categorical variable in our model, and test for statistical significance of the differences between hospitals. Such an analysis can be performed without adjustment for predictors (“unadjusted” or “crude” comparison), or with adjustment for important predictors of outcome (“adjusted” comparison). The differences that remain after adjustment are obviously of most interest, both from the viewpoint of the applicability of a prediction model across centres, and from the viewpoint of provider profiling (the comparison of the quality across centres).<sup>47</sup>

A theoretical objection to this “fixed-effect” approach is that “hospital” is actually measured at a higher level than at the patient level. This argues for using a multi-level (or “mixed”) model, where hospital is at the first level, and patients are considered within hospitals.<sup>95</sup> The hospital is defined as a random factor, and patient

characteristics are considered as fixed factors (within hospitals). We then estimate the distribution of the random effects, and can test for significance of this distribution, i.e. that the distribution is wider than expected based on chance alone.

### **21.1.2 Illustration of Heterogeneity in GUSTO-I**

Several prediction models can be considered for application in patients suffering from an acute myocardial infarction (MI). We focus on the TIMI-II model, as defined before (Chap. 20). This model includes eight dichotomous predictors.<sup>302</sup> We apply the TIMI-II model in patients from the GUSTO-I trial, with special attention to the validity in geographic groups.<sup>405</sup> Patients were entered in GUSTO-I between 1990 and 1993 at 1 of 1,082 participating hospitals in 14 countries. We distinguished 16 geographical regions within the GUSTO-I trial: 8 in the United States, 6 in Europe (based on combinations of neighboring countries), and 2 other regions (Canada and Australia/New Zealand). These regions included on average 2,552 patients and 178 deaths. Furthermore, we performed more detailed analyses based on geographically related groups of hospitals. The number of patients per hospital was too low for meaningful analyses at the hospital level (average,  $n = 38$ , 2.4 deaths). Grouping resulted in 121 small and 48 large groups, consisting of on average 9 and 23 hospitals and at least 20 and 50 deaths, respectively. The distinction in 16 regions, 48 large groups, and 121 small groups was considered to study regional heterogeneity.

We first test for regional differences in logistic regression models that included dummy variables for each region or group of hospitals. All such tests were highly statistically significant, indicating that the regional differences in 30-day mortality could not reasonably be explained by chance (Table 21.1). We used the TIMI-II model in two ways: as an offset variable, and with refitting of the regression coefficients. With an offset, the regression coefficients were kept at the values as estimated in TIMI-II, and the intercept and centre effects were the free parameters. We found slightly higher  $\chi^2$  statistics if the original TIMI-II coefficients were used (as shown in Table 21.1) rather than refitted coefficients.

Second, we test for regional differences in a random effects logistic regression model, where region or groups of hospitals are considered as a random factor, and

**Table 21.1** Testing for heterogeneity in mortality across groups in GUSTO-I, with adjustment according to the TIMI-II model

	Groups	Groups as fixed effect	Groups as random effect
Regions	16	$\chi^2 = 69$ , 15 df, $p < 0.0001$	$\tau^2 = 0.025$ , $\chi^2 = 28$ , 1 df, $p < 0.0001$
Large subsamples	48	$\chi^2 = 102$ , 47 df, $p < 0.0001$	$\tau^2 = 0.023$ , $\chi^2 = 18$ , 1 df, $p < 0.0001$
Small subsamples	121	$\chi^2 = 197$ , 120 df, $p < 0.0001$	$\tau^2 = 0.033$ , $\chi^2 = 17$ , 1 df, $p < 0.0001$

the TIMI-II coefficients are considered in an offset variable. We compare models with the random effect to models without and confirm that the random effect is statistically significant. The likelihood ratio test gives a one-sided  $p$  value; the two-sided  $p$  value is obtained by dividing by 2.<sup>390</sup>

An advantage of the random effects model is that we can interpret the values of the heterogeneity between groups (variance,  $\tau^2$ ). The standard deviation ( $\tau$ ) reflects differences between groups on the original scale, corrected for random noise. We find that  $\tau^2$  is around 0.025 ( $\tau$  around 0.16). The heterogeneity was similar between small or large subsamples and between the 16 regions.

### **21.1.3 Updating for Better Calibration-in-the Large**

If differences in outcome between centres are relevant (e.g. statistically significant and with substantial magnitude), we may want to update the prediction model with centre-specific estimates of the intercept.<sup>458</sup> In the traditional, fixed effects, approach we could simply use the intercepts per centre after adjusting for patient characteristics as centre-specific estimates (Table 21.2). These estimates may often be quite unstable, and show a relatively wide distribution. This will especially occur when many small centres are considered, i.e. with relatively few patients and/or events.

Preferably, we consider the hospitals as parts of an underlying superpopulation, making them to some extent related. This leads to Empirical Bayes (EB) estimation (Table 21.2).<sup>458</sup> The formula for EB adjusted centre effects is<sup>300</sup>:

$$\alpha_{\text{EB}} = \mu + \tau^2 / (\tau^2 + \sigma_i^2) * (\alpha_i - \mu),$$

where  $\mu$  is the overall mean estimate;  $\tau^2$  is the variance between settings (“heterogeneity”); and  $\alpha_i$  and  $\sigma_i^2$  are the estimated intercepts and their variances. The traditional fixed effect estimates  $\alpha_i$  are shrunken towards the overall mean  $\mu$ . The extent of shrinkage depends on  $\tau^2$  and  $\sigma_i^2$ . A relatively large sampling uncertainty (large  $\sigma_i^2$ ) implies substantial shrinkage for  $\alpha_i$  towards the overall mean  $\mu$ . In contrast, large heterogeneity (large  $\tau^2$ ) implies that  $\alpha_i$  is not much shrunken towards the overall mean  $\mu$ . As mentioned in Chap. 20, an infinite value for  $\tau^2$  implies that the fixed effect estimates  $\alpha_i$  are used as estimates for  $\alpha_{\text{EB}}$ . Every setting is then considered as unique and may have any intercept.

**Table 21.2** Approaches for testing and estimation of differences between settings to correctly estimate average outcomes

Approach	Testing	Estimation
Fixed effects	Setting as categorical variable	Adjusted intercepts
Random effects	Heterogeneity across settings	Empirical Bayes (direct or two steps)

### ***21.1.4 Empirical Bayes Estimates***

There are nowadays two approaches possible to EB estimation: a direct and a two-step approach. The direct approach is to use a random effects model, where the distribution of random effects and the updated intercepts are estimated in one step. This direct approach sounds attractive, but has not yet been completely worked out for non-linear models such as the Cox survival model. Especially, the model may have difficulties in the joint estimation of random effects for multiple differences between centres. For example, we may try to estimate heterogeneity in both intercept and calibration slope, but find that the model estimation does not converge.

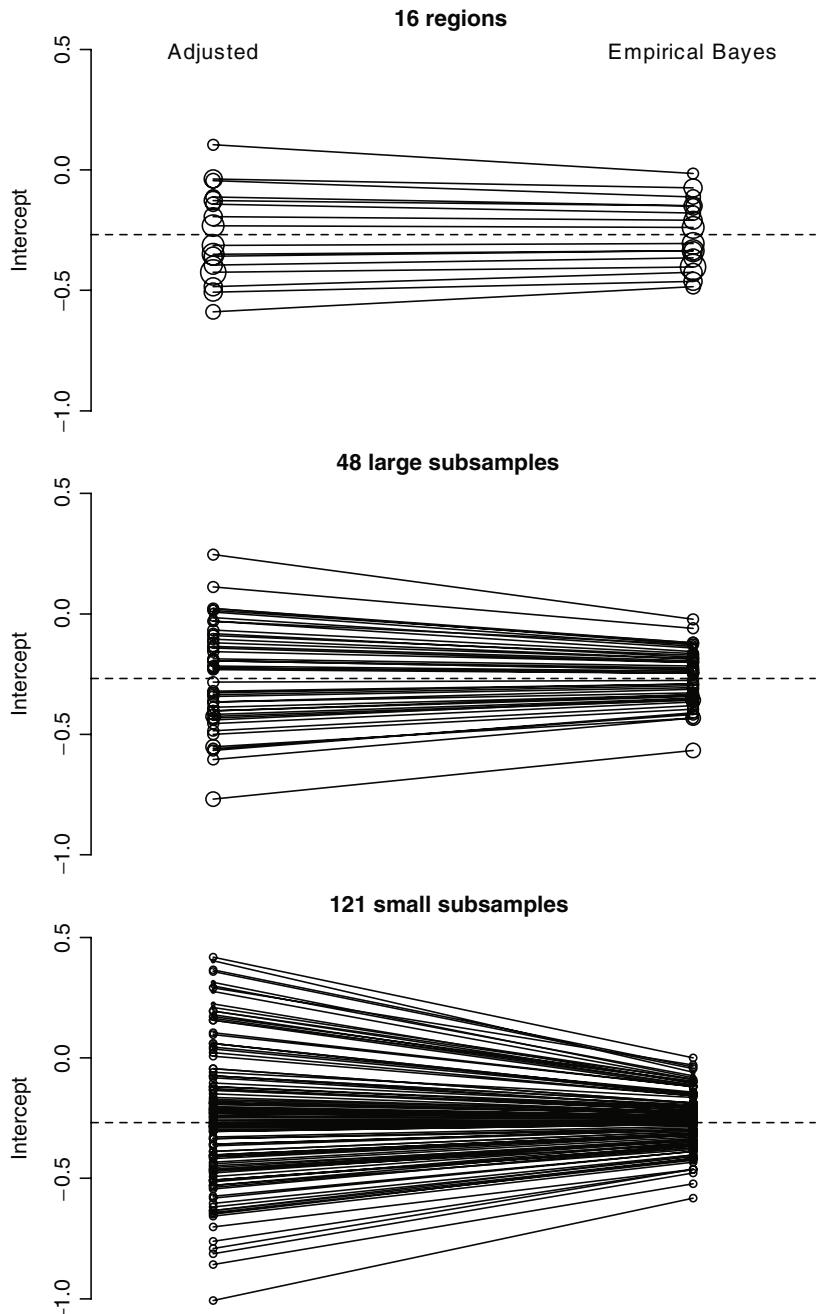
The two-step approach starts with a traditional fixed effect analysis of between centre differences. We may choose one large centre as the reference category for comparison of intercepts, but preferably we compare differences to the average outcome. Technically, this can be achieved by studying each centre while including an offset variable based on predictions for all centres. For each centre, we obtain an estimate of the difference to the average outcome, and a standard error (SE). For the second step, we use the centre-specific estimates as outcomes in a linear random effects model for continuous outcomes, with weights according to the variance of the fixed effect estimates. With this second step we estimate the heterogeneity between centres for use in the EB formula. The uncertainty in determining the heterogeneity is ignored in this two-step procedure, while it is included in the direct approach. Several examples of the two-step approach are available.<sup>405,458,390</sup>

### ***21.1.5 Illustration of Updating in GUSTO-I***

We wonder how large the differences between centres are relative to each other, and whether a simple overall update of the intercept from the TIMI-II model would be sufficient in GUSTO-I.<sup>405</sup> We use the TIMI-II model as an offset variable for updating in the GUSTO-I data.

With a fixed effects approach, we estimate the difference in intercept for each group within GUSTO-I compared with the predicted logodds from TIMI-II. We also obtain SEs for these differences. The R code is shown in Sect. 21.1.8. With a random effects model (`lmer` function in R), we can directly obtain EB estimates of the intercepts per group.

We find that the overall intercept should be updated with the value  $-0.27$ . Regional differences as estimated with traditional fixed effect methods were substantially reduced in the EB estimation, whether 16 regions, 48 large subsamples or 121 small subsamples were considered (Fig. 21.1). The EB estimates for the 2 extreme regions are  $-0.49$  and  $-0.02$ , while the fixed effect estimates are  $-0.59$  and  $+0.11$ . Hence, we would traditionally estimate that one region had a much



**Fig. 21.1** Updating of intercepts of the TIMI-II model for subsamples in GUSTO-I. The overall intercept adjustment is  $-0.27$  (dotted line). We note a substantial variability in fixed effect adjusted intercept estimates for the smaller groups (121 small subsamples), which are shrunk towards the average with Empirical Bayes estimation in a random effects model. Also among the 16 regions we note that smaller regions have more shrinkage of the intercept

lower mortality than observed in TIMI-II ( $-0.59$ ), and one region a slightly higher mortality ( $+0.11$ ). With EB estimation, these estimates are shrunk towards the average of  $-0.27$ . Not surprisingly, these 2 extremes had the smallest and second smallest sample size among the regions. The same patterns are observed for groups of hospitals, with shrinkage to values between  $-0.56$  and  $-0.02$  for large and between  $-0.58$  and  $-0.002$  for small subsamples. We can conclude that a substantial part of the variability in adjusted intercepts of the smaller groups can be attributed to chance.

### 21.1.6 Testing and Updating of Predictor Effects

Next to the intercept, an obvious question is whether the effects of predictors differ by setting. A simple approach is to test for interactions between predictor effects by setting. This is the traditional fixed effects approach. We can also consider the effect of one or more predictors as having distributions across settings in a random effects model.

It is more parsimonious to study interactions by setting for the linear predictor of the prediction model, since the linear predictor summarizes the effects of predictors. We can also study differences in overall effects in a random effects model.

### 21.1.7 Heterogeneity of Predictor Effects in GUSTO-I

We study the calibration slope for the (logodds of the) TIMI-II predictions of 30-day mortality for regional groups in GUSTO-I. We hereto use the linear predictor based on the TIMI-II model as the only predictor for updating in the GUSTO-I data.

We find that the calibration slope should be updated to the value 1.00; overall, there is no need for updating of the slope. In a fixed effect analysis we test interactions with groups and find that there is overall no such interaction within GUSTO-I (Table 21.3). This finding is confirmed in random effect models, where a very small distribution is estimated around the overall recalibration slope. If we consider the EB estimates of the slopes, these appear very close to the overall slope of 1.00.

In addition, we tested for fixed effect interactions of effects of individual predictors by group, e.g. age \* group, and shock \* group.<sup>405</sup> None of these overall tests

**Table 21.3** Testing for heterogeneity in calibration slope of the TIMI-II model across groups in GUSTO-I

	Groups	Fixed effect	Random effect
Regions	16	$\chi^2 = 18, 15 \text{ df}, p = 0.24$	$\tau^2 = 0.000, \chi^2 = 0, 1 \text{ df}, p = 0.49$
Large subsamples	48	$\chi^2 = 53, 47 \text{ df}, p = 0.25$	$\tau^2 = 0.000, \chi^2 = 0, 1 \text{ df}, p = 0.50$
Small subsamples	121	$\chi^2 = 117, 120 \text{ df}, p = 0.56$	$\tau^2 = 0.000, \chi^2 = 0, 1 \text{ df}, p = 0.50$

for interaction are statistically significant, suggesting that it is reasonable to assume a single effect of each predictor across the geographical areas in GUSTO-I.

We conclude that the variability in effects of predictors is extremely small in GUSTO-I. Hence no updating by group is necessary beyond the simple update of the model intercept (with  $-0.27$ ). This small variability may potentially be explained by the fact that predictors were registered according to uniform definitions, were relatively objective characteristics with limited measurement error (e.g. age), and that the quality of data collection was controlled well in this trial. Comparisons across less-controlled settings may show less consistency with respect to the effects of predictors.

### \*21.1.8 R Code for Random Effect Analyses

The essential R code for some of the random analyses in GUSTO-I is shown below, with a full script at the Web.

```
library(lme4)    # linear and generalized linear random effect models
timi8.par <- c(-4.465, 1.79, 0.99, 0.92, 0.74, 0.69, 0.59, 0.53, 0.47)
full18 <- lrm(DAY30~SHO+A65+HIG+DIA+HYP+HRT+TTR+SEX, data=gusto, x=T)
lp1 <- full18$x %*% timi8.par[-1] + timi8.par[1] # lp1 based on TIMI-II
```

#### *Test differences between regions (Table 21.1)*

Fixed and random effects with lp1 (including TIMI-II coefficients) as offset for 16 regions:

```
full18.REGL.o <- lrm(DAY30~as.factor(REGL), offset=lp1, data=gusto)
fullr.REGL.o <- lmer(gusto$DAY30~1+(1 | gusto$REGL), offset=lp1,
                      family=binomial, method="Laplace")
```

Likelihood ratio test for REGL effect, compare to deviance with offset only:

```
pchisq(q=deviance(full18.REGL.o) [2] - deviance (fullr.REGL.o),
       df=1, lower.tail=F) / 2 # divide 2-sided p-value
# Result: x2=28, df=1, p=6.67E-8
```

#### *Estimate calibration slopes between centres (Table 21.3)*

Fixed effects with lp1 (including TIMI-II coefficients) as predictor, interacting with region:

```
full18.REGL.lp <- lrm(DAY30~as.factor(gusto$REGL) * lp1, data=gusto)
```

Random coefficient model for lp1 #

```
fullr.lp.REGL <- lmer(DAY30~lp1 + (1 | REGL) + (0 + lp1 | REGL),
                       family=binomial, method="Laplace," data=gusto)
```

```
# Examine Empirical Bayes estimates
ranef(fullr.lp.REGL)
```

We note no heterogeneity in calibration slopes by region.

## 21.2 Provider Profiling

Whether prediction models are applicable across centres requires an assessment of differences between centres. Differences between centres are also central in comparisons of the quality of centres as part of provider profiling. Provider profiling often includes outcomes such as mortality and morbidity, but may also include measures such as patient satisfaction, and organizational issues such as procedures and processes of delivering care.<sup>99,219</sup> In addition to testing and estimation of differences between centres, a specific aspect of provider profiling is that we may want to rank centres according to their performance in league tables. Such ranking would enable patients (or “consumers”) to choose the best provider for their health problem. Moreover, a relative poor performance might be an incentive for a provider to critically review the processes of care delivery, and stimulate improvements. Such feedback should lead to a continuous quality improvement.

Provider profiling according to outcome is surrounded by many methodological problems.<sup>382</sup> Observational data are analyzed, which generally need to be interpreted with more caution than an experimental study. Some argue that we should concentrate on direct measurement of adherence to clinical and managerial standards.<sup>260</sup> If we aim to compare outcome across centres, two methodological issues are essential:

1. case-mix adjustment
2. dealing with uncertainty

Case-mix adjustment should appropriately capture differences between centres in patient characteristics that are outside the influence of actions in the centre. Instead of predictors, we now consider these patient characteristics as confounders, since they may be both related to setting and outcome. Some centres may treat more severe patients, which hampers a fair comparison with a centre with less severe patients. We want to compare centres after adjusting for confounding factors. Choosing an appropriate adjustment model is not easy, and may be limited by the type of data that is available. For example, administrative data bases may not include all potential confounders, and have problems in coding. For example, postoperative complications may be miscoded as comorbidities.<sup>173</sup> Moreover, end point assessment is often non-standardized.<sup>383</sup> Prediction model development was discussed at a technical level in Part II (Chaps. 7–18). Risk adjustment is discussed in a broader context by Iezzoni.<sup>208</sup>

Second, substantial differences may appear in traditional analysis, with or without adjustment for confounders. But this picture is noisy. We have seen that EB estimation is a more conservative solution, compensating for the randomness in the fixed effect analysis. EB estimates hence allow for a better interpretation of any differences between centres that remain after adjustment for case-mix.<sup>136,260,279,382,384,390,405</sup>

### 21.2.1 *Indicators for Differences Between Centres*

A simple indicator of differences between centres is obtained by comparing the observed outcomes to the expected outcomes for each centre. The expected outcomes can be estimated with a regression model that includes relevant confounders.

For dichotomous outcomes, we can express the absolute difference in the  $W$  statistic:  $W_i = (N_{\text{Obs},i} - N_{\text{exp},i})/N_i$ , or  $\text{mean}(\text{observed}_i) - \text{mean}(\text{expected}_i)$ , calculated for each centre  $i$ .<sup>432,324</sup> The weighted sum of the absolute  $W$  statistics can be interpreted as the percentage of patients that had a different outcome than that expected.

The alternative is to use indicator variables for centres in a regression model, such that relative centre effects are estimated in comparison with the mean outcome over all centres. The intercepts  $\alpha_i$  can be obtained from traditional fixed effect models. EB estimates of  $\alpha_{\text{EB}}$  can be obtained directly with random effects models, or with a two-step approach as discussed before:

$$\alpha_{\text{EB}} = \mu + \tau^2 / (\tau^2 + \sigma_i^2) * (\alpha_i - \mu),$$

The approximate relationship between  $\alpha_i$  and  $W_i$  is:  $\alpha_i \approx W_i/(p(1-p))$ , or  $W_i \approx \alpha_i * (p(1-p))$ , where  $p = \text{mean}(\text{observed}_i)$ . A more exact calculation of  $W$  statistics is possible by comparing the mean predictions per centre, including the centre effect  $\alpha_i$ , to the mean predictions from a model without centre (and only the overall intercept  $\alpha$ ). This calculation can also be done with  $\alpha_{\text{EB}}$  instead of  $\alpha_i$ .<sup>432</sup>

The SE of the  $W$  statistic is  $(1/N_i) * \text{sqrt}(\text{var}_i)$ ; for fixed effect estimates we obtain the SE from the regression model. For EB estimation with the two-step approach, the variance is:

$$\text{var}(\alpha_{\text{EB}}) = \tau^2 * \sigma_i^2 / (\tau^2 + \sigma_i^2),$$

where  $\tau^2$  is the heterogeneity between settings and  $\sigma_i^2$  is the estimated variance for  $\alpha_i$ . We note that if all centres have the same  $\sigma_i^2$ , the fixed estimate  $\alpha_i$  and EB estimate  $\alpha_{\text{EB}}$  differ only in scale, but the ranking nor uncertainty about the ranking have changed. However, centres may have different sample sizes reflected in different  $\sigma_i^2$ . For centres with a small variance  $\sigma_i^2$ ,  $\alpha_{\text{EB}}$  is close to  $\alpha_i$ , and the  $\text{SE}(\alpha_{\text{EB}})$  close to  $\text{SE}(\alpha_i)$ , while centres with large variance have effects near the overall outcome  $\mu$ , with SE close to  $\tau$ .

The SE can be used to calculate 95% confidence intervals, or in the case of EB estimates, “posterior probability intervals”:

$$\alpha_{\text{EB} \pm 1.96} * \sqrt{\text{var}(\alpha_{\text{EB}})}.$$

### 21.2.2 Ranking of Centres

The first attempts of provider profiling already included league tables: rankings were made for physician-specific mortality after coronary-artery bypass grafting surgery in New York State.<sup>150</sup> Ranking is also very popular in the lay press.<sup>476</sup>

Many argue that the uncertainty in differences between centres needs to be reflected in such league tables. The key problem with ranking is that one centre has to be first and one has to be last. One approach was illustrated for league tables of in vitro fertilization clinics, where the uncertainty in rank was indicated with a 95% confidence interval around the rank.<sup>279</sup> If ranking is very noisy, the confidence

intervals are very wide. Van Houwelingen advocates to use expected ranks as was also proposed before,<sup>457,385</sup> and is similar to the idea of using a median rank from a distribution of ranks.<sup>143</sup>

The expected rank is determined by the probability that the performance at centre  $i$  is better than at another centre  $j$ :  $P(\alpha_{EB,i} > \alpha_{EB,j})$ . We use the EB estimates of differences  $\alpha_i$  and  $\alpha_j$ , since these are considered better reflections of any true differences between centres. In practice, we can calculate this probability from the standardized difference in performance estimates:

$$(\alpha_{EB,i} - \alpha_{EB,j}) / \sqrt{(\text{var}(\alpha_{EB,i}) + \text{var}(\alpha_{EB,j}))}$$

We take the sum of these probabilities over all comparisons with centres  $j$ :  $\sum P(\alpha_{EB,i} > \alpha_{EB,j})$ , where  $j \neq i$ . The expected rank ER is estimated as:

$$ER_i = \sum P(\alpha_{EB,i} > \alpha_{EB,j}) = 1 + \sum \Phi((\alpha_{EB,i} - \alpha_{EB,j}) / \sqrt{(\text{var}(\alpha_{EB,i}) + \text{var}(\alpha_{EB,j}))})$$

where  $j \neq i$ , and  $\Phi$  is the normal distribution function. We assume that low values of  $\alpha_{EB}$  are good; if the low value is quite certain, the rank should be close to 1. Indeed, we note that if the summed probability that centre  $i$  has worse outcomes than any other centre  $j$  is low, the rank remains close to 1. If this probability is high, the rank becomes high (poor performance). Such ranking is possible if the differences  $\alpha_{EB,i} - \alpha_{EB,j}$  are large relative to the SE of this difference. If the standardized differences are close to zero, this corresponds to overlap between posterior probability intervals, and expected ranks are around the mid-rank.

For better interpretation we can scale the expected ranks ER between 0 and 100%:

$$PCER_i = 100 * (ER_i - 0.5) / Ncenters,$$

where PCER stands for percentiles based on expected ranks. The  $PCER_i$  can be interpreted as the probability that the performance in centre  $i$  is better than in any randomly selected other centre. If the ER is 1 for a centre, this indicates a much better performance than the other centres. If the comparison is with nine other centres ( $Ncenters=10$ ), the PCER becomes 5%; if the comparison is with 99 other centres ( $Ncenters=100$ ), the PCER becomes 0.5%. The definition hence accounts for the discrete nature of the number of centres instead of the simpler scaling as  $PCER_i = 100 * (ER_i - 1) / (Ncenters - 1)$ .

In summary, the ER and PCER incorporate both the magnitude of the difference of a particular centre compared with other centres and the uncertainty in this difference. These measures for ranking are still relatively new, and need further empirical support for their applicability. We will illustrate the ranking of centres in a case study of outcome after stroke, after considering traditional and EB estimation for between centre differences.

### 21.2.3 Example: Provider Profiling in Stroke

We consider differences in outcome between ten hospitals in The Netherlands.<sup>261</sup> All patients who were admitted to the neurology department with suspected acute brain ischemia between October 2002 and May 2003 were considered for enrollment in the Netherlands Stroke survey.<sup>373</sup> The participating sites comprised one small (<400 beds), four intermediate (400–800 beds) and five large centres (>800 beds). Two centres were University hospitals. All centres had a neurology department, a neurologist with expertise in stroke, and a multi-disciplinary stroke team. All but one hospital had a stroke unit, eight were participating in a regional stroke service, and nine were equipped for thrombolytic therapy.

Data were collected by trained research assistants. The primary outcome was whether patients were dead or disabled at 1 year after admission, i.e. a score on the modified Rankin scale ≥3. Potential confounders included demographics (age, sex), stroke subtype (brain infarction vs. transient ischemic attack), vascular risk factors (e.g. ischemic heart disease, diabetes, hypertension), history characteristics (previous stroke, pre-stroke living condition), presenting characteristics (consciousness level according to Glasgow Coma Scale, arrival at hospital < 48 h). In total 12 confounders were considered for a “full” logistic regression model for adjustment of differences in outcome between hospitals.

### 21.2.4 Testing of Differences Between Centres

The sample consisted of 505 patients with complete data on potential confounders and outcome. The lowest numbers enrolled were 22 and 24 patients in hospitals 5 and 6, and the highest numbers 92 and 99 in hospitals 2 and 7 respectively (Table 21.4). The mean age was 71 years; 279 were male (55%) and 451 (89%) had a brain infarction. The distributions of age and stroke subtype varied significantly by hospital (Table 21.4).

**Table 21.4** Characteristics of 10 hospitals treating 505 patients with acute brain ischemia

Hospital	n	Age (years)	Sex (male)	Stroke subtype (brain infarction)	Poor outcome
1	39	77	46%	97%	59%
2	92	73	54%	95%	72%
3	31	69	61%	97%	35%
4	41	65	59%	80%	44%
5	22	74	55%	91%	73%
6	24	65	67%	63%	29%
7	99	68	65%	94%	39%
8	37	70	41%	92%	78%
9	50	71	56%	88%	54%
10	70	72	47%	81%	46%
Total	505	71	55%	89%	53%

**Table 21.5** Testing for heterogeneity between ten hospitals providing care for stroke patients

	Fixed effect	Random effect
Unadjusted	$\chi^2 = 48, 9 \text{ df}, p < 0.0001$	$\tau^2 = 0.34, \chi^2 = 22, 1 \text{ df}, p < 0.0001$
Age adjusted	$\chi^2 = 39, 9 \text{ df}, p < 0.0001$	$\tau^2 = 0.29, \chi^2 = 15, 1 \text{ df}, p < 0.0001$
12 confounders	$\chi^2 = 23, 9 \text{ df}, p = 0.0056$	$\tau^2 = 0.17, \chi^2 = 3.3, 1 \text{ df}, p = 0.035$

**Table 21.6** Traditional fixed effects and Empirical Bayes (EB) estimates for differences between ten hospitals, adjusted for 12 confounders

Hospital	<i>n</i>	Unadjusted	Adjusted	EB
1	39	0.24	-0.36	-0.18
2	92	0.81	0.45	0.34
3	31	-0.72	-1.04	-0.49
4	41	-0.37	-0.39	-0.21
5	22	0.86	0.91	0.33
6	24	-1.01	-0.47	-0.20
7	99	-0.55	-0.15	-0.12
8	37	1.17	1.16	0.56
9	50	0.04	0.00	0.01
10	70	-0.29	-0.09	-0.05

Values are logistic regression coefficients

At 1 year, 268 (53%) patients had a poor outcome (dead,  $n = 143$ ; Rankin 3, 4, or 5;  $n = 125$ ). The fraction of patients with a poor outcome varied substantially between centres in unadjusted analysis, with apparently best results in hospital 6 (29% poor outcome) and worst in hospital 8 (78%, Table 21.4). These differences were highly significant in a traditional fixed effects analysis of differences between hospitals ( $\chi^2 = 48, 9 \text{ df}, p < 0.0001$ , Table 21.5), but were partly explained by a higher age of patients in hospitals with worse outcome. For example, hospitals 2, 5, and 8 had over 70% poor outcome, but mean ages of 73, 74, and 70 years (Table 21.4). Adjusting for all 12 potential confounders led to halving of the differences seen in unadjusted analysis ( $\chi^2 = 23$  instead of 48, Table 21.5). This pattern was also seen in the random effects analysis, where the estimated  $\tau^2$  (indicating heterogeneity between centres) with adjustment for 12 confounders was half that of the unadjusted  $\tau^2$  (0.17 vs. 0.34, Table 21.5).

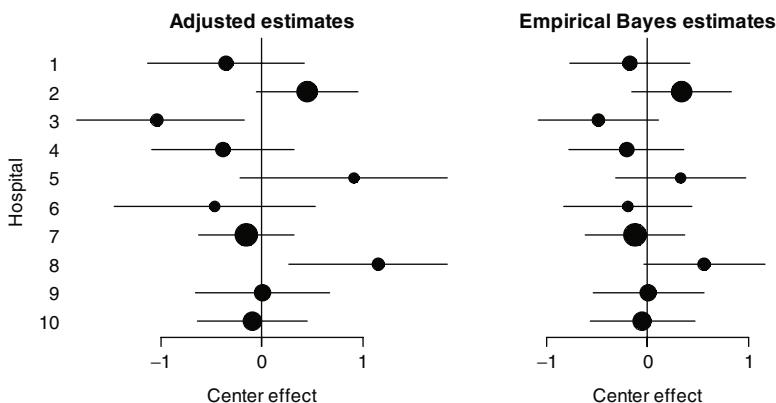
### 21.2.5 Estimation of Differences Between Centres

We can estimate the differences between centres in logistic regression models, where we compare each centre to the average. The traditional fixed effects change considerably between an unadjusted and an adjusted analysis with 12 confounders (Table 21.6). Hospital 1 seems to perform relatively poor in unadjusted analysis (positive coefficient), while an adjusted analysis indicates that this hospital performs relatively good (negative coefficient). Changes for other hospitals were only noted quantitatively, without changing sign, with adjusted differences generally closer to zero. Further changes were seen with EB estimation of differences. All differences were reduced, especially for smaller centres (e.g hospital 5: from 0.91 to 0.33).

**Table 21.7** W statistics for differences between ten hospitals treating patients with stroke

Hospital	<i>n</i>	Unadjusted	Adjusted	EB
1	39	6	-6	-3
2	92	19	8	5
3	31	-18	-19	-10
4	41	-9	-7	-4
5	22	20	14	5
6	24	-24	-8	-3
7	99	-14	-3	-2
8	37	25	18	9
9	50	1	0	0
10	70	-7	-2	-1

The *W* statistic represents the number of patients with a worse or better than average outcome per 100 patients. Values are quite similar in relative effect to the logistic regression coefficients with fixed effect estimation



**Fig. 21.2** Differences between ten centres with traditionally adjusted, fixed effect, estimates, and Empirical Bayes estimates. We note that estimates of relatively small centres (e.g. 5, 6, and 8) are shrunk towards to average with EB estimation

Very similar patterns are noted for the *W* statistics (Table 21.7). The *W* statistics indicate the absolute percentages of patients that have a worse than expected outcome (positive *W* statistic) and better than expected outcome (negative *W* statistic, Table 21.7). The weighted sum of the absolute *W* statistics was 13.3% in unadjusted analysis, 6.7% in adjusted analysis, and only 3.7% in EB analysis.

### \*21.2.6 Uncertainty in Differences

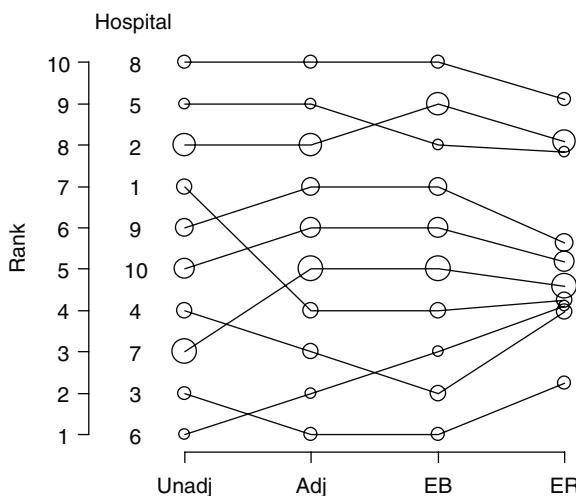
The uncertainty around the estimated differences between centres is indicated in Fig. 21.2 for the adjusted and EB analyses. We note that EB estimation does not affect the point estimate nor the confidence interval for the larger centres, such as

hospital 2 and 7. For smaller centres, such as hospitals 5, 6, and 8, the point estimates for the deviation from the average are shrunken, and the confidence intervals smaller. None of the centres have a deviation that is significantly away from zero in the EB estimation, while the overall heterogeneity is statistically significant (Table 21.5).

### 21.2.7 Ranking of Centres

We can simply rank hospitals in unadjusted, adjusted, and EB analyses (Fig. 21.3). The EB analyses are preferable for estimation of the magnitude of differences between hospitals. Ranking of hospitals based on EB estimates does however not circumvent the problem that one hospital has to be at the top and one at the bottom of a league table. We should also incorporate the uncertainty in the ranking, since there can still be substantial variability in the EB estimates of differences between hospitals. We therefore consider the expected rank (ER) and percentile expected rank (PCER) of each hospital (formulas in Sect. 21.2.2).

The ER can be calculated with consideration of the probability that a hospital is worse than any other hospital. Figure 21.3 shows that this approach leads to shrinkage of the ranks towards the median rank of 5.5 for the ten hospitals. Hospital 8 has rank 10 (poorest performance) in unadjusted, adjusted, and EB analyses, but the ER or EPC is 9.1 or 9.2 respectively, meaning that 1 of 10 centres is expected to be worse than this centre (Fig. 21.3). Hospital 6 seemed to do best in unadjusted analysis (rank 1), shifted to rank 2 in adjusted analysis, to rank 3 in EB analysis, and has an ER around 4.



**Fig 21.3** Ranks of the ten hospitals in unadjusted, adjusted, Empirical Bayes (EB) analyses, and the expected rank (ER). Dot size is based on the square root of the sample size per hospital. According to all analyses, hospital 8 ranks the poorest. Hospital 6 seemed to do best in unadjusted analysis (rank 1), shifted to rank 2 in adjusted analysis, to rank 3 in EB analysis, and has an ER around 4

We can also express these shrunk ranks on a 0–100% scale in the PCER. Hospital 8 has a PCER of 86%, which means that there is 86% probability that the performance in hospital 8 is worse than any randomly selected other centre. Hospital 6 has PCER 36%. Hospital 3 ranks highest, with PCER 17%, meaning that there is only 17% probability that any randomly selected other centre is better than this centre.

### 21.2.8 Essential R Code for Provider Profiling

Some of the R code for the analyses in the stroke example is shown below.

#### *Estimate differences between centres (Table 21.6)*

# Adjusted EB differences with random effects estimation

```
fullr.ZH.Laplace <- lmer(RANKI1J2~AGE+ ..11 vars.. +(1|ZHCLUCO),
  family=binomial, method="Laplace", data=cva, x=T, model=T)
rZH <- ranef(fullr.ZH.Laplace, postVar=T) # EB estimates and variance
EB.ZH <- cbind(as.vector(rZH[[1]]), as.vector(sqrt(rZH[[1]]@postVar)))
names(EB.ZH) <- c("Coef", "SE")
> EB.ZH
```

	Coef	SE
1	-0.17579	0.301
2	0.34016	0.252
3	-0.48861	0.305

#### *Ranking of centres (Fig. 21.3)*

Simple ranking of unadjusted, adjusted and EB estimates of between centre differences, e.g.:

```
rank(unadj.ZH[, "Coef"])
```

#### **Expected rank (ER) and percentile expected rank (PCER):**

```
ER <- rep(NA, 10)
tau2 <- as.numeric(VarCorr(fullr.ZH.Laplace) [[1]])
for (i in 1:10) {
  ER[i] <- 1 + sum(pnorm((EB.ZH[i, 1] - EB.ZH [-i,1])/
    sqrt(EB.ZH[i, 2]^2 + EB.ZH[-i,2]^2)))} # end loop
PCER <- 100*(ER-0.5)/10
> cbind(rank(unadj.ZH[, "Coef"]), rank(adj.ZH[, "Coef"]),
  rank(EB.ZH[, "Coef"]), ER, PCER) # Fig 21.3
```

	Unadj	Adj	EB	ER	PCER
1	7	4	4	4.24	37.4
2	8	8	9	8.09	75.9
3	2	1	1	2.23	17.3
...					

### 21.2.9 Guidelines for Provider Profiling

Some guidelines have recently been suggested for statistical methods for public reporting of health outcomes. These list seven preferred attributes of the statistical modelling.<sup>244</sup>

1. clear and explicit definition of patient sample
2. clinical coherence of model variables
3. sufficiently high-quality and timely data
4. designation of a reference time before which covariates are derived and after which outcomes are measured
5. use of an appropriate outcome and a standardized period of outcome assessment
6. application of an analytical approach that takes into account the multi-level organization of data
7. disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples.

Attributes 1–5 are more general in nature than attributes 6 and 7. We have focused in this chapter on the latter 2 attributes, especially attribute 6 (multi-level organization of data, implying random effects analysis and EB estimation).

## 21.3 Concluding Remarks

We started this chapter with some considerations on the local applicability of prediction models. Specifically, we studied the influence of centre on calibration of predictions. Calibration-in-the-large can be improved by adjusting the intercept in a regression model. The intercept is equivalent to the baseline hazard function in a survival model. The two main approaches to updating of the intercept are a fixed effect and a random effects approach.

If we consider only one specific setting, a fixed effect approach is most natural, although we might also attempt to perform an EB update of the intercept (see Chap. 20). If we consider multiple settings, such as multiple hospitals, EB updating has many advantages, as illustrated with the GUSTO-I and stroke examples. EB estimation is advisable whether we update the intercept, calibration slope, or effects of individual predictor effects per centre.

There may be some confusion about naming and notation in traditional and random effect models. Van Houwelingen refers to “crude” and “adjusted” estimates where we refer to “adjusted” and “EB” estimates. The latter is closer to standard epidemiological nomenclature, where crude estimates are synonymous to unadjusted, fixed effect estimates. Furthermore, random effects models are also known as mixed effect models, or multi-level models. A random effects model for between centre differences may also be labelled a random intercept model.

The methodologic issues around applicability of prediction models are very similar to issues in provider profiling. Note that we have to assume that the predictor effects are identical across settings for provider profiling, similar to traditional confounder correction in epidemiology. If predictor effects differ by setting, the comparison between settings becomes conditional on the specific values of the predictor, similar to the interpretation of predictive effects in the presence of interaction. Again, differences between centres can best be quantified with EB estimates rather than with fixed effect estimates. The randomness of estimates per setting can also be included in the ranking, as was illustrated with the Expected Rank and related measures.

### ***21.3.1 Bibliographic Notes***

The reliability of registration and case-mix adjustment have received substantial attention in debates around provider profiling. The issues of estimation of differences and ranking under uncertainty have only more recently received more attention, while we have seen that the uncertainty in estimates per centre has a large impact on the interpretation of provider profiling attempts. Individual centres are often too small to reliably determine whether they are an outlier (either good or bad).<sup>311</sup> Various graphical possibilities have emerged to indicate performance while taking into account uncertainty. One example is the funnel plot, which can be used in meta-analysis to check for publication bias.<sup>110</sup> Funnel plots avoid spurious ranking of centres into “league tables,” by plotting control limits around the estimated performance based on the precision of the estimates.<sup>396</sup> The performance of a centre over time can be monitored in a CUSUM graph.<sup>156</sup> Finally, we used EB estimation in a direct or two-step approach. A full Bayesian approach is possible with the Gibbs sampler,<sup>134</sup> as e.g. implemented in WinBUGS software (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>).

## Questions

### 21.1 Heterogeneity in across GUSTO-I (Table 21.1)

- (a) In Table 21.1, we note that estimate of regional variability,  $\tau^2$ , is larger when we consider the smaller subsamples (0.033 vs. 0.025 for regions). What could be an explanation for this increase in  $\tau^2$ ?
- (b) The corresponding  $\chi^2$  is 17, which is smaller than the value 28 for regions. How do you explain this?
- (c) What are the 95% probability intervals for true differences between regions and for true differences between the 121 small subsamples? First make the calculations at the logodds scale and verify that these estimates are consistent with Fig. 21.1. Next calculate the 95% probability intervals as odds ratios.

### 21.2 Adjusted vs. Empirical Bayes estimates in GUSTO-I (Fig. 21.1)

- (a) How do you explain the much larger spread between adjusted intercepts between the 121 small subsamples compared with the 16 regions? Why are these shrunk more?
- (b) Consider a subsample where we estimate a logistic regression coefficient of 0.4 for the difference to the (adjusted) average outcome (SE of estimate 0.5, traditional fixed effect analysis). What is the EB estimate if the heterogeneity  $\tau^2$  across centres is 0.2, 0.5, or 2? Use the formula from Sect. 21.1.3 for  $\alpha_{EB}$ .

### 21.3 Provider profiling (Sect. 21.2)

- (a) Mention at least two key problems of ranking providers, such as hospitals.
- (b) Why is ranking especially difficult for relatively small hospitals?

### 21.4 Case-mix adjustment (Tables 21.4–21.7)

Verify (a) that centres with a many good outcomes of stroke had mostly lower aged patients in Table 21.4 and (b) that case-mix adjustment halves the apparent heterogeneity between centres in Tables 21.5 to 21.7.

### 21.5 Rankability of stroke outcomes

- (a) The heterogeneity in the stroke outcomes is substantial and statistically significant (Table 21.5). Nevertheless, the expected ranks of many centres are close to the median rank of 5.5 in Fig. 21.3. How do you explain this modest rankability?
- (b) Calculate the rankability according to the formula  $\rho = \tau^2 / (\tau^2 + \text{median}(s^2))$ , with  $s^2$  indicating the between centre variance. Use the  $\tau^2$  estimate from Table 21.5. Use Fig. 21.2 to determine the median(s) (and median( $s^2$ )).

## **Part IV**

# **Applications**

# Chapter 22

## Prediction of a Binary Outcome: 30-Day Mortality After Acute Myocardial Infarction

**Background** Binary outcomes are encountered in many medical prediction problems, including diagnostic problems (presence of disease) and prognostic outcomes (occurrence of complications, short-term mortality). In this book, one key example of a binary outcome is 30-day mortality in patients suffering from an acute myocardial infarction. A prediction model was developed in over 40,000 patients from the GUSTO-I trial. We review the development of this model according to the seven steps of the checklist for developing valid prediction models presented in Part II. In addition we discuss design and results of a number of methodological studies that were performed in the GUSTO-I data set.

### 22.1 GUSTO-I Study

#### 22.1.1 *Acute Myocardial Infarction*

Acute myocardial infarction (heart attack) is caused by the formation of a clot in one of the coronary arteries that supply blood to the heart muscle. Acute MI is a major public health problem. The age-adjusted incidence of hospitalization for myocardial infarction is around 2 per 1,000 women and 4 per 1,000 for men in the United States.<sup>350</sup> Mortality is substantial in the period immediately after the event, and also during the years after surviving the initial infarction. Some patients die before reaching hospital. Patients seen in hospitals are reported to have an average mortality within 30 days around 6–15%.

The risk of 30-day mortality strongly depends on various prognostic factors (Table 22.1). In younger patients, risks are much lower than in older patients. Other patient demographics are also important (gender, length, weight), as well as the presence of risk factors (hypertension, diabetes, smoking, family history) and the history of previous cardiovascular events (previous MI, angina, stroke, bypass surgery). Relevant presenting characteristics include the location of the infarction and the extent of ECG

**Table 22.1** Categories of prognostic factors predicting 30-day mortality in acute MI

Categories	Examples
Demographics	Age, sex, weight, height, geographical site
Risk factors	Diabetes, hypertension, smoking status, hypercholesterolemia, family history of MI
Other history	Previous MI, angina, cerebrovascular disease (e.g. stroke), bypass surgery, angioplasty
Cardiac state	Location of infarction, electrocardiogram abnormalities
Presenting characteristics	Systolic and diastolic blood pressure, heart rate, left ventricular function (e.g. presence of shock, Killip class)

abnormalities. Very important is the acute state of the patient as reflected by blood pressure, heart rate, and left ventricular function (Killip class, presence of shock).

### **22.1.2 Treatment Results from GUSTO-I**

Various drugs and treatments are available for acute MI, including drugs that attack the clot (thrombolytics) and acute revascularization, such as percutaneous interventions (PTCA). GUSTO-I is one of the major randomized controlled trials that compared alternative treatments for acute MI. Specifically, the comparison was on efficacy of four intravenous thrombolytic regimens.<sup>2</sup> Earlier studies had shown that a new and more expensive thrombolytic drug, tissue plasminogen activator (tPA), restored blood flow through the coronary arteries more quickly and more often than alternative drug regimens. The hypothesis in GUSTO-I was that tPA would show a 1% absolute reduction in 30-day mortality.<sup>2</sup> Treatments in the three other arms included streptokinase (SK), an older and less-expensive thrombolytic drug, which was given with two different regimens of heparin (a drug that helps keep the coronary artery open after the initial break-up of the clot by a thrombolytic drug), and a combination of tPA and SK. The trial enrolled 41,021 patients admitted to 1,081 hospitals in 15 countries. The trial convincingly showed a benefit of tPA treatment ( $p<0.001$ ).

The GUSTO-I trial provides a rich and unique source of information. Various substudies have been reported, often in major general and cardiovascular journals. The large number of patients from all over the world makes for a good base to draw reliable conclusions. GUSTO-I has hence contributed to major progress in knowledge of acute MI.

### **22.1.3 Prognostic Modelling in GUSTO-I**

In the GUSTO-I trial, a comprehensive set of prognostic factors was collected, which was first used for prognostic modelling by Dr. Kerry Lee, representing a

team of GUSTO-I investigators.<sup>255</sup> A detailed analysis of predictors for 30-day mortality was presented. The Lee et al. paper in *Circulation* is quite extensive compared with other prognostic studies published in medical journals. It provides many statistical details on several predictive modelling issues for logistic regression.<sup>255</sup> The paper is freely available from the *Circulation* Web site<sup>254</sup>; the abstract is given in Box 22.1. We review the Lee et al. paper with the model development checklist (Table 22.2).

**Box 22.1 Abstract of the paper by Lee et al. in *Circulation* (1995;91: 1659–1668)<sup>254,255</sup>**

**Predictors of 30-Day Mortality in the Era of Reperfusion for Acute Myocardial Infarction: Results From an International Trial of 41 021 Patients**

Kerry L. Lee, PhD; Lynn H. Woodlief, MS; Eric J. Topol, MD; W. Douglas Weaver, MD; Amadeo Betriu, MD; Jacques Col, MD; Maarten Simoons, MD; Phil Aylward, MD; Frans Van de Werf, MD; Robert M. Califf, MD; for the GUSTO-I Investigators

**Background** Despite remarkable advances in the treatment of acute myocardial infarction, substantial early patient mortality remains. Appropriate choices among alternative therapies and the use of clinical resources depend on an estimate of the patient's risk. Individual patients reflect a combination of clinical features that influence prognosis, and these factors must be appropriately weighted to produce an accurate assessment of risk. Prior studies to define prognosis either were performed before widespread use of thrombolytic therapy or were limited in sample size or spectrum of data. Using the large population of the GUSTO-I trial, we performed a comprehensive analysis of relations between baseline clinical data and 30-day mortality and developed a multivariable statistical model for risk assessment in candidates for thrombolytic therapy.

**Methods and Results** For the 41 021 patients enrolled in GUSTO-I, a randomized trial of four thrombolytic strategies, relations between clinical descriptors routinely collected at initial presentation, and death within 30 days (which occurred in 7% of the population) were examined with both univariate and multivariable analyses. Variables studied included demographics, history and risk factors, presenting characteristics, and treatment assignment. Risk modeling was performed with logistic multiple regression and validated with bootstrapping techniques. Multivariable analysis identified age as the most significant factor influencing 30-day mortality, with rates of 1.1% in the youngest decile (<45 years) and 20.5% in patients >75 (adjusted  $\chi^2=717$ ,  $P<.0001$ ). Other factors most significantly associated with

(continued)

**Box 22.1** (continued)

increased mortality were lower systolic blood pressure ( $\chi^2=550, P<.0001$ ), higher Killip class ( $\chi^2=350, P<.0001$ ), elevated heart rate ( $\chi^2=275, P<.0001$ ), and anterior infarction ( $\chi^2=143, P<.0001$ ). Together, these five characteristics contained 90% of the prognostic information in the baseline clinical data. Other significant though less important factors included previous myocardial infarction, height, time to treatment, diabetes, weight, smoking status, type of thrombolytic, previous bypass surgery, hypertension, and prior cerebrovascular disease. Combining prognostic variables through logistic regression, we produced a validated model that stratified patient risk and accurately estimated the likelihood of death.

**Conclusions** The clinical determinants of mortality in patients treated with thrombolytic therapy within 6 hours of symptom onset are multifactorial and the relations complex. Although a few variables contain most of the prognostic information, many others contribute additional independent prognostic information. Through consideration of multiple characteristics, including age, medical history, physiological significance of the infarction, and medical treatment, the prognosis of an individual patient can be accurately estimated.

**Table 22.2** Checklist for developing valid prediction models applied to the GUSTO-I analysis by Lee et al. in *Circulation*<sup>254,255</sup>

Step	Specific issues	GUSTO-I model
<i>General considerations</i>		
Research question	Aim: predictors/prediction?	Both
Intended application	Clinical practice/research?	Clinical practice
Outcome	Clinically relevant?	30-day mortality
Predictors	Reliable measurement? Comprehensiveness	Standard clinical work-up; extensive set of candidate predictors
Study design	Retrospective/prospective? Cohort; case-control	RCT data: Prospective cohort
Statistical model	Appropriate for research question and type of outcome?	Logistic regression
Sample size	Sufficient for aim?	>40,000 patients; 2,851 events: excellent
<i>7 modelling steps</i>		
Data inspection	Distribution of data Missing values	Table 1 (here: Table 22.3) Single imputation
Coding of predictors	Continuous predictors	Extensive checks of transformations for continuous predictors

(continued)

**Table 22.2** (continued)

Step	Specific issues	GUSTO-I model
Model specification	Combining categorical predictors	Categories kept separate
	Combining predictors with similar effects	
	Appropriate selection of main effects? Assessment of assumptions (distributional, linearity and additivity)?	Stepwise selection Additivity checked with interaction terms, one included
Model estimation	Shrinkage included? External information used?	Not necessary No
Model performance	Appropriate measures used?	Calibration and discrimination
Model validation	Internal validation, including model specification and estimation? External validation?	Bootstrap and ten-fold cross-validation
Model presentation	Format appropriate for audience	No; formula in appendix; later paper focused on clinical application
<i>Validity</i>		
Internal: overfitting	Sufficient attempts to limit and correct for overfitting?	Large sample size, predictors from literature
External: generalizability	Predictions valid for plausibly related populations?	Large set of predictors, representing important domains; not assessed in this study

## 22.2 General Considerations of Model Development

### 22.2.1 Research Question and Intended Application

The title of the paper mentions “Predictors of 30-day mortality ...,” and indeed insight in prognostic effects is an aspect that receives much attention in this paper. But the text also states that the goal of the study was to develop a multivariable statistical model “with patient data routinely collected at initial presentation that would be clinically useful in managing patients who are candidates for thrombolytic therapy.”<sup>255</sup> So, research questions relate both to insight into the relevance of predictors and to obtaining predictions. “Managing patients with acute MI” likely refers to making appropriate decisions among alternative therapies, including the more expensive thrombolytic drug (tPA) and the cheaper drug (SK). The authors argue rightly that these choices should depend on an estimate of the patient’s risk.

This issue is further expanded on in a subsequent paper by Califf et al.<sup>63</sup> Decision-making based on risk is also illustrated in two other publications.<sup>138,229</sup>

The authors provide statements on the requirements for such a prognostic model:

To be broadly useful, a risk-assessment algorithm should include all clinically relevant prognostic indicators and should be derived from a population that represents the types of patients seen in clinical practice so that stable estimates of true risk relations can be assessed. A useful model should appropriately weight clinically relevant predictors and be validated in a population with a broad spectrum of patients and hospital settings.

According to the authors, the GUSTO-I trial data set fulfills these requirements.

### ***22.2.2 Outcome and Predictors***

The outcome was 30-day mortality. This is a “hard” end point, and it was the primary end point in the analysis of treatment efficacy in this trial.<sup>255</sup> For decision making on therapy, mortality and quality of life in the longer term may be more relevant. The gain by using a more expensive thrombolytic drug (tPA) is then reflected in a better (quality-adjusted) life-expectancy.<sup>44</sup>

The study considers many characteristics with potential predictive value. A comprehensive set was considered, loosely based on subject matter knowledge (input from expert clinicians, literature). An overview of the main characteristics is provided in Table 22.3, with their relationship to 30-day mortality in univariate and multivariable analyses.

### ***22.2.3 Study Design and Analysis***

The data come from an RCT. Data collection was prospective, with rigorous quality control on predictor information and outcome assessment. The inclusion criteria for GUSTO-I were relatively liberal, making the findings probably well generalizable to other acute MI patients.

The choice of the statistical model does not receive much attention in the paper; logistic regression is assumed to be suitable for this situation with a dichotomous outcome (dead or alive). This is in agreement with our overview in Chap. 4, where we noted that the logistic regression model is more flexible than some other methods, and can approach non-linear models by including interactions and non-linear terms.

A total of 2,851 patients had died by 30 days. Thirty-nine percent of the deaths (1,125) occurred within 24 h; more than half (55%) occurred within 48 h of randomization. This number of events provides an exceptional and excellent basis for prognostic modelling.

**Table 22.3** Illustration of the effects of prognostic factors in predicting 30-day mortality in acute MI

Predictor	Overall ( <i>N</i> =40,830)		Deaths ( <i>N</i> =2,851)		Unadjusted and adjusted $\chi^2$	
	median [25–75p]		median [25–75p]		2,099	717 (1 <i>df</i> )
Age (years)	62 [52–70]		72 [64–78]		733	550 (1 <i>df</i> )
Systolic BP (mm Hg)	130 [112–144]		120 [100–140]			
Killip	N	col%	N	row%	2,343	350 (3 <i>df</i> )
I	34,825	85%	1,773	5.1%		
II	5,141	13%	716	14%		
III	551	1.3%	181	33%		
IV	313	0.6%	181	58%		
Location of infarction					361	143 (2 <i>df</i> )
Anterior	15,900	39%	1,582	9.9%		
Inferior	23,704	58%	1,181	5.0%		
Other	1,226	3%	88	7.2%		
Previous MI	6,726	16%	807	12%	293	64 (1 <i>df</i> )
Diabetes <sup>a</sup>	6,005	15%	653	11%	146	21 (1 <i>df</i> )
Smoking <sup>a</sup>					483	22 (2 <i>df</i> )
Current	17,543	43%	736	4.2%		
Ex-smoker	11,210	27%	805	7.2%		
Never smoked	12,077	30%	1,310	11%		
Thrombolytic therapy <sup>a</sup>					15	15 (3 <i>df</i> )
SK+i.v. hep	10,377	25%	770	7.4%		
SK+subcut hep	9,796	24%	705	7.2%		
tPA+SK	10,504	26%	723	7.0%		
tPA+i.v. hep	10,344	25%	653	6.3%		
Total	40,830	100%	2,851	7.0%		

The  $\chi^2$  statistics are based on the difference in  $-2 \log$  likelihood between a logistic regression model with one (unadjusted) or more (adjusted) predictors and a model without the predictor  
 MI: myocardial infarction; SK Streptokinase; tPA: tissue plasminogen activator; hep heparin

<sup>a</sup>Data from later version of data set compared with the original publication<sup>255</sup>

## 22.3 Seven Modelling Steps in GUSTO-I

### 22.3.1 Data Inspection

An overview of the data is provided in Table 22.3 (based on Table 1 of the original paper).<sup>255</sup> Outcome (30-day mortality) was complete for 40,830 of the 41,021 patients (99.5%). Distributions of some candidate predictors was quite skewed, e.g.

for Killip class (a measure of left ventricular function). Categories III or IV were present in only 2% of the patients; these categories represent patients in shock.

Missing values occurred for various candidate predictors, but usually only in a small fraction. Missing values were imputed for further statistical analysis (“single imputation,” see Chap. 7 and 8). Imputation was based on the correlation among predictors, which were exploited with flexible functions (`transcan` function in R / S+). Details on the missing values were not provided, nor were analyses repeated with complete cases only.

### ***22.3.2 Coding of Predictors***

Much attention was given to the transformations of continuous predictors. Linear and restricted cubic spline functions were used to describe the relationships between predictor and mortality (see Chap. 10). For further analysis, some simplifying transformations were chosen, including truncation of values (for example for systolic blood pressure).

For categorical variables, detailed categorizations were kept as such for statistical analysis, which was reasonable given the large sample size. For example, many studies consider location of infarction as anterior vs. other. In GUSTO-I, the coding was as anterior (39%), inferior (58%), or other (3%), where “other” included posterior, lateral, and apical locations. Also, Killip class was considered as a categorical variable, despite that classes, III and IV each contained only 1% of the patients. The ordinal nature of this predictor was ignored in the analyses.

### ***22.3.3 Model Specification***

The authors state that they aimed to identify which variables were most strongly related to short-term mortality. This answers a research question related to hypothesis testing, rather than prediction per se. The specific technique used for selection is not explicitly stated, but likely only statistically significant variables were considered as predictors ( $p < 0.05$ ).

The authors tested interactions among the predictors, i.e. they examined whether the prognostic relation of a predictor differed by levels of other predictors (“additivity assumption,” Chap. 12). Linearity of predictors was assessed in detail; transformations chosen at univariate analysis were also used in multivariable analysis.

### ***22.3.4 Model Estimation***

Regression coefficients were estimated with standard logistic regression analysis, which maximizes the log-likelihood of the fit of the model to the data. More advanced

methods are available (Chap. 13), but these modern estimation methods are less relevant in very large data sets such as GUSTO-I.

### 22.3.5 *Model Performance*

Discrimination and calibration were studied to indicate model performance. The area under the receiver operating characteristic curve (AUC, equivalent to the  $c$  statistic) was used to study discrimination. The authors explain that the AUC measures the concordance of predictions with actual outcomes (how well the predictions rank order patients with respect to their outcomes) and that AUC is a simple transformation of Somer's Dxy rank correlation between the model predictions and actual outcomes.

Calibration of the model predictions was assessed graphically and by comparison of the average model prediction to the observed mortality rate across deciles of risk. The latter grouping procedure is often used in the Hosmer–Lemeshow goodness of fit test (Chap. 15). Further, the authors compared predictions and observed mortality within specific subgroups of patients with different risk levels. This method is not often performed to study calibration of prediction models. First, it is only reasonable with large numbers of patients in the subgroups, as in GUSTO-I. More importantly, it is only a check on marginals of predictions according to predictor values. The comparison with observed outcomes will only show violations of non-linearity for continuous variables, and is insensitive to having missed interactions in the model. We discussed various other measures for calibration in Chap. 15. Clinical usefulness was not evaluated explicitly.

### 22.3.6 *Model Validation*

GUSTO-I is a very large data set. This makes that the performance of the model can be assessed reliably in a simple and direct way, i.e. on the same patients that were used to develop the model. Optimism in performance would be a problem in relatively small data sets, i.e. either that many predictors were considered, or that relatively few events were available for the logistic regression analysis. Both are not the case in GUSTO-I. Nevertheless, the authors embarked on attempts to validate the predictive performance of the model, especially the AUC. The authors explain their approach clearly:

First, 10-fold cross validation was performed: the model was fitted on a randomly selected subset of 90% of the study patients, and the resulting fit was tested on the remaining 10%. This process was repeated 10 times to estimate the extent to which the predictive accuracy of the model (based on the entire sample) was overoptimistic. Second, for each of 100 bootstrap samples (samples of the same size as the original population but with patients drawn randomly, with replacement, from the full study population), the model was refitted and then tested on the original sample, again to estimate the degree to which the predictive

accuracy of the model would be expected to deteriorate when applied to an independent sample of patients.<sup>255</sup>

A more extensive description of these validation techniques was provided in Chap. 17. As expected, model optimism was negligible, both in the ten fold cross-validation procedure and with bootstrapping.

### 22.3.7 *Presentation*

#### 22.3.7.1 Predictor Effects

Results of the modelling process were presented in various ways. The relevance of each predictor was shown by an ANOVA table, where the contribution of each predictor was indicated by the drop in an adjusted  $\chi^2$  statistic (Table 22.3, last column). It appears that age is associated with a contribution to the  $\chi^2$  statistic of 717, systolic blood pressure 550, and Killip class (a measure for left ventricular function) 350. The contribution to the multivariable model is much smaller for most predictors than in univariate analysis. This is explained by correlations between predictors. Such correlation is also reflected in the odds ratios (OR). For example, the OR for an increase of 10 years in age was 2.3 in univariate analysis, but 2.1 in multivariable analysis.

Interestingly, the choice of thrombolytic therapy had an adjusted  $\chi^2$  of only 15, which is small compared with the importance of the other predictors. This phenomenon is observed in many prognostic studies: Treatment has a statistically significant impact on outcome, but its relevance is small compared with other prognostic factors.

ORs for the effect of predictors were shown graphically.<sup>255</sup> ORs are calculated from the logistic regression coefficients as  $\text{exp}(\text{coef})$ :  $\text{OR} = e^{\text{coef}}$ . An OR larger than 1 indicates that the risk of mortality is increased, while an OR smaller than 1 indicates that the risk of mortality is decreased (e.g. for tPA treatment vs. SK treatment). For continuous variables, the ORs were presented as the odds of death for patients at the 75 percentile of the distribution of the predictor vs. patients at the 25 percentile. Unfortunately, the graph showed ORs on a linear scale rather than a log scale, which makes it hard to compare the relative magnitude of effects. On a log scale, an OR of 4 would be twice as far away from 1 as an OR of 2, consistent with a doubling in prognostic effect.

#### 22.3.7.2 Predictions

The Appendix lists a formula that can be used to calculate the probability of 30-day mortality for an individual patient.<sup>255</sup> Note however that the formula is difficult to follow. It does not clearly indicate that some transformed variables (height) need a cubic transformation ( $X^3$ ). Also, it may seem remarkable that height is included as a linear term and five transformed variables, while it is stated in the text that 4 *df*

were used to model height. This is because the restricted cubic spline function was re-formulated. The coefficients are un-normalized and two coefficients are added that are linearly dependent on the other coefficients. This makes the model easier to use in a formula or spreadsheet program (see the book's Web site).

## 22.4 Validity

### 22.4.1 Internal Validity: Overfitting

Overfitting was of limited relevance, because of the very large data set. Quite extensive checks of assumptions were performed for a substantial number of candidate predictors, but this "data hungry" approach was reasonable in such a huge data set. Overfitting was assessed by cross-validation and bootstrapping, and found to be irrelevant.

### 22.4.2 External Validity: Generalizability

Will predictions be valid for plausibly related populations? External validity was not assessed in the paper. We note however that a large set of predictors was considered and included in the model, representing important domains of predictors.

Various other models have been developed to predict short-term mortality after acute MI, some before and some after the development of the GUSTO-I model. Usually, large sample sizes were available, such that model development could start de novo. Examples of models developed earlier were the TIMI-II model,<sup>302</sup> the GISSI-II model,<sup>274</sup> and a model from a Belgium centre.<sup>104</sup> More recent models have been developed,<sup>437,7,168</sup> which have not explicitly considered results from the GUSTO-I model. Different predictors were chosen, but the main factors have always included age, infarct location, and measures of ventricular function (such as Killip class).

Interestingly, we found that different models for acute MI may have a similar performance, e.g. an AUC around 0.8, but provide very different predictions for individual patients.<sup>406</sup> These differences were attributable to choice of predictors rather than to differences in regression coefficients, highlighting the importance of model selection issues.

### 22.4.3 Summary Points

- The Lee et al. paper is an excellent illustration of many of the essential steps in developing a valid prediction model
- Nowadays, we could readily deal with missing values in a slightly more sophisticated way than single imputation, although single imputation comes

close to multiple imputation, and is much better than a complete case analysis (Chap. 7).

- A limitation of the model is the translation into clinical practice, where no easily applicable format was used.
- Moreover, generalizability to current clinical practice is doubtful since the overall mortality may have decreased since the years that patients were enrolled in GUSTO-I (early 1990s). We expect a need for model updating, at least of the model intercept.

## 22.5 Translation into Clinical Practice

The model presented in *Circulation* is not easily applicable in the presented form. Many predictors were included, while it was found that 90% of the prognostic information was contained in five variables:

A perspective on the overall contribution of various components of the baseline clinical data to the prediction of mortality can be obtained by use of the global  $\chi^2$  statistic from the logistic model as an index of prognostic information. This index from the full model can be compared with reduced models containing a smaller number of variables. The likelihood ratio  $\chi^2$  statistic for the full model containing all of the prognostic factors was 4379. In contrast, this statistic for a model containing age alone was 2099, meaning that age provides nearly half the prognostic information. Adding other variables provides an increased proportion of information; combining age, systolic blood pressure, Killip class, heart rate, infarct location, and age-by-Killip-class interaction provides approximately 90% of the total prognostic information contained in this array of baseline clinical characteristics.<sup>255</sup>

Further, the presentation in the Appendix as a formula is probably frightening to most clinicians. A simpler format was required. Both issues were addressed in a later publication, which focused on decision making on thrombolytic therapy.<sup>63</sup>

### 22.5.1 Score Chart for Choosing Thrombolytic Therapy

Five predictors were considered and presented in a table to derive a summary score for a patient (see Chap. 18). Age and Killip class were included as main effects and with interaction terms. The interaction effect is well illustrated in Table 22.4. At younger ages, Killip class makes a substantial difference. Equivalently, age matters among those with Killip class I, but less among those with higher Killip classes. At the end of the age range (100 years), some strange patterns arise, with Killip class I patients having a higher score than those with Killip class II or III. This is a biologically implausible pattern. It illustrates that even in a huge data set such as GUSTO-I, artifacts can show up. These artifacts may be due to the specification of the logistic model with a linear interaction term, or to the specific sample. The

**Table 22.4** Score chart to estimate 30-day mortality after acute MI<sup>63</sup>

Predictor	Units	Points	Killip class			
			I	II	III	IV
Age (years)						
			1	II	III	IV
	40	28	42	53	59	
	50	38	49	59	65	
	60	47	56	64	70	
	70	57	63	70	76	
	80	66	70	75	82	
	90	75	77	81	88	
	100	94	91	92	100	
Systolic BP (mm Hg)	40	34				
	80	17				
	120+	0				
Heart rate (beats/min)	10	10				
	30	5				
	50	0				
	90	8				
	130	16				
Infarct location	Anterior	6				
	Inferior	0				
	Other	3				
Previous MI	Yes	5				
Total	Add points	...				

implausible pattern could have been prevented by placing some restrictions on the interactions, as was done for a prediction model for renal artery stenosis<sup>243</sup> and illustrated in Chap. 12.

### \*22.5.2 Predictions for Choosing Thrombolytic Therapy

The score from Table 22.4 corresponds to a probability of 30-day mortality (Table 22.5). We can also determine the benefit of administering tPA instead of SK from this table. A substantial benefit should be estimated before treating with tPA since this drug is expensive and has a substantial risk of side effects (especially bleeding).<sup>138,229</sup> Note that the tPA reduction shows an increase with the score on an absolute scale. The relative reduction was however more or less constant at 15% on the odds scale (OR, 0.85). So, the same relative benefit leads to substantially different absolute benefits. This observation has been made for many other diseases as well (see Chap. 2).

As an example, we consider the score for a hypothetical 65-year-old male. The score would be 60 points for the combination of age 65 and Killip class II, 8 points for a systolic blood pressure of 100 mm Hg, 5 points for heart rate 75 bpm, 6 points for anterior infarct location, and no points for previous MI. The total is 79 points.

**Table 22.5** Translation of score from Table 22.4 into estimated mortality with SK or tPA treatment<sup>63</sup>

Score	SK mortality (%)	tPA mortality (%)	tPA reduction (%)
30	0.4	0.4	–
40	0.8	0.8	0.01
50	1.7	1.4	0.3
60	3.5	2.8	0.8
70	10	8.3	1.7
80	20	17	3
90	40	35	5

When treated with tPA, his 30-day mortality risk is estimated as ~16%, while SK would be predicted to lead to a mortality of ~19%. So, tPA would reduce mortality by ~3%.

### \*22.5.3 Covariate Adjustment in GUSTO-I

The effects of adjustment for predictors have been described for the GUSTO-I data in two methodological studies.<sup>256,403</sup> Both studies considered the effect of tPA vs. SK. The first study considered adjustment for age or a comprehensive set of 17 predictors (age plus 16 other baseline characteristics).<sup>403</sup> The second study used another approach and adjusted for the five most important predictors.<sup>63,256</sup>

In the first analysis, it was found that patients were 0.17 years older in the tPA group (61.03 years,  $n = 10,348$ ) than in the two SK groups (60.86 years,  $n = 20,162$ ).<sup>403</sup> This difference should be fully attributed to chance, and a formal test to compare the ages makes no sense if a proper randomization procedure was followed.<sup>380</sup> However, we know that age is a very strong predictor. The univariate regression coefficient for age was 0.082 per year. We estimated the difference in treatment effect that was attributable to age imbalance by multiplying the difference in mean age with the regression coefficient:  $0.17 \times 0.082 = 0.014$ . The 0.17 years older age of the tPA group made that the treatment effect was underestimated by a factor 0.014 on the logistic scale. The adjusted treatment effect corrects for this imbalance. But it also provides a stratified estimate, which has an expectation further from zero.<sup>133,348</sup> This stratification effect was calculated as the remaining part of the difference between unadjusted and adjusted treatment effect.<sup>403</sup>

The unadjusted treatment effect was an OR of 0.853 (coefficient, -0.1586), and the adjusted estimate was an OR 0.829 (coefficient, -0.1878, 18% more extreme). Age imbalance explained -0.014 or 9% of the difference, leaving another 9% attributable to stratification. Some argue that unadjusted treatment effects are hence biased in a certain sense.<sup>133,182</sup>

It was estimated that an adjusted analysis with 26,900 patients would have the same power as the original unadjusted analysis of 30,510 patients. Such a 12% reduction in sample size is a major argument in favour of adjusted analyses to test

for treatment effect. Either sample sizes could be reduced, or the sample size could be kept at the number based on a traditional, unadjusted, analysis, while the actual analysis would give more power.

Much more can be said on adjustment of treatment effects in randomized clinical trials, which is however beyond the scope of this book. Adjusted analyses were the primary analysis in about half of recently reported RCTs.<sup>18</sup> Advantages are that adjusted analyses have more power, and that adjusted treatment effects may be more relevant for clinical practice. Note that adjusted *p*-values of a particular trial do not necessarily have to be more extreme than those from an unadjusted analysis.<sup>256</sup> However, since we are more interested in the adjusted than the unadjusted effect, the adjusted *p*-value is arguably preferable. The actual gain in power depends on the strength of the prognostic relationships of predictors to outcome. Some argue that adjusted analyses make sense once a specific type of correlation gets larger than 0.2.<sup>339</sup> Finally, any adjustment procedure should be pre-specified in the trial protocol, to prevent a search for the adjustment model that gives the most impressive effect estimate or most extreme *p* value for the treatment effect.

## 22.6 Concluding Remarks

The GUSTO-I case study by Lee et al. illustrates many of the steps that need to be considered in the development of a valid prediction model.<sup>255</sup> It is fortunate that the paper is freely accessible,<sup>254</sup> and that we can make parts of this rich data set available for practical experience in prediction modelling (Chap. 24, data courtesy: the GUSTO Investigators and Duke Clinical Research Institute, Durham, NC).

## Questions

### 22.1 Estimate 30-day mortality (Table 22.4 and spreadsheet)

Consider a male patient with Killip class I, a systolic blood pressure of 100 mm Hg, heart rate 80 bpm, anterior infarct location, and with a previous MI. Use the simple table (Table 22.4) to estimate 30-day mortality, and compare this estimate to the more exact calculation with the full regression formula (spreadsheet at [www.clinicalpredictionmodels.org](http://www.clinicalpredictionmodels.org)).

- (a) What is the risk of mortality from acute MI if this patient is 55-years old?
- (b) What if he were 75-years old?

Now consider decision-making on tPA treatment.

- (c) What is the impact of age on prioritizing tPA treatment based on the reduction in 30-day mortality?
- (d) What might be the priority if we consider gain in life-expectancy instead of 30-day mortality?
- (e) What is the threshold for the ratio between life-expectancies of a 75-vs. 55-year old patient in this example?

### 22.2 Stratification and treatment effects

We study the effect of a hypothetical treatment, with and without stratification for gender. The Table with results is presented here. We compare 30-day mortality (“dead”) between treatments A and B.

Table: hypothetical treatment effect in a randomized controlled trial, with stratification by gender

Treatment	Men		Women	
	Dead	Survived	Dead	Survived
A	10	80	72	18
B	18	72	80	10

- (a) What is the odds ratio for the treatment effect (A vs. B) among men?
- (b) And among women?
- (c) What is the OR for treatment if we do not stratify by gender?
- (d) Is treatment balanced by gender?
- (e) How do you explain these findings?
- (f) What is the OR of gender, ignoring treatment?
- (g) What is the OR of gender, conditional on treatment?
- (h) What would happen if gender had no prognostic effect, i.e. the OR for gender was 1?
- (i) How do these results explain the impact of covariate adjustment in GUSTO-I? Specifically, the unadjusted OR was 0.853 and the adjusted OR 0.829, while imbalance only accounted for a difference of -0.014 on the logodds scale<sup>403</sup>?

# Chapter 23

## Case Study on Survival Analysis: Prediction of Secondary Cardiovascular Events

**Background** Survival is an important long-term outcome in prognostic research, including medical areas such as cardiovascular disease and oncology. We consider a model for the occurrence of vascular events in patients with symptomatic cardiovascular disease. Patient data were from the second manifestations of arterial disease (SMART) study. We go through the seven steps of the checklist for developing valid prediction models, as presented in Part II. The final model looks promising, but needs external validation to prove its actual value. The data set and R code is made available at the book's Web site.

### 23.1 Prognosis in the SMART Study

The SMART study is an ongoing prospective cohort study at the University Medical Centre Utrecht, the Netherlands, initiated and led by Prof Van der Graaf and colleagues. The study was designed to

- (a) establish the prevalence of concomitant arterial diseases and risk factors for cardiovascular disease in a high-risk population;
- (b) identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease.<sup>388</sup>

Currently available prediction models include the Framingham risk score, PROCAM, and SCORE.<sup>17,441,487</sup> These were all developed with data from subjects without clinically manifest atherosclerosis and cannot reasonably be used for patients with clinically manifest cardiovascular disease. These models may be able to rank patients with clinically manifest disease according to risk, but would be expected to underestimate absolute risk.<sup>113</sup>

Assessment of absolute risk is important for secondary prevention. According to the current guidelines all patients who experienced a symptomatic cardiovascular event should be considered as at high risk (more than 20% absolute risk on a future event in the next 10 years). No further categorization is available.

Relevant outcomes in patients with cardiovascular disease (coronary artery disease, cerebral artery disease, peripheral arterial disease and abdominal aortic aneurysm

(AAA)) include stroke, myocardial infarction, or cardiovascular death (Table 23.1). Other end points can be considered depending on the research question, e.g. including cardiovascular interventions. Hard outcomes are generally preferred because they lead to better comparability between studies and hence a better generalizability. The aim in the current study was to develop a prediction rule for patients with cardiovascular disease. We estimate the 1-, 3-, and 5-year risks on the occurrence of vascular events (stroke, myocardial infarction, or cardiovascular death).

### **23.1.1 Patients in SMART**

We consider 3,873 patients who were enrolled in the study in the period September 1996 to March 2006. Patients had a clinical manifestation of atherosclerosis (transient ischaemic attack, ischaemic stroke, peripheral arterial disease, AAA, or coronary heart disease). After informed consent, they underwent a standardized vascular screening, including a health questionnaire for clinical information, laboratory assessment, and anthropometric measurements at enrolment. During follow-up, patients were biannually asked to fill in a questionnaire on hospitalizations and

**Table 23.1** Definitions of fatal and non-fatal vascular events in the SMART study

Event	Definition
Ischaemic stroke	Definite: Relevant clinical features that have caused an increase in impairment of at least one grade on the modified Rankin scale, accompanied by a fresh ischaemic infarction on a repeat brain-scan  Probable: Clinical features that have caused an increase in impairment of at least one grade on the modified Rankin scale; without a fresh ischaemic infarction on a repeat brain-scan
Myocardial infarction	Fatal or non-fatal myocardial infarction: at least two of the following criteria <ol style="list-style-type: none"> <li>1. chest pain for at least 20 min, not disappearing after administration of nitrates</li> <li>2. ST-elevation &gt; 1 mm in two following leads or a left bundle branch block on the ECG</li> <li>3. CK elevation of at least two times the normal value of CK and a MB-fraction &gt; 5% of the total CK</li> </ol>
Vascular death	Sudden death: Unexpected cardiac death occurring within 1 h after onset of symptoms, or within 24 h given convincing circumstantial evidence  Death from ischaemic stroke  Death from intracerebral haemorrhage (haemorrhage on CT-scan)  Death from congestive heart failure  Death from myocardial infarction  Death from rupture of abdominal aortic aneurysm (AAA)  Vascular death from other cause, such as sepsis following stent placement

outpatient clinic visits. When a possible event was reported by a participant, correspondence and relevant data were collected (discharge letters, laboratory and radiology results). Based on all obtained information, every event was audited by three physicians from different departments. The end points of interest for the present study were (acute) vascular death, (non-)fatal ischaemic stroke or (non-)fatal myocardial infarction and the composite end point of any of these vascular events (Table 23.1). If a patient had multiple events, the first recorded event was used for analysis. Data were available on 14,530 person-years collected during a mean follow-up of 3.8 years (range, 0–9 years). A total of 460 events occurred, corresponding to 1-, 3-, and 5-year cumulative incidences of 4.0%, 8.4%, and 14.1% respectively.

## 23.2 General Considerations in SMART

### 23.2.1 *Research Question and Intended Application*

The aim was to develop a prediction model for long-term outcome. Given the available follow-up, 1-, 3-, and 5-year risks could be assessed. Achieving adequate predictions was more prominent than insight in the predictor effects per se (Table 23.2). The intended application was in patient counseling; a high absolute risk might motivate patients to change inappropriate lifestyles and to comply with their medication regimens.

### 23.2.2 *Outcome and Predictors*

The primary outcome was any cardiovascular event, comprising cardiovascular death, non-fatal stroke and non-fatal myocardial infarction. Combining different events is a common approach in cardiovascular research to increase statistical power. A cardiovascular event occurred in 460 patients during follow-up.

The selection of predictors was motivated by characteristics included in Framingham and SCORE models. The relation with future events has also been established for several traditional risk factors, including hyperhomocysteinemia, intima media thickness (IMT), and creatinin.<sup>98,165</sup> Other candidate predictors were demographics (sex and age) and risk factors for vascular events in the general population (smoking, alcohol use, body mass index (BMI), diastolic and systolic blood pressure, lipids, and diabetes). It is well conceivable that indicators of the extent of atherosclerosis are very relevant to predict events in patients with symptomatic atherosclerosis. Such indicators are the location of symptomatic vascular disease (cerebral, coronary, peripheral arterial disease, or AAA), and markers of the extent of atherosclerosis (homocysteine, creatinin, albumin, IMT, and presence of a carotid artery stenosis, Table 23.3). In sum, a relatively limited set of well-defined predictors was studied.

**Table 23.2** Checklist for developing a valid prediction model in the SMART study

Step	Specific issues	SMART model
<i>General considerations</i>		
Research question	Aim: predictors/prediction?	Emphasis on prediction
Intended application	Clinical practice/research?	Clinical practice
Outcome	Clinically relevant?	Hard cardiovascular events
Predictors	Reliable measurement?	Detailed work-up; comprehensive set of candidate predictors
Study design	Comprehensiveness Retrospective/prospective? Cohort; case-control	Prospective cohort
Statistical model	Appropriate for research question and type of outcome?	Cox regression
Sample size	Sufficient for aim?	3,873 patients, 460 events: Very good
<i>Seven modeling steps</i>		
Data inspection	Distribution of data Missing values	Table 23.3 Multiple and single imputation
Coding of predictors	Continuous predictors Combining categorical predictors Combining predictors with similar effects	Truncation and spline transformations for continuous predictors; sum scores for cardiovascular history
Model specification	Appropriate selection of main effects? Assessment of assumptions (distributional, linearity, and additivity)?	Stepwise selection with high <i>p</i> value and Lasso Additivity checked with interaction terms, one included Proportional hazards checked Penalized estimation with Lasso
Model estimation	Shrinkage included?	No
Model performance	External information used?	Focus on discrimination
Model validation	Appropriate measures used? Internal validation, including model specification and estimation?	Bootstrap
Model presentation	External validation? Format appropriate for audience	No external validation Nomogram
<i>Validity</i>		
Internal: Overfitting	Sufficient attempts to limit and correct for overfitting?	Large sample size, predictors from literature, Lasso for selection and shrinkage
External: Generalizability	Predictions valid for plausibly related populations?	Large set of predictors, representing important domains; not assessed in this study

**Table 23.3** Potential predictors in the SMART study data set ( $n = 3873$ )

Characteristics	
<i>Demographics</i>	
Female sex (“SEX,” $n$ , 0 missing)	975 (25%)
Age (“AGE,” in years, 0 missing)	60 [52–68]
<i>Classical risk factors</i>	
Smoking (“SMOKING,” $n$ (%), 25 missing)	
Never	693 (18%)
Former	2711 (70%)
Current	444 (12%)
Packyears (“PACKYRS,” in years, 21 missing)	20 [6–34]
Alcohol (“ALCOHOL,” $n$ (%), 25 missing)	
Never	751 (20%)
Former	408 (11%)
Current	2,689 (69%)
Body mass index (“BMI,” in kg/m <sup>2</sup> , 3 missing)	26.7 (24–29)
Diabetes (“DIABETES,” $n$ (%), 40 missing)	846 (22%)
<i>Blood pressure</i>	
Systolic, by hand (“SYSTH,” in mm Hg, 1,498 missing)	140 (126–155)
Systolic, automatic (“SYSTBP,” in mm Hg, 1,223 missing)	139 (127–154)
Diastolic, by hand (“DIASTH,” in mm Hg, 1,499 missing)	82 (75–90)
Diastolic, automatic (“DIASTBP,” in mm Hg, 1,221 missing)	79 (73–86)
<i>Lipid levels</i>	
Total cholesterol (“CHOL,” in mmol/L, 18 missing)	5.1 [4.4–5.9]
High-density lipoprotein cholesterol (“HDL,” mmol/L, 30 missing)	1.17 [0.96–1.42]
Low-density lipoprotein cholesterol (“LDL,” mmol/L, 216 missing)	3.06 [2.39–3.83]
Triglycerides (“TRIG,” mmol/L, 28 missing)	1.54 [1.12–2.23]
<i>Previous symptomatic atherosclerosis</i>	
Cerebral (“CEREBRAL,” $n$ (%), 0 missing)	1,147 (30%)
Coronary (“CARDIAC,” $n$ (%), 0 missing)	2,160 (56%)
Peripheral (“PERIPH,” $n$ (%), 0 missing)	940 (24%)
Abdominal aortic aneurysm (“AAA,” $n$ (%), 0 missing)	416 (11%)
<i>Markers of atherosclerosis</i>	
Homocysteine (“HOMOC,” µmol/L, 463 missing)	12.8 [10.3–15.7]
Glutamine (“GLUT,” µmol/L, 19 missing)	5.7 [5.3–6.5]
Creatinine clearance (“CREAT,” mL/min, 17 missing)	89 [78–101]
Albumin (“ALBUMIN,” $n$ (%), 207 missing)	
No	2,897 (79%)
Micro	655 (18%)
Macro	114 (3%)
Intima media thickness (“IMT,” mm, 98 missing)	0.88 [0.75–1.07]
Carotid artery stenosis > 50% (“STENOSIS,” $n$ (%), 93 missing)	722 (19%)

Numbers are  $n$  (%) or median (25–75 percentile)

### 23.2.3 Study Design and Analysis

The SMART study is designed as an ongoing, prospective dynamic cohort study. Patients are enrolled when presenting at the hospital, with follow-up starting from study inclusion. We used the Cox regression model, which is the default statistical model for survival outcomes. This model is appropriate for prediction of an outcome at relatively short-term such as 5-year cumulative incidence of cardiovascular events. For long-term predictions (e.g. 10-year incidences), a parametric model might be preferable such as a Weibull model. A Weibull model provides more stable estimates at the end of the follow-up.<sup>312,65</sup>

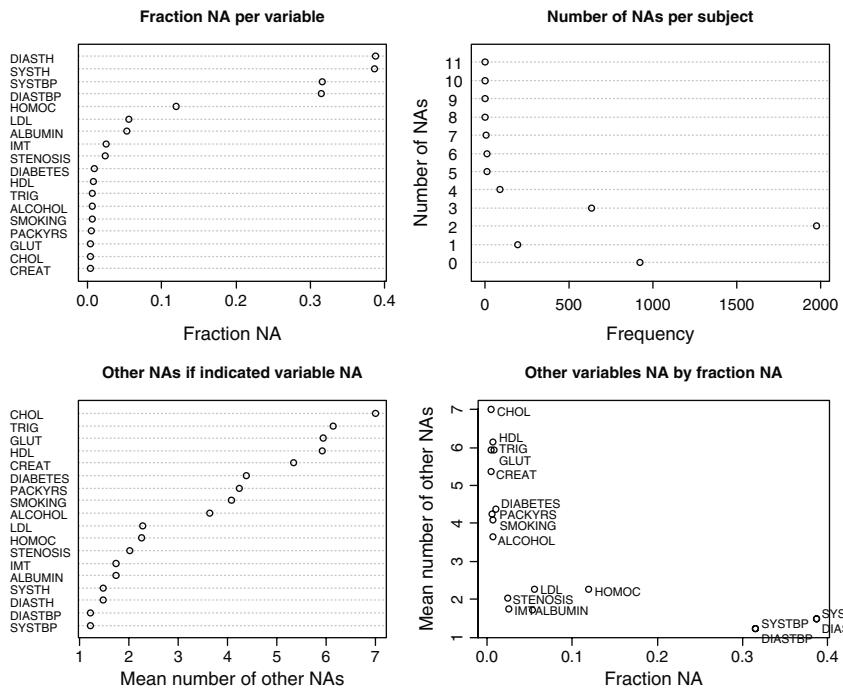
With respect to sample size, the balance of 460 events and ~25 candidate predictors is reasonable (Table 23.3). At least 10–20 events per candidate predictor have been proposed in previous guidelines for sensible development of a prediction model.<sup>175,326,410</sup>

## 23.3 Data Inspection in the SMART Cohort

It appeared that the number of missing values was rather limited for most of the 18 potential predictors (<1%, Table 23.3 and Fig. 23.1). Many missings were however noted among four variables that relate to blood pressure measurements (two for diastolic and two for systolic pressure). In the first years of the study, blood pressure was measured combined with measurement of the distensibility of the carotid artery wall (“SYSTBP” and “DIASTBP” variables). Four years after the start of the study it was decided to measure blood pressure with the conventional sphygmomanometry as well. This measurement is considered in most current guidelines. Hence, conventional diastolic and systolic measurements are obvious candidate predictors for our model rather than the automated measurements. Nearly all patients had at least one type of blood pressure measurement, but many had missing values for conventional sphygmomanometry ( $n=1,498$ , “SYSTH” and “DIASTH” variables). Pearson correlation coefficients were 0.69 and 0.59 for systolic and diastolic blood pressure measurements in 1,155 and 1,156 patients with conventional as well as automatic measurements available, respectively.

The variable homocysteine (“HOMOC”) had 463 missings (12%, Table 23.3, Fig. 23.1, upper left panel). This was related to the fact that homocysteine was not routinely measured in the first years of the study. This is a typical “missing completely at random” (MCAR) situation. Also for the other variables we assume that missingness was more related to logistic reasons, because all patients underwent the same screening protocol. The decision to measure variables was not obviously dependent on other observations (MAR mechanism), the values of the characteristic itself, or characteristics not available in our dataset (MNAR mechanisms).

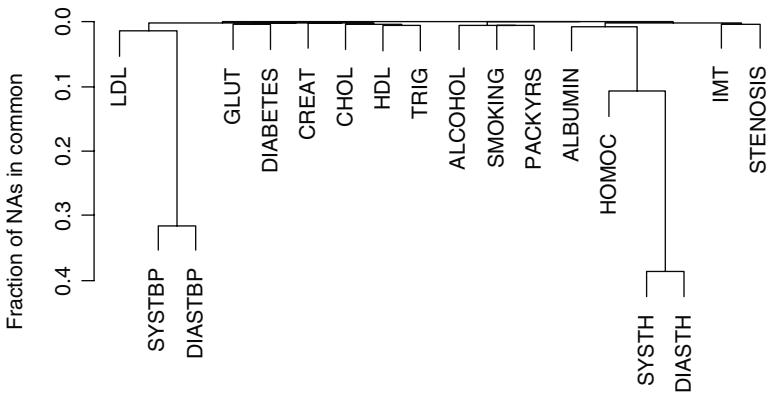
A total of 925 patients had no missing values among the 18 potential predictors, and 1,975 patients had 2 missing values (mostly: 1 type of blood pressure measurement not performed). Few patients had many missings (18 with 7 or more missings,



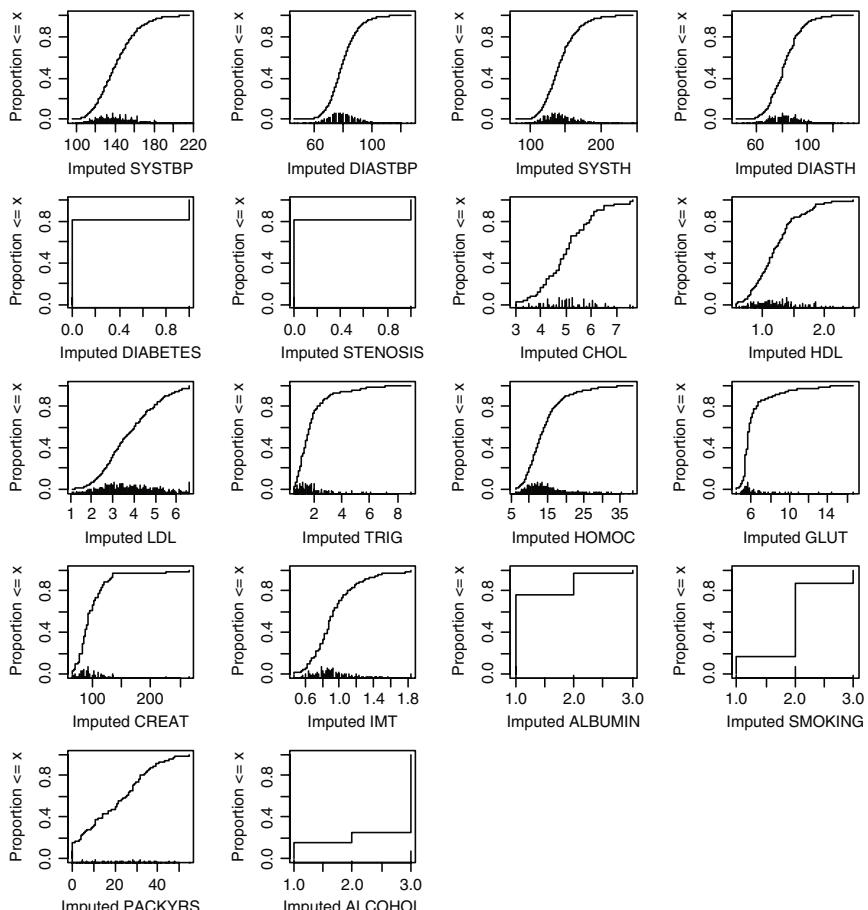
**Fig. 23.1** Patterns of missing data in the SMART study ( $n=3873$ , na.plot2 function). NA: not available (“missing”)

Fig. 23.1, upper right panel). If one type of blood pressure measurement was missing, few other variables had missing values. If cholesterol or triglycerides were missing (which was very rare), many other predictors were also missing (Fig. 23.1, lower left and lower right panels). Further details on the combinations of missing values are shown in Fig. 23.2. Again we note that the diastolic and systolic blood pressure measurements are always jointly missing. In the early years of the study, both homocysteine (“HOMOC”) and conventional sphygmomanometry blood pressure measurements (“SYSTH” and “DIASTH” variables) were not performed, leading to some correlation of missingness between these variables.

Missing data per predictor would lead to a substantial loss of information if only complete cases were used in the multivariable model. We therefore used multiple imputation techniques (aregImpute function) to replace the missing values (Fig. 23.3). The set of first imputations was used for further analyses (“single imputation”). Although multiple imputation is preferable from a theoretical view point, single imputation was considered more practical and sufficient to obtain reasonable predictions (Chap. 7). Final models were also constructed with multiple imputed data sets to check for any relevant differences in point estimates, and widening of confidence intervals.



**Fig. 23.2** Cluster analysis of patterns of missingness in the SMART study ( $n=3,873$ , naclus function)



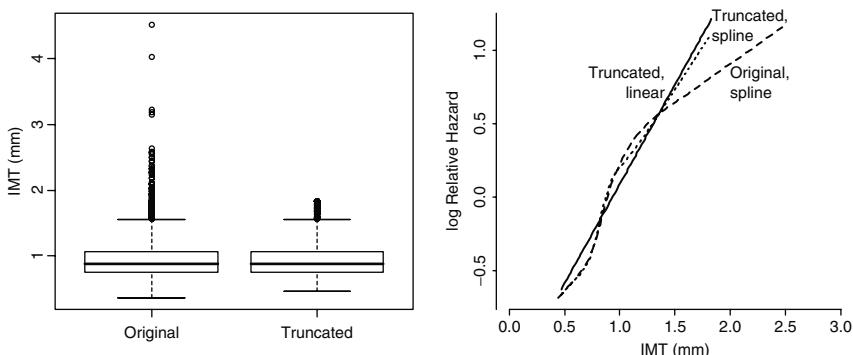
**Fig. 23.3** Distribution of imputed values for the 18 most relevant predictors, which had missing values in the SMART study

## 23.4 Coding of Predictors

### 23.4.1 Extreme Values

Before any modelling started, the distributions of all potential predictors were carefully examined for extreme values. Preferably data are checked with the source documents but sometimes such decisions have to be made on common sense. Biologically implausible values were set to missing values, and remaining extreme values were truncated by shifting the values below the 1 centile and above the 99 centile to “truncation points” (Chap. 9). Such truncation may prevent distortion of the relationship between predictor and outcome due to high leverage of the extreme values, which is not desirable.<sup>356</sup>

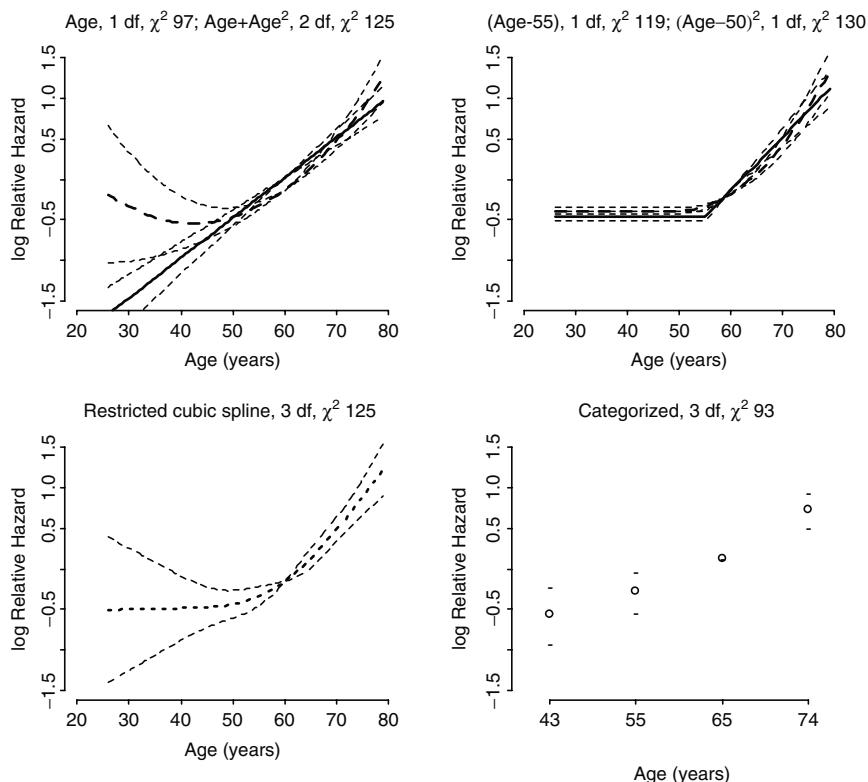
We truncated extreme values for IMT (Fig. 23.4). The mean IMT was 0.94 mm, but some patients had measurements as high as 4 mm. These high values are the result of plaque formation in the carotid artery, and may have an unduly large influence on estimates of cardiovascular event risk. A total of 51 values higher than 1.83 were shifted to 1.83 (the upper truncation point), and 13 values below 0.47 were shifted to 0.47 (the lower truncation point). We note a substantial effect of truncation on the relationship between IMT and outcome (Fig. 23.4, right panel). A restricted cubic spline based on the original IMT values flattens off with high IMT ( $>1.5$  mm), while a restricted cubic spline based on the truncated IMT values is very close to a straight line. This finding illustrates that truncation may obviate the need for a non-linear transformation.<sup>174</sup> Before truncation the Cox regression coefficient for a linear IMT variable was 0.91, while it was 1.36 after truncation. The univariate model  $\chi^2$  improved from 61 before to 75 (1 df) after truncation. Similarly we truncated BMI, lipids (cholesterol, HDL, LDL, triglycerides), homocysteine, and creatinin levels by shifting values below 1 centile and above 99 centile to the truncation points.



**Fig. 23.4** Boxplot of intima media thickness (IMT, in mm, left panel) before and after truncation, and a plot of the effect of IMT on cardiovascular events in a univariate Cox regression model (right panel). The original IMT values are sometimes extremely high, leading to a spline that flattens off with high IMT values. The truncated IMT values have a smaller range and lead to a quite linear relationship (solid line, linear term; dotted line, spline)

### 23.4.2 Transforming Continuous Predictors

Age is an important predictor of cardiovascular events. We considered several age transformations (Fig. 23.5, Table 23.4). In our cohort the Wald  $\chi^2$  of the linear fit was 97. Adding age increased the  $\chi^2$  to 125, but there was a biologically implausible increased risk below age 40 years. Based on visual inspection (Fig. 23.5), it may be judged reasonable to assume no age effect till age 55, and a linear effect for age >55 (“(Age–55) $_{+}^{}$ ” variable,  $\chi^2$  119). A transformation such as  $(\text{Age}-50)_{+}^{2}$  led to an even better model ( $\chi^2$  130, Fig. 23.5). A restricted cubic spline with 3 df (4 knots) did not describe the relationship of age to outcome better ( $\chi^2$  125). Categorizing in quartiles has a clearly lower performance ( $\chi^2$  93). Such categorization should not be used because jumps in predictions are unnatural. Dichotomizing at age 60 (close to the median of 61 years) led to a substantial decrease in performance ( $\chi^2$  72, Table 23.4), illustrating that dichotomization is “a bad idea.”<sup>355</sup>



**Fig. 23.5** Transformations of age in univariate analysis of the SMART study. *Upper left:* age linear and age plus age squared; *upper right:* age linear after 55 years (“(Age–55) $_{+}^{}$ ”) and age squared above 50 years (“(Age–50) $_{+}^{2}$ ”); *lower left:* restricted cubic spline, 4 knots, 3 df; *Lower right:* age categorized in four groups.

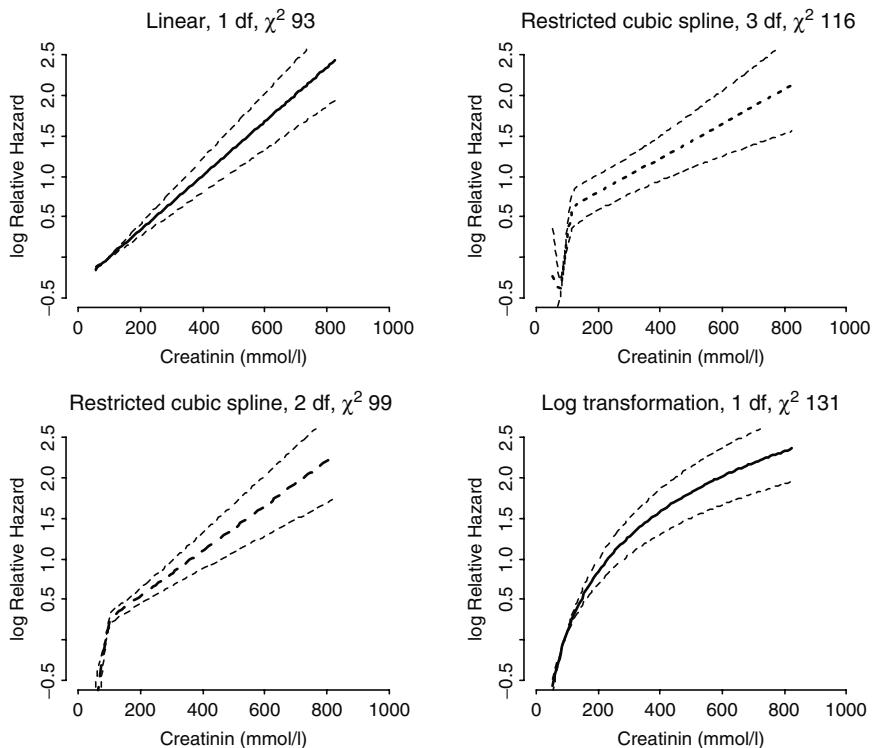
**Table 23.4** Impact of various codings of predictors in a univariate Cox regression models for the SMART study

Predictor	Coding	Wald $\chi^2$	df
Age	Linear	97	1
	Squared	125	2
	(Age–55) $^2$ : Linear effect after age 55	119	1
	(Age–50) $^2$ : Square effect after age 50	130	2
	Restricted cubic spline, 3 df	125	3
	<50, 50–59.9, 60–69.9, $\geq$ 70	93	3
Creatinine	<60, $\geq$ 60	72	1
	Linear	93	1
	Restricted cubic spline, 3 df	116	3
	Restricted cubic spline, 2 df	99	2
Blood pressure (conventional reading)	Log	131	1
	Linear systolic	15	1
	Restricted cubic spline systolic, 2 df	15	2
	Linear diastolic	0.7	1
	Restricted cubic spline diastolic, 2 df	2	2
Previous symptomatic atherosclerosis	Sumscore 0–4	96	1
	Sumscore 0–5 (AAA=2)	119	1
	Separate terms	123	4
	Cerebral	36	1
	Coronary	19	1
	Peripheral	23	1
	Abdominal aneurysm aorta	96	1

Other continuous predictor variables were examined in a similar way; some examples are shown in Table 23.4. For creatinine, a log transformation gave the best fit (Fig. 23.6). A linear coding of systolic blood pressure was reasonable, and diastolic blood pressure had no effect when we analyzed the conventional sphygmomanometry blood pressure measurements (“SYSTH” and “DIASTH” variables). All analyses were repeated with multiply imputed data sets, with largely similar results.

### 23.4.2 Combining Predictors with Similar Effects

Combining predictors with similar effects can be an effective way to limit the degrees of freedom of predictors in a model (Chap. 10). In atherosclerotic patients several variables reflect the extent of atherosclerosis. The affected organs reflect the load of atherosclerosis in one particular patient. The location of symptomatic events (cerebral, coronary, AAA, peripheral artery disease) can be entered separately in the model. For each parameter we would spend 1 df, resulting in a model  $\chi^2$  of 123 (4 df, Table 23.4). If we combine the presence of previous



**Fig. 23.6** Transformations of creatinine in univariate analysis of the SMART study

vascular events in one variable, simply by assuming equal weights for each condition, the model  $\chi^2$  is 96 (1 *df*). The difference of the two models is a  $\chi^2$  of 27, which is highly significant at 3 *df*. Separate terms hence lead to a much better fit. When we test for the separate contributions of each localization it appears that the contribution of an AAA is considerably higher than the contribution of the other localizations. If we attribute 2 points for the presence of an AAA, the sumscore performs remarkably better (range, 0–5; model  $\chi^2$ , 119, close to 123 for separate terms, Table 23.4).

## 23.5 Model Specification

A full, main effects model was defined, which included the common demographics age and sex, important classical risk factors (smoking status, alcohol use, BMI, blood pressure, lipid levels, and diabetes), the sum score for previous symptoms of

atherosclerosis, and finally markers of the extent of the atherosclerotic process (including hyperhomocysteinemia, creatinin, IMT of the carotid artery, carotid artery stenosis, and albuminuria). We focused on systolic blood pressure since recent publications stress the more important role of systolic rather than diastolic blood pressure in predicting cardiovascular events.<sup>428</sup> The full model consisted of 14 predictors, with several having rather limited contributions (Table 23.5). Predictors with a large prognostic strength were age ( $\chi^2$  39), the localization of the symptom of atherosclerosis (sumscore  $\chi^2$  37), and the marker of renal damage creatinin ( $\chi^2$  24). Other characteristics had much smaller prognostic relevance, with some impact of the general marker of atherosclerosis IMT ( $\chi^2$  9.9), but a minor contribution of homocysteine. The classical risk factors had at most a  $\chi^2$  of 6 (for HDL) and hence hardly contributed to the model predictions.

**Table 23.5** Hazard ratios (HRs) and contribution to Cox regression model ( $\chi^2$  and  $df$ ) of the predictors in a full model for cardiovascular events in the SMART study

Predictor	HR [95% CI] <sup>a</sup>	$\chi^2$	$df$
(Age–50) <sup>2</sup> (years above 50)	1.5 [1.3–1.7]	39	1
Gender (male)	0.9 [0.7–1.2]	0.1	1
<i>Classical risk factors</i>			
Smoking		1.1	2
Never	0.9 [0.7–1.2]		
Former	1		
Current	1.1 [0.7–1.6]		
Alcohol		1.1	2
Never	1.2 [0.8–1.6]		
Former	1		
Current	1.1 [0.8–1.4]		
Body mass index (kg/m <sup>2</sup> )	0.9 [0.8–1.0]	3.2	1
Systolic blood pressure (mm Hg)	1.0 [0.9–1.2]	0.3	1
HDL	0.8 [0.7–1.0]	5.4	1
Diabetes	1.3 [1.0–1.8]	4.5	1
<i>Previous symptomatic atherosclerosis</i>			
Sumscore (AAA 2 points)	1.4 [1.3–1.6]	37	1
<i>Markers of atherosclerosis</i>			
Homocysteine (mmol/L)	1.0 [0.9–1.1]	0.2	1
Creatinin (mmol/L)	1.2 [1.1–1.3]	24	1
Albumin		5.2	2
No	0.8 [0.6–1.0]		
Micro	1		
Macro	1.1 [0.7–1.7]		
Intima media thickness (mm)	1.2 [1.1–1.3]	10	1
Carotid artery stenosis > 50%	1.2 [1.0–1.5]	3.6	1

A single imputed data set was used with  $n=3,873$ . Hazard ratio [95% confidence interval] refers to interquartile range for continuous predictors

We tested interactions between the predictors and gender by including cross-product terms with predictors in the selected model (overall  $\chi^2 15$ , 10  $df$ ,  $p=0.14$ ). The strongest interaction was between sex and the sumscore for previous symptomatic atherosclerosis ( $\chi^2 8.1$ , 1  $df$ ,  $p=0.004$ ). In all, the interactions were not considered relevant enough to include an interaction term with sex in the model. We also tested proportionality of hazards. The overall test was not significant (overall  $\chi^2 12$ ,  $df 10$ ,  $p=0.27$ , `cox.zph` function). Detailed results of the assessment of interactions are provided at the Web.

### 23.5.1 Selection

We judged our sample size as large enough to allow for some model reduction for easier practical application (460 events, full model with 17 degrees of freedom, ignoring that the coding of predictors also consumed some degrees of freedom). One approach was to apply a backward selection procedure with a higher than standard  $p$  value. We used Akaike's Information Criterion (AIC), which implies a  $p$  value  $< 0.157$  for selection of predictors with 1  $df$ .<sup>14</sup>

A promising alternative is to apply the Lasso method, which achieves selection of predictors by shrinking some coefficients to zero by setting a constraint on the sum of the absolute standardized coefficients.<sup>435</sup> The Lasso model was found to be optimal with ten predictors, but in this model, the coefficient of homocysteine was close to zero. With more shrinkage, this predictor was dropped, and the same set of nine predictors was selected as in the stepwise selection procedure with AIC (Table 23.6).

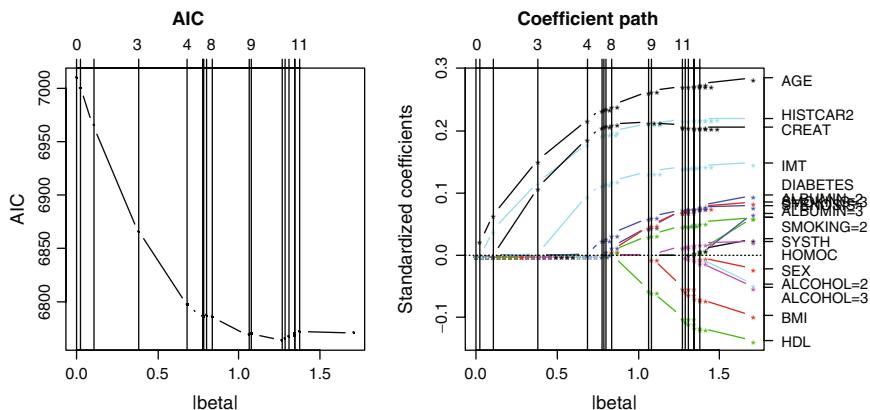
## 23.6 Model Estimation, Performance, Validation, and Presentation

### 23.6.1 Model Estimation

Regression coefficients were first estimated as default with Cox regression analysis, i.e. by maximizing the log-likelihood of the fit of the model to the data. The coefficients of the nine predictors in the stepwise backward selected model were rather similar to their corresponding coefficients in the full model (Table 23.6). In contrast, the Lasso model shrunk coefficients of weaker predictors such as BMI, HDL, diabetes, and albumin considerably towards zero. The effects of strong predictors, such as age, sumscore for atherosclerosis, creatinin, IMT, and carotid artery stenosis, were comparable with the maximum likelihood estimates (Fig. 23.7).

Table 23.6 Cox regression coefficients in the full model, a stepwise selected model (using Akaike's Information Criterion), and in the Lasso model

Predictor	Full	Stepwise (AIC)	Lasso
(Age-50) <sup>2</sup> (years above 50)	0.0013	0.0013	0.0012
Gender (male)	-0.049	Not selected	Not selected
Smoking		Not selected	Not selected
Never	0		
Former	0.13		
Current	0.21		
Alcohol		Not selected	Not selected
Never	0		
Former	-0.15		
Current	-0.11		
Body mass index ( $\text{kg}/\text{m}^2$ )	-0.025	-0.026	-0.001
Blood pressure (mm Hg)	0.0012	Not selected	Not selected
HDL	-0.37	-0.39	-0.16
Diabetes	0.23	0.23	0.11
Previous vascular disease	0.34	0.35	0.33
Homocysteine (mmol/L)	0.0042	Not selected	Not selected
Log(creatinin) (mmol/L)	0.68	0.71	0.71
Albumin			
No	0	0	0
Micro	0.22	0.24	0.13
Macro	0.35	0.35	0.20
Intima media thickness (mm)	0.55	0.56	0.50
Carotid artery stenosis > 50%	0.20	0.22	0.16



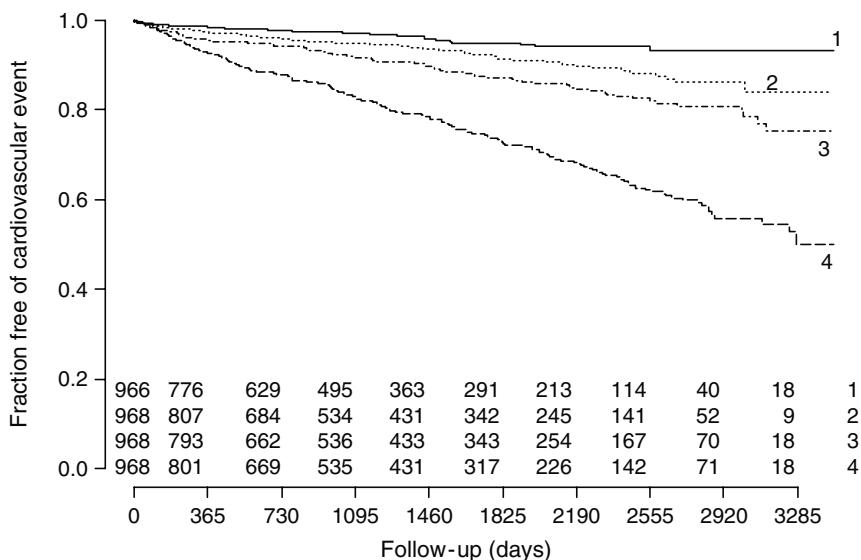
**Fig. 23.7** Lasso path with increasing sum of the absolute standardized coefficients ( $|\beta|$ ). The optimal AIC is obtained with 11 predictors, but differences are small between models with 9–12 predictors. The coefficient path shows that predictors have effects other than zero with higher  $|\beta|$

### 23.6.2 Model Performance

Discrimination of the final model was indicated by the  $c$  statistic, which was 0.693 (95% CI, 0.65–0.73). Discrimination was further illustrated by dividing the predictions in quartiles, and plotting the Kaplan-Meier curves of these four groups (Fig. 23.8). We note that patients in the lower quartile had a considerably poorer chance of being free of cardiovascular events during follow-up: Around 75% at 5 years of follow-up, and near 50% at 9 years of follow-up.

### 23.6.3 Model Validation: Stability

We used a bootstrap re-sampling procedure to study the stability of our stepwise selected model, and to quantify the optimism of our modelling strategy. We found that age and localization of symptoms were strong predictors and were always selected when we repeated the stepwise selection procedure in 200 bootstraps (Table 23.7). In contrast, sex, smoking, alcohol, and systolic blood pressure were selected in only 26%, 40%, 36%, and 57% of the bootstrap samples respectively, consistent with their exclusion from the stepwise model. Albumin, HDL, and IMT were selected in the majority of the bootstraps, but not in all.



**Fig. 23.8** Fraction free of cardiovascular event according to quartiles of the linear predictor. Numbers at risk are indicated for the upper to lower quartile (numbered 1–4)

**Table 23.7** Frequency of selection of predictors from the full model in the backward stepwise selected model (with AIC)

	AGE	SEX	SMOKING	ALCOHOL	SYSTH	BMI	HDL	DIABETES	HISTCAR2	HOMOC	CREAT	ALBUMIN	STENOSI	IMT
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
100	26	40	36	57	60	78	68	100	38	99	84	63	95	

Results are shown for the first 20 bootstrap samples

There was a clear association between the estimated effect of a predictor according to the Lasso and the frequency of selection in the bootstrap procedure. The coefficients for BMI, HDL, and diabetes were considerably reduced according to the Lasso, and indeed these were not selected in 40%, 22%, and 32% of the bootstrap samples. Instead of excluding the predictor, which is equivalent to setting the coefficient to zero, the coefficient was shrunk towards zero. The coefficients of age, localization of symptoms, creating, and IMT were virtually not affected by the Lasso, consistent with their selection in over 95% of the bootstraps.

### ***23.6.4 Model Validation: Optimism***

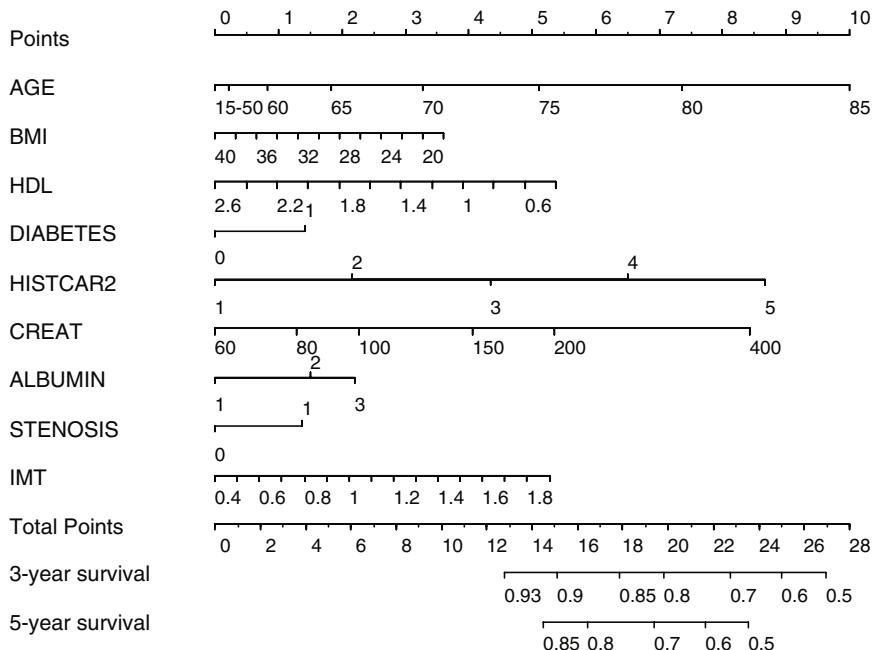
The  $c$  statistic was expected to decrease from 0.693 to 0.680, or a decrease of 0.013, in a bootstrap procedure with repeated selection of predictors in every bootstrap sample. We estimated the required shrinkage for the coefficients in the stepwise selected model as 0.94, suggesting that each coefficient should be reduced by 6% to correct for optimism of the modelling process. Instead of using this shrinkage factor for the final model, we used the Lasso coefficients, which reduce coefficients for weak predictors more than for strong predictors. In all, the bootstrap validation procedure showed some instability of the model specification, but a modest amount of optimism in the final model.

### ***23.6.5 Model Presentation***

The results of the modelling process can be presented in various ways. From Table 23.4 we learn about the relative contributions of each predictor to the model. For a survival model such as the SMART prediction model, an attractive way is to present the model as a nomogram (Fig. 23.9). In the nomogram, we can judge the relative importance of each predictor by the number of points attributed over the range of the predictor, and we can calculate 3-year and 5-year survival estimates. Survival relates to the probability of being free of a cardiovascular event.

## **23.7 Concluding Remarks**

This case study illustrates how a prediction model can be developed and internally validated for a survival analysis problem. We recognize that not all modelling steps could be considered in the bootstrap procedure for internal validation. Further external validation is necessary in the same setting (with more recent patients) and in other settings (to assess transportability).



**Fig. 23.9** Nomogram to calculate predicted 3-year and 5-year survival (probability of being free of a cardiovascular event). Coefficients are based on the Lasso model. For example, a 75-year-old patient, with a BMI of 28, HDL 1, no diabetes, previous aortic aneurysm but no other symptoms of atherosclerosis (HISTCAR2=2), a creatinin value of 100, low albumin, no carotid stenosis, IMT of 1 mm, has a total points score of  $5 + 2 + 4 + 0 + 2 + 2 + 0 + 0 + 2 = 17$ . This corresponds to predicted 3- and 5-year survival of 87% and 79% respectively

We note a distinction between risk factors in the general population (without cardiovascular disease) and prognostic factors in patients with symptomatic disease. Classical risk factors such as smoking, alcohol use, BMI, blood pressure, HDL and diabetes, had very limited prognostic value in the clinical setting. These characteristics are hence not useful to predict future events once cardiovascular disease has developed. Indicators of previous symptomatic cardiovascular disease and the extent of atherosclerosis were more useful. This finding is similar to findings in the GUSTO-I sample, where e.g. smoking was associated with a better outcome after acute MI.

## Questions

### 23.1 Composite outcomes (Sect. 23.2.2 and Table 23.1)

Outcomes were combined in the presented analyses.

- (a) What does this imply about the effects of the predictors for each outcome?
- (b) How could this be tested? See Glynn and Rosner<sup>140</sup>

### 23.2 Missing values (Fig. 23.1)

- (a) Some might argue to exclude patients with many missing values. What would be a reasonable number as maximum of missing values per patient in this analysis?
- (b) We note that missing values occur together for some predictors. We could also choose to exclude patients with missing values (NA) in specific predictors. Which would you choose?

### 23.3 Effects of Lasso vs. stepwise selection (Table 23.6)

We select the same predictors with a Lasso procedure as with stepwise selection using AIC.

- (a) How is it possible to obtain the same selection with these very different methods?
- (b) The effect of age is similar with both methods, while the effect of BMI is very weak according to the Lasso. How is this possible? Consider also the validation in Table 23.7.

# Chapter 24

## Lessons from Case Studies

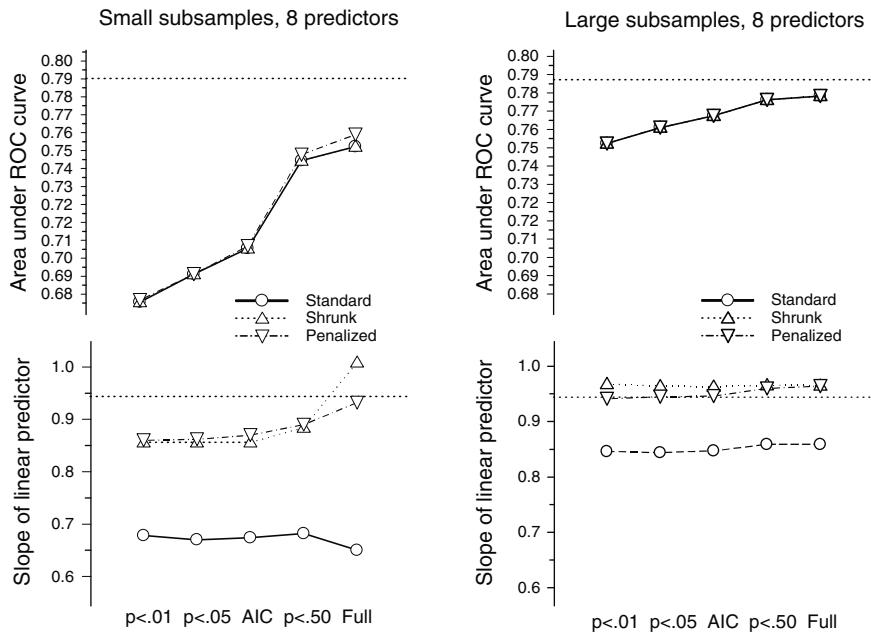
**Background** In this final chapter, we review some practical issues of development, validation, and updating of prediction models, based on the empirical experience from case studies as described in this book, and modelling experience in other medical prediction problems. We consider the essential elements to successful modelling: sufficient sample size; emphasis on validation; using, not ignoring, subject matter knowledge. Tentative recommendations are made, recognizing that specific circumstances may ask for specific approaches. We end this chapter with a description of the case studies used throughout this book, where data sets are available through the book's Web site.

### 24.1 Sample Size

Developing a valid prediction model from a relatively small data set has proven to be hard. Empirical examples discussed by Altman and Royston all show a poor performance at external validation.<sup>13</sup> Overfitting is a severe problem; it is common to ask too much from a small sample. Asking many questions is natural: Data collection in empirical studies is costly, and we are curious what patterns emerge from our precious data. Small data sets hence usually should serve to explore rather than to derive firm relationships. Yet, we need such firm relationships for accurate predictions. Also, we need strong predictors<sup>332</sup>, hence, when we have only a few relatively weak predictors it is tempting to search further for additional predictors.<sup>210</sup> An honest internal validation procedure should reveal the optimism that is associated with the full modelling procedure, including any searches for interesting patterns.<sup>401</sup> Harrell has observed that model uncertainty usually is more important for optimism in model performance than parameter uncertainty.<sup>174</sup> Hence, this step should never be forgotten. See Chatfield for a more theoretical but well-readable discussion.<sup>69</sup>

#### 24.1.1 Example: Sample Size and Number of Predictors

Simulations in GUSTO-I have highlighted the relevance of sample size to derive well-performing models.<sup>409,410</sup> When we study an 8 predictor model in samples with



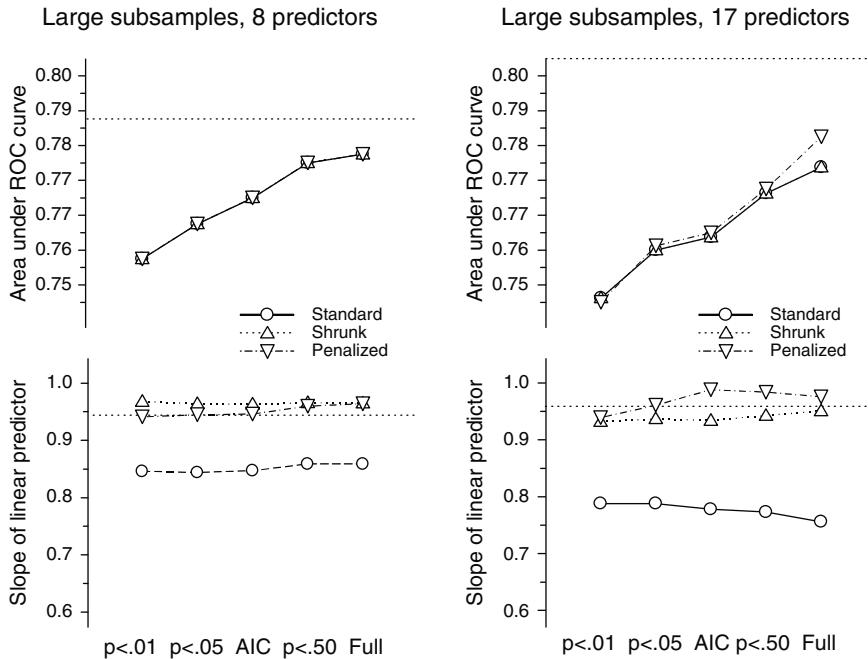
**Fig. 24.1** Influence of sample size on model performance.<sup>409</sup> Graphs show performance of an 8 predictor model in simulations from the GUSTO-I trial. Small and large subsamples included on average 23 and 62 events respectively (30-day mortality). Models were created with stepwise selection ( $p < 0.01$ ,  $p < 0.05$ , AIC ( $p < 0.157$ ),  $p < 0.50$ ) and with fitting a full model, with estimation by standard maximum likelihood, shrinkage of regression coefficients, and penalized maximum likelihood. Models were tested in a large independent test part (part B,  $n = 20,318$ ). Performance criteria included the area under the ROC curve (or  $c$  statistic, to indicate discrimination), and the slope of the linear predictor (to indicate calibration). We note that better models were identified with shrinkage or penalized estimation, with no (full) or limited selection ( $p < 0.50$ ). Substantially better performance was noted for models derived from the larger subsamples; using  $p < 0.05$  for model selection in the larger subsamples led to better performance than the full models from the smaller subsamples

on average 23 events, discriminative ability is clearly below the maximum possible with a very large sample size ( $n = 20,512$ , 1,423 deaths, Fig. 24.1). Moreover, stepwise selection is a poor-performing strategy, which is explained by the lack of power to select important predictors. Calibration of standard maximum likelihood estimates was poor, either with stepwise selection or with a full model. Shrinkage or penalized estimation largely resolved this miscalibration.

In larger samples (62 deaths on average), everything looks somewhat better: The performance was closer to the maximum, and stepwise methods were less detrimental than in smaller samples.

### 24.1.2 Number of Predictors

When we study more predictors, we would expect that we could obtain better performing models. Remarkably this was not the case in simulations in GUSTO-I.<sup>409</sup>



**Fig. 24.2** Influence of number of predictors on model performance (8 vs. 17 predictors). See legend of Fig. 24.1. We note that models estimated with standard maximum likelihood were worse with 17 rather than 8 predictors considered: Discrimination was 0.01 lower, and the calibration slope with further below 1 (indicating more need for shrinkage)

A full model with 17 predictors had at most similar performance to a full 8 predictor model, when we applied penalized maximum likelihood estimation. But backward stepwise selection with  $p < 0.05$  led to poorer models when 17 predictors were considered instead of 8 (Fig. 24.2). Hence, when we start with too many predictors, stepwise selection methods may not be able to save us. The balance between number of predictors and number of events should be for candidate predictors, not the number of selected predictors (Chap. 4).

### 24.1.3 Potential Solutions

A potential solution for small sample size is to perform collaborative studies (Table 24.1). For example, instead of analyzing a single centre retrospective cohort study, we may try to collect data from multiple centres, leading to a multicentre cohort study. Apart from simply increasing sample size other advantages occur. The multiple centres may be slightly different from each other, in local protocols for diagnostic work-up, treatment choices, definition of predictors, etc. Such heterogeneity is beneficial for the generalizability of the resulting model. If it were derived from a single centre, the results might be typical for that setting, rather than represent

**Table 24.1** Problem areas with prognostic modelling, and potential solutions with their benefits

Problem	Characterization	Potential solutions	Benefits
Sample size	Asking too much from the data relative to its size	Balance research question with available information	Less overfitting
	Particularistic, single centre samples used	Collaborative efforts	Statistical and epidemiological advantages (standard errors decrease with larger sample size; generalizability increases; cross-validation possible for external validation)
Validation	Internal validity is a minimum requirement;	Bootstrap validation;	Honest impression of model performance for similar patients
	External validity important as a second aim	Multi-centre/international studies for external validation	Impression of model performance in plausibly related settings
Subject	Use rather than ignore matter knowledge	Literature review; expert opinion	Model stability, hence less overfitting; better estimation possible

“current practice.” Also, cross-validation becomes possible, where we leave out one centre to test a model that was developed on other centres (Chaps. 17 and 19).

## 24.2 Validation

Internal and external validations deserve our full attention in prediction modelling. Describing patterns in a data set have no meaning if these patterns are invalid outside the specific data set. First, we need to check internal validity. The bootstrap is a very useful tool for this purpose, but we should be careful to apply it honestly, i.e. not forgetting some model specification steps.<sup>401</sup> Second, we are concerned about external validity; if a model is only applicable in strict settings, we are astray from serious science.<sup>222</sup>

Sample size is important both for development and validation samples. If sample size is insufficient at model development, overfitting will occur. If sample size is insufficient at model validation, we may falsely conclude that a model performs satisfactorily, while substantial invalidity may in fact exist.

### 24.2.1 Examples of Internal and External Validation

The practical experience with validation is mixed: Some models may generalize well if developed according to the principles outlined in Part II, but some models

require at least an adjustment for the average, case-mix adjusted incidence of the outcome. In GUSTO-I, we noted that the variability by subsample or region was largely attributable to chance, but this was in the context of a randomized trial, with a specific protocol (Chap. 22). A previously developed model (TIMI-II) required updating of the intercept.<sup>402,406</sup> In the testicular cancer example, we noted some differences between centres, but the sample sizes were not large enough to draw a firm conclusion on similarity of the intercept across settings (Chap. 19).<sup>467</sup> In the stroke example, substantial differences between centres were noted that were beyond chance (Chap. 21). Systematic differences can sometimes be attributed to specific circumstances; for example, we found systematically poorer than predicted outcome in patients from the CRASH trial, which included many patients with traumatic brain injury from developing and middle-income countries.

Konig et al. recently reported on the internal, temporal, and geographic validity of outcome prediction models in stroke.<sup>240</sup> They noted that internal validation was not enough, and that some form of external validation was necessary for a good impression of model performance in new patients. This was partly caused by problems to fully capture all modelling steps in the internal validation procedure, which hence resulted in still too optimistic estimates of model performance. A study in children with fever also suggested that external validation was necessary beyond internal validation.<sup>40</sup>

### 24.3 Subject Matter Knowledge

Throughout this book, using subject matter knowledge has been emphasized. Examples of valid models that were built from scratch are rare. Most successful models combine well-known predictors, and limit the use of the data set to some fine-tuning of the model specification. For example, we eliminate some main effects that do not contribute to outcome prediction. On the other hand we include some non-linear terms that are important to capture the relationship of a continuous predictor with the outcome. We may include some interactions, if these are very strong. The main role of the data set then is to quantify the predictor–outcome relationship, and provide an impression of the performance of the model. As discussed in Chap. 1, we aim to avoid the situation that we develop a model without some knowledge on which predictors to include, in what functional form, and unknown effects (see Table 1.1). A huge sample size would be required for this situation.

Model updating is a formal approach to use prior knowledge (Chaps. 20 and 21). We start with assuming that a prior model is valid for a new setting, and modify coefficients and add other predictors if indicated by the data under study. Such model updating is only possible if a reasonable prior multivariable model exists. We are back at standard model development if only univariate associations are known, or qualitative statements on the strength of a predictive effects.

Several disadvantages can be mentioned for modelling with subject knowledge. First, we may miss important new predictors. We should be prepared to take this

risk, since searching for new predictors has many risks of its own, including testimation bias and instability of the search. Second, we do not discover new knowledge. We only combine what is known already. This is however precisely the role of prediction models in medicine: They quantify what is already known. Knowledge discovery is a phase before we can start serious prediction modelling. Prediction models may have a role beyond systematic review of prognostic factors, as is starting to be promoted by the Cochrane collaboration. Systematic reviews may provide summaries of relative effects; prediction models provide absolute effects.

We may be interested in a prediction model that includes new predictors, such as a genetic marker or other type of biomarker. We first would need robust evidence on the univariate effect of the marker, and preferably also on its effect adjusted for other important predictors.<sup>285</sup> If this evidence is sufficient, we could study the performance of the marker when integrated in a prediction model. Of interest is the incremental value of the marker.<sup>225</sup> Several performance criteria can be used, such as increase in discriminative ability, re-classification, and decision-curve analysis (see Chap. 16).

## 24.4 Data Sets

We considered many examples throughout the text. For some case studies, empirical data are available through the book's Web site (Table 24.2). These case studies are discussed below in a simple format. First we list the abstract of the key publication of the study, if relevant. We then list the contents of the data sets. The data sets

**Table 24.2** Summary of case studies with data sets available at the book's Web site

Case study	Characterization	<i>N</i> patients (outcome); predictors
GUSTO-I	Prediction of 30-day mortality in acute myocardial infarction	Original: <i>n</i> =40,830 (2,851). West region <i>n</i> =2,188 (135); Sample4, <i>n</i> =785 (52); Sample5, <i>n</i> =429 (24); 17 predictors
SMART	Prediction of secondary cardiovascular events	<i>N</i> =3,873 (460); 26 predictors
Testicular cancer	Diagnosis of residual mass histology (benign vs. other, or in three categories)	Development, <i>n</i> =544 (245 benign); six predictors validation, <i>n</i> =273 (76 benign); five predictors
Abdominal aortic aneurysm	Prediction of peri-operative mortality after elective surgery	<i>N</i> =238 (18); seven predictors
Traumatic brain injury	Prediction of 6-month outcome	<i>N</i> =2,159; 503 deaths, 851 unfavorable outcome; 14 predictors

are made available for didactic purposes only. If publication by any means is pursued, investigators are required to contact the authors of the original publication and the author of this book.

#### **24.4.1 *GUSTO-I Prediction Models***

The key publication is by Lee *et al.* (*Circulation*, 1995, see Box 22.1).<sup>255</sup> Many other publications are available that use the GUSTO-I data, including a practical prediction tool by Califf *et al.*).<sup>63</sup> Small parts of the GUSTO-I data set are made available here: sample5 contains 429 patients, sample4 785 patients, and the West region 2,188 patients (Table 24.3). The patients partly overlap, which can be identified by matching on the 17 predictors and the outcome in the data set.

#### **24.4.2 *Modern Learning Methods in GUSTO-I***

Several simulation studies have been performed with the GUSTO-I data base. Ennis et al. compared a variety of modern learning methods, including logistic regression, Tree, GAM, and MARS methods (see Chap. 6).<sup>115</sup> For evaluation purposes, the data set was randomly divided into two parts: two-thirds of the data form the training set ( $n=27,220$ ), and the rest form the test set ( $n=13,610$ ). The training set was used for model development, with a smaller training set ( $n=18,147$ ) and a validation set ( $n=9,073$ ) if necessary, leaving the test set ( $n=13,610$ ) for final assessment of predictive performance. Performance measures included the log-likelihood and AUC of predictions in the test set.

#### **24.4.3 *Modelling Strategies in Small Data Sets from GUSTO-I***

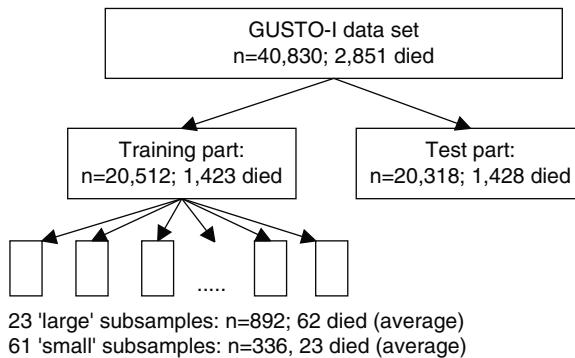
The GUSTO-I data set has also been instrumental to compare various aspects of predictive modelling strategies in small data sets.<sup>402,405,406,407,409,410,413</sup> The large size of GUSTO-I makes that subsamples can be created where models can be developed, which can subsequently be tested on an independent part of the data set. This approach has been followed to empirically test many aspects of logistic regression modelling. Especially we have focused on aspects of selection of predictors in a prognostic model and estimation of regression coefficients.<sup>407,409,410,413</sup> Detailed results were presented in many chapters.

The design of a key study for this book is shown in Fig. 24.3. The GUSTO-I data set consists of patients from 1,081 hospitals in 14 countries. At a higher level, eight regions could be defined within the United States, and from neighbouring countries outside the United States (another eight regions). As a first split, the 16 regions

**Table 24.3** Data description of various subsamples from the GUSTO-I trial as considered in this book

Name	Description (coding: no/yes is coded as 0/1)	GUSTO-I (2,851/40,830)	US (1,565/23,034)	West region (1352/188)	Sample4 (52/785)	Sample5 (24/429)
AGE	Age in years (range: 19–110)	61	61	60	62	60
AGE	Age >65 years (0/1)	40%	39%	38%	42%	37%
SEX	Gender (male=0, female=1)	25%	27%	25%	26%	27%
KILLIP	Killip class (1–4): A measure for left ventricular function	85/13/1/1%	87/11/1/1%	89/10/1/1%	78/19/3/0%	86/13/1/1%
SHO	Shock: Killip class 3/4 vs. 1/2 (0/1)	2%	2%	1%	3%	2%
DIA	Diabetes (0/1)	15%	17%	14%	11%	13%
HYP	Hypotension: Systolic BP<100 (0/1)	8%	10%	10%	5%	11%
HRT	Heart rate: Pulse>80 (“tachycardia,” 0/1)	33%	34%	33%	27%	35%
ANT	Anterior infarct location (0/1)	39%	37%	37%	36%	36%
PMI	Previous myocardial infarction (0/1)	16%	17%	17%	18%	15%
HIG	High risk: ANT or PMI (0/1)	49%	48%	49%	46%	47%
HEI	Height in cm (range: 140–212)	171	172	172	169	172
WEI	Weight in kg (range: 36–213)	79	82	83	75	83
SMK	Smoking (1=never; 2=exsmoker; 3=current smoker)	43/27/30%	43/28/29%	41/31/28%	33/39/28%	43/30/28%
HTN	Hypertension history (0/1)	38%	43%	40%	38%	39%
LIP	Lipids: Hypercholesterolaemia (0/1)	34%	38%	40%	39%	35%
PAN	Previous angina pectoris (0/1)	37%	35%	34%	38%	31%
FAM	Family history of MI (0/1)	42%	49%	48%	41%	48%
STE	ST elevation on ECG: Number of leads (range: 0–11)	4.1	4.0	4.0	4.3	4.0
ST4	ST elevation on ECG: >4 leads (0/1)	38%	37%	36%	41%	36%
TR	Time to relief of chest pain > 1 h (0/1)	65%	66%	61%	50%	61%

The full GUSTO-I trial contained 40,830 patients of whom 2,851 died



**Fig. 24.3** Simulation design for GUSTO-I, with creation of 23 large and 61 small subsamples

were used to create a training part four US regions, four other) and a test part (four US regions, four other). So, the split was not at random but based on geographical balance. Care was taken that the mortality was 7% in both parts. Each part contained eight regions, and within regions subsamples were created, labelled “large” and “small.” Large subsamples contained at least 50 patients who died, and small subsamples at least 20. The grouping was based on merging patients from nearby hospitals. This process mimicked the real-life situation that a prognostic model is developed in cooperation with a number of centres. Note that the large and small subsamples contained partly the same patients, and hence were not independent.

In the training part, 23 large subsamples were created, which contained on average 892 patients, of whom 62 had died by 30 days. A total of 61 small subsamples was created, with 23 deaths among 336 patients on average. All models were tested in the independent test part of 20,318 patients. As a gold standard we could use models based on the full training part. Because of the still very large sample size ( $n=20,512$ , 1,423 deaths), optimism was not a concern in this training part. Sample4 and sample5 were chosen for illustration in this book since the results of modelling in these samples was representative for the average pattern over the small and large subsamples.

#### 24.4.4 SMART Case Study

The SMART (second manifestations of arterial disease) study is discussed in detail in Chap. 23. The seven modeling steps from part II were followed, and R computer code is available to perform the described analyses (Table 24.4).

#### 24.4.5 Testicular Cancer Case Study

The key publication for clinicians is a paper in the *Journal of Clinical Oncology* in 1995, with a validation study in 1998.<sup>112,417</sup> The methodological aspects are discussed in a paper in *Statistics in Medicine* in 2001 (see Box 24.1).<sup>425</sup> See also Table 24.5.

**Table 24.4** SMART study data set (*n*=3,873)

Name	Description (coding: no/yes is coded as 0/1)	Development (460/3,873)
Tevent	Time to cardiovascular event (days)	1370
Event	Cardiovascular event (clinical, 0/1)	460
Sex	1=male, 2=female sex	25%
Age	Age (years)	60
Diabetes	Ever diabetes (0/1)	22%
Cerebral	Ever cerebrovascular disease (0/1)	30%
Cardiac	Ever cardiovascular disease (0/1)	56%
AAA	Ever abdominal aortic aneurysm (0/1)	11%
Periph	Ever peripheral vascular disease (0/1)	24%
Stenosis	Carotid stenosis >= 50% by duplex (0/1)	19%
Systbp	Systolic blood pressure (automatic, in mm Hg)	141
Diastbp	Diastolic blood pressure (automatic, in mm Hg)	80
Systh	Systolic blood pressure (by hand, in mm Hg)	142
Diasth	Diastolic blood pressure (by hand, in mm Hg)	82
Length	Length (m)	1.74
Weight	Weight (kg)	81
BMI	Body mass index (kg/m <sup>2</sup> )	26.7
Chol	Cholesterol level (mmol/L)	5.2
HDL	High-density lipoprotein cholesterol (mmol/L)	1.2
LDL	Low-density lipoprotein cholesterol (mmol/L)	3.1
Trig	Triglycerides level (mmol/L)	1.9
Homoc	Homocysteine level (μmol/L)	13.8
Glut	Glutamine (μmol/L)	6.3
Creat	Creatinine clearance (mL/min)	98
IMT	Intima media thickness (mm)	0.93
Albumin	Albumin in urine: 1=no; 2=low; 3=high	79%/18%/3%
Smoking	Smoking status: 1=no; 2=former; 3=current	18%/70%/12%
Packyrs	packyears smoked	23
Alcohol	Alcohol consumption: 1=no; 2=former; 3=current	20%/11%/70%

The primary outcome was a cardiovascular event, which occurred in 460 patients during follow-up (5-year cumulative incidence, 14%)

**Table 24.5** Description of testicular cancer development (*n*=544) and validation set (*n*=273)

Name	Description (coding: no/yes is coded as 0/1)	Development (245/544 (45%))	Validation (76/273 (28%))
patkey	Patient ID	–	–
hosp	Institution ID	–	–
orchyr	Year of orchidectomy (surgical removal of primary tumor)	1985	1993
histr3	Histology at resection: 1=necrosis; 2=teratoma; 3=viable cancer	45%/42%/13%	28%/58%/13%
ter	primary tumor teratoma-negative? (0–1)	46%	38%
preafp	Prechemotherapy AFP normal? (0–1)	34%	25%
prehcg	Prechemotherapy HCG normal? (0–1)	38%	27%

(continued)

**Table 24.5** (continued)

Name	Description (coding: no/yes is coded as 0/1)	Development (245/544 (45%))	Validation (76/273 (28%))
lnldhst	Ln of standardized prechemotherapy LDH (LDH/upper limit of local normal value)	0.46 (LDHst 2.0)	NA
sqpost	Square root of post-chemotherapy mass size (original mass size in mm)	5.1 (33 mm)	7.8 (70 mm)
reduc10	Reduction in mass size per 10%: (pre-post)/pre*10	4.5 (=45%)	1.4 (=14%)
nec	Necrosis at resection (0–1)	45%	28%
matter	Mature teratoma vs. cancer, if not necrosis (0–1)	77%	82%
dev	Part of data set: 1=development (n=544); 0=validation (n=273)	1	0

The primary outcome was a benign histology at post-chemotherapy resection, which occurred in 45% and 28% respectively

**Box 24.1** Abstract of the methodological paper on prediction of residual mass histology in testicular cancer patients<sup>425</sup>

### **Residual mass histology in testicular cancer: development and validation of a clinical prediction rule**

Ewout W. Steyerberg; Yvonne Vergouwe; H. Jan Keizer and J. Dik F. Habbema for the ReHiT study group

After chemotherapy for metastatic non-seminomatous testicular cancer, surgical resection is a generally accepted treatment to remove remnants of the initial metastases, since residual tumour may still be present (mature teratoma or viable cancer cells). In this paper, we review the development and external validation of a logistic regression model to predict the absence of residual tumour.

Three sources of information were used. A quantitative review identified six relevant predictors from 19 published studies (996 resections).<sup>420</sup> Second, a development data set included individual data of 544 patients from six centres.<sup>417</sup> This data set was used to assess the predictive relationships of five continuous predictors, which resulted in dichotomization for two, and a log, square root, and linear transformation for three other predictors. The multiple logistic regression coefficients were reduced with a shrinkage factor (0.95) to improve calibration, based on a bootstrapping procedure. Third, a validation data set included 172 more recently treated patients.<sup>412</sup> The model showed adequate calibration and good discrimination in the development and in the validation sample (areas under the ROC curve 0.83 and 0.82).

This study illustrates that a careful modeling strategy may result in an adequate predictive model. Further study of model validity may stimulate application in clinical practice.

PMID: 11782038

#### **24.4.6 Abdominal Aortic Aneurysm Case Study**

The Leiden cohort contains patients undergoing elective surgery for an abdominal aortic aneurysm (Table 24.6). Results are described in detail in a PhD thesis by Dr. Alexander de Mol van Otterloo (currently working as a surgeon in The Hague). The publication that presents the prediction rule based on the combination of the Leiden data and literature data is in *Archives of Internal Medicine* in 1995 (see Box 22.2).<sup>421</sup>

**Table 24.6** Aortic aneurysm data set ( $n = 238$ )

Name	Description (coding: no/yes is coded as 0/1)	Development (18/238 (8%))
Sex	Female (0/1)	9%
Age10	Age in decades	6.6 (66 years)
MI	Infarction on ECG (0/1)	24%
CHF	Congestive heart failure (0/1)	34%
Ischaemia	Ischaemia on ECG (0/1)	35%
Lung	Lung comorbidity (0/1)	19%
Renal	Renal comorbidity (0/1)	6%
Status	Peri-operative mortality (0/1)	8%

The primary outcome was surgical mortality, which occurred in only 18 patients (7.6%).

**Box 24.2** Abstract of the paper on prediction of perioperative mortality in AAA<sup>421</sup>

**Perioperative mortality of elective abdominal aortic aneurysm surgery. A clinical prediction rule based on literature and individual patient data**

Steyerberg EW, Kievit J, de Mol Van Otterloo JC, van Bockel JH, Eijkemans MJ, Habbema JD.

**BACKGROUND:** Abdominal aortic aneurysm surgery is a major vascular procedure with a considerable risk of (mainly cardiac) mortality. **OBJECTIVE:** To estimate elective perioperative mortality, we developed a clinical prediction rule based on several well-established risk factors: age, gender, a history of myocardial infarction, congestive heart failure, ischemia on the electrocardiogram, pulmonary impairment, and renal impairment.

**METHODS:** Two sources of data were used: (1) individual patient data from 246 patients operated on at the University Hospital Leiden (the Netherlands) and (2) studies published in the literature between 1980 and 1994. The Leiden data were analyzed with univariate and multivariable logistic regression. Literature data were pooled with meta-analysis techniques.

(continued)

**Box 24.2** (continued)

The clinical prediction rule was based on the pooled odds ratios from the literature, which were adapted by the regression results of the Leiden data.

**RESULTS:** The strongest adverse risk factors in the literature were congestive heart failure and cardiac ischemia on the electrocardiogram, followed by renal impairment, history of myocardial infarction, pulmonary impairment, and female gender. The literature data further showed that a 10-year increase in age more than doubled surgical risk. In the Leiden data, most multivariable effects were smaller than the univariate effects, which is explained by the positive correlation between the risk factors. In the clinical prediction rule, cardiac, renal, and pulmonary comorbidity are the most important risk factors, while age per se has a moderate effect on mortality.

**CONCLUSIONS:** A readily applicable clinical prediction rule can be based on the combination of literature data and individual patient data. The risk estimates may be useful for clinical decision making in individual patients.

PMID: 7575054

#### **24.4.7 Traumatic Brain Injury Data Set**

Prognostic studies based on patients included in the Tirilazad trials are described in detail in a PhD thesis by Chantal Hukkelhoven. The publication that presents a prognostic model is in a neurosurgical journal (*J Neurotrauma* 2005) (see Box 24.3).<sup>203</sup> More extensive data became later available through the IMPACT project as described in Chap. 10, with several publications led by Prof. Dr. Andrew Maas. See Table 24.7.

### **24.5 Concluding Remarks**

The described data sets are made available to promote practical experience with the described techniques in this book. Many other medical data sets are publicly available nowadays, which can be used to train researchers in prediction modelling, and readers are encouraged to examine these. The author welcomes any comments and suggestions for improvement of the text of this book, the questions at the end of each chapter, the practical exercises at the Web, and usefulness of data sets.

**Table 24.7** Traumatic brain injury data set ( $n=2,159$ ). Patients are from the International and US Tirilazad trials. The primary outcome was 6 months Glasgow Outcome Scale (range 1 for dead to 5 for good recovery)

Name	Description (coding: no/yes is coded as 0/1)	Development $n=2,159$
trial	Tirilazad international ( $n=1,118$ ) / US ( $n=1,041$ )	–
d.gos	GOS at 6 months: 1 = dead 2 = vegetative 3 = severe disability 4 = moderate disability 5 = good recovery*	23% 4% 12% 16% 44%
d.mort	Mortality at 6 months (0/1)	23%
d.unfav	Unfavorable outcome at 6 months (0/1)	39%
age	Age (in years, median [interquartile range])	29 [21–42]
d.motor	Admission motor score (1–6, median)	4
d.pupil	Pupillary reactivity (1=both reactive/2=one reactive/ 3=no reactive pupils)	70%/14%/16%
pupil.i	Single imputed pupillary reactivity (1/2/3)	70%/14%/16%
hypoxia	Hypoxia before/at admission (0/1)	22%
hypotens	Hypotension before/at admission	19%
ctclass	Marshall CT classification (1–6, median)	2
tsah	tSAH at CT (0/1)	46%
edh	EDH at CT (0/1)	13%
cisterns	Compressed cisterns at CT (0=no/1=slightly compressed/ 2=fully compressed)	57%/26%/10%
shift	Midline shift > 5 mm at CT (0/1)	18%
glucose	Glucose at admission (mmol/l, median [interquartile range])	8.2 [6.7–10.4]
glucoset	Truncated glucose values (median [interquartile range])	8.2 [6.7–10.4]
ph	pH (median [interquartile range])	7.4 [7.3–7.5]
sodium	Sodium (mmol/l, median [interquartile range])	140 [137–142]
sodiumt	Truncated sodium (median [interquartile range])	140 [137–142]
hb	Hb (g/dL, median [interquartile range])	12.8 [10.9–14.3]
hbt	Truncated hb (median [interquartile range])	12.8 [10.9–14.3]

\*d. variables denote “derived”.

**Box 24.3** Abstract of the paper on prediction of outcome in traumatic brain injury<sup>203</sup>

**Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics**

Hukkelhoven CW, Steyerberg EW, Habbema JD, Farace E, Marmarou A, Murray GD, Marshall LF, Maas AI.

The early prediction of outcome after traumatic brain injury (TBI) is important for several purposes, but no prognostic models have yet been developed with proven generalizability across different settings. The objective of this

(continued)

**Box 24.3** (continued)

study was to develop and validate prognostic models that use information available at admission to estimate 6-month outcome after severe or moderate TBI. To this end, this study evaluated mortality and unfavorable outcome, that is, death, and vegetative or severe disability on the Glasgow Outcome Scale (GOS), at 6 months post-injury.

Prospectively collected data on 2269 patients from two multi-centre clinical trials were used to develop prognostic models for each outcome with logistic regression analysis. We included seven predictive characteristics: age, motor score, pupillary reactivity, hypoxia, hypotension, computed tomography classification, and traumatic subarachnoid hemorrhage. The models were validated internally with bootstrapping techniques. External validity was determined in prospectively collected data from two relatively unselected surveys in Europe ( $n = 796$ ) and in North America ( $n = 746$ ). We evaluated the discriminative ability, that is, the ability to distinguish patients with different outcomes, with the area under the receiver operating characteristic curve (AUC). Further, we determined calibration, that is, agreement between predicted and observed outcome, with the Hosmer-Lemeshow goodness-of-fit test.

The models discriminated well in the development population (AUC 0.78–0.80). External validity was even better (AUC 0.83–0.89). Calibration was less satisfactory, with poor external validity in the North American survey ( $p < 0.001$ ). Especially, observed risks were higher than predicted for poor prognosis patients. A score chart was derived from the regression models to facilitate clinical application.

Relatively simple prognostic models using baseline characteristics can accurately predict 6-month outcome in patients with severe or moderate TBI. The high discriminative ability indicates the potential of this model for classifying patients according to prognostic risk.

PMID: 16238481

## Questions

### 24.1 Number of predictors and sample size (Fig. 24.2)

In Fig. 24.2, we note that the discriminative ability (area under ROC curve, or  $c$  statistic) does not increase by considering 17 rather than 8 predictors with standard maximum likelihood or shrunk estimation.

- (a) How is it possible that considering more predictors does not increase the discriminative ability?
- (b) What is the slope of the linear predictor (or calibration slope) with standard estimation, with 17 or 8 predictors?
- (c) And what is the slope with shrinkage or penalized estimation, with 17 or 8 predictors?
- (d) What would you do if 17 candidate predictors were available in a data set with ~50 events, and the aim was to make a model for predictions in individual patients?

## References

1. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results of the thrombolysis in myocardial infarction (TIMI) phase II trial. The TIMI Study Group. *N Engl J Med* 1989;320(10):618-27.
2. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. The GUSTO investigators. *N Engl J Med* 1993;329(10):673-82.
3. A predictive model for aggressive non-Hodgkin's lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. *N Engl J Med* 1993;329(14):987-94.
4. A comparison of reteplase with alteplase for acute myocardial infarction. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO III) Investigators. *N Engl J Med* 1997;337(16):1118-23.
5. International germ cell consensus classification: a prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. *J Clin Oncol* 1997;15:594-603.
6. Surgery for colorectal cancer in elderly patients: a systematic review. Colorectal Cancer Collaborative Group. *Lancet* 2000;356(9234):968-74.
7. Addala S, Grines CL, Dixon SR, Stone GW, Boura JA, Ochoa AB, et al. Predicting mortality in patients with ST-elevation myocardial infarction treated with primary percutaneous coronary intervention (PAMI risk score). *Am J Cardiol* 2004;93(5):629-32.
8. Altman DG. Practical statistics for medical research. 1st ed. London; New York: Chapman and Hall, 1991.
9. Altman DG. ROC curves and confidence intervals: getting them right. *Heart* 2000;83(2):236.
10. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Stat Med* 1989;8(7):771-83.
11. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *Bmj* 1995;311(7003):485.
12. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86(11):829-35.
13. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453-73.
14. Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* 2002;21(24):3803-22.
15. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type I error rate. *J Stat Comput Simul* 2001;69:89-108.
16. Ammar KA, Kors JA, Yawn BP, Rodeheffer RJ. Defining unrecognized myocardial infarction: a call for standardized electrocardiographic diagnostic criteria. *Am Heart J* 2004;148(2):277-84.
17. Assmann G, Schulte H. The Prospective Cardiovascular Munster (PROCAM) study: prevalence of hyperlipidemia in persons with hypertension and/or diabetes mellitus and the relationship to coronary heart disease. *Am Heart J* 1988; 116:1713-24.
18. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355(9209):1064-9.

19. Austin PC. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Stat Med* 2008;27:3286-3300.
20. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 2004;57(11):1138-46.
21. Austin PC, Tu JV. Bootstrap methods for developing predictive models in cardiovascular research. *Am Stat* 2004;58:131-137.
22. Aylin P, Alves B, Best N, Cook A, Elliott P, Evans SJ, et al. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: was Bristol an outlier? *Lancet* 2001;358(9277):181-7.
23. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004;66(3):411-21.
24. Bach PB, Guadagnoli E, Schrag D, Schussler N, Warren JL. Patient demographic and socioeconomic characteristics in the SEER-Medicare database applications and limitations. *Med Care* 2002;40(8 Suppl):IV-19-25.
25. Balmana J, Stockwell DH, Steyerberg EW, Stoffel EM, Deffenbaugh AM, Reid JE, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. *Jama* 2006;296(12):1469-78.
26. Bancroft TA, Han CP. Inference based on conditional specification: a note and a bibliography. *Int Statist Rev* 1977;45:117-127.
27. Bao Y. Predicting the use of outpatient mental health services: do modeling approaches make a difference? *Inquiry* 2002;39(2):168-83.
28. Barnetson RA, Tenesa A, Farrington SM, Nicholl ID, Cetnarskyj R, Porteous ME, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N Engl J Med* 2006;354(26):2751-63.
29. Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 1984;71(1):11-8.
30. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39(1):207-15.
31. Begg CB, Satagopan JM, Berwick M. A new strategy for evaluating the impact of epidemiologic risk factors for cancer with application to melanoma. *JASA* 1998;93:415-26.
32. Bellacosa A, Genuardi M, Anti M, Viel A, Ponz de Leon M. Hereditary nonpolyposis colorectal cancer: review of clinical, molecular genetics, and counseling aspects. *Am J Med Genet* 1996;62(4):353-64.
33. Benichou J, Gail MH, Mulvihill JJ. Graphs to estimate an individualized risk of breast cancer. *J Clin Oncol* 1996;14(1):103-10.
34. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. *Ann Epidemiol* 2006;16(7):540-4.
35. Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KG. Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol* 2008;61(2):125-34.
36. Birim O, Kappetein AP, Waleboer M, Puvimanasinghe JP, Eijkemans MJ, Steyerberg EW, et al. Long-term survival after non-small cell lung cancer surgery: development and validation of a prognostic model with a preoperative and postoperative mode. *J Thorac Cardiovasc Surg* 2006;132(3):491-8.
37. Birkmeyer JD, Marrin CA, O'Connor GT. Should patients with Bjork-Shiley valves undergo prophylactic replacement? *Lancet* 1992;340(8818):520-3.
38. Blattenberger G, Lad F. Separating the Brier score into calibration and refinement components: a graphical exposition. *Am Stat* 1985;39(1):26-32.
39. Bleeker S, Derkxen-Lubsen G, Grobbee D, Donders A, Moons K, Moll H. Validating and updating a prediction rule for serious bacterial infection in patients with fever without source. *Acta Paediatr* 2007;96(1):100-4.
40. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derkxen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003;56(9):826-32.

41. Bleeker SE, Moons KG, Derkxen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. *Acta Paediatr* 2001;90(11):1226-32.
42. Blot WJ, Omar RZ, Kallewaard M, Morton LS, Fryzek JP, Ibrahim MA, et al. Risks of fracture of Bjork-Shiley 60 degree convexo-concave prosthetic heart valves: long-term cohort follow up in the UK, Netherlands and USA. *J Heart Valve Dis* 2001;10(2):202-9.
43. Boersma E, Pieper KS, Steyerberg EW, Wilcox RG, Chang WC, Lee KL, et al. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. The PURSUIT Investigators. *Circulation* 2000;101(22):2557-67.
44. Boersma E, Steyerberg EW, van der Vlugt MJ, Simoons ML. Reperfusion therapy for acute myocardial infarction. Which strategy for which patient? *Drugs* 1998;56(1):31-48.
45. Boissel JP, Gueyffier F, Cucherat M, Bricca G. Pharmacogenetics and responders to a therapy: theoretical background and practical problems. *Clin Chem Lab Med* 2003;41(4):564-72.
46. Bos JM, Rietveld E, Moll HA, Steyerberg EW, Luytjes W, Wilschut JC, et al. The use of health economics to guide drug development decisions: determining optimal values for an RSV-vaccine in a model-based scenario-analytic approach. *Vaccine* 2007;25(39-40):6922-9.
47. Boscarino JA, Adams RE. Public perceptions of quality care and provider profiling in New York: implications for improving quality care and public health. *J Public Health Manag Pract* 2004;10(3):241-50.
48. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138(1):40-4.
49. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138(1):W1-12.
50. Bower M, Gazzard B, Mandalia S, Newsom-Davis T, Thirlwell C, Dhillon T, et al. A prognostic index for systemic AIDS-related non-Hodgkin lymphoma treated in the era of highly active antiretroviral therapy. *Ann Intern Med* 2005;143(4):265-73.
51. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, editors. *Robustness in statistics: proceedings of a workshop*. New York: Academic Press, 1979:xvi, 296p.
52. Box GEP, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962;4:531-550.
53. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. *Trauma Score and the Injury Severity Score*. *J Trauma* 1987;27(4):370-8.
54. Brady AR, Fowkes FG, Greenhalgh RM, Powell JT, Ruckley CV, Thompson SG. Risk factors for postoperative death following elective surgical repair of abdominal aortic aneurysm: results from the UK Small Aneurysm Trial. On behalf of the UK Small Aneurysm Trial participants. *Br J Surg* 2000;87(6):742-9.
55. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002;137(8):693-5.
56. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules "in the real world": how a widely disseminated rule is used in everyday practice. *Acad Emerg Med* 2005;12(10):948-56.
57. Breiman L. Classification and regression trees. Belmont, CA: Wadsworth International Group, 1984.
58. Breiman L. Better subset regression using the nonnegative Garrote. *Technometrics* 1995;37:373-384.
59. Breiman L. Bagging predictors. *Machine Learning* 1996;24:123-140.
60. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;163(12):1149-56.
61. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004;91(1):4-8.

62. Butcher I, McHugh GS, Lu J, Steyerberg EW, Hernandez AV, Mushkudiani N, *et al.* Prognostic value of cause of injury in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):281-6.
63. Calif RM, Woodlief LH, Harrell FE, Jr., Lee KL, White HD, Guerci A, *et al.* Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. *Am Heart J* 1997;133(6):630-9.
64. Campbell MJ. Statistics at square two: understanding modern statistical applications in medicine. 2nd ed. Malden, MA: Blackwell, 2006.
65. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Control Clin Trials* 2003;24(6):682-701.
66. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158(3):280-7.
67. Chan SF, Deeks JJ, Macaskill P, Irwig L. Three methods to construct predictive models using logistic regression and likelihood ratios to facilitate adjustment for pretest probability give similar results. *J Clin Epidemiol* 2008;61(1):52-63.
68. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40(5):373-83.
69. Chatfield C. Model uncertainty, data mining and statistical inference. *J R Stat Soc Ser A* 1995;158(3):419-66.
70. Chen CH, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Stat Med* 1985;4(1):39-46.
71. Chen S, Wang W, Lee S, Nafa K, Lee J, Romans K, *et al.* Prediction of germline mutations and cancer risk in the Lynch syndrome. *Jama* 2006;296(12):1479-87.
72. Chun FK, Karakiewicz PI, Briganti A, Gallina A, Kattan MW, Montorsi F, *et al.* Prostate cancer nomograms: an update. *Eur Urol* 2006;50(5):914-26; discussion 926.
73. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003;56(1):28-37.
74. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer* 1994;73(3):643-51.
75. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *JASA* 1988;83:596-610.
76. Collins JA, Burrows EA, Wilan AR. The prognosis for live birth among untreated infertile couples. *Fertil Steril* 1995;64(1):22-8.
77. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001;6(4):330-51.
78. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115(7):928-35.
79. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med* 2006;145(1):21-9.
80. Cooper GS, Virnig B, Klabunde CN, Schussler N, Freeman J, Warren JL. Use of SEER-Medicare data for measuring cancer surgery. *Med Care* 2002;40(8 Suppl):IV-43-8.
81. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc Ser B* 1983;45(3):311-354.
82. Copas JB, Long T. Estimating the residual variance in orthogonal regression with variable selection. *Statistician* 1991;40:51-59.
83. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, *et al.* Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* 1999;91(18):1541-8.
84. Coulter A. Partnerships with patients: the pros and cons of shared clinical decision-making. *J Health Serv Res Policy* 1997;2(2):112-21.
85. Cox D. Regression models and life-tables (with discussion). *J R Stat Soc Ser B* 1972;34:187-220.

86. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958;45:562-5.
87. Croft RP, Nicholls PG, Steyerberg EW, Richardus JH, Cairns W, Smith S. A clinical prediction rule for nerve-function impairment in leprosy patients. *Lancet* 2000;355(9215):1603-6.
88. Cucciare MA, O'Donohue W. Predicting future healthcare costs: how well does risk-adjustment work? *J Health Organ Manag* 2006;20(2-3):150-62.
89. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;17(19):2265-81.
90. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *Jama* 2001;286(2):180-7.
91. De Coster C, Quan H, Finlayson A, Gao M, Halfon P, Humphries KH, et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.
92. De Dombal FT. Diagnosis of acute abdominal pain. Edinburgh: Churchill Livingstone, 1980.
93. De Ridder MA, Stijnen T, Hokken-Koelega AC. Validation and calibration of the Kabi Pharmacia International Growth Study prediction model for children with idiopathic growth hormone deficiency. *J Clin Endocrinol Metab* 2003;88(3):1223-7.
94. Decarli A, Calza S, Masala G, Specchia C, Palli D, Gail MH. Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation into Cancer and Nutrition cohort. *J Natl Cancer Inst* 2006;98(23):1686-93.
95. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Stat Med* 1997;16(23):2645-64.
96. Derkzen S, Keselman H. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* 1992;45:265-82.
97. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials* 2007;28(2):105-14.
98. Dijk JM, van der Graaf Y, Bots ML, Grobbee DE, Algra A. Carotid intima-media thickness and the risk of new vascular events in patients with manifest atherosclerotic disease: the SMART study. *Eur Heart J* 2006;27(16):1971-8.
99. Donabedian A. The definition of quality and approaches to its assessment. Ann Arbor, MI: Health Administration Press, 1980.
100. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59(10):1087-91.
101. Draper D. Assessment and propagation of model uncertainty. *J R Stat Soc Ser B* 1995;57:45-97.
102. Draper NR, Smith H. Applied regression analysis. 3rd ed. New York: Wiley, 1998.
103. Dresser GK, Bailey DG. A basic conceptual and practical overview of interactions with highly prescribed drugs. *Can J Clin Pharmacol* 2002;9(4):191-8.
104. Dubois C, Pierard LA, Albert A, Smeets JP, Demoulin JC, Boland J, et al. Short-term risk stratification at admission based on simple clinical data in acute myocardial infarction. *Am J Cardiol* 1988;61(4):216-9.
105. Earle CC, Nattinger AB, Potosky AL, Lang K, Mallick R, Berger M, et al. Identifying cancer relapse using SEER-Medicare data. *Med Care* 2002;40(8 Suppl):IV-75-81.
106. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat* 2004;32:407-99.
107. Efron B, Morris C. Stein's paradox in statistics. *Scientific American* 1977;236(5):119-127.
108. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall, 1993.
109. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *JASA* 1997;92:548-60.

110. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Bmj* 1997;315(7109):629-34.
111. Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Loosman CW, Habbema JD. The prediction of the chance to conceive in subfertile couples. *Fertil Steril* 1994;61(1):44-52.
112. Elmore JG, Fletcher SW. The risk of cancer risk prediction: "What is my risk of getting breast cancer"? *J Natl Cancer Inst* 2006;98(23):1673-5.
113. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J* 2003;24(21):1903-11.
114. Enders CK. A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosom Med* 2006;68(3):427-36.
115. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. *Stat Med* 1998;17(21):2501-8.
116. Euhus DM, Smith KC, Robinson L, Stucky A, Olopade OI, Cummings S, et al. Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *J Natl Cancer Inst* 2002;94(11):844-51.
117. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol* 2006;59(8):798-801.
118. Fanning J, Gangestad A, Andrews SJ. National Cancer Data Base/Surveillance Epidemiology and End Results: potential insensitive-measure bias. *Gynecol Oncol* 2000;77(3):450-3.
119. Faraway JJ. On the cost of data analysis. *J Comp Graph Stat* 1992;1(3):213-29.
120. Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. *Am J Manag Care* 2006;12(2):110-9.
121. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF, Jr., et al. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail* 2004;10(6):460-6.
122. Figueiras A, Domenech-Massons JM, Cadarso C. Regression models: calculating the confidence interval of effects in the presence of interactions. *Stat Med* 1998;17(18):2099-105.
123. Fine AM, Nigrovic LE, Reis BY, Cook EF, Mandl KD. Linking surveillance to action: incorporation of realtime regional data into a medical decision rule. *J Am Med Inform Assoc* 2007;14(2):206-11.
124. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *JASA* 1999;94:496-509.
125. Finlayson EV, Birkmeyer JD. Operative mortality with elective surgery in older adults. *Eff Clin Pract* 2001;4(4):172-7.
126. Fortin M, Hudon C, Dubois MF, Almirall J, Lapointe L, Soubhi H. Comparative assessment of three different indices of multimorbidity for studies on health-related quality of life. *Health Qual Life Outcomes* 2005;3:74.
127. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, et al. Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. *J Clin Oncol* 2002;20(6):1480-90.
128. Friedenson B. Assessing and managing breast cancer risk: clinical tools for advising patients. *MedGenMed* 2004;6(1):8.
129. Friedman JH. Multivariate adaptive regression splines. *Ann Stat* 1991;19:1-141.
130. Gail MH. The estimation and use of absolute risk for weighing the risks and benefits of selective estrogen receptor modulators for preventing breast cancer. *Ann N Y Acad Sci* 2001;949:286-91.
131. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81(24):1879-86.
132. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005;6(2):227-39.
133. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted variables. *Biometrika* 1984;71:431-44.

134. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *JASA* 1990;85:398-409.
135. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *Bmj* 2003;327(7417):741-4.
136. Glance LG, Dick A, Osler TM, Li Y, Mukamel DB. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State cardiac surgery report card. *Med Care* 2006;44(4):311-9.
137. Glantz SA, Slinker BK. Primer of applied regression & analysis of variance. 2nd ed. New York: McGraw-Hill, Medical Pub. Division, 2001.
138. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *Bmj* 1995;311(7016):1356-9.
139. Gloeckler Ries LA, Reichman ME, Lewis DR, Hankey BF, Edwards BK. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist* 2003;8(6):541-52.
140. Glynn RJ, Rosner B. Methods to evaluate risks for composite end points and their individual components. *J Clin Epidemiol* 2004;57(2):113-22.
141. Goeman JJ, le Cessie S. A goodness-of-fit test for multinomial logistic regression. *Biometrics* 2006;62(4):980-5.
142. Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *N Engl J Med* 1982;307(10):588-96.
143. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc Ser A* 1996;159:385-443.
144. Goodacre S, Sutton AJ, Sampson FC. Meta-analysis: the value of clinical assessment in the diagnosis of deep venous thrombosis. *Ann Intern Med* 2005;143(2):129-39.
145. Gotz HM, van Bergen JE, Veldhuijzen IK, Broer J, Hoebe CJ, Steyerberg EW, et al. A prediction rule for selective screening of *Chlamydia trachomatis* infection. *Sex Transm Infect* 2005;81(1):24-30.
146. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18(17-18):2529-45.
147. Graham ID, Stiell IG, Laupacis A, O'Connor AM, Wells GA. Emergency physicians' attitudes toward and use of clinical decision rules for radiography. *Acad Emerg Med* 1998;5(2):134-40.
148. Grambsch PM, O'Brien PC. The effects of transformations and preliminary tests for non-linearity in regression. *Stat Med* 1991;10(5):697-709.
149. Gray RJ. Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis. *JASA* 1992;87:942-51.
150. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. *N Engl J Med* 1995;332(18):1229-32.
151. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev* 1987;9:1-30.
152. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med* 1993;12(8):717-36.
153. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol* 2007;36:195-202.
154. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008;167(5):523-9; discussion 530-1.
155. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 1995;142(12):1255-64.
156. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Methods Med Res* 2003;12(2):147-70.
157. Grundy SM, Cleeman JL, Merz CN, Brewer HB, Jr., Clark LT, Hunnighake DB, et al. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. *Circulation* 2004;110(2):227-39.

158. Grunkemeier GL, Anderson RP, Miller DC, Starr A. Time-related analysis of nonfatal heart valve complications: cumulative incidence (actual) versus Kaplan-Meier (actuarial). *Circulation* 1997;96(9 Suppl):II-70-4; discussion II-74-5.
159. Grunkemeier GL, Jin R, Eijkemans MJ, Takkenberg JJ. Actual and actuarial probabilities of competing risks: apples and lemons. *Ann Thorac Surg* 2007;83(5):1586-92.
160. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, *et al.* Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. *Jama* 2000;284(10):1290-6.
161. Habbema JD, Hilden J. The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. *Methods Inf Med* 1981;20(2):80-96.
162. Habbema JD, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. *Methods Inf Med* 1978;17(4):217-26.
163. Habbema JD, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. V. General recommendations. *Methods Inf Med* 1981;20(2):97-100.
164. Hahn GJ, Raghunathan TE. Combining information from various sources: a prediction problem and other industrial applications. *Technometrics* 1988;30(1):41-52.
165. Hajeer GR, van der Graaf Y, Olijhoek JK, Verhaar MC, Visseren FL. Levels of homocysteine are increased in metabolic syndrome patients but are not associated with an increased cardiovascular risk, in contrast to patients without the metabolic syndrome. *Heart* 2007;93(2):216-20.
166. Hakulinen T. On long-term relative survival rates. *J Chronic Dis* 1977;30(7):431-43.
167. Hakulinen T, Abeywickrama KH. A computer program package for relative survival analysis. *Comput Programs Biomed* 1985;19(2-3):197-207.
168. Halkin A, Singh M, Nikolsky E, Grines CL, Tcheng JE, Garcia E, *et al.* Prediction of mortality after primary percutaneous coronary intervention for acute myocardial infarction: the CADILLAC risk score. *J Am Coll Cardiol* 2005;45(9):1397-405.
169. Hall WH, Jani AB, Ryu JK, Narayan S, Vijayakumar S. The impact of age and comorbidity on survival outcomes and treatment patterns in prostate cancer. *Prostate Cancer Prostatic Dis* 2005;8(1):22-30.
170. Hand DJ. Statistical methods in diagnosis. *Stat Methods Med Res* 1992;1(1):49-67.
171. Hand DJ. Classifier technology and the illusion of progress. *Statist Sci* 2006;21(1):1-14.
172. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
173. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? *Health Serv Res* 1997;31(6):659-78.
174. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001.
175. Harrell FE, Jr., Lee KL, Califff RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3(2):143-52.
176. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87.
177. Harrell FE, Jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst* 1988;80(15):1198-202.
178. Harrell FE, Jr., Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, *et al.* Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. *Stat Med* 1998;17(8):909-44.
179. Harrison RF, Kennedy RL. Automatic covariate selection in logistic models for chest pain diagnosis: a new approach. *Comput Methods Programs Biomed* 2008;89(3):301-12.

180. Hastie T, Tibshirani R. Generalized additive models. Boca Raton, FL: Chapman & Hall/CRC, 1999.
181. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2001.
182. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials?. *Control Clin Trials* 1998;19(3):249-56.
183. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;61:92-105.
184. Hemingway H. Prognosis research: why is Dr. Lydgate still waiting? *J Clin Epidemiol* 2006;59(12):1229-38.
185. Henry NL, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. *Oncologist* 2006;11(6):541-52.
186. Hermanek P, Hutter RV, Sobin LH. Prognostic grouping: the next step in tumor classification. *J Cancer Res Clin Oncol* 1990;116(5):513-6.
187. Hermans J, Krol AD, van Groningen K, Kluin PM, Kluin-Nelemans JC, Kramer MH, et al. International Prognostic Index for aggressive non-Hodgkin's lymphoma is valid for all malignancy grades. *Blood* 1995;86(4):1460-3.
188. Hernandez AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power?. *Ann Epidemiol* 2006;16(1):41-8.
189. Hernandez AV, Steyerberg EW, Butcher I, Mushkudiani N, Taylor GS, Murray GD, et al. Adjustment for strong predictors of outcome in traumatic brain injury trials: 25% reduction in sample size requirements in the IMPACT study. *J Neurotrauma* 2006;23(9):1295-303.
190. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004;57(5):454-60.
191. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods Inf Med* 1978;17(4):227-37.
192. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. *Methods Inf Med* 1978;17(4):238-46.
193. Hoaglin DC, Mosteller F, Tukey JW. Understanding robust and exploratory data analysis. Wiley classics library ed. New York: Wiley, 2000.
194. Hoeting J, Madigan D, Raftery A, Volinsky C. Bayesian model averaging: a tutorial. *Statist Sci* 1999;14:382-401.
195. Hokken RB, Steyerberg EW, Verbaan N, van Herwerden LA, van Domburg R, Bos E. 25 years of aortic valve replacement using mechanical valves. Risk factors for early and late mortality. *Eur Heart J* 1997;18(7):1157-65.
196. Hollander N, Augustin NH, Sauerbrei W. Investigation on the improvement of prediction by bootstrap model averaging. *Methods Inf Med* 2006;45(1):44-50.
197. Homs MY, Steyerberg EW, Eijkenboom WM, Tilanus HW, Stalpers LJ, Bartelsman JF, et al. Single-dose brachytherapy versus metal stent placement for the palliation of dysphagia from oesophageal cancer: multicentre randomised trial. *Lancet* 2004;364(9444):1497-504.
198. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16(9):965-80.
199. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley, 2000.
200. Hu G, Root MM. Building prediction models for coronary heart disease by synthesizing multiple longitudinal research findings. *Eur J Cardiovasc Prev Rehabil* 2005;12(5):459-64.
201. Hudon C, Fortin M, Vanasse A. Cumulative Illness Rating Scale was a reliable and valid index in a family practice context. *J Clin Epidemiol* 2005;58(6):603-8.
202. Hukkelhoven CW, Rampen AJ, Maas AI, Farace E, Habbema JD, Marmarou A, et al. Some prognostic models for traumatic brain injury were not valid. *J Clin Epidemiol* 2006;59(2):132-43.

203. Hukkelhoven CW, Steyerberg EW, Habbema JD, Farace E, Marmarou A, Murray GD, et al. Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics. *J Neurotrauma* 2005;22(10):1025-39.
204. Hukkelhoven CW, Steyerberg EW, Rampen AJ, Farace E, Habbema JD, Marshall LF, et al. Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. *J Neurosurg* 2003;99(4):666-73.
205. Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Hum Reprod* 2004;19(9):2019-26.
206. Hunault CC, te Velde ER, Weima SM, Macklon NS, Eijkemans MJ, Klinkert ER, et al. A case study of the applicability of a prediction model for the selection of patients undergoing in vitro fertilization for single embryo transfer in another centre. *Fertil Steril* 2007;87(6):1314-21.
207. Hurria A, Leung D, Trainor K, Borgen P, Norton L, Hudis C. Factors influencing treatment patterns of breast cancer patients age 75 and older. *Crit Rev Oncol Hematol* 2003;46(2):121-6.
208. Iezzoni LI. Risk adjustment for measuring health care outcomes. 3rd ed. Chicago: Health Administration Press, 2003.
209. Imani B, Eijkemans MJ, Faessen GH, Bouchard P, Giudice LC, Fauser BC. Prediction of the individual follicle-stimulating hormone threshold for gonadotropin induction of ovulation in normogonadotropic anovulatory infertility: an approach to increase safety and efficiency. *Fertil Steril* 2002;77(1):83-90.
210. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2(8):e124.
211. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol* 2006;164(7):609-14.
212. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation* 1999;99(16):2098-104.
213. Jamieson GG, Mathew G, Ludemann R, Wayman J, Myers JC, Devitt PG. Postoperative mortality following oesophagectomy and problems in reporting its rate. *Br J Surg* 2004;91(8):943-7.
214. Janssen-Heijnen ML, Houterman S, Lemmens VE, Brenner H, Steyerberg EW, Coebergh JW. Prognosis for long-term survivors of cancer. *Ann Oncol* 2007;18(8):1408-13.
215. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61(1):76-86.
216. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 2006;8(7):395-400.
217. Janssens AC, Deng Y, Borsboom GJ, Eijkemans MJ, Habbema JD, Steyerberg EW. A new logistic regression approach for the evaluation of diagnostic test results. *Med Decis Making* 2005;25(2):168-77.
218. Jennett B, Snoek J, Bond MR, Brooks N. Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. *J Neurol Neurosurg Psychiatry* 1981; 44(4):285-93.
219. Jennings BM, Staggers N, Brosch LR. A classification scheme for outcome indicators. *Image J Nurs Sch* 1999;31(4):381-8.
220. Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Stat Med* 2007;26(29):5320-34.
221. Jiang W, Varma S, Simon R. Calculating confidence intervals for prediction error in microarray classification using resampling. *Stat Appl Genet Mol Biol* 2008;7:Article8.
222. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):515-24.

223. Kallewaard M, Algra A, Defauw J, Grobbee D, van der Graaf Y. Long-term survival after valve replacement with Bjork-Shiley CC valves. Bjork-Shiley Study Group. Am J Cardiol 2000;85(5):598-603.
224. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. JASA 1958;53:457-81.
225. Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;95(9):634-5.
226. Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst 1998;90(10):766-71.
227. Kattan MW, Eastham JA, Wheeler TM, Maru N, Scardino PT, Erbersdobler A, *et al.* Counseling men with prostate cancer: a nomogram for predicting the presence of small, moderately differentiated, confined tumors. J Urol 2003;170(5): 1792-7.
228. Kattan MW, Heller G, Brennan MF. A competing-risks nomogram for sarcoma-specific death following local recurrence. Stat Med 2003;22(22):3515-25.
229. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007;298(10): 1209-12.
230. Kerkhof M, van Dekken H, Steyerberg EW, Meijer GA, Mulder AH, de Bruine A, *et al.* Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. Histopathology 2007;50(7):920-7.
231. Kertai MD, Steyerberg EW, Boersma E, Bax JJ, Vergouwe Y, van Urk H, *et al.* Validation of two risk models for perioperative mortality in patients undergoing elective abdominal aortic aneurysm surgery. Vasc Endovascular Surg 2003;37(1):13-21.
232. Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Malden, MA: Blackwell Science, 2003.
233. Klabunde CN, Warren JL, Legler JM. Assessing comorbidity using claims data: an overview. Med Care 2002;40(8 Suppl):IV-26-35.
234. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.
235. Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, *et al.* Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004;57(12):1223-31.
236. Knaus WA, Harrell FE, Jr., Lynn J, Goldman L, Phillips RS, Connors AF, Jr., *et al.* The SUPPORT prognostic model Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. Ann Intern Med 1995;122(3):191-203.
237. Knol MJ, van der Tweel I, Grobbee DE, Numans ME, Geerlings MI. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. Int J Epidemiol 2007;36(5):1111-8.
238. Knottnerus JA. Evidence base of clinical diagnosis. Oxford: Blackwell BMJ Books, 2002.
239. Kollmannsberger C, Nichols C, Meisner C, Mayer F, Kanz L, Bokemeyer C. Identification of prognostic subgroups among patients with metastatic 'IGCCCG poor-prognosis' germ-cell cancer: an explorative analysis using cart modeling. Ann Oncol 2000; 11(9):1115-20.
240. Konig IR, Malley JD, Weimar C, Diener HC, Ziegler A. Practical experiences on the necessity of external validation. Stat Med 2007;26(30):5499-511.
241. Krijnen P, van Jaarsveld BC, Deinum J, Steyerberg EW, Habbema JD. Which patients with hypertension and atherosclerotic renal artery stenosis benefit from immediate intervention? J Hum Hypertens 2004;18(2):91-6.
242. Krijnen P, van Jaarsveld BC, Hunink MG, Habbema JD. The effect of treatment on health-related quality of life in patients with hypertension and renal artery stenosis. J Hum Hypertens 2005;19(6):467-70.
243. Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in 't Veld AJ, Schalekamp MA, Habbema JD. A clinical prediction rule for renal artery stenosis. Ann Intern Med 1998;129(9): 705-11.

244. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. *Circulation* 2006;113(3):456-62.
245. Krumholz HM, Wang Y, Mattera JA, Han LF, Ingber MJ, Roman S, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation* 2006;113(13):1683-92.
246. Kwakkel G, Wagenaar RC, Kollen BJ, Lankhorst GJ. Predicting disability in stroke—a critical review of the literature. *Age Ageing* 1996;25(6):479-89.
247. Kyrgidis A, Kountouras J, Zavos C, Chatzopoulos D. New molecular concepts of Barrett's esophagus: clinical implications and biomarkers. *J Surg Res* 2005;125(2):189-212.
248. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.
249. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *Jama* 1994;272(3):234-7.
250. le Cessie S, van Houwelingen HC. Testing the fit of a regression model via score tests in random effects models. *Biometrics* 1995;51(2):600-14.
251. Le CT, Lindgren BL. Computational implementation of the conditional logistic regression model in the analysis of epidemiologic matched studies. *Comput Biomed Res* 1988;21(1):48-52.
252. Leclaire S, Di Fiore F, Antonietti M, Ben Soussan E, Hellot MF, Grigioni S, et al. Undernutrition is predictive of early mortality after palliative self-expanding metal stent insertion in patients with inoperable or recurrent esophageal cancer. *Gastrointest Endosc* 2006;64(4):479-484.
253. Lee KI, Koval JJ. Determinants of the best significance level in forward stepwise logistic regression. *Comm Stat Sim Comp* 1997;26(2):559-75.
254. Lee KL. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction:<http://circ.ahajournals.org/cgi/content/full/91/6/1659>, 2006.
255. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. *Circulation* 1995;91(6):1659-68.
256. Lesaffre E, Bogaerts K, Li X, Bluhmki E. On the variability of covariate adjustment: experience with Koch's method for evaluating the absolute difference in proportions in randomized clinical trials. *Control Clin Trials* 2002;23(2):127-42.
257. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;360(9349):1903-13.
258. Liao JG, Chin KV. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics* 2007;23(15):1945-51.
259. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *Jama* 1999;282(11):1061-6.
260. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;363(9415):1147-54.
261. Lingsma HF, Dippel DW, Hoeks S, Steyerberg EW, Franke CL, van Oostenbrugge RJ, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: data from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry* 2008.
262. Lipton LR, Johnson V, Cummings C, Fisher S, Risby P, Eftekhar Sadat AT, et al. Refining the Amsterdam Criteria and Bethesda Guidelines: testing algorithms for the prediction of

- mismatch repair mutation status in the familial cancer clinic. *J Clin Oncol* 2004;22(24):4934-43.
263. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, NJ: Wiley, 2002.
264. Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial cohort study. *Jama* 2004;291(21):2591-9.
265. Loeb M, Walter SD, McGeer A, Simor AE, McArthur MA, Norman G. A comparison of model-building strategies for lower respiratory tract infection in long-term care. *J Clin Epidemiol* 1999;52(12):1239-48.
266. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;144(11):850-5.
267. Lorenz MO. Methods of measuring the concentration of wealth. *JASA* 1905;9:209-19.
268. Lubien E, DeMaria A, Krishnasamy P, Clopton P, Koon J, Kazanegra R, et al. Utility of B-natriuretic peptide in detecting diastolic dysfunction: comparison with Doppler velocity recordings. *Circulation* 2002;105(5):595-601.
269. Lubsen J, Pool J, van der Does E. A practical device for the application of a diagnostic or prognostic function. *Methods Inf Med* 1978;17(2):127-9.
270. Lynch HT, de la Chapelle A. Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* 1999;36(11):801-18.
271. Maas AI, Marmarou A, Murray GD, Teasdale SG, Steyerberg EW. Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. *J Neurotrauma* 2007;24(2):232-8.
272. Maas AI, Steyerberg EW, Butcher I, Dammers R, Lu J, Marmarou A, et al. Prognostic value of computerized tomography scan characteristics in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):303-14.
273. Maas AI, Steyerberg EW, Murray GD, Bullock R, Baethmann A, Marshall LF, et al. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. *Neurosurgery* 1999;44(6):1286-98.
274. Maggioni AP, Maseri A, Fresco C, Franzosi MG, Mauri F, Santoro E, et al. Age-related increase in mortality among patients with first myocardial infarctions treated with thrombolysis. The Investigators of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2). *N Engl J Med* 1993;329(20):1442-8.
275. Malek MH, Berger DE, Coburn JW. On the inappropriateness of stepwise regression analysis for model building and testing. *Eur J Appl Physiol* 2007;101(2):263-4; author reply 265-6.
276. Marmarou A, Lu J, Butcher I, McHugh GS, Murray GD, Steyerberg EW, et al. Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. *J Neurotrauma* 2007;24(2):270-80.
277. Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, et al. IMPACT database of traumatic brain injury: design and description. *J Neurotrauma* 2007;24(2):239-50.
278. Marquand A, Hanon O, Fauvel JP, Mounier-Vehier C, Equine O, Girerd X. [Validity of the clinical prediction rule for the diagnosis of renal arterial stenosis in hypertensive patients resistant to treatment]. *Arch Mal Coeur Vaiss* 2000;93(8):1041-5.
279. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *Bmj* 1998;316(7146):1701-4; discussion 1705.
280. Marshall LF, Marshall SB, Klauber MR, Van Berkum Clark M, Eisenberg H, Jane JA, et al. The diagnosis of head injury requires a classification based on computed axial tomography. *J Neurotrauma* 1992;9 (Suppl 1):S287-92.
281. Matthews DE, Farewell VT. Using and understanding medical statistics. 4th, completely rev. and enl. ed. Basel: Karger, 2007.
282. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *Jama* 2000;284(1):79-84.

283. McHugh GS, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, Marmarou A, *et al.* Statistical approaches to the univariate prognostic analysis of the IMPACT database on traumatic brain injury. *J Neurotrauma* 2007;24(2):251-8.
284. McHugh GS, Engel DC, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, *et al.* Prognostic value of secondary insults in traumatic brain injury: results from the IMPACT study. *J. Neurotrauma* 2007;24(2):287-93.
285. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005;97(16):1180-4.
286. Michel P, Domecq S, Salmi LR, Roques F, Nashef SAM. Confidence intervals for the prediction of mortality in the logistic EuroSCORE. *Eur J Cardiothorac Surg* 2005;27(6):1129-32.
287. Michel P, Roques F, Nashef SA. Logistic or additive EuroSCORE for high-risk patients? *Eur J Cardiothorac Surg* 2003;23(5):684-7.
288. Michie D, Spiegelhalter DJ, Taylor CC. Machine learning, neural and statistical classification. New York: Ellis Horwood, 1994.
289. Miller AJ. Subset selection in regression. 2nd ed. Boca Raton: Chapman & Hall/CRC, 2002.
290. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991;10(8):1213-26.
291. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Med Decis Making* 1993;13(1):49-58.
292. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005;21(15):3301-7.
293. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 2004;57(12):1262-70.
294. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;59(10):1092-101.
295. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. *J Epidemiol Community Health* 2002;56(5):337-8.
296. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol* 2002;55(7):633-6.
297. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *J Clin Epidemiol* 2002;55(10):1054-5.
298. Morgan TM, Elashoff RM. Effect of categorizing a continuous covariate on the comparison of survival time. *JASA* 1986;81:917-21.
299. Morris AP, Diamond GA, Detrano R, Bobbio M, Gunel E. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making* 1996;16(2):133-42.
300. Morris CN. Parametric Empirical Bayes inference: theory and applications. *JASA* 1983;78:47-55.
301. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *Bmj* 2003;326(7398):1083-4.
302. Mueller HS, Cohen LS, Braunwald E, Forman S, Feit F, Ross A, *et al.* Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. Analyses of patient subgroups in the Thrombolysis in Myocardial Infarction (TIMI) trial, phase II. *Circulation* 1992;85(4):1254-64.
303. Murphy AH. A new vector partition of the probability score. *J Appl Meteor* 1973;12(4):595-600.
304. Murray GD, Barer D, Choi S, Fernandes H, Gregson B, Lees KR, *et al.* Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. *J Neurotrauma* 2005;22(5):511-7.
305. Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AI, *et al.* Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):329-37.

306. Mashkudiani NA, Engel DC, Steyerberg EW, Butcher I, Lu J, Marmarou A, et al. Prognostic value of demographic characteristics in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):259-69.
307. Mashkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *J Clin Epidemiol* 2008;61(4):331-43.
308. Nab HW, Hop WC, Crommelin MA, Kluck HM, van der Heijden LH, Coebergh JW. Changes in long term prognosis for breast cancer in a Dutch cancer registry. *Bmj* 1994;309(6947):83-6.
309. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika* 1991;78:691-2.
310. Narayan RK, Michel ME, Ansell B, Baethmann A, Biegton A, Bracken MB, et al. Clinical trials in head injury. *J Neurotrauma* 2002;19(5):503-57.
311. Normand SL, Wolf RE, Ayanian JZ, McNeil BJ. Assessing the accuracy of hospital clinical performance measures. *Med Decis Making* 2007;27(1):9-20.
312. Odell PM, Anderson KM, Kannel WB. New models for predicting cardiovascular events. *J Clin Epidemiol* 1994;47(6):583-92.
313. Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol* 2003;56(6):501-6.
314. Ormsby AH, Petras RE, Henricks WH, Rice TW, Rybicki LA, Richter JE, et al. Observer variation in the diagnosis of superficial oesophageal adenocarcinoma. *Gut* 2002;51(5):671-6.
315. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). *Pract Assess Res Eval* 2004;9(6).
316. Ott K, Weber W, Siewert JR. The importance of PET in the diagnosis and response evaluation of esophageal cancer. *Dis Esophagus* 2006;19(6):433-42.
317. Oxford Centre for Evidence-Based Medicine. Glossary of terms in Evidence-Based Medicine. <http://www.cebm.net/glossary.asp>. Accessed Jan 2, 2007.
318. Palatini P. Reliability of ambulatory blood pressure monitoring. *Blood Press Monit* 2001;6(6):291-5.
319. Papworth DG, Lloyd RA. Cancer survival in the USA, 1973-1990: a statistical analysis. *Br J Cancer* 1998;78(11):1514-5.
320. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *J R Stat Soc Ser B* 2007;69(4):659-77.
321. Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. *Biostatistics* 2007;8(2):212-27.
322. Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Stat Med* 1993;12(18):1723-32.
323. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am J Hum Genet* 1998;62(1):145-58.
324. Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group. *Bmj* 1998;316(7149):1931-5.
325. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;302(20):1109-17.
326. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;48(12):1503-10.
327. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49(12):1373-9.
328. Peirce CS. The numerical measure of success of predictions. *Science* 1884;4:453-4.
329. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Comments on "Integrated discrimination and net reclassification improvements-Practical advice." *Stat Med* 2008;27(2):207-12.

330. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2): 157-72; discussion 207-12.
331. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.
332. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Stat Med* 2005;24(24):3687-96.
333. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
334. Perry JJ, Stiell IG. Impact of clinical decision rules on clinical care of traumatic injuries to the foot and ankle, knee, cervical spine, and head. *Injury* 2006;37(12):1157-65.
335. Petrina M, Goodman SG, Eagle KA. The 12-lead electrocardiogram as a predictive tool of mortality after acute myocardial infarction: current status in an era of revascularization and reperfusion. *Am Heart J* 2006;152(1): 11-8.
336. Picard RR, Cook RD. Cross-validation of regression models. *JASA* 1984;79:575-83.
337. Piccirillo JF, Tierney RM, Costas I, Grove L, Spitznagel EL, Jr. Prognostic importance of comorbidity in a hospital-based cancer registry. *Jama* 2004;291(20):2441-7.
338. Pocock SJ. Clinical trials: a practical approach. Chichester [West Sussex]: Wiley, 1983.
339. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21(19):2917-30.
340. Poldermans D, Bax JJ, Kertai MD, Krenning B, Westerhout CM, Schinkel AF, et al. Statins are associated with a reduced incidence of perioperative mortality in patients undergoing major noncardiac vascular surgery. *Circulation* 2003; 107(14): 1848-51.
341. Powers CA, Meyer CM, Roebuck MC, Vaziri B. Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Med Care* 2005;43(11): 1065-72.
342. Raftery AE, Madigan D, Hoeting J. Bayesian model averaging for linear regression models. *JASA* 1997;92:179-91.
343. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299(17):926-30.
344. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; 144(3):201-9.
345. Rietveld E, De Jonge HC, Polder JJ, Vergouwe Y, Veeze HJ, Moll HA, et al. Anticipated costs of hospitalization for respiratory syncytial virus infection in young children at risk. *Pediatr Infect Dis J* 2004;23(6):523-9.
346. Rietveld E, Vergouwe Y, Steyerberg EW, Huysman MW, de Groot R, Moll HA. Hospitalization for respiratory syncytial virus infection in young children: development of a clinical prediction rule. *Pediatr Infect Dis J* 2006;25(3):201-7.
347. Rigatelli G. Assessing the appropriateness and increasing the yield of renal angiography in patients undergoing coronary angiography: a scoring system. *Int J Cardiovasc Imaging* 2006;22(2): 135-9.
348. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Review* 1991;59(2):227-40.
349. Rodriguez-Bigas MA, Boland CR, Hamilton SR, Henson DE, Jass JR, Khan PM, et al. A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. *J Natl Cancer Inst* 1997;89(23):1758-62.
350. Rosamond WD, Chambliss LE, Folsom AR, Cooper LS, Conwill DE, Clegg L, et al. Trends in the incidence of myocardial infarction and in mortality due to coronary heart disease, 1987 to 1994. *N Engl J Med* 1998;339(13):861-7.
351. Ross PL, Scardino PT, Kattan MW. A catalog of prostate cancer nomograms. *J Urol* 2001; 165(5): 1562-8.
352. Roukema J, van Loenhout RB, Steyerberg EW, Moons KG, Bleeker SE, Moll HA. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. *J Clin Epidemiol* 2008;61(2):135-41.

353. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Stat Med* 2000;19(14):1831-47.
354. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl Stat* 1994;43(3):429-67.
355. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25(1):127-41.
356. Royston P, Sauerbrei W. Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. *Computational Stat Data Anal* 2007;51(9):4240-53.
357. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581-92.
358. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
359. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; 127(8 Pt 2):757-63.
360. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007; 11(50):iii, ix-51.
361. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies *CMAJ* 2006;174(4):469-76.
362. Sackett DL, Rosenberg WM. On the need for evidence-based medicine. *J Public Health Med* 1995;17(3):330-4.
363. Sanada H, Sugama J, Kitagawa A, Thigpen B, Kinoshita S, Murayama S. Risk factors in the development of pressure ulcers in an intensive care unit in Pontianak, Indonesia. *Int Wound J* 2007;4(3):208-15.
364. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005;23(9):2020-7.
365. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *J R Stat Soc Ser C* 1999;48(3):313-29.
366. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational Stat Data Anal* 2006;50(12):3464-85.
367. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc Ser A* 1999; 162(1):71-94.
368. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* 1992;11(16):2093-109.
369. Scarvelis D, Wells PS. Diagnosis and treatment of deep-vein thrombosis. *CMAJ* 2006;175(9):1087-92.
370. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall, 1997.
371. Schapire RE. The strength of weak learnability. *Machine Learning* 1990;5(2): 197-227.
372. Schemper M. Predictive accuracy and explained variation. *Stat Med* 2003;22(14):2299-308.
373. Scholte op Reimer WJ, Dippel DW, Franke CL, van Oostenbrugge RJ, de Jong G, Hoeks S, et al. Quality of hospital and outpatient care after stroke or transient ischemic attack: insights from a stroke survey in the Netherlands. *Stroke* 2006;37(7):1844-9.
374. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics* 2007; 23:1768-74.
375. Schumacher M, Graf E, Gerds T. How to assess prognostic models for survival data: a case study in oncology. *Methods Inf Med* 2003;42(5):564-71.
376. Schwarzer G, Schumacher M. Artificial neural networks for diagnosis and prognosis in prostate cancer. *Semin Urol Oncol* 2002;20(2):89-95.
377. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000;19(4):541-61.
378. Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *Am J Cardiol* 2003;92(12): 1447-51.

379. Segar RW, Wilson JH, Habbema JD, Malchow-Moller A, Hilden J, van der Maas PJ. Transferring a diagnostic decision aid for jaundice. *Neth J Med* 1988;33(1-2):5-15.
380. Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994;13(17):1715-26.
381. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;58(6):550-9.
382. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001;72(6):2155-68.
383. Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SL. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation* 2007;115(12): 1518-27.
384. Shahian DM, Torchiana DF, Shemin RJ, Rawn JD, Normand SL. Massachusetts cardiac surgery report card: implications of statistical methodology. *Ann Thorac Surg* 2005;80(6):2106-13.
385. Shen W, Louis TA. Triple-goal estimates in two-stage hierarchical models. *J R Stat Soc Ser B* 1998;60:455-71.
386. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994;69(6):979-85.
387. Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y. Design and analysis of DNA microarray investigations. New York: Springer, 2003.
388. Simons PC, Algra A, van de Laak MF, Grobbee DE, van der Graaf Y. Second manifestations of ARTerial disease (SMART) study: rationale and design. *Eur J Epidemiol* 1999;15(9):773-81.
389. Skacel M, Petras RE, Gramlich TL, Sigel JE, Richter JE, Goldblum JR. The diagnosis of low-grade dysplasia in Barrett's esophagus and its implications for disease progression. *Am J Gastroenterol* 2000;95(12):3383-7.
390. Smits JM, De Meester J, Deng MC, Scheld HH, Hummel M, Schoendube F, et al. Mortality rates after heart transplantation: how to compare centre-specific outcome data? *Transplantation* 2003;75(1):90-6.
391. Smits M, Dippel DW, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, et al. Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. *Ann Intern Med* 2007; 146(6):397-405.
392. Snee RD. Validation of regression models: methods and examples. *Technometrics* 1977; 19:415-28.
393. Snick HK, Snick TS, Evers JL, Collins JA. The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod* 1997; 12(7):1582-8.
394. Solomon DH, Chibnik LB, Losina E, Huang J, Fossel AH, Husni E, et al. Development of a preliminary index that predicts adverse events after total knee replacement. *Arthritis Rheum* 2006;54(5): 1536-42.
395. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986;5(5):421-33.
396. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24(8):1185-202.
397. Spiegelhalter DJ, Crean GP, Holden R, Knill-Jones RP. Taking a calculated risk: predictive scoring systems in dyspepsia. *Scand J Gastroenterol Suppl* 1987;128:152-60.
398. Stein CM. Estimation of the mean of a multivariate normal distribution. *Ann Stat* 1981;9(6):1135-51.
399. Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* 2005;104(2):290-8.
400. Steyerberg EW, Balmana J, Stockwell DH, Syngal S. Data reduction for prediction: robust coding of age and family history for the risk of having a genetic mutation. *Stat Med* 2007; 26(30): 5545-56.

401. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* 2003;56(5):441-7.
402. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23(16):2567-86.
403. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *Am Heart J* 2000;139(5):745-51.
404. Steyerberg EW, Earle CC, Neville BA, Weeks JC. Racial differences in surgical evaluation, treatment, and outcome of locoregional esophageal cancer: a population-based analysis of elderly patients. *J Clin Oncol* 2005;23(3):510-7.
405. Steyerberg EW, Eijkemans MJ, Boersma E, Habbema JD. Applicability of clinical prediction models in acute myocardial infarction: a comparison of traditional and empirical Bayes adjustment methods. *Am Heart J* 2005;150(5):920.
406. Steyerberg EW, Eijkemans MJ, Boersma E, Habbema JD. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logistic regression models. *J Clin Epidemiol* 2005;58(4):383-90.
407. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 1999;52(10):935-42.
408. Steyerberg EW, Eijkemans MJ, Habbema JD. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl* 2001;55:76-88.
409. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000;19(8):1059-79.
410. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001;21(1):45-56.
411. Steyerberg EW, Eijkemans MJ, van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Stat Med* 2000;19(2):141-60.
412. Steyerberg EW, Gerl A, Fossa SD, Sleijfer DT, de Wit R, Kirkels WJ, et al. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. *J Clin Oncol* 1998;16(1):269-74.
413. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54(8):774-81.
414. Steyerberg EW, Horns MY, Stokvis A, Essink-Bot ML, Siersema PD. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: a prognostic model to guide treatment selection. *Gastrointest Endosc* 2005;62(3):333-40.
415. Steyerberg EW, Kallewaard M, van der Graaf Y, van Herwerden LA, Habbema JD. Decision analyses for prophylactic replacement of the Bjork-Shiley convexo-concave heart valve: an evaluation of assumptions and estimates. *Med Decis Making* 2000;20(1):20-32.
416. Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Bajorin DF, Donohue JP, et al. Resection of residual retroperitoneal masses in testicular cancer: evaluation and improvement of selection criteria. The ReHiT study group. Re-analysis of histology in testicular cancer. *Br J Cancer* 1996;74(9):1492-8.
417. Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Toner GC, Schraffordt Koops H, et al. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic non-seminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. *J Clin Oncol* 1995;13(5):1177-87.
418. Steyerberg EW, Keizer HJ, Habbema JD. Prediction models for the histology of residual masses after chemotherapy for metastatic testicular cancer. ReHiT study group. *Int J Cancer* 1999;83(6):856-9.

419. Steyerberg EW, Keizer HJ, Sleijfer DT, Fossa SD, Bajorin DF, Gerl A, *et al.* Retroperitoneal metastases in testicular cancer: role of CT measurements of residual masses in decision making for resection after chemotherapy. *Radiology* 2000;215(2):437-44.
420. Steyerberg EW, Keizer HJ, Stoter G, Habbema JD. Predictors of residual mass histology following chemotherapy for metastatic non-seminomatous testicular cancer: a quantitative overview of 996 resections. *Eur J Cancer* 1994;30A(9):1231-9.
421. Steyerberg EW, Kievit J, de Mol Van Otterloo JC, van Bockel JH, Eijkemans MJ, Habbema JD. Perioperative mortality of elective abdominal aortic aneurysm surgery. A clinical prediction rule based on literature and individual patient data. *Arch Intern Med* 1995;155(18):1998-2004.
422. Steyerberg EW, Marshall PB, Keizer HJ, Habbema JD. Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. *Cancer* 1999;85(6):1331-41.
423. Steyerberg EW, Neville BA, Koppert LB, Lemmens VE, Tilanus HW, Coebergh JW, *et al.* Surgical mortality in patients with esophageal cancer: development and validation of a simple risk score. *J Clin Oncol* 2006;24(26):4277-84.
424. Steyerberg EW, Roobol MJ, Kattan MW, van der Kwast TH, de Koning HJ, Schroder FH. Prediction of indolent prostate cancer: validation and updating of a prognostic nomogram. *J Urol* 2007;177(1):107-12.
425. Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JD. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. *Stat Med* 2001;20(24):3847-59.
426. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008;28(1):146-9.
427. Stiell I, Wells G, Laupacis A, Brison R, Verbeek R, Vandemeulen K, *et al.* Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. Multicentre Ankle Rule Study Group. *Bmj* 1995;311(7005):594-7.
428. Strandberg TE, Pitkala K. What is the most important component of blood pressure: systolic, diastolic or pulse pressure? *Curr Opin Nephrol Hypertens* 2003;12(3):293-7.
429. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;59(5):437-47.
430. Teasdale G, Jennett B. Assessment and prognosis of coma after head injury. *Acta Neurochir (Wien)* 1976;34(1-4):45-55.
431. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;56(8):721-9.
432. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med* 1994;13(9):889-903.
433. Thompson KM. Risk in perspective: insight and humor in the age of risk management. Newton Centre, MA: AORM, 2004.
434. Tibshirani R. Regression and shrinkage via the Lasso. *J R Stat Soc Ser B* 1996;58:267-88.
435. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997;16(4):385-95.
436. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49(11):1225-31.
437. Tu JV, Austin PC, Walld R, Roos L, Agras J, McDonald KM. Development and validation of the Ontario acute myocardial infarction mortality prediction rules. *J Am Coll Cardiol* 2001;37(4):992-7.
438. Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The Steering Committee of the Cardiac Care Network of Ontario. *Med Decis Making* 1998;18(2):229-35.

439. Tuma RS. Trial and error: prognostic gene signature study design altered. *J Natl Cancer Inst* 2005;97(5):331-3.
440. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol* 2002;55(4):329-37.
441. Ulmer H, Kollerits B, Kelleher C, Diem G, Concin H. Predictive accuracy of the SCORE risk function for cardiovascular disease in clinical practice: a prospective evaluation of 44 649 Austrian men and women. *Eur J Cardiovasc Prev Rehabil* 2005;12(5):433-41.
442. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, et al. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst* 2004;96(4):261-8.
443. Vach K, Sauerbrei W, Schumacher M. Variable selection and shrinkage: comparison of some approaches. *Stat Neerl* 2001;55:53-75.
444. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol* 1991;134(8):895-907.
445. Vach W, Blettner M. Missing data in epidemiologic studies. *Encyclopedia of Biostatistics*. New York: Wiley, 1998:2641-54.
446. van Beek JG, Mushkudiani NA, Steyerberg EW, Butcher I, McHugh GS, Lu J, et al. Prognostic value of admission laboratory parameters in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):315-28.
447. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999;18(6):681-94.
448. van der Graaf Y, de Waard F, van Herwerden LA, Defauw J. Risk of strut fracture of Bjork-Shiley valves. *Lancet* 1992;339(8788):257-61.
449. van der Meulen JH, Steyerberg EW, van der Graaf Y, van Herwerden LA, Verbaan CJ, Defauw JJ, et al. Age thresholds for prophylactic replacement of Bjork-Shiley convexo-concave heart valves. A clinical and economic evaluation. *Circulation* 1993;88(1):156-64.
450. van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross-validation. *Stat Decis* 2006;24:351-71.
451. van Dijk MR, Steyerberg EW, Habbema JD. A decision-analytic approach to define poor prognosis patients: a case study for non-seminomatous germ cell cancer patients. *BMC Med Inform Decis Mak* 2008;8:1.
452. van Dijk MR, Steyerberg EW, Stenning SP, Habbema JD. Identifying subgroups among poor prognosis patients with nonseminomatous germ cell cancer by tree modelling: a validation study. *Ann Oncol* 2004;15(9):1400-5.
453. van Dijk MR, Steyerberg EW, Stenning SP, Habbema JD. Survival estimates of a prognostic classification depended more on year of treatment than on imputation of missing values. *J Clin Epidemiol* 2006;59(3):246-53.
454. van Gorp MJ, Steyerberg EW, Kalleswaard M, van der Graaf Y. Clinical prediction rule for 30-day mortality in Bjork-Shiley convexo-concave valve replacement. *J Clin Epidemiol* 2003;56(10):1006-12.
455. van Gorp MJ, Steyerberg EW, van der Graaf Y. Decision guidelines for prophylactic replacement of Bjork-Shiley convexo-concave heart valves: impact on clinical practice. *Circulation* 2004;109(17):2092-6.
456. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000;19(24):3401-15.
457. van Houwelingen HC, Brand R, Louis TA. Empirical Bayes methods for monitoring health care quality. Technical Report 2006.
458. van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 1995;14(18):1999-2008.
459. van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9(11):1303-25.
460. van Jaarsveld BC, Krijnen P, Pieterman H, Derkx FH, Deinum J, Postma CT, et al. The effect of balloon angioplasty on hypertension in atherosclerotic renal-artery stenosis. *Dutch Renal*

- Artery Stenosis Intervention Cooperative Study Group. N Engl J Med 2000; 342(14):1007-14.
461. van Koningsveld R, Steyerberg EW, Hughes RA, Swan AV, van Doorn PA, Jacobs BC. A clinical prognostic scoring system for Guillain-Barre syndrome. Lancet Neurol 2007;6(7):589-94.
462. van Westreenen HL, Westerterp M, Sloof GW, Groen H, Bossuyt PM, Jager PL, *et al.* Limited additional value of positron emission tomography in staging oesophageal cancer. Br J Surg 2007;94(12):1515-20.
463. Vandenbroucke JP. Prospective or retrospective: what's in a name? BMJ 1991;302(6771): 249-50.
464. Vapnik VN. The nature of statistical learning theory. New York: Springer, 1995.
465. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58(5):475-83.
466. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol 2001;165(1):84-8.
467. Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fossa SD, Gerl A, *et al.* Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. Eur Urol 2007;51(2):424-32.
468. Verweij PJ, van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med 1994;13(23-24):2427-36.
469. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.
470. Vickers AJ, Kramer BS, Baker SG. Selecting patients for randomized trials: a systematic approach based on risk group. Trials 2006;7:30.
471. Virnig BA, Warren JL, Cooper GS, Klabunde CN, Schussler N, Freeman J. Studying radiation therapy using SEER-Medicare-linked data. Med Care 2002;40(8 Suppl):IV-49-54.
472. Vittinghoff E. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. New York: Springer, 2005.
473. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165(6):710-8.
474. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. Biometrics 2000;56(1):256-62.
475. Wang D, Lertsithichai P, Nanchahal K, Yousufuddin M. Risk factors of coronary heart disease: a Bayesian model averaging approach. J Appl Stat 2003;30:813-26.
476. Wang OJ, Wang Y, Lichtman JH, Bradley EH, Normand SL, Krumholz HM. "America's Best Hospitals" in the treatment of acute myocardial infarction. Arch Intern Med 2007;167(13):1345-51.
477. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357(21):2189-94.
478. Warren JL, Harlan LC, Fahey A, Virnig BA, Freeman JL, Klabunde CN, *et al.* Utility of the SEER-Medicare data to identify chemotherapy use. Med Care 2002;40(8 Suppl):IV-55-61.
479. Wehberg S, Schumacher M. A comparison of nonparametric error rate estimation methods in classification problems. Biometrical J 2004;46(1):35-47.
480. Weimar C, Ziegler A, Konig IR, Diener HC. Predicting functional outcome and survival after acute ischemic stroke. J Neurol 2002;249(7):888-95.
481. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, *et al.* Value of assessment of pretest probability of deep-vein thrombosis in clinical management. Lancet 1997;350(9094):1795-8.
482. Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, *et al.* Accuracy of clinical assessment of deep-vein thrombosis. Lancet 1995;345(8961):1326-30.
483. Whitworth A. NCI launches an innovative design for a breast cancer clinical trial. J Natl Cancer Inst 2006;98(17):1178-9.

484. Wijesinha A, Begg CB, Funkenstein HH, McNeil BJ. Methodology for the differential diagnosis of a complex data set. A case study using data from routine CT scan examinations. *Med Decis Making* 1983;3(2):133-54.
485. Wijnen JT, Vasen HF, Khan PM, Zwinderman AH, van der Klift H, Mulder A, et al. Clinical findings with implications for genetic testing in families with clustering of colorectal cancer. *N Engl J Med* 1998;339(8):511-8.
486. Wikipedia contributors. Bootstrapping. <http://en.wikipedia.org/w/index.php?title=Bootstrapping&oldid=95537267> 2006:Accessed Dec 27, 2006.
487. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
488. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med* 2008;27:3227-46.
489. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004;1(4):368-76.
490. Wood SN. Generalized additive models: an introduction with R. Boca Raton, FL: Chapman & Hall/CRC, 2006.
491. Wright J. The Glasgow Outcome Scale. The Centre for Outcome Measurement in Brain Injury. <http://www.tbims.org/combi/gos> (accessed September 30, 2006).
492. Wynne R. Variable definitions: implications for the prediction of pulmonary complications after adult cardiac surgery. *Eur J Cardiovasc Nurs* 2004;3(1):43-52.
493. Yang Y. Consistency of cross validation for comparing regression procedures. *Ann Stat* 2007;35(6):2450-73.
494. Ye J. On measuring and correcting the effects of data mining and model selection. *JASA* 1998;93:120-31.
495. Yuan S, Yu Y, Chao KS, Fu Z, Yin Y, Liu T, et al. Additional value of PET/CT over PET in assessment of locoregional lymph nodes in thoracic esophageal squamous cell cancer. *J Nucl Med* 2006;47(8):1255-9.
496. Zelen M. The randomization and stratification of patients to clinical trials. *J Chronic Dis* 1974;27(7-8):365-75.
497. Zhou XH. Effect of verification bias on positive and negative predictive values. *Stat Med* 1994;13(17):1737-45.

# Index

## A

Abdominal aortic aneurysm (AAA), 430, 440, 459–461  
Accelerated failure time (AFT) models, 78  
Acute appendicitis diagnosis, 39  
Acute myocardial infarction (AMI) mortality  
  recursive partitioning, 67  
  treatment cost-effectiveness, 18  
Acute myocardial infarction (MI) and  
  GUSTO-I study  
  30-day mortality in  
    prediction models development  
      checklist, 416–417  
      prognostic modelling, 414–417  
      risk and prognostic factors, 413–414  
      tPA and streptokinase (SK), 414  
  covariate adjustment and tPA vs. SK  
    advantages, 427  
    analysis, adjusted and unadjusted, 426–427  
model development  
  outcome and predictors, 418–419  
  research question and intended  
    application, 417–418  
  study design and analysis, 419  
modelling steps in  
  coding of predictors and model  
    specification, 420  
  discrimination and calibration, 421  
  estimation methods, 420–421  
  odds ratios (OR), 422  
  presentation, 422–423  
  prognostic factor effects, 418–419  
  validation, 421–422  
thrombolytic therapy  
  prediction-based selection, 425–426  
  score chart for, 424–425  
validity, internal and external, 423

Adaptation methods, multivariable regression  
  coefficients

  clinical results, 250–251  
  description, 244  
  estimation, 245  
  improving calibration, 247–248  
  predictive performance, 247  
  simulation results, 245

Additivity and interaction terms  
  model specifications, 213–214  
  potential interactions, 214–216  
  time and survival, 216

Akaike information criterion (AIC), 195

Aneurysm mortality, 309, 310

ANOVA table, 422

Aortic aneurysms study, Bayesian  
  analysis, 252

Apparent validation, 300–301

## B

Bacterial meningitis, prediction models, 282

Barrett's oesophagus, 42

Bayes rule  
  and logistic regression, 65  
  Naïve Bayes estimation, 63–65  
  prior probability of disease, 61–62

Bayes' theorem, 13

Bayesian information criterion (BIC), 195

Bayesian methods  
  aneurysm study, 252  
  predicting neonatal death, 252  
  regression modelling, 251–252  
Bayesian model averaging (BMA), 208–210  
Björk–Shiley convexo–concave  
  (BScc) valves  
Poisson regression model, 80–81  
risk of fracture, 21–23

- Bootstrap resampling  
applications of, 93  
definition, 92  
optimism-corrected performance, 94–96  
regression coefficients, 93–94  
stepwise selection, 96–97
- Bootstrap validation, 98, 303–304
- Boxplot, 265–266. *See also* Concordance statistic
- Breast cancer susceptibility gene (BRCA), 13
- Breast cancer, prediction models, 294
- Brier score, 257
- C**
- C-reactive protein (CRP), 39
- Candidate predictors, prediction model  
averaging effects, 180–181  
*Chlamydia trachomatis* infection risk, 180  
heart valve replacement patients, 180–181  
comorbidity coding, 177–178  
distribution, 177  
equal weights assumption assessment, 178–179  
in testicular cancer, 176–177  
logical weighting, 179  
meta-analysis for, 176  
principal component analysis, 180
- Categorical predictors, GUSTO-I study  
coding impact, 159–160  
in acute myocardial infarction, 160–161
- Classification and regression tree (CART) methods  
advantages, 67–68  
binary tree construction, 67  
disadvantages, 68–69  
logistic regression model, 69–70
- Claus model, 13
- Clinical practice, prediction models  
delaying treatment, 19  
diagnostic algorithm, 16–17  
gold standard test, 13–14  
heart valve replacement  
BScc valves, 21  
decision analysis, 22–23  
risk of fracture, 22  
short-term *vs.* long-term risk, 21–22
- renal artery stenosis, 14–15
- spontaneous pregnancy  
in vitro fertilization, 21  
prognostic index score, 19–20
- surgical decision-making, 21  
treatment cost-effectiveness, 18–19  
treatment intensity  
benefit and harm, 16–18  
risks and side effects, 16  
treatment threshold, 15–16
- Clinical prediction models  
confidence intervals, 316–317  
interaction terms  
estimation, 217–218  
GUSTO-I interactions, 217  
nomograms, 317–319  
score charts, 319  
specific formats, 321  
tabular presentation, 320–321
- Cohort study. *See* Prognostic cohort study
- Collinear models, 192
- Colorectal cancers (CRC)  
adenoma polyps, 184–185  
age effect and diagnosis, 185–186  
clinical background and patient data, 181–184  
prediction model, 186–187
- Comorbidity definition, 40
- Complete case (CC) analysis, missing data  
inefficiency of, 116  
multiple predictors, 127–128  
regression coefficients  
MAR and MNAR, 117–118  
MCAR, 117  
simulation of, 124
- Concordance statistic, discriminative ability  
box plots and discrimination slope, 264  
Lorenz curve, 264–267  
receiver operating characteristic (ROC) curve, 260–262  
sensitivity and specificity, 260  
survival data, discrimination, 267–268  
testicular cancer discrimination, 268–269  
verification bias and discriminative ability, 269  
*vs.* explained variation ( $R^2$ ), 262–264
- Confidence intervals, regression models, 316–318, 320
- Continuation ratio model, 77
- Continuous predictors  
categorization, 327, 329  
prediction models and decision rules, 314  
regression formula, 315  
score chart, 319, 324, 325  
table format, 320
- Continuous predictors, GUSTO-I study  
categorization, 162–163  
coding impact, 161–162

- non-linear functions for  
fractional polynomials (FPs), 164  
polynomials, 164  
spline function, 165–166  
outliers and extreme values of, 167–168  
predictor effect interpretation, 170
- Covariate adjustment in GUSTO-I  
advantages, 427  
analysis, adjusted and unadjusted, 426–427
- Cox proportional hazards regression, 77–78
- Cox regression analysis, 439, 442–443
- Creatinine transformations, coding of predictors, 440
- Cross-validation  
classical, 301, 302  
jack-knife procedure, 301–303  
vs. split-sample validation, 302
- D**
- Data sets, 454  
abdominal aortic aneurysm (AAA), 460–461
- GUSTO-I prediction models  
data description of, 456  
modelling strategies in, 455  
modern learning methods, 455  
simulation design for, 457
- SMART case study, 457, 458
- testicular cancer case study, 457–459
- traumatic brain injury (TBI), 461
- Decision curves, prediction models, 284–285, 295
- testicular cancer, 287–288
- treat all strategy, 285–286
- Decision rules  
meta-model, 329–330
- score chart rule, 328–329
- survival analyses, 329
- vs. prediction model, 313–314
- Decision threshold, prediction models, 281, 295
- classifications, false-negative and false-positive, 282–283
- decision curve, 284
- definition, 283
- net benefit, 286
- validation setting, 288
- Deep venous thrombosis (DVT), 16–17
- Diagnostic study design  
case-control studies, 39
- cross-sectional study, 38
- renal artery stenosis, 38–39
- E**
- Empirical Bayes (EB) estimation, 395, 396, 404
- Epidural haematoma (EDH), IMPACT study  
imputed values, 150
- missingness patterns, 144–146
- predictor missingness quantification, 143–144
- rank correlations, 148
- European study on prostate cancer (ERSPC), 308
- EuroSCORE, 323, 327
- Expected rank (ER), 402, 406, 407
- Explained variation ( $R^2$ ), 255–256
- External validation, prediction models, 304  
calibration, discrimination, and clinical usefulness  
calibration slope and calibration-in-the-large, 342–345, 347–349, 354–358
- invalidity patterns, 336–337
- performance measures, 339–340
- simulation designs, 338–339
- case-control design outcomes, 344–345
- determinants of  
case-mix, 335–336
- regression coefficients, 336–337
- differences between populations, 336
- fully independent validation, 308–309
- geographic validation, 307–308
- invalidity patterns, 348–349
- missed predictors  
heterogeneous case-mix, 344
- more-or-less-severe case-mix, 342–344
- model performance reference values  
calculation, 349–350
- model refitting, 350–351
- R code, 350
- testicular cancer and TBI, 351–352
- observed predictors  $X$   
heterogeneous case-mix, 341–342
- more-or-less-severe case-mix, 340–341
- performance estimation  
standard errors, 354–355
- testicular cancer and TBI, 351–352
- validation uncertainty, 352–354
- power calculations, 356
- regression coefficients  
calibration, 348
- clinical usefulness, 349
- linear predictor, 345–346
- R code, 346–347
- sample size requirement, 356–357

**F**

- Fixed-effect approach
  - centre-specific estimates, 395
  - GUSTO-I and TIMI-II, 396
  - outcome differences, 393
  - predictors effects, 398
- Fractional polynomials (FP), 221–222.
  - See also* Testicular cancer
- Framingham risk functions, 12

**G**

- GAM. *See* Generalized additive model
- Generalizability, prediction model. *See* Prediction models, external validity patterns
- Generalized additive logistic regression model (GAM), 108–109
- Generalized additive model (GAM)
  - binary outcomes, 65–66
  - continuous outcomes, 55–57
- Generalized additive models (GAM), 220
- Generalized degrees of freedom (GDF), 98
- Genetic mutation, prediction modelling
  - colorectal cancers (CRC)
    - adenoma polyps, 184–185
    - age effect and diagnosis, 185–186
    - clinical background and patient data, 181–184
    - prediction model, 186–187
- Geographic validation, 307–308
- Gini index, 266
- Glasgow coma scale, 328
- Glasgow outcome scale, 74–75
  - imputation model, 147, 149
  - imputed values distribution, 149
  - missing values, 142
  - missingness patterns, 144–146
  - patient selection, 140
  - predictors
    - coding and time dependency, 141–142
    - missingness quantification, 143–144
    - selection, 140–141
- Goodness-of-fit tests
  - Goeman-Le Cessie test, 276
  - Hosmer-Lemeshow (H-L) test, 274–275
- GUSTO-I data
  - acute myocardial infarction, 103
  - modelling
    - GAM and MARS, 109
    - logistic regression and classification tree, 108
    - predictive performance, 109–110
- GUSTO-I interactions, 217
- GUSTO-I prediction models, data sets

**H**

- description of, 456
- modern learning methods, 455
- simulation design for, 457

**H**

- Hazard ratios (HRs), SMART study, 441
- Heterogeneity testing, TIMI-II model, 398–399
- High-dose chemotherapy (HD-CT), 18
- Homocysteine (HOMOC), SMART study, 434–435
- Hosmer-Lemeshow test, 278. *See also* Goodness-of-fit tests

**I**

- Imputation
  - auxiliary variables, 124–125
  - imputation, single and multiple methods, 122–123
  - missing outcomes
    - analysis, univariate and adjusted, 131–132
    - missing indicator, 130
    - predictor effect, 131
    - prognostic study guidelines, 133–134
    - stochastic SI, 132–133
  - multiple imputation, 123–124
  - multiple predictor simulations, 127–128
  - principle, 121–122
  - variables transformations, 125
- Indolent prostate cancer. *See* Prostate cancer, prediction models
- Internal validation, prediction models
  - validation techniques
    - apparent, 300–301
    - bootstrap, 303–304
    - cross-validation, 301–303
    - vs. external validation, 299–300
- International Germ Cell Classification (IGCC), 294
- International Mission on Prognosis and Analysis of Clinical Trials (IMPACT), TBI
  - missing values, predictors
    - adjusted effects estimation, 149–155
    - imputation model, 147, 149
    - missingness patterns, 144–146
    - missingness quantification, 141–142
    - multivariable effects, 155
    - outcome, 142
    - rank correlations, 147
  - predictor selection, 140–141
- Intima media thickness (IMT), SMART study, 437

**J**

Jack-knife cross-validation, 301–303

**K**

Kaplan–Meier analysis, 79

**L**

Lasso model

- estimation, 238
- GUSTO sample4 analyses, 239
- model performance after shrinkage, 240
- predictions after shrinkage, 239

Linear regression model, 53–54

- generalized additive model, 55–57
- optimism-corrected performance, 92
- outcome transformation, 54–55
- variance estimation, 55

Linear regression models

- Lasso model
  - estimation, 238
  - GUSTO sample4 analyses, 239
  - model performance after shrinkage, 240
  - predictions after shrinkage, 239
- neural networks, 231–232
- penalized maximum likelihood (PML)
  - estimation, 234–235
  - GUSTO sample4 analyses, 235–237
  - shrinkage and selection methods, 238
- shrinkage
  - characteristics, 232
  - uniform shrinkage, 233

Literature-adapted model, 247

Logistic regression analysis

- and Bayes rule, 65
- linear discriminant analysis, 60–61
- log likelihood scale
  - maximum likelihood techniques, 58
- R<sup>2</sup> definition, 60
- neural network, hidden layer, 66
- vs. tree characteristics, 69–70

Logistic regression model

- age-related mortality
  - acute myocardial infarction, 103
  - GUSTO-I data set, 104
- data differences between centres, 404–405
- fixed-effect approach, 393–394
- GUSTO-I modelling
  - GAM and MARS, 109
  - logistic regression and classification tree, 108
  - predictive performance, 109–110
- operative mortality

and age effect, 104–105

definition, 104

logistic regression models, 106

random effects, 394–396

Logistic regression model, external

validation, 337

calibration slope and calibration-in-the-large, 338

discrimination and calibration estimation, 352–354

power calculations, 356

standard error estimation, 354–355

stratification, 338

Lorenz curve, discriminative ability

advantages and disadvantages, 267

cumulative distribution of population, 264–265

Gini index in, 266

Lynch syndrome prediction model,

305, 306, 310

**M**

Medical research, prediction models

cardiac outcome ranking, 29–30

covariate adjustment

gain in power, 26–27

in RCT, 25–26

GUSTO-III trial analysis, 27

observational studies, 27–28

propensity scores, 28

provider profiling, 29

statin treatment effects, 28–29

stratification, RCT, 23–24

TBI trial selection, 24

Meta-analysis. *See* Univariate logistic regression coefficients

Meta-model, decision rules, 314, 329–330

Missing at random (MAR)

data mechanism, 117–118

imputation

missing outcomes, 130

model, 124–125

multiple predictors, 127–128

simulations of, 126

R code, 119–120

Missing completely at random (MCAR), SMART study, 434

Missing data outcomes

missing indicator, 130

predictor effect, 131

prognostic studies guidelines, 133–134

stochastic SI, 132–133

univariate and adjusted analysis, 131–132

- Missing not at random (MNAR )  
 data mechanisms, 117–118  
 imputation  
   missing outcomes, 130  
   multiple predictors, 128  
   simulations of, 126  
 regression coefficients, 119–120
- Missing completely at random (MCAR)  
 data mechanisms, 117  
 imputation  
   missing outcomes, 130  
   model, 125  
   multiple predictors, 128  
   simulations of, 126  
 regression coefficients, 119
- Missing values, TBI  
 outcomes, 142  
 predictors  
   adjusted effects estimation, 149–155  
   imputation model, 147, 149  
   imputed values distribution, 149  
   missingness patterns, 144–146  
   missingness quantification, 143–144  
   multivariable effects, 155  
   rank correlations, 147
- Modern selection methods  
 bagging and boosting, 208  
 Bayesian model averaging (BMA), 208–210  
 bootstrapping, 208  
 regression coefficients shrinkage, 210
- Multivariable analysis, non-linearity  
 fractional polynomials (FP), 221–222  
 generalized additive models (GAM), 220  
 restricted cubic splines (RCS), 220–221  
 splines in GAM, 222
- Multivariable diagnostic univariate log, 65
- Multivariable logistic regression coefficients, 249–250
- Multivariable regression coefficients  
 adaptation methods  
   clinical results, 250–251  
   description, 244  
   estimation, 245  
   improving calibration, 247–248  
   predictive performance, 247  
   simulation results, 245
- Bayesian methods  
 aneurysm study, 252  
 predicting neonatal death, 252  
 regression modelling, 251–252
- Multivariate additive regression splines (MARS), 110–111
- Myocardial infarction  
 age-related mortality, 103
- N**
- Naïve Bayes estimation  
 calibration, 65  
 prediction, 63–64
- Nerve-function impairment (NFI), 79–80
- Net benefit (NB), prediction models  
 definition, 284  
 discrimination and calibration, 289–290  
 formula and interpretation, 285  
 in decision curves, 285–286
- testicular cancer  
 decision curves, 287–288  
 model development and validation, 286–287
- Noise variables, 205
- Nomogram presentation, SMART study, 444–445
- Nomograms, prediction models, 314  
 benign histology, 317, 318  
 testicular cancer, 324, 328
- Non-linear prediction models  
 multivariable analysis  
   fractional polynomials (FP), 221–222  
   generalized additive models (GAM), 220  
   restricted cubic splines (RCS), 220–221  
   splines in GAM, 222  
 testicular cancer case study  
   analysis, univariate and multivariable, 224–226  
   fractional polynomials, multivariable, 224  
   LDH effect, 222–223  
   predictive performance, 226–227
- Non-significant variables, 193
- O**
- Odds ratios (OR), 422
- Odds threshold, prediction models, 283–285
- Oesophageal cancer  
 early mortality, 35  
 long-term mortality, 36  
 surgical mortality, 37
- Oesophagectomy, surgical mortality, 84
- Outcome and predictors  
 acute MI and GUSTO-I study, 416–417  
 SMART study, 429, 431
- Outcome, prediction models  
 cancer registry, 45–46  
 composite end points, 46  
 diagnostic end points, 47

- survival endpoints, 45  
types of, 46–47
- Overfitting and optimism, prediction models  
bootstrap resampling  
applications of, 95  
definition, 94  
optimism-corrected performance, 96–98  
regression coefficients, 95–96  
stepwise selection, 98–99
- data analysis cost  
outcome prediction, 99  
tree model and practical implications, 98
- definition, 83–84
- mortality variability  
noise *vs.* true heterogeneity, 85–87  
within single centre, 84–85
- regression models  
model performance, 90–91  
model uncertainty, 87–88  
optimal cut-point bias, 89  
optimism-corrected performance, 91  
parameter uncertainty, 90  
testimation bias, 88–89
- shrinkage  
mortality prediction, 87  
parameter uncertainty, 90
- P**
- Parsimony, 193
- Penalized coefficients, regression models  
formula, 323  
LDH values, 325  
post-chemotherapy histology, 322, 324
- Penalized maximum likelihood (PML)  
estimation  
Akaike's Information Criterion (AIC), 234–235  
effective degrees of freedom, 235  
shrinkage and selection methods, 238
- Percentile expected rank (PCER), 406, 407
- Performance measures, prediction model  
Brier score, 257  
calibration  
calibration-in-the-large model, 271  
discrimination and, 278  
estimation, 274–275  
goodness-of-fit tests, 274–276  
Hosmer–Lemeshow test, 276  
in testicular cancer, 276–277  
Kaplan–Meier method, 271  
outcomes and probabilities, 273–274
- regression coefficients, 272  
smoothing techniques, 270–271  
statistical tests, 274  
survival predictions, 276
- concordance statistic  
box plots and discrimination slope, 264  
Lorenz curve, 264–267  
R code, 269–270  
receiver operating characteristic (ROC) curve, 260–262  
sensitivity and specificity, 260  
survival data discrimination, 267–268  
testicular cancer discrimination, 268–269  
verification bias and discriminative ability, 269  
*vs.* explained variation ( $R^2$ ), 262–264
- discrimination and calibration, 259  
explained variation, 255–256  
for testicular cancer, 257–258  
survival outcomes, 258–259
- Poisson regression model, 80–81, 216
- Polytomous logistic regression analysis  
residual masses histology, 72–73  
*vs.* multivariable dichotomous logistic models, 73–75
- Positron-emission tomography (PET) scans, 47
- Predicted probabilities, regression models, 316  
benign histology, 319  
formula, 323  
necrosis, 319–320  
pre-and post-chemotherapy mass size, 321–322
- Prediction intervals, regression models, 316–317
- Predictive biomarkers, 47–48
- Predictor selection  
collinearity, 192  
non-significant variables, 193  
parsimony, 193  
reduction modelling, 191–192
- Predictors, prediction models  
categories, 40  
choice of, 44  
costs, 40–41  
determinants, 41  
prognosis in oncology, 41–42  
reliability  
biological variability, 43  
observer variability, 42  
regression dilution bias, 43–44  
strength, 39–40

- Presentation formats  
 decision rules  
   meta-model, 329–330  
   score chart rule, 328–329  
   survival analyses, 329
- prediction models  
   confidence intervals, 316–317  
   nomograms, 317–319  
   score charts, 319  
   specific formats, 321  
   tabular presentation, 320–321
- Prognosis  
   medicine, 1  
   oncology, 41–42
- Prognostic cohort study  
   nested case-control studies, 37–38  
   prospective designs, 35–38  
   registry data, 36–37  
   retrospective designs, 35  
   types, 33–35
- Prognostic factors, acute MI, 413–414, 419–420
- Proportional odds logistic regression  
   advantages of, 75–76  
   GOS dichotomous categorizations, 75
- Prostate cancer  
   fully independent external validation, 308, 309  
   prediction models, 284, 295
- Provider profiling, prediction models  
   case-mix adjustment, 400  
   centres' ranking, 401–402, 406–407  
   data differences between centres  
     analyses, adjusted and EB, 405–406  
     estimation of, 404–405  
     fixed effects and EB estimates for, 404  
     testing of, 403–404  
   ER and PCER in, 402, 407  
   guidelines for, 408  
   in stroke, 403  
   standard error (SE), 401  
   W statistic for, 400–401
- Public health, prediction models  
   breast cancer incidence, 12–13  
   preventive interventions, 12
- R**
- Random effects model  
   EB estimation, 306, 401  
   heterogeneity in, 402  
   predictor effects in, 398
- Randomized controlled trial (RCT)  
   different case-mix, 336
- inclusion and exclusion criteria, 341–342  
 reference values, 351  
 vs. surveys, 348
- Randomized controlled trials (RCTs)  
   covariate adjustment  
     linear regression analysis, 25–26  
     logistic regression, 25  
   inclusion and stratification, 23–24  
   prospective study, 36  
   traumatic brain injury, 24  
   treatment benefits and harms, 16–18
- Receiver operating characteristic (ROC)  
   curve, 282  
   and consecutive cutoffs, 260–261  
   confidence intervals, 262  
   probability outcomes, 261
- Recursive partitioning. *See* Classification and regression tree (CART) methods
- Reduction models, 191–192
- Reference values  
   performance with refitting, 350–351  
   R code, 350  
   testicular cancer and TBI, 351–352
- Regression models. *See also* Logistic regression model, external validation  
   confidence intervals, 316–318, 320  
   missed predictors  $Z$   
     heterogeneous case-mix, 344  
     more-or-less-severe case-mix, 342–344  
   observed predictors  $X$   
     heterogeneous case-mix, 341–342  
     more-or-less-severe case-mix, 340–341  
   penalized coefficients  
     formula, 323  
     LDH values, 325  
     post-chemotherapy histology, 322, 324  
   predicted probabilities  
     benign histology, 319  
     formula, 323  
     necrosis, 319–320  
     pre-and post-chemotherapy mass size, 321–322  
   regression coefficients  
     adjusted and unadjusted, 337–339, 343  
     calibration, 348  
     clinical usefulness, 349  
     defining risk factors, 328–329  
     linear predictor, 345–346  
     post-chemotherapy histology, 322  
     rescaled predictor, 326  
     testicular cancer, 324  
     variance of, 323  
   regression formula, 314

- heuristic shrinkage and uses, 316  
logistic model, 321–323  
prediction steps involved, 315–316  
shrunk coefficients, 323
- Regression models, additivity and interaction terms  
model specifications, 213–214  
potential interactions, 214–216  
time and survival, 216
- Renal artery stenosis, 14–15
- Respiratory syncytial virus (RSV), 54
- Restricted cubic spline (RCS), prediction models  
extrapolation and robustness, 166–167  
spline functions, 165–166  
vs. fractional polynomials, 164–165
- Restricted cubic splines (RCS), 220–221
- S**
- Second manifestations of arterial disease (SMART) study  
bootstrap procedure and model validation, 444–446  
case study, 457, 458  
coding of predictors  
combining similar effects predictors, 439–440  
Cox regression models for, 439  
creatinine transformations in, 440  
transforming continuous predictors, 438–439  
truncated extreme values for IMT, 437
- diastolic and systolic measurements, 434–435
- discrimination of, 444
- fatal and non-fatal vascular events in, 430–431
- model estimation  
Cox regression coefficients, 442–443  
Lasso model shrunk coefficients, 443
- model specification  
and hazard ratios (HRs), 441  
predictor selection, 442
- nomogram presentation, 446–447
- outcome and predictors, 431, 432
- prediction model development checklist, 432
- preliminary modeling steps  
cluster analysis, missing data, 436  
missing completely at random (MCAR) situation, 434  
missing data patterns in, 434–435  
multiple imputation techniques, 435–436
- prognosis in, 429–431  
research question and intended applications, 431  
study design and analysis, 432
- Shrinkage of regression coefficients  
characteristics, 232  
Lasso model, 239–240  
PML model, 238  
uniform shrinkage, 233
- Shrunk coefficients, regression models, 323
- Socio-economic status (SES), 37, 40
- Split-sample validation  
drawbacks of, 301–302  
vs. cross-validation, 302
- Statistical model prediction  
age-related mortality  
acute myocardial infarction, 105  
GUSTO-I data set, 106  
GUSTO-I modelling  
GAM and MARS, 111  
logistic regression and classification tree, 110  
predictive performance, 111–112
- model type, 104–105
- operative mortality  
and age effect, 106–107  
definition, 106  
logistic regression models, 108
- StatLog project, 109–110  
testing, 104
- Statistical modelling  
estimation problem and hypothesis testing, 2–3  
model uncertainty, 3  
sample size, 4–5
- Statistical prediction models  
binary outcomes  
Bayes rule, 61–62  
classification and regression tree methods, 67–70  
generalized additive model, 65–66  
likelihood ratio calculations, 62–63  
log likelihood scale, R<sup>2</sup> calculation, 58–60  
logistic regression analysis, 57–58  
MARS and SVM, 70–71  
multivariable diagnostic univariate log, 65  
Naïve Bayes estimation, 63–65
- categorical outcomes  
differential diagnoses, 71  
multivariable dichotomous logistic models, 73–75  
polytomous logistic regression, 72

- Statistical prediction models (*cont.*)
- residual masses histology, 72–73
  - continuous outcomes
    - cost prediction, 54
    - generalized additive model, 55–57
    - linear regression model, 53–54
    - transformation, 54–55
    - variance estimation, 55
  - economic outcomes, 54
  - estimation problem and hypothesis testing, 2–3
  - model uncertainty, 3
  - ordinal outcomes
    - continuation ratio model, 77
    - Glasgow outcome scale, 74–75
    - proportional odds logistic regression model, 75–77
  - sample size, 4–5
  - survival outcomes
    - Cox proportional hazards regression model, 77–78
    - Kaplan–Meier analysis, 79
    - nerve–function impairment, 79–80
    - parametric models, 80
    - proportionality assumption, 78–79
    - risky heart valves replacement, 80–81
- Stein's paradox, 90
- Stepwise selection, prediction models
- advantages of, 196
  - disadvantages of
    - biased coefficients, 199
    - event variable bias, 199–201
    - p*-values exaggeration, 204
    - selection instability, 197–199
    - variability misspecification, 201–204
  - stopping rules, 195–196
  - variants, 194–195
- Streptokinase (SK), 414
- Stroke, provider profiling, 403
- Support vector machine (SVM), 70–71
- T**
- Traumatic brain injury (TBI) trials, 24
- Temporal validation, 305–307
- Testicular cancer
- candidate predictors, 176–177
  - non-linearity case study
    - analysis, univariate and multivariable, 224–226
    - fractional polynomials, multivariable, 224
    - LDH effect, 222–223
    - predictive performance, 226–227
  - prediction model
- Brier score, 257
- calibration method, 276–277
- discriminative ability, 260
- receiver operating characteristic (ROC) curve, 260–262
- Testicular cancer case study, 457–459
- Testicular cancer, prediction models, 307–308
- clinical usefulness, 286–287
  - decision curves, 287–288
  - decision rule, 330
  - more–or less-severe patients, 349
  - nomogram, 324
  - performance, 289–290
  - pre–and post-chemotherapy mass size, 321–322, 326–327
  - predictor categorization, 327
  - presentation formats, 330
  - reference values, 351–352
  - regression formula, 321–323
  - score chart, 319, 324–325
  - tabular presentation, 320–321
  - validity interpretation, 341
- Thrombolytic therapy, acute MI
- predictions for selection of, 423–424
  - score chart for, 422–423
- TIMI-II models and GUSTO-I trial
- outcome differences in
    - centre-specific estimates, 395
    - intercepts updating, 397
    - logistic regression models, 394
    - R code for random effect analyses, 399
    - random effect models, 394, 395
    - standard error (SE), 396
  - predictor effects in
    - heterogeneity in calibration slope, 398–399
    - testing and updating, 398
- Tissue plasminogen activator (tPA), 414
- Transforming continuous predictors, 438–439
- Transportability. *See* Prediction model, generalizability
- Traumatic brain injury (TBI)
- adjusted effects estimation
    - complete predictors, 151–154
    - incomplete predictors, 154–155
    - logistic regression coefficients, 151
  - age and outcome, 108–109
  - continuous predictors
    - coding impact, 161
    - effects of, 171
  - Glasgow outcome scale (GOS)
    - imputation model, 147, 149
    - imputed values distribution, 149

- missing values, 142  
missingness patterns, 144–146  
patient selection, 140  
predictor coding and time dependency, 141–142  
predictors missingness quantification, 143–144  
predictors selection, 140–141  
glucose values and outcome, 168–169  
haematocrit (ht) and haemoglobin (Hb) correlation, 122  
multivariable analyses, 155  
Traumatic brain injury (TBI) data set, 461
- U**  
Univariate analyses, model specification advantages of, 207  
multivariable modelling, 206  
screening of, 207  
Univariate logistic regression coefficients, 248–249  
Updating prediction model centre-specific estimates and EB estimation, 395  
in GUSTO-I, 396–398  
intercept updating, 397  
predictor effects in
- heterogeneity in calibration slope, 398–399  
testing and updating of, 398
- V**  
Validation. *See also* Prediction models, external validity patterns case-mix, 337–338 missed predictors  $Z$  heterogeneous case-mix, 344 more-or-less-severe case-mix, 342–344 observed predictors  $X$  heterogeneous case-mix, 341–342 more-or-less-severe case-mix, 340–341 performance measures, 339 power calculations, 356 regression coefficients, 336–337, 341  
Validation, internal and external prediction models, 450–451  
Variance inflation factors (VIF). *See* Collinear models
- W**  
W statistic, differences between centres, 400–401