



THIRD EDITION

# STATISTICAL THINKING FOR NON-STATISTICIANS IN DRUG REGULATION

RICHARD KAY

WILEY Blackwell



# **Statistical Thinking for Non-Statisticians in Drug Regulation**



# **Statistical Thinking for Non-Statisticians in Drug Regulation**

**THIRD EDITION**

**Richard Kay, PhD**

Statistical Consultant, RK Statistics Ltd  
Honorary Visiting Professor, School of Pharmacy, Cardiff University, UK

**WILEY Blackwell**

This edition first published 2023  
© 2023 John Wiley & Sons Ltd

#### **Edition History**

First edition John Wiley & Sons, Ltd. (2007); Second edition John Wiley & Sons, Ltd. (2015)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Richard Kay to be identified as the author of this work has been asserted in accordance with law.

#### *Registered Offices*

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at [www.wiley.com](http://www.wiley.com).

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

**Trademarks:** Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

#### *Limit of Liability/Disclaimer of Warranty*

The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting scientific method, diagnosis, or treatment by physicians for any particular patient. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

#### *Library of Congress Cataloging-in-Publication Data*

Names: Kay, Richard, 1949– author.

Title: Statistical thinking for non-statisticians in drug regulation /

Richard Kay.

Description: Third edition. | Hoboken, NJ : Wiley-Blackwell, 2023. |

Includes bibliographical references and index.

Identifiers: LCCN 2022032464 (print) | LCCN 2022032465 (ebook) | ISBN

9781119867388 (hardback) | ISBN 9781119867395 (adobe pdf) | ISBN

9781119867401 (epub)

Subjects: MESH: Clinical Trials as Topic--methods | Drug Approval |

Statistics as Topic | Drug Industry

Classification: LCC RM301.27 (print) | LCC RM301.27 (ebook) | NLM QV

771.4 | DDC 615.1072/4–dc23/eng/20220916

LC record available at <https://lccn.loc.gov/2022032464>

LC ebook record available at <https://lccn.loc.gov/2022032465>

Cover Design: Wiley

Cover Image: © WHYFRAME/Shutterstock

Set in 9.5/13pt MeridienLTStd by Straive, Pondicherry, India

# Contents

Preface to the third edition,	xv
Preface to the second edition,	xvii
Preface to the first edition,	xix
Abbreviations,	xxiii
<b>1 Basic ideas in clinical trial design,</b>	<b>1</b>
1.1 Historical perspective,	1
1.2 Control groups,	2
1.3 Placebos and blinding,	3
1.4 Randomisation,	4
1.4.1 Unrestricted randomisation,	4
1.4.2 Block randomisation,	4
1.4.3 Unequal randomisation,	6
1.4.4 Stratified randomisation,	6
1.4.5 Central randomisation,	8
1.4.6 Dynamic allocation and minimisation,	9
1.4.7 Cluster randomisation,	10
1.5 Bias and precision,	10
1.6 Between- and within-patient designs,	12
1.7 Crossover trials,	13
1.8 Signal, noise and evidence,	14
1.8.1 Signal,	14
1.8.2 Noise,	15
1.8.3 Signal-to-noise ratio,	16
1.9 Confirmatory and exploratory trials,	16
1.10 Superiority, equivalence and non-inferiority trials,	17
1.11 Endpoint types,	18
1.12 Choice of endpoint,	20
1.12.1 Primary endpoints,	20
1.12.2 Secondary endpoints,	21
1.12.3 Surrogate endpoints,	21
1.12.4 Global assessment endpoints,	22
1.12.5 Composite endpoints,	23
1.12.6 Categorisation,	23

<b>2 Sampling and inferential statistics,</b>	25
2.1 Sample and population,	25
2.2 Sample statistics and population parameters,	26
2.2.1 Sample and population distribution,	26
2.2.2 Median and mean,	27
2.2.3 Standard deviation,	27
2.2.4 Notation,	28
2.2.5 Box plots,	29
2.3 The normal distribution,	30
2.4 Sampling and the standard error of the mean,	33
2.5 Standard errors more generally,	36
2.5.1 The standard error for the difference between two means,	38
2.5.2 Standard errors for proportions,	39
2.5.3 The general setting,	39
<b>3 Confidence intervals and <i>p</i>-values,</b>	40
3.1 Confidence intervals for a single mean,	40
3.1.1 The 95% confidence interval,	40
3.1.2 Changing the confidence coefficient,	42
3.1.3 Changing the multiplying constant,	42
3.1.4 The role of the standard error,	43
3.2 Confidence intervals for other parameters,	44
3.2.1 Difference between two means,	44
3.2.2 Confidence interval for proportions,	45
3.2.3 General case,	46
3.2.4 Bootstrap confidence interval,	47
3.3 Hypothesis testing,	47
3.3.1 Interpreting the <i>p</i> -value,	48
3.3.2 Calculating the <i>p</i> -value,	50
3.3.3 A common process,	53
3.3.4 The language of statistical significance,	56
3.3.5 One-sided and two-sided tests,	56
<b>4 Tests for simple treatment comparisons,</b>	58
4.1 The unpaired t-test,	58
4.2 The paired t-test,	59
4.3 Interpreting the t-tests,	62
4.4 The chi-square test for binary endpoints,	63
4.4.1 Pearson chi-square,	63
4.4.2 The link to a ratio of the signal to the standard error,	66
4.5 Measures of treatment benefit,	66
4.5.1 Odds ratio,	67
4.5.2 Relative risk,	67
4.5.3 Relative and absolute risk reduction,	68
4.5.4 Number needed to treat,	69

4.5.5	Confidence intervals,	69
4.5.6	Interpretation,	71
4.6	Fisher's exact test,	72
4.7	Tests for categorical and ordered categorical endpoints,	73
4.7.1	Categorical endpoints,	73
4.7.2	Ordered categorical (ordinal) endpoints,	75
4.7.3	Measures of treatment benefit,	76
4.8	Count endpoints,	77
4.9	Extensions for multiple treatment groups,	77
4.9.1	Continuous endpoints,	77
4.9.2	Binary, categorical and ordered categorical endpoints,	78
4.9.3	Dose-ranging studies,	79
4.9.4	Further discussion,	79
5	Adjusting the analysis,	80
5.1	Objectives for adjusted analysis,	80
5.2	Comparing treatments for continuous endpoints,	80
5.3	Least squares means,	84
5.4	Evaluating the homogeneity of the treatment effect,	85
5.4.1	Treatment-by-factor interactions,	85
5.4.2	Quantitative and qualitative interactions,	87
5.5	Methods for binary and ordered categorical endpoints,	88
5.6	Multi-centre trials,	89
5.6.1	Adjusting for centre,	89
5.6.2	Significant treatment-by-centre interactions,	89
5.6.3	Combining centres,	90
6	Regression and analysis of covariance,	92
6.1	Adjusting for baseline factors,	92
6.2	Simple linear regression,	92
6.3	Multiple regression,	95
6.4	Logistic regression for binary endpoints,	97
6.4.1	Negative binomial regression for count endpoints,	97
6.5	Analysis of covariance for continuous outcomes,	98
6.5.1	Main effect of treatment,	98
6.5.2	Treatment-by-covariate interactions,	100
6.5.3	A single model,	102
6.5.4	Connection with adjusted analyses,	102
6.5.5	Advantages of ANCOVA,	103
6.5.6	Least squares means,	104
6.5.7	Random element,	105
6.6	Other endpoint types,	105
6.6.1	Binary endpoints and extensions,	105
6.6.2	Count endpoints,	108
6.7	Mixed models,	109

6.8	Regulatory aspects of the use of covariates,	110
6.9	Baseline testing,	113
6.10	Correlation and regression,	113
<b>7</b>	Intention-to-treat, analysis sets and missing data,	115
7.1	The principle of intention-to-treat,	115
7.2	The practice of intention-to-treat,	119
7.2.1	Full analysis set,	119
7.2.2	Per-protocol set,	120
7.2.3	Further aspects of ITT,	120
7.3	Missing data,	121
7.3.1	Introduction,	121
7.3.2	Complete cases analysis,	122
7.3.3	Last observation carried forward (LOCF),	123
7.3.4	Baseline observation carried forward (BOCF),	123
7.3.5	Success/failure classification,	123
7.3.6	Worst-case/best-case classification,	124
7.3.7	Sensitivity,	124
7.3.8	Avoidance of missing data,	125
7.3.9	Classification of missing data,	126
7.3.10	Multiple imputation,	127
7.4	Intention-to-treat and time-to-event data,	129
7.5	General questions and considerations,	131
<b>8</b>	Estimands,	134
8.1	ICH E9 (R1)	134
8.2	Attributes of an estimand,	134
8.2.1	Population,	135
8.2.2	Variable,	135
8.2.3	Intercurrent event (ICE),	136
8.2.4	Statistic for treatment effect,	136
8.3	Estimand strategies,	136
8.3.1	Five strategies,	136
8.3.2	Treatment policy, composite and hypothetical strategies,	137
8.3.3	While on treatment,	139
8.3.4	Principal stratification,	139
8.4	Sensitivity and supplementary analyses,	141
8.4.1	Main estimator,	141
8.4.2	Sensitivity analyses,	142
8.4.3	Supplementary analyses,	143
<b>9</b>	Power, sample size and clinical relevance,	144
9.1	Type I and type II errors,	144
9.2	Power,	145
9.3	Calculating sample size,	148
9.4	Impact of changing the parameters,	152
9.4.1	Standard deviation,	152

---

9.4.2	Event rate in the control group,	152
9.4.3	Clinically relevant difference,	153
9.5	Regulatory aspects,	154
9.5.1	Power $\geq 80\%$ ,	154
9.5.2	Sample size adjustment,	154
9.6	Reporting the sample size calculation,	155
9.7	Post hoc power,	156
9.8	Link between <i>p</i> -values and confidence intervals,	157
9.9	Confidence intervals for clinical importance,	159
9.10	Misinterpretation of the <i>p</i> -value,	160
9.10.1	Conclusions of similarity,	160
9.10.2	The problem with 0.05,	161
9.11	Single pivotal trial and 0.05,	161
<b>10</b>	Multiple testing,	164
10.1	Inflation of the type I error,	164
10.1.1	False positives,	164
10.1.2	A simulated trial,	164
10.2	How does multiplicity arise?,	165
10.3	Regulatory and scientific view,	166
10.4	Methods for adjustment,	167
10.4.1	Bonferroni correction,	167
10.4.2	Holm correction,	168
10.4.3	Hochberg correction,	169
10.4.4	Interim analyses,	170
10.5	Avoiding adjustment,	171
10.5.1	Co-primary endpoints,	171
10.5.2	Composite endpoints,	172
10.5.3	Hierarchical testing,	173
10.6	Fallback procedure,	176
10.7	Multiple comparisons of treatments,	177
10.8	Subgroup testing,	178
10.9	Other aspects of multiplicity,	181
10.9.1	Using different statistical tests,	181
10.9.2	Different analysis sets and methods for missing data,	182
10.9.3	Pre-planning,	182
10.9.4	Nominal significance,	183
<b>11</b>	Non-parametric and related methods,	184
11.1	Assumptions underlying the t-tests and their extensions,	184
11.2	Homogeneity of variance,	184
11.3	The assumption of normality,	185
11.4	Non-normality and transformations,	187
11.5	Non-parametric tests,	190
11.5.1	The Mann–Whitney U-test,	190

11.5.2	The Wilcoxon signed rank test,	192
11.5.3	General comments,	193
11.6	Advantages and disadvantages of non-parametric methods,	194
11.7	Outliers,	195
<b>12</b>	Equivalence and non-inferiority,	196
12.1	Demonstrating similarity,	196
12.2	Confidence intervals for equivalence,	198
12.3	Confidence intervals for non-inferiority,	199
12.4	A <i>p</i> -value approach,	201
12.5	Assay sensitivity,	202
12.6	Analysis sets,	204
12.7	The choice of $\Delta$ ,	205
12.7.1	Bioequivalence,	206
12.7.2	Therapeutic equivalence, biosimilars,	206
12.7.3	Non-inferiority,	207
12.7.4	The 10% rule for cure rates,	208
12.7.5	The synthesis method,	209
12.8	Biocreep and constancy,	210
12.9	Sample size calculations,	211
12.10	Switching between non-inferiority and superiority,	213
12.11	Biosimilars,	215
<b>13</b>	The analysis of survival data,	217
13.1	Time-to-event data and censoring,	217
13.2	Kaplan-Meier curves,	218
13.2.1	Plotting Kaplan-Meier curves,	218
13.2.2	Event rates and relative risk,	220
13.2.3	Median event times,	221
13.3	Treatment comparisons,	222
13.4	The hazard ratio,	225
13.4.1	The hazard rate,	225
13.4.2	Constant hazard ratio,	225
13.4.3	Non-constant hazard ratio,	226
13.4.4	Link to survival curves,	227
13.4.5	Calculating Kaplan-Meier curves,	228
13.5	Restricted mean survival time,	229
13.6	Adjusted analyses,	230
13.6.1	Stratified methods,	230
13.6.2	Proportional hazards regression,	231
13.6.3	Accelerated failure time model,	232
13.7	Independent censoring,	233
13.8	Crossover,	234
13.8.1	Rank Preserving Structural Failure Time Model,	234
13.8.2	Regulatory position,	236

- 
- 13.9 Composite time-to-event endpoints, 237
    - 13.9.1 Cumulative incidence functions, 237
    - 13.9.2 Regulatory position, 238
  - 13.10 Sample size calculations, 241
- 14** Interim analysis and data monitoring committees, 243
- 14.1 Stopping rules for interim analysis, 243
  - 14.2 Stopping for efficacy and futility, 245
    - 14.2.1 Efficacy, 245
    - 14.2.2 Futility and conditional power, 246
    - 14.2.3 Some practical issues, 247
    - 14.2.4 Point estimates and confidence intervals, 249
  - 14.3 Monitoring safety, 249
  - 14.4 Data monitoring committees, 250
    - 14.4.1 Introduction and responsibilities, 250
    - 14.4.2 Structure and process, 252
    - 14.4.3 Meetings and recommendations, 253
- 15** Bayesian statistics, 255
- 15.1 Introduction, 255
  - 15.2 Prior and posterior distributions, 256
    - 15.2.1 Prior beliefs, 256
    - 15.2.2 Prior to posterior, 256
    - 15.2.3 Bayes theorem, 258
  - 15.3 Bayesian inference, 259
    - 15.3.1 Frequentist methods, 259
    - 15.3.2 Posterior probabilities, 260
    - 15.3.3 Credible intervals, 261
  - 15.4 Case study, 262
  - 15.5 History and regulatory acceptance, 263
  - 15.6 Discussion, 265
- 16** Adaptive designs, 266
- 16.1 What are adaptive designs?, 266
    - 16.1.1 Advantages and drawbacks, 266
    - 16.1.2 Restricted adaptations, 267
    - 16.1.3 Flexible adaptations, 268
  - 16.2 Minimising bias, 268
    - 16.2.1 Control of type I error, 268
    - 16.2.2 Estimation, 271
    - 16.2.3 Operational bias, 272
  - 16.3 Unblinded sample size re-estimation, 273
    - 16.3.1 Product of  $p$ -values, 273
    - 16.3.2 Weighting the two parts of the trial, 274
    - 16.3.3 Rationale, 275
  - 16.4 Seamless phase II/III studies, 275

16.4.1	Standard framework,	275
16.4.2	Multiplicity,	276
16.4.3	Incorporating the phase II data,	277
16.4.4	Logistical challenges,	278
16.5	Other types of adaptation,	278
16.5.1	Changing the primary endpoint,	278
16.5.2	Enrichment,	279
16.5.3	Dropping the placebo arm in a non-inferiority trial,	280
16.6	Further regulatory considerations,	281
<b>17</b>	Observational studies,	283
17.1	Introduction,	283
17.1.1	Non-randomised comparisons,	283
17.1.2	Study types,	284
17.1.3	Sources of bias,	285
17.1.4	An empirical investigation,	286
17.1.5	Selection bias in concurrently controlled studies,	287
17.1.6	Selection bias in historically controlled studies,	288
17.1.7	Some conclusions,	289
17.2	Guidance on design, conduct and analysis,	290
17.2.1	Regulatory guidance,	290
17.2.2	Strengthening the Reporting of Observational Studies in Epidemiology,	291
17.3	Assessing the presence of baseline balance,	291
17.4	Adjusting for selection bias: stratification and regression,	292
17.5	Adjusting for selection bias: propensity scoring,	293
17.5.1	Defining propensity scores,	293
17.5.2	Propensity score stratification, regression and matching,	295
17.6	Comparing methods that correct for selection bias,	297
17.7	Inverse propensity score weighting,	300
17.8	Case-control studies,	302
17.8.1	Background,	302
17.8.2	Odds ratio and relative risk,	304
<b>18</b>	Meta-analysis and network meta-analysis,	306
18.1	Definition,	306
18.2	Objectives,	307
18.3	Statistical methodology,	309
18.3.1	Methods for combination,	309
18.3.2	CIs,	310
18.3.3	Fixed and random effects,	310
18.3.4	Graphical methods,	310
18.3.5	Detecting heterogeneity,	312
18.3.6	Robustness,	313
18.3.7	Rare events,	313
18.3.8	Individual patient data,	314

- 18.4 Case study, 314
  - 18.5 Ensuring scientific validity, 316
    - 18.5.1 Planning, 316
    - 18.5.2 Assessing the risk of bias, 317
    - 18.5.3 Publication bias and funnel plots, 318
    - 18.5.4 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), 320
  - 18.6 Regulatory aspects of meta-analysis, 320
  - 18.7 Introduction to network meta-analysis, 321
  - 18.8 Case Study, 322
  - 18.9 Indirect treatment comparisons, 323
    - 18.9.1 Cross-trial calculations, 323
    - 18.9.2 Effect modifiers, 324
    - 18.9.3 Critique, 325
  - 18.10 Bayesian rank analysis, 326
- 19** Methods for safety analysis, safety monitoring and assessment of benefit-risk, 328
- 19.1 Introduction, 328
    - 19.1.1 Methods for safety data, 328
    - 19.1.2 The rule of three, 329
  - 19.2 Routine evaluation in clinical studies, 330
    - 19.2.1 Types of data, 331
    - 19.2.2 Adverse events, 332
    - 19.2.3 Laboratory data, 335
    - 19.2.4 ECG data, 338
    - 19.2.5 Vital signs, 339
    - 19.2.6 Safety summary across trials, 339
    - 19.2.7 Specific safety studies, 340
  - 19.3 Data monitoring committees, 340
  - 19.4 Assessing benefit-risk, 341
    - 19.4.1 Current approaches, 341
    - 19.4.2 Multi-criteria decision analysis, 342
    - 19.4.3 Quality-adjusted time without symptoms or toxicity, 348
  - 19.5 Pharmacovigilance, 350
    - 19.5.1 Post-approval safety monitoring, 350
    - 19.5.2 Proportional reporting ratios, 351
    - 19.5.3 Bayesian neural networks, 354
- 20** Diagnosis, 356
- 20.1 Introduction, 356
  - 20.2 Measures of diagnostic performance, 357
    - 20.2.1 Sensitivity and specificity, 357
    - 20.2.2 Positive and negative predictive value, 358
    - 20.2.3 False positive and false negative rates, 358
    - 20.2.4 Prevalence, 358

20.2.5 Likelihood ratio,	359
20.2.6 Predictive accuracy,	359
20.2.7 Choosing the correct cutpoint,	360
20.3 Receiver operating characteristic curves,	360
20.3.1 Receiver operating characteristic,	360
20.3.2 Comparing ROC curves,	361
20.4 Diagnostic performance using regression models,	362
20.5 Aspects of trial design for diagnostic agents,	364
20.6 Assessing agreement,	365
20.6.1 The kappa statistic,	365
20.6.2 Other applications for kappa,	367
20.7 Companion diagnostics,	367
<b>21</b> The role of statistics and statisticians,	370
21.1 The importance of statistical thinking at the design stage,	370
21.2 Regulatory guidelines,	371
21.3 The statistics process,	372
21.3.1 The statistical methods section of the protocol,	372
21.3.2 The statistical analysis plan,	373
21.3.3 The data validation plan,	373
21.3.4 The blind review,	374
21.3.5 Statistical analysis,	374
21.3.6 Reporting the analysis,	375
21.3.7 Pre-planning,	375
21.3.8 Sensitivity and robustness,	378
21.4 The regulatory submission,	378
21.5 Publications and presentations,	379
References,	383
Index,	394

# Preface to the third edition

There have been numerous regulatory guidelines issued since the publication of the second edition of this book, and many of these impact the application of statistics to our clinical investigations and the development of new medicines and devices. Outside of the regulatory arena, the introduction of new methodologies also feeds through to affect the application of statistics in pharmaceutical medicine. The key milestone from a regulatory guideline perspective has been the publication of the ICH E9 Addendum on Estimands. The need for this guideline has grown out of concerns from regulators and the statistical community more widely regarding several elements of the way we design our trials and analyse our data. Primarily, these concerns have been driven by the developing sophistication and availability of methods for dealing with missing data and the use of these methods without really understanding their purpose or properties in relation to the clinical question being addressed. In addition, there have been concerns regarding how we use data collected following events such as withdrawal of consent, taking rescue medication, withdrawal from trial medication due to serious adverse events and also the role of sensitivity analyses, which have often been seen as simply a tick-box exercise that we undertake without really understanding what questions they are addressing. The estimand framework starts with the clinical question(s) of interest and sets down methods of statistical analysis that address those questions in a systematic and structured way – and this leads to a more considered position on the aspects that I have mentioned. A new chapter (Chapter 8) in this third edition covers estimands.

Both the EMA and the FDA have revised (in the case of EMA) and introduced (in the case of the FDA) guidelines on multiplicity. Multiple testing remains a major concern within the regulatory framework, and these guidelines provide additional insight into the issues and methodologies for dealing with the problems. We have also seen a shift in the way journal editors view the issues associated with multiplicity, with tighter control on the reporting of  $p$ -values. The chapter on multiplicity (Chapter 10) has been revised to incorporate developments in this area. We have also seen the publication by CPMP of a guideline on the evaluation of subgroups. Multiplicity plays a role in the way we consider subgroups, and discussions that relate to this topic are also included in the revised Chapter 10.

In 2016, the FDA updated their guideline on non-inferiority, and some of the points raised in that revision have been incorporated in Chapter 12. I have also introduced a specific section on biosimilars (Section 12.11). This in part is

reflective of my own experience in that area, which to a large extent is a consequence of the increasing development of biosimilars.

Three new sections in the chapter on the analysis of survival data (Chapter 13) cover restricted mean survival time, cross-over in oncology studies and cumulative incidence functions. The restricted mean survival time is an additional summary measure calculated from the Kaplan-Meier curve and can be the basis for a comparison of those curves when the assumption of proportional hazards is violated. This topic is covered in Section 13.5. Cross-over occurs in patients who progress in the control arm and are offered a switch to the experimental treatment. If the experimental treatment has efficacy, this switch will likely impact overall survival in the context of comparing the pure effect of the experimental treatment. This topic is covered in Section 13.8. Cumulative incidence functions, now discussed in Section 13.9, are used for composite time-to-event endpoints and allow the evaluation of the separate components of the composite.

The use of network meta-analysis for incorporating indirect non-randomised evidence in treatment comparisons has increased in recent years. This is not primarily in regulatory submissions but more in the context of health technology assessment and promotional activities. This topic is introduced and discussed in Chapter 18.

I would like to thank readers who have provided feedback to the first two editions of this book. Please do continue to give your feedback and pull me up on inconsistencies, mistakes and, indeed, areas where you disagree with the position I have taken. I very much learn from these interactions.

**Richard Kay**

Bakewell

May 2022

# Preface to the second edition

The first edition of this book was submitted for publication over seven years ago. As predicted, there have been numerous developments, both in the world of pharmaceutical statistics and within the regulatory environment, which need to be presented and explained. In the intervening years, the FDA has published guidelines on Adaptive Designs and on Non-Inferiority Trials. The CHMP have updated their guidance on Missing Data, produced a guideline on the Clinical Evaluation of Diagnostic Agents and have undertaken a major exercise on the evaluation of the benefit–risk balance. In recent years, we have seen greater application of the use of observational studies within regulatory applications, particularly when dealing with orphan drugs, and also a willingness to consider the use of Bayesian methods. Although there is still considerable concern expressed about various elements of adaptive designs, we are beginning to understand where the boundaries are and what can be done without compromising the scientific integrity of the study. All of these aspects and more have led to me write this new edition.

There are five new chapters and several other chapters have been restructured. Chapter 15 looks at Bayesian Statistics, contrasting the Bayesian methodology with classical methods, presenting the advantages and concerns and discussing the use of these methods in pharmaceutical applications. Adaptive designs are discussed in Chapter 16 with sections on minimising bias and covering various types of adaptations. Some practicalities and recommendations regarding the circumstances under which such designs can be considered are also presented. We all recognise that the randomised controlled trial is the gold standard in terms of evaluating the efficacy and also the safety of a new medicine, but in some settings running such trials is not possible. Observational studies offer an alternative, but their ability to provide valid conclusions is heavily dependent on good design and conduct and careful analysis. Such designs are discussed in Chapter 17. In recent years, we have become much more formal in the statistical evaluation of safety data. Chapter 19 covers various aspects of safety data analysis, including the use of graphical methods, and goes on to detail potential approaches to the quantification of the benefit–risk balance both inside and outside of the regulatory submission. This chapter concludes with a discussion on methods for post-approval safety monitoring. Statistical methods for the evaluation of diagnostic methods and method comparison are discussed in Chapter 20.

Chapter 5 in the first edition of the book was entitled ‘Multi-Centre Trials’. This chapter has been restructured and is now entitled ‘Adjusting the Analysis’. This restructuring is based on my recent teaching experiences in explaining the rationale around adjusting the statistical analysis for baseline factors in a general way. Several other chapters contain sections similarly restructured as a result of my teaching experience. I hope that these changes make things clearer than maybe they were before. Chapter 18 entitled ‘Meta-Analysis’ is a major restructuring of the initial chapter on this topic and reflects the current uses of this methodology within the pharmaceutical industry.

I have received many encouraging comments on the first edition of this book and I would like to thank everyone who has given feedback. I hope that this revision is also well-received. My aim is to make statistical thinking and methods used within the pharmaceutical industry accessible to non-statisticians so that they are better able to communicate using statistical language, better able to understand statistical methods used in reports and publications and what can and cannot be concluded from the resulting analyses, and are better equipped to contribute to statistical arguments used within regulatory submissions and beyond. I continue to teach courses on statistics for non-statisticians, and many of the changes and additions that I make to my teaching materials and have made to this book have come out of my experiences on those courses. I would like to thank my students for their challenging questions; they make me think about better ways to explain things. Finally I would like to thank all of those who got in touch to point out mistakes in the first edition. I have corrected those but some may remain and indeed the new material may contain some more. I encourage the reader to provide feedback for this second edition and please do not hesitate to point out any mistakes, for which I am solely responsible.

**Richard Kay**

Great Longstone

March 2014

# Preface to the first edition

This book is primarily concerned with clinical trials planned and conducted within the pharmaceutical industry. Much of the methodology presented is in fact applicable on a broader basis and can be used in observational studies and in clinical trials outside of the pharmaceutical sector; nonetheless, the primary context is clinical trials and pharmaceuticals. The development is aimed at non-statisticians and will be suitable for physicians, investigators, clinical research scientists, medical writers, regulatory personnel, statistical programmers, senior data managers and those working in quality assurance. Statisticians moving from other areas of application outside of pharmaceuticals may also find the book useful in that it places the methods that they are familiar with, in context in their new environment. There is substantial coverage of regulatory aspects of drug registration that impact on statistical issues. Those of us working within the pharmaceutical industry recognise the importance of being familiar with the rules and regulations that govern our activities, and statistics is a key aspect of this.

The aim of the book is not to turn non-statisticians into statisticians. I do not want you to go away from this book and ‘do’ statistics. It is the job of the statistician to provide statistical input to the development plan, to individual protocols, to write the statistical analysis plan, to analyse the data and to work with medical writing in producing the clinical report, and also to support the company in its interactions with regulators on statistical issues.

The aims of the book are really threefold. Firstly, to aid communication between statisticians and non-statisticians; secondly, to help in the critical review of reports and publications; and finally, to enable the more effective use of statistical arguments within the regulatory process. We will take each of these points in turn.

In many situations, the interaction between a statistician and a non-statistician is not a particularly successful one. The statistician uses terms such as power, odds ratio,  $p$ -value, full analysis set, hazard ratio, non-inferiority, type II error, geometric mean, last observation carried forward and so on, of which the non-statistician has a vague understanding, but maybe not a good enough understanding to be able to get an awful lot out of such interactions. Of course, it is always the job of a statistician to educate and every opportunity should be taken for imparting knowledge about statistics, but in a specific context, there may not be time for that. Hopefully this book will explain, in ways that are understandable, just what these terms mean and provide some insight into their interpretation and the context in which they are used. There is also a lot of

confusion between what on the surface appear to be the same or similar things: significance level and *p*-value, equivalence and non-inferiority, odds ratio and relative risk, relative risk and hazard ratio (by the way this is a minefield!) and meta-analysis and pooling to name just a few. This book will clarify these important distinctions.

It is unfortunately the case that many publications, including some in leading journals, contain mistakes with regard to statistics. Things have improved over the years with the standardisation of the ways in which publications are put together and reviewed. For example, the CONSORT statement (see Section 16.5 [this is Section 21.5 in the 2nd edition]) has led to a distinct improvement in the quality of reporting. Nonetheless mistakes do slip through, in terms of poor design, incorrect analysis, incomplete reporting and inappropriate interpretation – hopefully not all at once! It is important therefore when reading an article that the non-statistical reader is able to make a judgement regarding the quality of the statistics and to notice any obvious flaws that may undermine the conclusions that have been drawn. Ideally, the non-statistician should involve their statistical colleagues in evaluating their concerns, but keeping a keen eye on statistical arguments within the publication may help to alert the non-statistician to a potential problem. The same applies to presentations at conferences, posters, advertising materials and so on.

Finally, the basis of many concerns raised by regulators, when they are reviewing a proposed development plan or assessing an application for regulatory approval, is statistical. It is important that non-statisticians are able to work with their statistical colleagues in correcting mistakes, changing aspects of the design, responding to questions about the data to hopefully overcome those concerns.

In writing this book, I have made the assumption that the reader is familiar with the general aspects of the drug development process. I have assumed knowledge of the phase I to phase IV framework, of placebos, control groups, and double-dummy together with other fundamental elements of the nuts and bolts of clinical trials. I have assumed however no knowledge of statistics! This may or may not be the correct assumption in individual cases, but it is the common denominator that we must start from, and also it is actually not a bad thing to refresh on the basics. The book starts with some basic issues in trial design in Chapter 1, and I guess most people picking up this book will be familiar with many of the topics covered there. But don't be tempted to skip this chapter; there are still certain issues, raised in this first chapter, that will be new and important for understanding arguments put forward in subsequent chapters. Chapter 2 looks at sampling and inferential statistics. In this chapter, we look at the interplay between the population and the sample, basic thoughts on measuring average and variability and then explore the process of sampling leading to the concept of the standard error as a way of capturing precision/reliability of the sampling process. The construction and interpretation of confidence intervals

are covered in Chapter 3 together with testing hypotheses and the (dreaded!) *p*-value. Common statistical tests for various data types are developed in Chapter 4 which also covers different ways of measuring treatment effect for binary data, such as the odds ratio and relative risk.

Many clinical trials that we conduct are multi-centre and Chapter 5 looks at how we extend our simple statistical comparisons to this more complex structure. These ideas lead naturally to the topics in Chapter 6 which include the concepts of adjusted analyses, and more generally, analysis of covariance which allows adjustment for many baseline factors, not just centre. Chapters 2–6 follow a logical development sequence in which the basic building blocks are initially put in place and then used to deal with more and more complex data structures. Chapter 7 moves a little away from this development path and covers the important topic of ‘intention-to-treat’ and aspects of conforming with that principle through the definition of different analysis sets and dealing with missing data. In Chapter 8, we cover the very important design topics of power and the sample size calculation which then leads naturally to a discussion about the distinction between statistical significance and clinical importance in Chapter 9.

The regulatory authorities, in my experience, tend to dig their heels in on certain issues and one such issue is multiplicity. This topic, which has many facets, is discussed in detail in Chapter 10. Non-parametric and related methods are covered in Chapter 11. In Chapter 12, we develop the concepts behind the establishment of equivalence and non-inferiority. This is an area where many mistakes are made in applications, and in many cases, these slip through into published articles. It is a source of great concern to many statisticians that there is widespread misunderstanding of how to deal with equivalence and non-inferiority. I hope that this chapter helps to develop a better understanding of the methods and the issues. If you have survived so far, then Chapter 13 covers the analysis of survival data. When an endpoint is time to some event, for example, death, the data are inevitably subject to what we call censoring and it is this aspect of so-called survival data that has led to the development of a completely separate set of statistical methods. Chapter 14 builds on the earlier discussion on multiplicity to cover one particular manifestation of that, the interim analysis. This chapter also looks at the management of these interim looks at the data through data monitoring committees. Meta-analysis and its role in clinical development is covered in Chapter 15, and the book finishes with a general Chapter 16 on the role of statistics and statisticians in terms of the various aspects of design and analysis and statistical thinking more generally.

It should be clear from the last few paragraphs that the book is organised in a logical way; it is a book for learning rather than a reference book for dipping into. The development in later chapters will build on the development in earlier chapters. I strongly recommend, therefore, that you start on page 1 and work through. I have tried to keep the discussion away from formal mathematics. There are formulas in the book but I have only included these where I think this

will enhance understanding; there are no formulas for formulas sake! There are some sections that are more challenging than others and I have marked with an asterisk those sections that can be safely sidestepped on a first (or even a second) run through the book.

The world of statistics is ever changing. New methods are being developed by theoreticians within university departments, and ultimately some of these will find their way into mainstream methods for design and statistical analysis within our industry. The regulatory environment is ever changing as regulators respond to increasing demands for new and more effective medicines. This book in one sense represents a snapshot in time in terms of what statistical methods are employed within the pharmaceutical industry and also in relation to current regulatory requirements. Two statistical topics that are not included in this book are Bayesian Methods and Adaptive (Flexible) Designs (although some brief mention is made of this latter topic in Section 14.5.2). Both areas are receiving considerable attention at the moment, and I am sure that within a fairly short period of time, there will be much to say about them in terms of the methodological thinking, examples of their application and possibly with regard to their regulatory acceptance but for the moment they are excluded from our discussions.

The book has largely come out of courses that I have been running under the general heading of ‘Statistical Thinking for Non-Statisticians’ for a number of years. There have been several people who have contributed from time to time and I would like to thank them for their input and support: Werner Wierich, Mike Bradburn and in particular Ann Gibb who gave these courses with me over a period of several years and enhanced my understanding through lively discussion and asking many challenging questions. I would also like to thank Simon Gillis who contributed to Chapter 16 [this is Chapter 21 in the 2nd edition] with his much deeper knowledge of the processes that go on within a pharmaceutical company in relation to the analysis and reporting of a clinical trial.

**Richard Kay**  
Great Longstone  
January 2007

# Abbreviations

<b>6MWD</b>	Six minute walking distance
<b>ADR</b>	adverse drug reaction
<b>AE</b>	adverse event
<b>AFT</b>	accelerated failure time
<b>AIDAC</b>	Anti-Infective Drugs Advisory Committee
<b>ALKPH</b>	alkaline phosphatase
<b>ALT</b>	alanine transaminase
<b>AMD</b>	age-related macular degeneration
<b>ANCOVA</b>	analysis of covariance
<b>ANOVA</b>	analysis of variance
<b>ARR</b>	absolute risk reduction
<b>AST</b>	aspartate transaminase
<b>AUC</b>	area under the curve
<b>BILTOT</b>	total bilirubin
<b>BMD</b>	bone mineral density
<b>BSC</b>	best supportive care
<b>CDER</b>	Center for Drug Evaluation and Research
<b>CFC</b>	Chlorofluorocarbon
<b>CHMP</b>	Committee for Medicinal Products for Human Use
<b>CI</b>	confidence interval
<b>CIF</b>	cumulative incidence function
<b>CMAX</b>	maximum concentration
<b>CMH</b>	Cochran-Mantel-Haenszel
<b>CNS</b>	central nervous system
<b>COPD</b>	chronic obstructive pulmonary disease
<b>CPMP</b>	Committee for Proprietary Medicinal Products
<b>CR</b>	complete response
<b>crd</b>	clinically relevant difference
<b>CRF</b>	Case Report Form
<b>CRO</b>	clinical research organisation
<b>CSR</b>	Clinical Study Report
<b>CTC(AE)</b>	Common Terminology Criteria (for Adverse Events)
<b>dBp</b>	diastolic blood pressure
<b>df</b>	degrees of freedom
<b>DILI</b>	drug-induced liver injury
<b>DLQI</b>	dermatology life quality index
<b>DMC</b>	Data Monitoring Committee
<b>DSMB</b>	Data and Safety Monitoring Board

<b>DSMC</b>	Data and Safety Monitoring Committee
<b>EBGM</b>	empirical Bayes geometric mean
<b>ECG</b>	Electrocardiogram
<b>ECOG</b>	Eastern Cooperative Oncology Group
<b>EMEA</b>	European Medicines Evaluation Agency
<b>ESRD</b>	end-stage renal disease
<b>FAS</b>	full analysis set
<b>FDA</b>	Food and Drug Administration
<b>FEV<sub>1</sub></b>	forced expiratory volume in one second
<b>FN(R)</b>	false negative (rate)
<b>FP(R)</b>	false positive (rate)
<b>FWER</b>	family-wise error rate
<b>GP</b>	General Practitioner
<b>HAMA</b>	Hamilton Anxiety Scale
<b>HAMD</b>	Hamilton Depression Scale
<b>HER2</b>	human epidermal growth factor receptor-2
<b>HIV</b>	human immunodeficiency virus
<b>HR</b>	Hazard Ratio
<b>HTA</b>	health technology assessment
<b>IC</b>	information component
<b>ICE</b>	intercurrent event
<b>ICH</b>	International Committee on Harmonisation
<b>ICU</b>	intensive care unit
<b>IPSW</b>	inverse propensity score weighting
<b>ISPOR</b>	International Society for Pharmacoeconomics and Outcomes Research
<b>ITT</b>	intention-to-treat
<b>IVRS</b>	Interactive Voice Response System
<b>IWRS</b>	Interactive Web Response System
<b>KM</b>	Kaplan-Meier
<b>LLN</b>	lower limit of normal
<b>LOCF</b>	last observation carried forward
<b>LR</b>	likelihood ratio
<b>MACE</b>	major cardiovascular event(s)
<b>MAR</b>	missing at random
<b>MCAR</b>	missing completely at random
<b>MCDA</b>	multi-criteria decision analysis
<b>mCRC</b>	metastatic colorectal cancer
<b>MedDRA</b>	Medical Dictionary for Regulatory Activities
<b>MFS</b>	metastases-free survival
<b>MH</b>	Mantel-Haenszel
<b>MI</b>	myocardial infarction
<b>mITT</b>	modified intention-to-treat
<b>MNAR</b>	missing not at random
<b>NICE</b>	National Institute for Health and Care Excellence
<b>NMA</b>	network meta-analysis
<b>NNH</b>	number needed to harm
<b>NNT</b>	number needed to treat
<b>NOAC</b>	non-vitamin K antagonist oral anticoagulants

---

<b>NPS</b>	nasopharyngeal cancer
<b>NPV</b>	negative predictive value
<b>NS</b>	not statistically significant
<b>OAB</b>	overactive bladder syndrome
<b>OR</b>	odds ratio
<b>ORR</b>	objective response rate
<b>OS</b>	overall survival
<b>PA</b>	predictive accuracy
<b>PASI</b>	psoriasis area and severity index
<b>PD</b>	progressive disease
<b>PEF</b>	peak expiratory flow
<b>PFS</b>	progression-free survival
<b>PGA</b>	physician's global assessment
<b>PHN</b>	post-hepatic neuralgia
<b>PHS</b>	public health service
<b>PPS</b>	per-protocol set
<b>PPV</b>	positive predictive value
<b>PR</b>	partial response
<b>PRR</b>	proportional reporting ratio
<b>PT</b>	preferred term
<b>QoL</b>	quality of life
<b>RECIST</b>	Response Evaluation Criteria in Solid Tumours
<b>RENAAL</b>	Reduction of Endpoints in NIDDM with the Angiotensin II Antagonist Losartan Study
<b>RMST</b>	restricted mean survival time
<b>RPSFTM</b>	rank preserving structural failure time model
<b>RR</b>	relative risk
<b>RRR</b>	relative risk reduction
<b>RT</b>	radiotherapy
<b>RTC</b>	radiotherapy plus chemotherapy
<b>SAE</b>	serious adverse event
<b>SAP</b>	Statistical Analysis Plan
<b>sBP</b>	systolic blood pressure
<b>SD</b>	stable disease
<b>sd</b>	standard deviation
<b>se</b>	standard error
<b>SFE</b>	summary of favourable effects
<b>SOC</b>	system organ class
<b>SRC</b>	safety review committee
<b>SUCRA</b>	surface under the cumulative ranking curve
<b>SUFE</b>	summary of unfavourable effects
<b>TLFs</b>	tables, listings and figures
<b>TN</b>	true negative
<b>TP</b>	true positive
<b>ULN</b>	upper limit of normal
<b>VAS</b>	visual analogue scale
<b>WHO</b>	World Health Organization



## CHAPTER 1

# Basic ideas in clinical trial design

### 1.1 Historical perspective

As many of us who are involved in clinical trials will know, the randomised controlled trial is a relatively new invention. As pointed out by Pocock (1983) and others, very few clinical trials of the kind we now regularly see were conducted prior to 1950. It took a number of high-profile successes plus the failure of alternative methodologies to convince researchers of their value.

#### **Example 1.1** The Salk Polio Vaccine trial

One of the largest trials ever conducted took place in the USA in 1954 and concerned the evaluation of the Salk polio vaccine. The trial has been reported extensively by Meier (1978) and is used by Pocock (1983) in his discussion of the historical development of clinical trials.

Within the project, there were essentially two trials, and these clearly illustrated the effectiveness of the randomised controlled design.

##### Trial 1: Original design: Observed control

1.08 million children from selected schools were included in this first trial. The second graders in those schools were offered the vaccine, while the first and third graders would serve as the control group. Parents of the second graders were approached for their consent, and it was noted that the consenting parents tended to have higher incomes. Also, this design was not blinded so that both parents and investigators knew which children had received the vaccine and which had not.

##### Trial 2: Alternative design: Randomised control

A further 0.75 million children in other selected schools in grades one to three were to be included in this second trial. All parents were approached for their consent, and those children where consent was given were randomised to receive either the vaccine or a placebo injection. The trial was double-blind with parents, children and investigators unaware of who had received the vaccine and who had not.

The results from the randomised controlled trial were conclusive. The incidence of paralytic polio, for example, was 0.057% in the placebo group compared to 0.016% in the active group, and there were four deaths in the placebo group compared to none in the active group. The results from the observed control trial, however, were less convincing, with a smaller observed difference (0.046% vs. 0.017%). In addition, in the cases where consent

could not be obtained, the incidence of paralytic polio was 0.036% in the randomised trial and 0.037% in the observed control trial, event rates considerably lower than those among placebo patients and in the untreated controls, respectively. This has no impact on the conclusions from the randomised trial, which is robust against this absence of consent; the randomised part is still comparing like with like. In the observed control part, however, the fact that the *no consent* (grade two) children have a lower incidence than those children (grades one and three) who were never offered the vaccine potentially causes some confusion in a non-randomised comparison; does it mean that grade two children naturally have a lower incidence than those in grades one and three? Whatever the explanation, the presence of this uncertainty reduced confidence in the results from the observed control part of the trial.

The randomised part of the Salk Polio Vaccine trial has all the hallmarks of modern-day trials – randomisation, control group and blinding – and it was experiences of these kinds that helped convince researchers that only under such conditions can clear, scientifically valid conclusions be drawn.

## **1.2 Control groups**

We invariably evaluate our treatments by making comparisons – active compared to control. It is very difficult to make absolute statements about specific treatments, and conclusions regarding the efficacy and safety of a new treatment are made relative to an existing treatment or placebo.

### ***ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'***

*'Control groups have one major purpose: to allow discrimination of patient outcomes (for example, changes in symptoms, signs, or other morbidity) caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment'.*

Control groups can take a variety of different forms; here are just a few examples of trials with alternative types of control group:

- Active versus placebo
- Active A versus active B (vs. active C)
- Placebo versus dose level 1 versus dose level 2 versus dose level 3 (dose finding)
- Active A + active B versus active A + placebo (add-on)

The choice will depend on the objectives of the trial.

Open trials with no control group can nonetheless be useful in an exploratory, maybe early phase setting, but it is unlikely that such trials will be able to

provide confirmatory, robust evidence regarding the performance of the new treatment.

Similarly, external concurrent or historical controls (groups of subjects external to the study either in a different setting or previously treated) cannot provide definitive evidence in most settings. We will discuss such trials in Chapter 17. The primary focus of this book, however, is the randomised controlled trial.

### 1.3 Placebos and blinding

It is important to have blinding of both the subject and the investigator wherever possible to avoid unconscious bias creeping in, either in terms of the way a subject reacts psychologically to treatment or in relation to the way the investigator interacts with the subject or records the subject outcome.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Blinding or masking is intended to limit the occurrence of conscious or unconscious bias in the conduct and interpretation of a clinical trial arising from the influence which the knowledge of treatment may have on the recruitment and allocation of subjects, their subsequent care, the attitudes of subjects to the treatments, the assessment of the endpoints, the handling of withdrawals, the exclusion of data from analysis, and so on'.*

Initially, it is important that the investigator responsible for deciding whether a subject should be entered into the study is unaware of the treatment that is to be allocated to that subject. This *concealment* is an essential element of blinding. Ideally, the trial should be *double-blind* with both the subject and the investigator being blind to the specific treatment allocation. If this is not possible for the investigator, for example, then the next best thing is to have an independent evaluation of outcome, both for efficacy and for safety. A *single-blind* trial arises when either the subject or investigator, but not both, is blind to treatment.

An absence of blinding can seriously undermine the validity of an endpoint in the eyes of regulators and the scientific community more generally, especially when the evaluation of that endpoint has an element of subjectivity. In situations where blinding is not possible, it is important to use hard, unambiguous endpoints and to have independent recording of those endpoints.

The use of placebos and blinding go hand in hand. The existence of placebos enables trials to be blinded and accounts for the placebo effect – *the change in a patient's condition that is due to the act of being treated but is not caused by the active component of that treatment*.

Note that having a placebo group does not necessarily imply that one group is left untreated. In many situations – and oncology is a good example – the experimental therapy/placebo is added to an established active drug regimen; this is the add-on study.

## 1.4 Randomisation

Randomisation is clearly a key element in the design of our clinical trials. There are two reasons why we randomise subjects to the treatment groups:

- To avoid any bias in the allocation of the patients to the treatment groups
- To ensure the validity of the statistical test comparisons

Randomisation lists are produced in a variety of ways, and we will discuss several methods later. Once the list is produced, the next patient entering the trial receives the next allocation within the randomisation scheme. In practice, this process is managed by *packaging* the treatments according to the predefined randomisation list.

There are a number of different possibilities when producing these lists:

- Unrestricted randomisation
- Block randomisation
- Unequal randomisation
- Stratified randomisation
- Central randomisation
- Dynamic allocation and minimisation
- Cluster randomisation

### 1.4.1 Unrestricted randomisation

*Unrestricted (or simple) randomisation* is simply a random list of, for example, As and Bs. In a moderately large trial, with, say,  $n = 200$  subjects, such a process will likely produce approximately equal group sizes. There is no guarantee, however, that this will automatically happen; and in small trials, in particular, this can cause problems.

### 1.4.2 Block randomisation

To ensure balance in terms of numbers of subjects, we usually undertake *block randomisation*, where a randomisation list is constructed by randomly choosing from the list of potential blocks (*permuted block randomisation*). For example, there are six ways of allocating two As and two Bs in a *block* of size four:

AABB, ABAB, ABBA, BAAB, BABA, BBAA

We choose at random from this set of six blocks to construct our randomisation list: for example,

ABBA BAAB ABAB ABBA, ...

Clearly, if we recruit a multiple of four patients into the trial, we will have perfect balance, and approximate balance (which is usually good enough) for any sample size.

In large trials, it could be argued that block randomisation is unnecessary. In one sense, this is true; overall balance will be achieved by chance with an unrestricted randomisation list. However, it is usually the case that large trials will be multi-centre trials, and not only is it important to have balance overall, but it is also important to have balance within each centre. In practice, therefore, we would allocate several blocks to each centre: for example, five blocks of size four if we are planning to recruit 20 patients from each centre. This will ensure balance within each centre and overall.

How do we choose block size? There is no magic formula, but, more often than not, the block size is equal to two times the number of treatments.

What are the issues with block size?

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'Care must be taken to choose block lengths which are sufficiently short to limit possible imbalance, but which are long enough to avoid predictability towards the end of the sequence in a block. Investigators and other relevant staff should generally be blind to the block length....'*

Shorter block lengths are better at producing balance. With two treatments, a block length of 4 is better at producing balance than a block length of 12. The block length of 4 gives perfect balance if there is a multiple of four patients entering, whereas with a block length of 12, perfect balance is only going to be achieved if there is a multiple of 12 patients in the study. However, the problem with the shorter block lengths is that this is an easy code to crack and inadvertent unblinding can occur. For example, suppose a block length of 4 was being used in a placebo-controlled trial, and assume also that experience with the active drug suggests that many patients receiving that drug will suffer nausea. Suppose the trial begins and the first two patients suffer nausea. The investigator is likely to conclude that both these patients have been randomised to active and that therefore the next two allocations are to placebo. This knowledge could influence his/her willingness to enter certain patients into the next two positions in the randomisation list, causing bias in the mix of patients randomised into the two treatment groups. Note the comment in the ICH guideline regarding keeping the investigator (and others) blind to the block length. While in principle this comment is sound, the drug is often delivered to a site according to the chosen block length, making it difficult to conceal information on block size. If the issue of inadvertent unblinding has the potential to cause problems, then more sophisticated methodologies can be used, such as having the block length itself vary, perhaps randomly chosen from 2, 4 or 6.

### 1.4.3 Unequal randomisation

All other things being equal, having equal numbers of subjects in the two treatment groups provides the maximum amount of information (the greatest power) on the relative efficacy of the treatments. There may, however, be issues that override statistical efficiency:

- It may be necessary to place more patients on active compared to placebo to obtain the required safety information.
- In a three-group trial with active A, active B and placebo (P), it may make sense to have a 2:2:1 randomisation to give more power for the A versus B comparison as that difference is likely to be smaller than the A versus P and B versus P differences.

*Unequal randomisation* is sometimes needed as a result of these considerations. To achieve this, the randomisation list will be designed for the second example with double the number of A and B allocations compared to placebo.

For unequal randomisation, we would choose the block size accordingly. For a 2:1 randomisation to A or P, we could randomly choose from 3 blocks:

AAP, APA, PAA

For a 2:2:1 randomisation, blocks of size 5 could be used, and we would randomly choose from the 30 possible permutations of the letters A, A, B, B, C for each of those blocks.

### 1.4.4 Stratified randomisation

Block randomisation therefore forces the required balance in terms of the numbers of patients in the treatment groups, but things can still go wrong. For example, let's suppose in an oncology study with time to death as the primary endpoint that we can measure baseline risk (say, ECOG 0 or 1 vs. 2) and classify patients as either high risk (H) or low risk (L); and further, suppose that the groups turn out as follows:

A: HHLHLHHHHLLHHHLHHLHHH ( $H = 15, L = 6$ )

B: LLHHLHHLLHLHLHLHLLHLL ( $H = 10, L = 12$ )

Note that there are 15 (71%) high-risk patients and six (29%) low-risk patients in treatment group A compared to a split of 10 (45%) high-risk and 12 (55%) low-risk patients in treatment group B.

Now suppose that the mean survival times are observed to be 21.5 months in group A and 27.8 months in group B. What conclusions can we draw? It is very difficult; the difference we have seen could be due to real treatment differences or could be caused by the imbalance in terms of differential risk across the groups, or a mixture of the two. Statisticians talk in terms of *confounding* (just a fancy way of saying *mixed up*) between the treatment effect and the effect of baseline risk. This situation is very difficult to unravel, and we avoid it by

*stratified randomisation* to ensure that the *case mix* in the treatment groups is comparable.

This means that we produce separate randomisation lists for the high-risk and low-risk patients, the strata in this case, and choose the next allocation from the list according to baseline risk. For example, the following lists (block size 4)

H: ABAAABBABABABABBAAABBAABABBAA

L: BAABBABAABBBAAABABABBAAABBAABAAB

will ensure firstly that we end up with balance in terms of treatment group sizes and also secondly that both the high- and low-risk patients will be equally split across those groups, that is, balance in terms of the mix of patients.

Having separate randomisation lists for the different centres in a multi-centre trial to ensure *equal* numbers of patients in the treatment groups within each centre is using *centre* as a stratification factor; this will ensure that we do not end up with treatment being confounded with centre.

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'It is advisable to have a separate random scheme for each centre, i.e. to stratify by centre or to allocate several whole blocks to each centre. Stratification by important prognostic factors measured at baseline (e.g. severity of disease, age, sex, etc.) may sometimes be valuable in order to promote balanced allocation within strata'.*

Note that allocating several whole blocks to each centre/site is, by definition, stratifying by centre. It is also the case that stratification by centre happens in many trials by default when trial medication is delivered to sites several blocks at a time.

Where the requirement is to have balance in terms of several factors, a stratified randomisation scheme would use all combinations of those factors to define the strata. For example, if balance is required for sex and age, then a scheme with four strata – males, <50 years; females, <50 years; males, ≥50 years; and females, ≥50 years – will achieve the required balance. Extending this example, if baseline risk (low/medium/high) is added as a third stratification factor, there will be 12 combinations of the various factors/levels and 12 randomisation lists. Note, however, that this can cause problems with empty strata or strata containing only a few patients, especially if the proportions across the levels of the stratification factors are very different. We do not want combinations of stratification factors where there are very few patients. Only when complete blocks are used up (filled) is balance ensured in terms of both overall numbers allocated to groups A and B and the mix of patients in those groups.

Here are some recommendations regarding stratification:

- 1 Avoid *over-stratification* and only choose a small number of stratification factors, maybe two or three at most, even if the sample size is large. Adjustment for additional factors thought to be predictive of outcome can be achieved through analysis of covariance type techniques at the data analysis stage (see Chapter 5).
- 2 If there is the need to stratify by several factors, be aware of the distribution of patients across the levels of those factors and how this will play out in relation to the combinations of the levels of those factors. If there is the potential to end up with several small strata, do not be afraid to use a small block size (for example 2) to force balance. Use this in conjunction with central randomisation (see the next section) to preserve blinding.

### **1.4.5 Central randomisation**

In *central randomisation*, the randomisation process is controlled and managed from a centralised point of contact. Each investigator makes contact through an *Interactive Voice Response System* (IVRS) or an *Interactive Web Response System* (IWRS) to this centralised point when they have identified a patient to be entered into the study and is given the next allocation, taken from the appropriate randomisation list. Blind can be preserved by simply specifying the number of the (pre-numbered) pack to be used to treat the particular patient; the computerised system keeps a record of which packs have been used already and which packs contain which treatment. Central randomisation has a number of practical advantages:

- It can provide a check that the patient about to be entered satisfies certain inclusion/exclusion criteria, thus reducing the number of protocol violations.
- It provides up-to-date information on all aspects of recruitment.
- It allows more efficient distribution and stock control of medication.
- It provides some protection against biased allocation of patients to treatment groups in trials where the investigator is not blind; the investigator knowing the next allocation could (perhaps subconsciously) select patients to include or not include based on that knowledge. With central randomisation, the patient is identified and information is given to the system before the next allocation is revealed to the investigator.
- It gives an effective way of managing multi-centre trials by providing ‘live’ information on recruitment patterns.
- It allows the implementation of more complex allocation schemes such as minimisation and dynamic allocation (but see comments later on these techniques).

Earlier, we discussed the use of stratified randomisation in multi-centre trials, and where the centres are large, this is appropriate. With small centres, however – for example, in general practitioner (GP) trials – this does not make sense, and stratified randomisation with *region* defining the strata may be more appropriate. Central randomisation would be essential to manage such a scheme.

### 1.4.6 Dynamic allocation and minimisation

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

'Dynamic allocation is an alternative procedure in which the allocation of treatment to a subject is influenced by the current balance of allocated treatments and, in a stratified trial, by the stratum to which the subject belongs and the balance within that stratum. Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomisation should be incorporated for each treatment allocation'.

*Dynamic allocation* moves away from having a pre-specified randomisation list, and the allocation of patients evolves as the trial proceeds. The method looks at the current balance in terms of the mix of patients and a number of pre-specified factors and allocates the next patient in an optimum way to help redress any imbalances that exist at that time.

For example, suppose we require balance in terms of sex and age ( $\geq 65$  vs.  $< 65$ ), and part way through the trial, we see a mix of patients as in Table 1.1.

Treatment group A contains proportionately more males (12 out of 25 vs. 10 out of 25) than treatment group B but fewer patients over 65 years (7 out of 25 vs. 8 out of 25). Further, suppose that the next patient to enter is male and aged 68 years. In terms of sex, we would prefer that this patient be placed in treatment group B, while for age, we would prefer this patient to enter group A. The greater imbalance, however, is in relation to sex, so our overall preference would be for treatment group B to help *correct* for the current imbalance. The method of *minimisation* would simply put this patient in group B. But ICH E9 recommends that we have a *random element* to that allocation, and so, for example, we would allocate this patient to treatment group A with, say, a probability of 0.7. Minimisation is the deterministic special case of dynamic allocation where the random assignment probability (0.7 in the example) is equal to one. With a small number of baseline factors, stratified randomisation will give good enough balance, and there is no need to consider the more complex dynamic allocation. On the other hand, this technique has been used when more factors are involved.

Since the publication of ICH E9, however, there has been considerable debate about the validity of dynamic allocation, even with the random element. One school of thought supports the view that the properties of standard statistical methodologies, notably *p*-values and confidence intervals, are not strictly valid when such allocation schemes are used, due to the lack of control of the type I error. As a result, regulators are somewhat cautious.

**Table 1.1** Current mix of patients

	A	B
Total	25	25
Male	12/25	10/25
Age $\geq 65$	7/25	8/25

**CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'**

'...techniques of dynamic allocation such as minimisation are sometimes used to achieve balance across several factors simultaneously. Even if deterministic schemes are avoided, such methods remain highly controversial. Thus applicants are strongly advised to avoid such methods'.

The updated CHMP guideline on baseline covariates published in 2015 raises the issue of the type I error.

**CHMP (2015): 'Guideline on adjustment for baseline covariates in clinical trials'**

'Possible implications of dynamic allocation methods on the analysis e.g. with regard to bias and Type I error control should be carefully considered, taking into account that for some situations . . . it has been shown that in case of dynamic treatment allocation conventional statistical methods do not always control the Type I error'.

If you are planning a trial, then the recommendation is to stick with stratification and avoid dynamic allocation. If you have an ongoing trial using dynamic allocation, then be prepared at the statistical analysis stage to supplement the standard methods of calculating *p*-values with more complex methods that take account of the dynamic allocation scheme. These methods go under the name of *re-randomisation tests*.

See Roes (2004) for a comprehensive discussion of dynamic allocation.

### **1.4.7 Cluster randomisation**

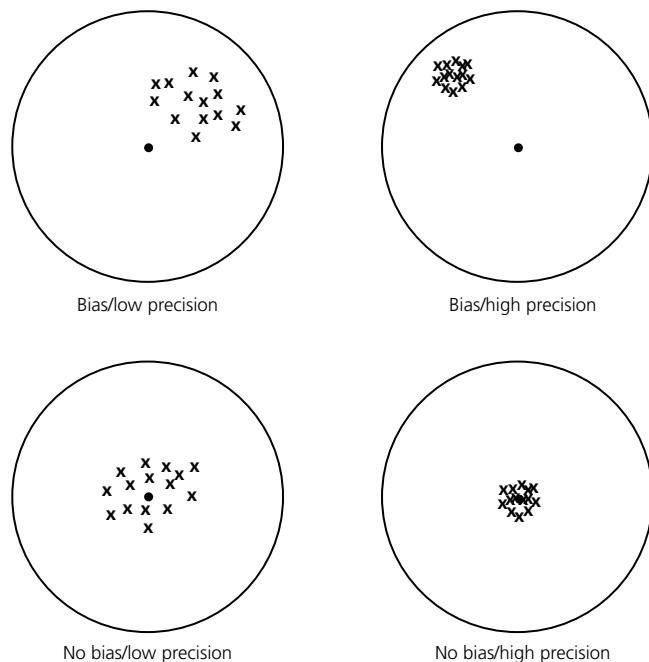
In some cases, it can be more convenient or appropriate not to randomise individual patients but to randomise groups of patients. The groups, for example, could correspond to GPs so that each GP enters, say, four patients, and it is the 100 GPs that are randomised, 50 giving treatment A and 50 giving treatment B. Such methods are used but are more suited to phase IV than the earlier phases of clinical development. Health interventions in third-world countries are frequently evaluated using cluster randomisation.

Bland (2004) provides a review and some examples of cluster randomised trials, while Campbell, Donner and Klar (2007) give a comprehensive review of the methodology.

## **1.5 Bias and precision**

When we are evaluating and comparing our treatments, we are looking for two things:

- An unbiased, correct view of how effective (and safe) the treatment is
- An accurate estimate of how effective (and safe) the treatment is



**Figure 1.1** Bias and precision

As statisticians, we talk in terms of *bias* and *precision*; we want to eliminate bias and have high precision. Imagine having 10 attempts at hitting the bull's-eye on a target board, as shown in Figure 1.1. Bias is about hitting the bull's-eye on average; precision is about being consistent.

These aspects are clearly set out in ICH E9.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Many of the principles delineated in this guidance deal with minimising bias and maximising precision. As used in this guidance, the term "bias" describes the systematic tendency of any factors associated with the design, conduct, analysis and interpretation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value'.*

What particular features in the design of a trial help to eliminate bias?

- Concurrent control group as the basis for a *treatment comparison*
- Randomisation to avoid bias in allocating subjects to treatments
- Blinding of both the subject and the investigator
- Pre-specification of the methods of statistical analysis

What particular features in the design of a trial help to increase precision?

- Large sample size
- Measuring the endpoints in a precise way
- Standardising aspects of the protocol that impact patient-to-patient variation
- Collecting data on key prognostic factors and including those baseline factors as covariates in the statistical analysis
- Stratifying the randomisation for the most important factors that are predictive of outcome
- Choosing the most appropriate design (e.g. using a crossover design rather than a parallel-group design where this is appropriate)

Several of the issues raised here may be unclear at this point; simply be aware that eliminating bias and increasing precision are the key issues that drive our statistical thinking from a design perspective. Also, be aware that if something should be sacrificed, it is precision rather than bias. High precision in the presence of bias is of no value; we are simply getting a more precise wrong answer! First and foremost, we require an unbiased view; increasing precision is then a bonus. Similar considerations are also needed when we choose the appropriate statistical methodology at the analysis stage.

One point to make clear – and this is a common misunderstanding – is that having a large sample size of itself does not remove bias. If there is a flaw in the trial design or in the planned methods of statistical analysis that causes bias, then beating the trial over the head with large patient numbers will not eliminate that bias, and we will still be misled regarding the true treatment difference – perhaps even more so, because the trial is large. As mentioned earlier, having a large sample size in a flawed clinical trial will just result in a more precise, incorrect answer!

## 1.6 Between- and within-patient designs

The simplest trial design, of course, is the *parallel-group design* assigning patients to receive either treatment A or treatment B. For example, suppose we have a randomised parallel-group design in hypertension with 50 patients per group and that the mean fall in diastolic blood pressure in each of the two groups is as follows:

$$\text{A: } \bar{x}_1 = 4.6 \text{ mmHg}$$

$$\text{B: } \bar{x}_2 = 7.1 \text{ mmHg}$$

One thing to note that will aid our discussion later is that it would be easy (but incorrect) to conclude in light of these data that B is a more effective treatment than A simply because we have seen a greater fall on average with treatment B than with treatment A. One thing we have to remember is that the 50 patients in group A are a different group of patients from the 50 patients in group B, and

patients respond differently; so, the observed difference between the treatments could simply be caused by patient-to-patient variation. As we will see later, unravelling whether the observed difference is reflective of a real treatment difference or simply a chance difference caused by inherent patient-to-patient variation with identical treatments is precisely the role of the *p*-value; but it is not easy.

This design is what we refer to as a *between-patient design*. The basis of the treatment comparison is the comparison between two independent groups of patients.

An alternative design is the *within-patient design*. Such designs are not universally applicable but can be very powerful under certain circumstances. One form of the within-patient design is the *paired design*:

- In ophthalmology – Use treatment A in the right eye and treatment B in the left eye.
- In a volunteer study in wound care – Create a wound on each forearm and use a dressing of type A on the right forearm and a dressing of type B on the left forearm.

Here, the 50 subjects receiving A will be the same 50 subjects who receive B, and the comparison of A and B in terms of, say, mean healing time in the second example, is a comparison based on identical *groups* of subjects. At least in principle, drawing conclusions regarding the relative effect of the two treatments and accounting for the patient-to-patient (or subject-to-subject) variation may be easier under these circumstances.

Another example of the within-patient design is the *crossover design*. Again, each subject receives each of the treatments but now sequentially in time, with some subjects receiving the treatments in the order A followed by B and some in the order B followed by A.

In both the paired design and the crossover design, there is, of course, randomisation; in the second paired design example earlier, it is according to which forearm receives A and which forearm receives B, and randomisation is according to treatment order, A/B or B/A, in the crossover design.

## 1.7 Crossover trials

The crossover trial was mentioned in the previous section as one example of a within-patient design. In order to discuss some issues associated with these designs, we will consider the simplest form of crossover trial – two treatments A and B and two treatment periods I and II.

The main problem with the use of this design is the possible presence of the so-called *carry-over effect*. This is the residual effect of one of the treatments in period I, influencing the outcome of the other treatment in period II. An extreme example of this would be the situation where one of the treatments, say, A, was

very efficacious, so much so that many of the patients receiving treatment A were cured of their disease, while B was ineffective and had no impact on the underlying disease. As a consequence, many of the subjects following the A/B sequence would give a good response at the end of period I (an outcome ascribed to A) but would also give a good response at the end of period II (an outcome ascribed to B) because they were cured by A. These data would give a false impression of the A versus B difference. In this situation, the B data obtained from period II is contaminated, and the data coming out of such a trial are virtually useless.

Therefore, it is important to only use these designs when we can be sure that carry-over effects will not be seen. Introducing a washout period between period I and period II can help to eliminate carry-over so that when the subject enters period II, their disease condition is similar to what it was at the start of period I. Crossover designs should not be used where there is the potential to affect the underlying disease state. ICH E9 is very clear on the use of these designs.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Crossover designs have a number of problems that can invalidate their results. The chief difficulty concerns carryover, that is, the residual influence of treatments in subsequent treatment periods. . . When the crossover design is used it is therefore important to avoid carryover. This is best done by selective and careful use of the design on the basis of adequate knowledge of both the disease area and the new medication. The disease under study should be chronic and stable. The relevant effects of the medication should develop fully within the treatment period. The washout periods should be sufficiently long for complete reversibility of drug effect. The fact that these conditions are likely to be met should be established in advance of the trial by means of prior information and data'.*

The crossover design is used extensively in phase I trials in healthy volunteers to compare different formulations in terms of their bioequivalence (where there is no underlying disease to affect). They can also be considered in diseases (for example, asthma) where the treatments are being used simply to relieve symptoms; once the treatments are removed, the symptoms return to their earlier level.

## **1.8 Signal, noise and evidence**

### **1.8.1 Signal**

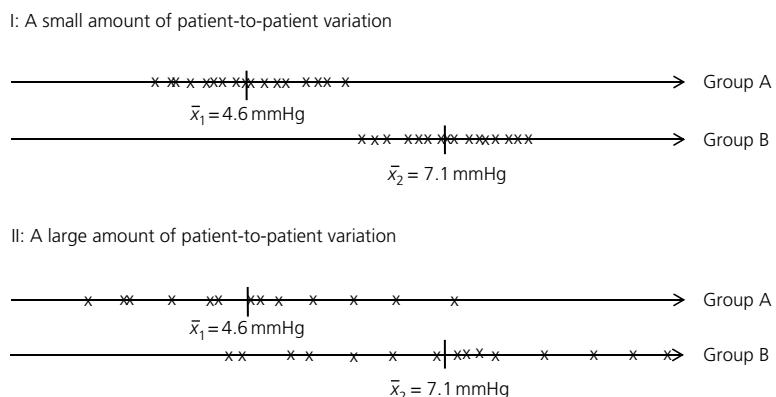
Consider the example in Section 1.6 comparing treatments A and B in a parallel-group trial. The purpose of this investigation was to detect differences in the mean reductions in diastolic blood pressure between the two groups. The observed difference between  $\bar{x}_1 = 4.6 \text{ mmHg}$  and  $\bar{x}_2 = 7.1 \text{ mmHg}$  is

2.5 mmHg. We will refer to this difference as the *signal*, and this captures in part the evidence that the treatments truly are different. Clearly, if the observed difference was larger, we would likely be more inclined to conclude differences. Large differences give strong signals, while small differences give weak signals.

### 1.8.2 Noise

The signal, however, is not the only aspect of the data that plays a part in our analysis and conclusions. If we were to see a large amount of patient-to-patient variation, then we would be less inclined to conclude differences than if all the patients in treatment group A had reductions tightly clustered around 4.6 mmHg, while all those in treatment group B had values tightly clustered around 7.1 mmHg. As can be seen in Figure 1.2, the evidence for a real treatment difference in situation I is much stronger than the evidence seen in situation II, although the mean values for both groups are the same in each case. We refer to the patient-to-patient variation as the *noise*, and clearly the extent of the noise will influence our willingness to declare differences between the treatments. An observed difference of 2.5 mmHg based on a small amount of noise is much stronger evidence for a true treatment difference than an observed difference of 2.5 mmHg in the presence of a large amount of noise.

The sample size plays an additional role in our willingness to conclude treatment differences and in a sense serves to compensate for the extent of the noise. If there is a large amount of patient-to-patient variation (a large amount of noise), then a large sample size is needed before we are able to *see what is happening on average* and conclude that the true means are indeed separated. In contrast, with a small amount of patient-to-patient variation, it is somewhat easier to recognise that the means truly are different, even with a small sample size.



**Figure 1.2** Differing degrees of patient-to-patient variation

### 1.8.3 Signal-to-noise ratio

These concepts of signal and noise provide a way of thinking for statistical experiments. In declaring differences, we look for strong signals and small amounts of noise: that is, a large *signal-to-noise ratio*. If either the signal is weak or the noise is large or both, then this ratio will be small, and we will have little evidence on which to declare differences. The sample size can then be added into this mix. The value of a signal-to-noise ratio based on a small sample size is less reliable than the value of a signal-to-noise ratio based on a large sample size, and clearly this is also going to influence our willingness to declare treatment differences.

In one sense, the signal is out of our control; it will depend entirely on what the true treatment difference is. Similarly, there is little we can do about patient-to-patient variability, although we can reduce this by having, for example, precise measures of outcome or a more homogeneous group of patients. The sample size, however, is very much under our control, and common sense tells us that increasing this will provide a more reliable comparison and make it easier for us to detect treatment differences when they exist.

Later, in Chapter 8, we will discuss power and sample size and see how to choose sample size in order to meet our objectives. We will also see in Section 3.3 how, in many circumstances, the calculation of the *p*-value is based on the signal-to-noise ratio, which when combined with the sample size allows us to numerically calculate the *evidence* in favour of treatment differences. We will see in that section that when comparing two treatment means with  $n$  subjects per group, for example, the *evidence* for treatment differences is captured by the square root of  $n/2$  multiplied by the signal-to-noise ratio.

## 1.9 Confirmatory and exploratory trials

ICH E9 makes a very clear distinction between *confirmatory* and *exploratory* trials. From a statistical perspective, this is an important distinction as certain aspects of the design and analysis of data – for example, the strict control of multiplicity of endpoints (see Chapter 10) – depend upon this confirmatory/exploratory distinction.

### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

'A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are needed to provide firm evidence of efficacy or safety'.

### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

'The rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out in a series of exploratory studies. Like all clinical trials, these

*exploratory studies should have clear and precise objectives. However, in contrast to confirmatory trials, their objectives may not always lead to simple tests of predefined hypotheses'.*

Typically, later phase trials tend to contain the confirmatory elements, while the earlier phase studies – proof of concept, dose finding, etc. – are viewed as exploratory. Indeed, an alternative word for *confirmatory* is *pivotal*. It is the confirmatory elements of our trials that provide the pivotal information from a regulatory perspective.

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'Any individual trial may have both confirmatory and exploratory aspects'.*

Usually, the primary and secondary endpoints provide the basis for the confirmatory claims. Additional endpoints may then provide the basis for exploratory investigations.

## **1.10 Superiority, equivalence and non-inferiority trials**

In a *superiority* trial, our objective is to demonstrate either that our treatment works by demonstrating superiority over placebo or that we are superior to some reference or standard treatment.

In an *equivalence* trial, we are looking to show that we are similar to some reference treatment; bioequivalence and trials developing biosimilar products are the most common examples of these types of trials.

Finally, in a *non-inferiority* trial, we are trying to demonstrate that we are no more than a certain, pre-specified, usually small amount worse than (clinically, *at least as good as*) some active reference treatment.

In therapeutic equivalence trials and in non-inferiority trials, we are often looking to demonstrate efficacy of our test treatment indirectly. It may be that for ethical or practical reasons, it is not feasible to show efficacy by undertaking a superiority trial against placebo. In such a case, we compare our test treatment to a control treatment that is known to be efficacious and demonstrate either strict equivalence or *at least as good as* (non-inferiority). If we are successful, then we can be confident that our test treatment works.

Alternatively, there may be commercial reasons why we want to demonstrate the non-inferiority of our treatment against an active control. Maybe our treatment potentially has fewer side effects than the active control, and we are prepared to pay a small price for this safety advantage in relation to efficacy. If this were the case, then of course we would need to show advantages in terms of a reduction in side effects, but we would also need to demonstrate that we do not lose much with regard to efficacy.

Non-inferiority trials have become more common as time has gone on. This is due in part to the constraints imposed by the revised Helsinki Declaration (2004) and the increasing concern in some circles regarding the ethics of placebo use. However, these trials require very careful design and conduct, and we will discuss this whole area in Chapter 12.

## 1.11 Endpoint types

It is useful to classify the types of outcomes or endpoints that we see in our clinical investigations.

The most common endpoint type that we see is *continuous*. Examples include change from baseline in cholesterol level, exercise duration, blood pressure or FEV<sub>1</sub> and so on. Each of these quantities is based on a continuum of potential values. In some cases, of course, our measurement technique may only enable us to record to the nearest whole number (e.g. blood pressure), but that does not alter the basic fact that the underlying scale is continuous.

Probably the second most common endpoint type is *binary*. Examples of binary endpoints include cured/not cured, responder/non-responder and died/survived. Here, the measure is based on a dichotomy.

Moving up from binary is a *categorical* endpoint where there are more than two categories that form the basis of the *outcome*. The following are examples of categorical endpoints:

- Death from cancer causes/death from cardiovascular causes/death from respiratory causes/death from other causes/survival
- Pain: none/mild/moderate/severe/very severe

The categories are non-overlapping, and each patient is placed into one and only one of the outcome categories. A binary endpoint is a special case where the number of categories is just two.

These two examples are different, however; in the first example, the categories are unordered, while in the second example, there is a complete ordering across the defined categories. In the latter case, we term the endpoint type either *ordered categorical* or *ordinal*.

Ordered categorical endpoints arise in many situations. In oncology (solid tumours), the Response Evaluation Criteria in Solid Tumours (RECIST) criteria place a patient in one of four response categories (National Cancer Institute, [www.cancer.gov](http://www.cancer.gov)):

- Complete response (CR) = disappearance of all target lesions
- Partial response (PR) = 30% decrease in the sum of the longest diameter of target lesions
- Progressive disease (PD) = 20% increase in the sum of the longest diameter of target lesions
- Stable disease (SD) = small changes that do not meet the aforementioned criteria

When analysing data, it is important of course that we clearly specify the appropriate order, and in this case, it is CR, PR, SD and PD.

Other endpoints arise as *scores*. These are frequently a result of the need to provide a measure of some clinical condition such as depression or anxiety. The Hamilton Depression (HAMD) Scale and the Hamilton Anxiety (HAMA) Scale provide measures in these cases. These scales contain distinct items that are scored individually, and the total score is then obtained as the sum of the individual scores. For the HAMD Scale, there are usually 17 items – depressed mood, self-depreciation, guilt feelings, etc. – each scored on a three-point to five-point scale. The five-point scales are typically scored 0 = absent, 1 = mild, 2 = moderate, 3 = severe and 4 = incapacitating, while the three-point scales are typically 0 = absent, 1 = doubtful or trivial and 2 = present.

From time to time, an endpoint arises as a *count* of items or events: number of epileptic seizures in a 6-month period, number of severe asthma exacerbations in a 12-month period and number of incontinence episodes recorded in a 3-day diary are just a few examples.

Finally, we have time-to-event endpoints, such as time to death in oncology, time to first seizure in epilepsy, time to rash healing in herpes zoster and so on, where time is measured from the point of randomisation. Such endpoints are quite common and are used increasingly across many therapeutic settings.

Note that we are using *endpoint* defined at the patient/subject level and not at the group level. The median survival time is not an endpoint in these considerations: it is a summary measure for a group of patients that may well be a focus of attention. But it is something that is not measured at the patient level and therefore does not fall within our definition of an endpoint. Similarly, the proportion of patients responding is not an endpoint in our definition: it is again a group-level summary measure. This can conflict with descriptions and definitions seen in other textbooks, and in publications. In Chapter 8, we will discuss the concept of estimands; moving forward, they will be central to the way we design our trials and formulate our hypotheses. The estimand framework makes a clear distinction between the endpoint (at the patient/subject level) and the summary measure (at the group level) that is the focus of our attention in quantifying treatment differences and effects.

As we shall see later, the endpoint type to a large extent determines the class of statistical tests that we undertake. For continuous endpoints, we commonly use the t-tests and their extensions – analysis of variance and analysis of covariance. For binary, categorical and ordered categorical endpoints, we use the class of chi-square tests (Pearson chi-square for categorical endpoints and the Mantel–Haenszel chi-square for ordinal endpoints) and their extension, logistic regression. For count endpoints, we use methods built around a negative-binomial distribution, while for time-to-event endpoints, the logrank test and its extension, the proportional hazards model, will be used to make treatment comparisons.

**Table 1.2** Categorisation

Group	Cigarettes per day
1	0
2	1–5
3	6–20
4	>20

Note also that we can move between endpoint types depending on the circumstances. In hypertension, we might be interested in:

- The fall in diastolic blood pressure (continuous)
- Success/failure, with success defined as a reduction of at least 10 mmHg in diastolic blood pressure and diastolic below 90 mmHg (binary)
- Complete success/partial success/failure with complete success = reduction of at least 10 mmHg and diastolic below 90 mmHg, partial success = reduction of at least 10 mmHg but diastolic 90 mmHg or above and failure = everything else (ordinal)

There are further links across the endpoint types. In some settings we may want to group continuous, score or count endpoints into ordered categories and analyse using techniques for ordered categorical endpoints. For example, in a smoking cessation study, we may reduce the basic data on cigarette consumption to just four groups (Table 1.2), accepting that there is little reliable information beyond that.

We will continue this discussion on endpoints in the next section.

## 1.12 Choice of endpoint

### 1.12.1 Primary endpoints

Choosing a single primary endpoint is part of a strategy to reduce multiplicity in statistical testing. We will leave discussion of the problems arising with multiplicity until Chapter 10 and focus here on the nature of endpoints both from a statistical and a clinical point of view.

Generally, the primary endpoint should be that endpoint that is the clinically most relevant endpoint from the patients' perspective.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The primary variable ("target" variable, primary endpoint) should be that variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial'.*

This choice should allow, among other things, a clear quantitative measure of benefit for patients. As we will see, identifying new treatments is not just about statistical significance but also about clinical importance, and the importance of the clinical finding can only ever be evaluated if we can quantify the clinical benefit for patients.

Usually, the primary endpoint will relate to efficacy, but not always. If the primary objective of the trial concerns safety or quality of life, then a primary endpoint relating to these issues would be needed.

The primary endpoint must be pre-specified in a confirmatory trial as specification after unblinding could clearly lead to bias. Generally, there would be only one primary endpoint, but in some circumstances, more than one primary endpoint may be needed to study the different effects of a new treatment. For example, in acute stroke, it is generally accepted that two primary endpoints are needed – one relating to survival free of disability and a second relating to improvement in neurological outcome. See CPMP (2001) 'Note for Guidance on Clinical Investigation of Medicinal Products for the Treatment of Acute Stroke' for further details on this.

### **1.12.2 Secondary endpoints**

Secondary endpoints may be defined that support a more detailed evaluation of the primary endpoint(s); or, alternatively, such variables/endpoints may relate to secondary objectives. These endpoints may not be critical to a claim but may help in understanding the nature of the way the treatment works. In addition, data on secondary endpoints may help to embellish a marketing position for the new treatment.

If the primary endpoint gives a negative result (not statistically significant), then secondary endpoints generally cannot recover a claim. However, if the primary endpoint has given a positive result, additional claims can be based on the secondary endpoints, provided these have been structured correctly within the confirmatory strategy. In Chapter 10, we will discuss hierarchical testing as one basis for such a strategy.

### **1.12.3 Surrogate endpoints**

Surrogate endpoints are generally used when it is not possible within the time-frame of the trial to measure true clinical benefit. Many examples exist, as seen in Table 1.3.

Unfortunately, many treatments that have shown promise in terms of surrogate endpoints have been shown not to provide subsequent improvement in terms of the clinical outcome. Fleming and DeMets (1996) provide several examples where we have been disappointed by surrogate endpoints and provide in each of these cases possible explanations for this failure of the surrogate. One common issue in particular is that a treatment may have an effect on a surrogate

**Table 1.3** Surrogate endpoints and clinical endpoints

Disease	Surrogate endpoint	Clinical endpoint
Congestive heart failure	Exercise tolerance	Mortality
Osteoporosis	Bone mineral density	Fractures
HIV	CD4 cell count	Mortality
Hypercholesterolemia	Cholesterol level	Coronary heart disease

through a particular pathway that is unrelated to the underlying disease process or the clinical outcome.

Treatment effects on surrogate endpoints therefore do not necessarily translate into treatment effects on clinical endpoints, and the validity of the surrogate depends not only on the variable itself but also on the disease area and the mode of action of the treatment. Establishing new valid surrogates is very difficult. Fleming and DeMets conclude that surrogates are extremely valuable in phase II *proof-of-concept* studies, but they question their general use in phase III confirmatory trials.

**Example 1.2** Bone mineral density and fracture risk in osteoporosis

Li, Chines and Meredith (2004) quote three clinical trials evaluating the effectiveness of alendronate, risedronate and raloxifene in increasing bone mineral density (BMD) and reducing fracture risk in osteoporosis. These treatments are seen to reduce fracture risk by similar amounts (47%, 49% and 46%, respectively), yet their effects on relevant increases in BMD are somewhat different (6.2%, 5.8% and 2.7%, respectively). Drawing conclusions on the relative effectiveness of these treatments based solely in terms of the surrogate BMD would clearly be misleading.

#### 1.12.4 Global assessment endpoints

Global assessment endpoints involve an investigator's overall impression of improvement or benefit. Usually, this is done in terms of an ordinal scale of categories. Clinical Global Impression (CGI) is an example of such an endpoint and is used in many therapeutic settings to enable the clinician to judge severity or improvement. While the guidelines allow such endpoints, experience shows that they should be accompanied by objective measures of benefit.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

'If objective variables are considered by the investigator when making a global assessment, then those objective variables should be considered as additional primary, or at least important secondary, variables'.

### 1.12.5 Composite endpoints

In some circumstances, it may be necessary to combine several events/endpoints to produce a combined or composite endpoint. The main purpose for doing this is to avoid multiple testing, and more will be said about ‘multiplicity’ in Chapter 10. In addition, combining endpoints/events will increase the absolute numbers of events observed, and this can increase sensitivity (power) for the detection of treatment effects.

### 1.12.6 Categorisation

In general, a variable measured on a continuous scale contains more information and is a better reflection of the effect of treatment than a categorisation of such a scale. For example, in hypertension, the clinical goal may be to reduce diastolic blood to below 90 mmHg; that is not to say that a reduction to 91 mmHg is totally unacceptable, while a reduction to 89 mmHg is a perfect outcome. Having a binary outcome that relates to achieving 90 mmHg is clearly a somewhat crude measure of treatment benefit. The CHMP recognise that the original variable contains more information; and although they support the presentation of the proportion of responders to gauge clinical benefit, they suggest that statistical testing be undertaken on the original scale.

#### ***CHMP (2017) ‘Guideline on multiplicity issues in clinical trials’***

*‘If the “responder” analysis is not the primary analysis it may be used after statistical significance has been established on the mean level of the required primary endpoint(s), to establish the clinical relevance of the observed differences in the proportion of “responders”. When used in this manner, the test of the null hypothesis of no treatment effect is better carried out on the original primary variable than on the proportion of responders’.*

Nonetheless, categorisation can be of benefit under some circumstances. In an earlier section, however, we discussed the categorisation of number of cigarettes to a four-point ordinal scale, accepting that measures on the original scale may be subject to substantial error and misreporting; in this case the additional information contained in original variable is in a sense spurious precision.

There may also be circumstances where a categorisation combines responses measured on different measurement domains: for example, to give a single dichotomous responder/non-responder outcome. There are connections here with global assessment endpoints. This approach is taken in Alzheimer’s disease, where the effect of treatment is in part expressed in terms of the ‘proportion of patients who achieve a meaningful benefit (response)’; see the CHMP (2016) ‘Draft guideline on medicinal products in the treatment of Alzheimer’s disease and other dementias’.

In oncology, the RECIST criteria may be used simply to give the proportion of patients who ‘respond’ – that is achieve a CR or a PR. This reduces the

sensitivity of the complete scale but may make it easier to quantify the clinical benefit in what is often termed a *responder analysis*. For an interesting exchange on the value of dichotomisation, see Senn (2003) and Lewis (2004). Royston, Altman and Sauerbrei (2006) are against categorisation for data analysis as this tends to waste information and consequently is less able to detect treatment differences should they exist. However, each of these publications recognises that such analyses can be beneficial in terms of data presentation and communication.

Finally, a few words about the use of the visual analogue scale (VAS). A value on this 10 mm line gives a continuous measure (the distance between the left-hand end and the marked value), and these scales are used successfully in several therapeutic settings. But the advantage of a VAS scale over an ordinal four- or five-point scale is questionable, as again there is an argument that the additional precision provided by VAS is of no value. A study by Jensen et al. (1989) on the measurement of post-operative pain showed that information relating to pain was best captured using an 11-point scoring scale (0, 1, 2, . . . , 10) – sometimes referred to as a *Likert scale* – or a verbal rating scale with five points (mild, discomforting, distressing, horrible, excruciating). In addition, around 10% of the patients were unable to understand the requirement for completion of the VAS for pain. These ordered categorical scales may well be as precise or more precise than the VAS and at the same time prove to be more effective because patients understand them better.

## CHAPTER 2

# Sampling and inferential statistics

### 2.1 Sample and population

Consider the comparison of a new treatment A to an existing treatment B for lowering blood pressure in mild to moderate hypertension in the context of a clinical trial conducted across Europe. The characteristics of the *population* of mild to moderate hypertensive patients to be studied will be defined by the inclusion (and exclusion) criteria and may well contain several millions of individuals. In another sense, this population will be infinite if we also include those patients satisfying the same inclusion/exclusion criteria in the future. Our clinical trial will, for example, involve selecting a *sample* of, say, 200 individuals from this population and randomly assigning 100 to treatment group A and 100 to treatment group B.

Each subject in the sample will be observed and provide a value for the fall in diastolic blood pressure, the primary endpoint. The mean fall in blood pressure in groups A and B will then be computed and compared. Suppose that the group A and group B means are, respectively,

$$\bar{x}_1 = 8.6 \text{ mmHg}$$

$$\bar{x}_2 = 3.9 \text{ mmHg}$$

The conclusion we draw will be based on a comparison of these means, and in general, there are three possibilities in relation to what we conclude:

- Treatment A is better than treatment B.
- Treatment B is better than treatment A.
- There are no differences.

Suppose in this case we conclude, on the basis of the data, that treatment A is better than treatment B. This statement of course is correct in terms of what we have seen on average in the sample data, but the statement we are making is in fact stronger than that; it is a statement about the complete population from which the sample was drawn. We are concluding that treatment A will, on average, work better than treatment B in this population; we are extrapolating from the sample to the population. Statisticians talk in terms of making *inferences*. On

the basis of the sample data, we are inferring things about the complete population.

In one sense, moving from the sample to the population in this way is a leap of faith! However, it should work, provided the sample is representative of the population. If it is not representative, but we can assume that the treatment difference is homogeneous across the population, then we will still obtain a valid estimate of that treatment difference.

To make any progress in understanding how inferential statistics works, we need to understand what happens when we take a sample from a population. In a later section, we will explore this through a computer simulation and see how all of this comes together in practical applications.

## 2.2 Sample statistics and population parameters

### 2.2.1 Sample and population distribution

The *sample histogram* in Figure 2.1 provides a visual summary of the distribution of total cholesterol in a group of 100 patients at baseline (artificial data). The  $x$ -axis is divided into intervals of width 0.5 mmol/l, and the  $y$ -axis counts the number of individuals with values within those intervals.

We will sometimes refer to the sample histogram as the *sample distribution*.

These data form the sample, and they have been taken from a well-defined population, which sits in the background. We can also envisage a corresponding histogram for the complete population, and this will have a smooth shape as a result of the size of that population; we use the terms *population histogram* or *population distribution* for this. Figure 2.2 shows the population histogram superimposed on the sample histogram. Provided the sample is representative of the population, the sample and population histograms should be similar. In practice,

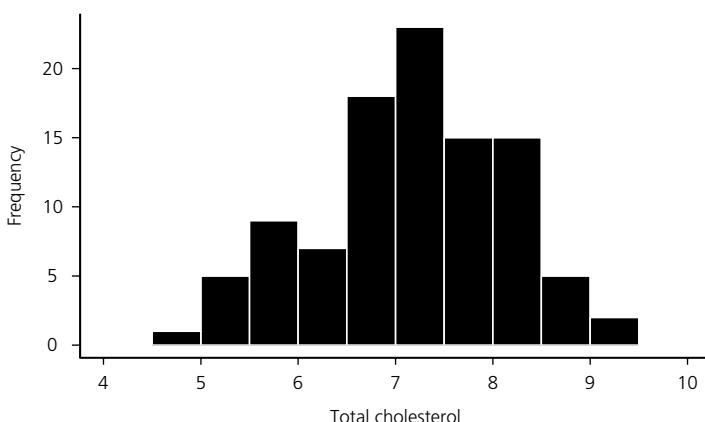
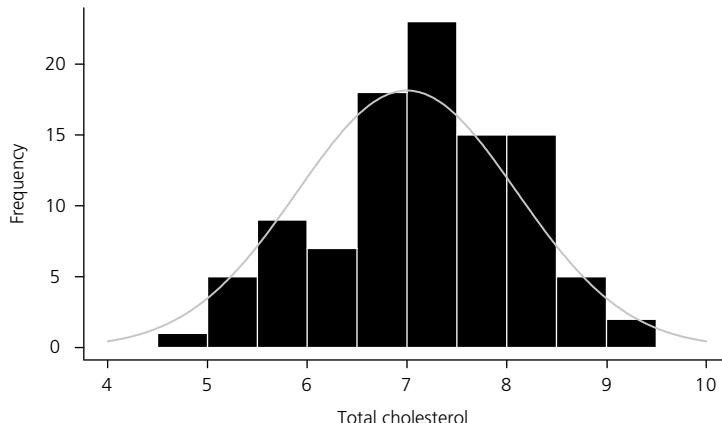


Figure 2.1 Histogram for total cholesterol ( $n = 100$ )



**Figure 2.2** Sample histogram ( $n = 100$ ) and population histogram

remember that we only see the sample histogram, and the population histogram is hidden from us; indeed, we want to use the sample distribution to tell us about the distribution in the population.

We are usually interested in two aspects of these histograms: what is happening on average and the patient-to-patient variation or spread of the data values. The average will be used as a basis for measuring the signal – and in particular we will be looking at differences between two averages for that – while the patient-to-patient variation relates directly to the noise as discussed in Section 1.8.2.

The measures of average that we commonly use are the mean and the median, while the standard deviation provides us with our measure of patient-to-patient variation.

### 2.2.2 Median and mean

The *median* (denoted by  $\tilde{x}$ ) is the middle value when the data values are ordered from smallest to largest. The median can only be defined in this way when there are an odd number of values (subjects). When the number of subjects is even, we define the median to be the average of the middle two values. For the data in Figure 2.1,  $n = 100$  and the median  $\tilde{x} = 7.20 \text{ mmol/l}$ . The *mean* (denoted by  $\bar{x}$ ) is the arithmetic average:  $\bar{x} = \frac{1}{n} \sum x$ . For the data in Figure 2.1,  $n = 100$  and  $\bar{x} = 7.16 \text{ mmol/l}$ .

### 2.2.3 Standard deviation

The *standard deviation* (denoted  $s$  or  $sd$ ) is a measure of patient-to-patient variation and provides our measure of noise when we are dealing with parallel-group studies. There are other potential measures, but this quantity is used as it possesses a number of desirable mathematical properties and appropriately captures the overall amount of variability within the sample of patients.

It is related to another quantity called the *variance*, and

$$\text{Variance} = (\text{standard deviation})^2$$

If the standard deviation is 3, the variance is 9; and if the variance is 25, the standard deviation is 5.

The method of calculation of the standard deviation seems, at least at face value, to have some arbitrary elements to it. There are several steps:

- 1 Calculate the mean of all values.
- 2 Calculate the difference between each individual value and the mean, and square each of those differences.
- 3 Take the average of these squared differences but with the *average* calculated by dividing by  $n - 1$ , not  $n$ ; the resulting quantity is called the *variance* (with units mmol/l<sup>2</sup> for the data in our example).
- 4 Take the square root of the variance to revert to the original units, mmol/l; this is the standard deviation.

For the example data,  $n = 100$  and  $s = 0.98$  mmol/l.

People often ask, why divide by  $n - 1$  rather than  $n$ ? Well, the answer is fairly technical. It can be shown mathematically that dividing by  $n$  gives a quantity that, on average, underestimates the true standard deviation, particularly in small samples, and dividing by  $n - 1$  rather than  $n$  corrects for this underestimation. Of course, for a large sample size, it makes very little difference – dividing by 99 is much the same as dividing by 100.

Another frequent question is, why square, average and then square root – why not simply take the average distance of each point from the mean without bothering about the squaring? Well, you could do this, and yes, you would end up with a measure of patient-to-patient variability; this quantity is referred to as the *mean absolute deviation* and is indeed sometimes used as a measure of spread. The standard deviation, however, has several strong theoretical properties that we will need in our subsequent development, and therefore we will go with that as our measure of variation.

## 2.2.4 Notation

To distinguish between quantities measured in the sample and corresponding quantities in the population, we use different symbols:

The mean in the sample is denoted  $\bar{x}$ .

The mean in the population is denoted  $\mu$ .

The standard deviation in the sample is denoted  $s$  or *sd*.

The standard deviation in the population is denoted  $\sigma$ .

Remember,  $\bar{x}$  and  $s$  are quantities that we calculate from our data, while  $\mu$  and  $\sigma$  are theoretical quantities (parameters) that are unknown to us but nonetheless exist in the context of the broader population from which the sample (and

therefore the data) is taken. If we had access to every single subject in the population, then yes, we could compute  $\mu$  and  $\sigma$ , but this is never going to be the case. We can also think of  $\mu$  and  $\sigma$  as the *true* mean and *true* standard deviation, respectively, in the population as a whole.

The calculation of mean and standard deviation only really makes sense when we are dealing with continuous and score endpoints. These quantities have little relevance when we are looking at binary or ordered categorical data. In such situations, we would tend to use proportions in the various categories as our summary statistics and population parameters of interest.

For a binary endpoint:

The sample proportion is denoted  $r$ .

The population proportion is denoted  $\theta$ .

## 2.2.5 Box plots

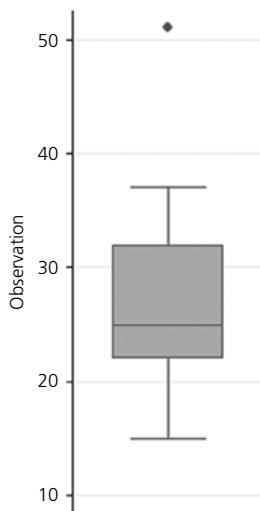
The *box plot* is a useful way to display the distribution of data in a sample. Figure 2.3 displays a box plot for some artificial data where the 11 observations have values as follows:

32 15 22 31 32 23 19 51 37 24 25

When ordered from the smallest to largest, these observations are

15 19 22 23 24 25 31 32 32 37 51

The median value is 25. The *upper quartile* is the value in general that cuts off the largest 25% of the observations, while the *lower quartile* in general is the value



**Figure 2.3** Box plot for artificial data

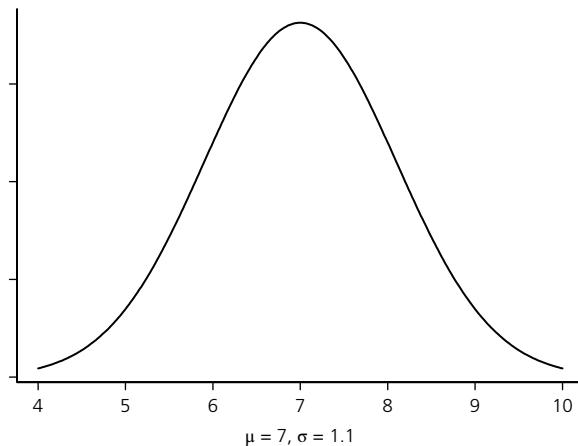
that cuts off the smallest 25% of the observations. The easiest way to obtain the upper quartile in this small data set is to look firstly at the values above the median and to cut this group of observations into two further halves. The values above the median are 31 32 32 37 51, and the value 32 cuts these into two halves, so 32 is the upper quartile. The calculation for the lower quartile is similar but looks at the data below the median. The lower quartile for these data is 22. The *interquartile (IQ) range* is the difference between the upper and lower quartiles,  $10 (= 32 - 22)$  in this case. *Whiskers* are then placed at the furthest observed values in the data that are within  $1.5 \times \text{IQ range}$  of the quartiles. In the example,  $1.5 \times \text{IQ range} = 15$ , and we look to find the largest value in the data that is within 47 ( $=\text{upper quartile} + 15$ ), which is 37, and the smallest value that is within 7 ( $=\text{lower quartile} - 15$ ), which is 15. The whiskers are then placed, respectively, at 15 and 37. All values outside of the whiskers are then marked individually.

These plots give a good visual impression about what is happening on average through the median, while the lower and upper quartiles show the spread of the *middle* 50% of the data. The positioning of the whiskers helps us to identify if the distribution of the data is skewed rather than symmetric. Not infrequently, we see the upper whisker further from the median than the lower whisker, indicating what we refer to as a *positively skewed* distribution. We refer to the opposite setting, the lower whisker further from the median than the upper whisker, as *negatively skewed*. Values outside of the whiskers can be considered as outlying values. We will discuss *outliers*, and action to be taken when we see such values, in Section 11.7.

## 2.3 The normal distribution

The *normal* or *Gaussian* distribution was first discovered by de Moivre, a French mathematician, in 1733. Gauss came upon it somewhat later, just after 1800, but from a completely different start point. Nonetheless, it is Gauss who has his name attached to this distribution.

The normal distribution is a particular form of population histogram. It is symmetric around the mean  $\mu$  and is bell shaped. It has been noted empirically that in many situations, data, particularly when collected from random systems, gives a histogram with this *normal distribution* shape. In fact, a very powerful theorem called the *central limit theorem* looks at the behaviour of data and says that under certain general conditions, data behave according to this distribution. An example of a normal distribution is given in Figure 2.4. The normal distribution is all about the behaviour of averages in random systems. I am going to make a very bold philosophical statement at this point, and that is, ‘at the helicopter level, randomness is entirely predictable’. To explain, consider a lottery. In the UK, the national lottery between its launch in 1994 and 2015 involved choosing six numbers in the range 1 to 49, with draws taking place on Wednesdays and



**Figure 2.4** Normal distribution for total cholesterol at baseline ( $\mu = 7$  mmol/l,  $\sigma = 1.1$  mmol/l)

Saturdays. The mechanism for choosing the six winning numbers involved a revolving drum containing the 49 numbered balls, with the chosen balls (numbers) rolling down a tube and coming to rest. If you had the ticket containing the six chosen numbers, then you won an obscene amount of money, something close to a banker's bonus! It has been shown by looking at the data over the 20 or so years that the lottery ran in that format that the process for choosing the numbers was entirely random. However, if we look at all the draws over the 20-year period and plot the frequency distribution according to the numbers 1 to 49, it can be seen that 1 occurred almost the same number of times as 2, and 2 occurred almost the same number of times as 3, and so on. This is precisely what you would expect, and in the long term with a completely random process, the frequency for all of the numbers would be exactly the same – 'randomness is entirely predictable'. If these frequencies were not the same, then the process would not be random! You might ask, where does the normal distribution fit into this? Well, suppose on each occasion that the lottery numbers were drawn we calculated the numerical average of the six numbers that were chosen and then plotted the histogram of these average values over the 20-year period. The histogram would look exactly like a normal distribution with mean value at 25. The mean is 25 because that is the average of the numbers 1 to 49, and the normal shape will occur because that is exactly how averages behave in a random system; this behaviour is completely predictable. The central limit theorem is in fact a very sophisticated law of averages. To extend this argument, it is invariably the case that each clinical outcome measure that we record – for example, blood pressure or cholesterol level – is the result of the influence or averaging of a number of different but interrelated physiological processes, and these considerations lead us to consider the normal distribution as a suitable distribution that can summarise the behaviour of such outcomes for a population of individuals.

As a consequence of both this theoretical base and the empirical evidence, we often assume that the data we are collecting have been drawn from a distribution with a normal shape; we assume that our data are *normally distributed*.

One further point relating to the normal distribution in the population is that, because of the symmetry, the median and the mean take the same value. This is a property of any population distribution for data that is symmetric.

Briefly returning to Gauss, one can gauge the importance of *his* discovery by observing the old German 10 Mark banknote in Figure 2.5. Here we have Gauss, and just above the 10 and to the right, we can see the normal distribution and its mathematical equation.

When the population does indeed follow this distribution, then the standard deviation,  $\sigma$ , has a more specific interpretation. If we move  $\sigma$  units below the mean to  $\mu - \sigma$  and  $\sigma$  units above the mean to  $\mu + \sigma$ , then that interval  $(\mu - \sigma, \mu + \sigma)$  will capture 68.3% of the population values. This is true whatever we are considering – diastolic blood pressure, fall in diastolic blood pressure over a six-month period, cholesterol level, FEV<sub>1</sub> etc. – and whatever the values of  $\mu$  and  $\sigma$ ; in all cases, 68.3% of the patients in the population will have data values in the range  $\mu - \sigma$  to  $\mu + \sigma$  provided the data are normally distributed.

Several further properties hold, as shown in Table 2.1.



**Figure 2.5** Gauss on the Deutsche 10 Mark note. Source: hansmuench/Adobe Stock

**Table 2.1** Probabilities for the normal distribution

Range	Percentage of patients
$\mu - 2\sigma$ to $\mu + 2\sigma$	95.4
$\mu - 3\sigma$ to $\mu + 3\sigma$	99.7
$\mu - 1.645\sigma$ to $\mu + 1.645\sigma$	90
$\mu - 1.960\sigma$ to $\mu + 1.960\sigma$	95
$\mu - 2.576\sigma$ to $\mu + 2.576\sigma$	99

Note that the normal distribution curve has a mathematical equation, and integrating the equation of this curve – for example, between  $\mu - 2\sigma$  and  $\mu + 2\sigma$  – irrespective of the values of  $\mu$  and  $\sigma$ , will always give the answer 0.954. Therefore, 95.4% of the area under the normal curve is contained between  $\mu - 2\sigma$  and  $\mu + 2\sigma$ , and it is this area calculation that also tells us that 95.4% of the individuals within the population will have data values in that range.

**Example 2.1** Normal distribution (Figure 2.4)

A population of patients in a cholesterol-lowering study have their total cholesterol measured at baseline. Assume that total cholesterol is normally distributed with mean of 7.0 mmol/l and standard deviation of 1.1 mmol/l so that the variance is 1.21 ( $=1.1^2$ ). We write this as  $N(7.0, 1.21)$ . For historical reasons, we put the variance as the second parameter here. Under these assumptions, the following results hold:

- 68.3% of the patients have total cholesterol in the range 5.9 mmol/l to 8.1 mmol/l.
- 90% of the patients have values in the range 5.19 mmol/l to 8.81 mmol/l.
- 95.4% of the patients have values in the range 4.8 mmol/l to 9.2 mmol/l.

## 2.4 Sampling and the standard error of the mean

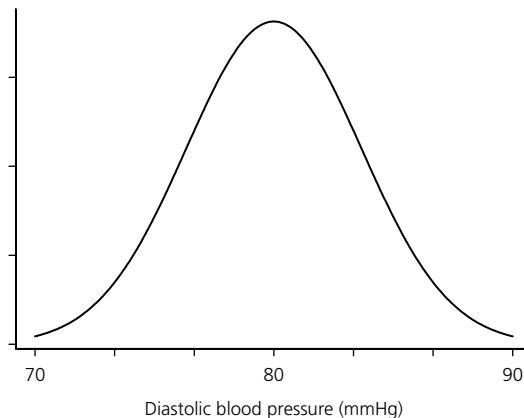
Earlier in this chapter, we spoke about the essence of inferential statistics, drawing conclusions about a population based upon a sample taken from that population. To understand how this is done, we need to understand what happens when we take a sample from the population:

- Do we always reach the correct conclusion about the population?
- Are we sometimes misled by the sample data?
- How big a sample do we need to be confident that we will end up with a correct conclusion?

To gain an understanding of the sampling process, we have undertaken a computer simulation. For this simulation, we have set up, on the computer, a very large population of patients whose diastolic blood pressures have been recorded. The population has been structured to be normally distributed with mean of 80 mmHg and a standard deviation of 4 mmHg,  $N(80, 16)$ , as shown in Figure 2.6.

Imagine that this is a real clinical trial setting, and our objective is to find out the value of the mean diastolic blood pressure in the population (but remember that because this is a computer simulation, we know the answer!). So, let's take a sample of size 50.

The mean,  $\bar{x}$ , from this sample turned out to be 80.218 mmHg. The one thing you will notice about this value is that it is not equal to  $\mu$ , which is 80 mmHg, so we see immediately that the sampling process does not necessarily hit the absolute truth. So, let's take a second sample. The second sample gave a mean of



**Figure 2.6** Normal distribution,  $N(80, 16)$

80.767 mmHg; again, this value is not equal to the true mean. Not only that, but the second sample has given a different answer to the first sample. We repeated this sampling process 100 times, going back to the population and taking further samples of size 50. The complete set of mean values is given in Table 2.2.

There are two things you will notice about this list of means. Firstly, not one of them has hit the true mean value of 80 mmHg. Secondly, all the values are different. The implication of this is as follows. Whenever you get an answer in a clinical trial, the only thing you know for certain is that it is the wrong answer! Not only that, but if you were to repeat the trial under identical circumstances with the same protocol, same investigators and so on, but with a new set of patients, you would get a different answer. These are simply aspects of sampling; it is by no means a perfect process. This so-called *sampling variation* is fundamentally a result of patient-to-patient variation; patients behave differently, and successive samples of 50 patients are going to give different results. To be able to work in this uncertain environment, we need to quantify the extent of the sampling variation.

The standard deviation of this list of  $\bar{x}$  values can be calculated using the method described earlier and gives a measure of the inherent variability in the sampling process. A large value for this standard deviation would indicate that the  $\bar{x}$  values are all over the place and we are in an unreliable situation in terms of estimating where the true mean ( $\mu$ ) lies; a small value for this standard deviation would indicate that the  $\bar{x}$  values are closely bunched together and the sampling process is giving a consistent, reliable value. This standard deviation for the list of  $\bar{x}$  values was calculated to be 0.626 and provides a measure of the variation inherent in the sampling process.

In practice, we will never have the luxury of seeing the behaviour of the sampling process in this way; remember, this is a computer simulation. However, there is a way of estimating the standard deviation associated with the sampling process through a mathematical expression applied to the data from a single

**Table 2.2** Mean values  $\bar{x}$  from 100 samples of size 50 from  $N(80, 16)$ ; units in mmHg

1	80.218	26	79.894	51	79.308	76	80.821
2	80.767	27	79.629	52	79.620	77	80.498
3	78.985	28	80.233	53	80.012	78	79.579
4	81.580	29	79.738	54	80.678	79	81.051
5	79.799	30	80.358	55	80.185	80	80.786
6	80.302	31	79.617	56	79.901	81	79.780
7	79.094	32	79.784	57	79.778	82	79.802
8	80.660	33	79.099	58	79.597	83	80.510
9	79.455	34	79.779	59	79.320	84	79.592
10	79.275	35	81.438	60	79.076	85	79.617
11	80.732	36	80.066	61	80.580	86	79.587
12	79.713	37	79.591	62	79.878	87	79.124
13	79.314	38	80.254	63	79.656	88	79.520
14	80.010	39	80.494	64	79.302	89	79.587
15	79.481	40	79.259	65	80.242	90	79.544
16	79.674	41	80.452	66	78.344	91	80.054
17	80.049	42	80.957	67	80.653	92	80.458
18	79.156	43	80.113	68	79.848	93	79.895
19	80.826	44	80.043	69	80.294	94	79.293
20	80.321	45	80.436	70	80.797	95	79.376
21	79.476	46	81.220	71	79.226	96	80.296
22	80.155	47	79.391	72	78.883	97	79.722
23	79.429	48	80.552	73	79.871	98	78.464
24	80.775	49	80.422	74	80.291	99	78.695
25	80.490	50	80.265	75	79.544	100	79.692

sample. This formula is given by  $s / \sqrt{n}$ , where  $s$  is the *sd* from the single sample and  $n$  is the sample size.

So, in practice, we calculate the mean from a sample (size  $n$ ) of data plus the corresponding standard deviation,  $s$ . We then divide the standard deviation by  $\sqrt{n}$ , and the resulting numerical value gives an estimate of the standard deviation associated with the sampling process: the standard deviation for the repeat  $\bar{x}$  values had we undertaken the sampling many times.

### Example 2.2 Sampling variation

In the first computer simulation,  $n = 50$ ,  $\bar{x} = 80.218$  and  $s = 4.329$ , the standard deviation of the 50 patient diastolic blood pressures in that sample.

The estimated standard deviation associated with the repeat  $\bar{x}$  values is then given by

$$\frac{s}{\sqrt{n}} = \frac{4.329}{\sqrt{50}} = 0.612$$

In other words, were we to repeat the sampling process, getting a list of  $\bar{x}$  values by repeatedly going back to the population and sampling 50 subjects, 0.612 gives us an estimate of the standard deviation associated with these  $\bar{x}$  values.

One potentially confusing issue here is that there are two standard deviations: one measures the patient-to-patient variability from the single sample/trial, while the second measures the mean-to-mean variation you would get by repeating the sampling exercise. To help distinguish the two, we reserve the term *standard deviation* for the first of these (patient-to-patient variation), and we call the second the *standard error* (*se*) of  $\bar{x}$  (mean-to-mean variation). In the earlier example, 0.612 is the standard error of  $\bar{x}$  from the first computer simulation sample, an estimate of the standard deviation of mean values under repeated sampling. Note that this value is close to 0.626, the *standard error* calculated directly via the computer simulation through repetition of the sampling process.

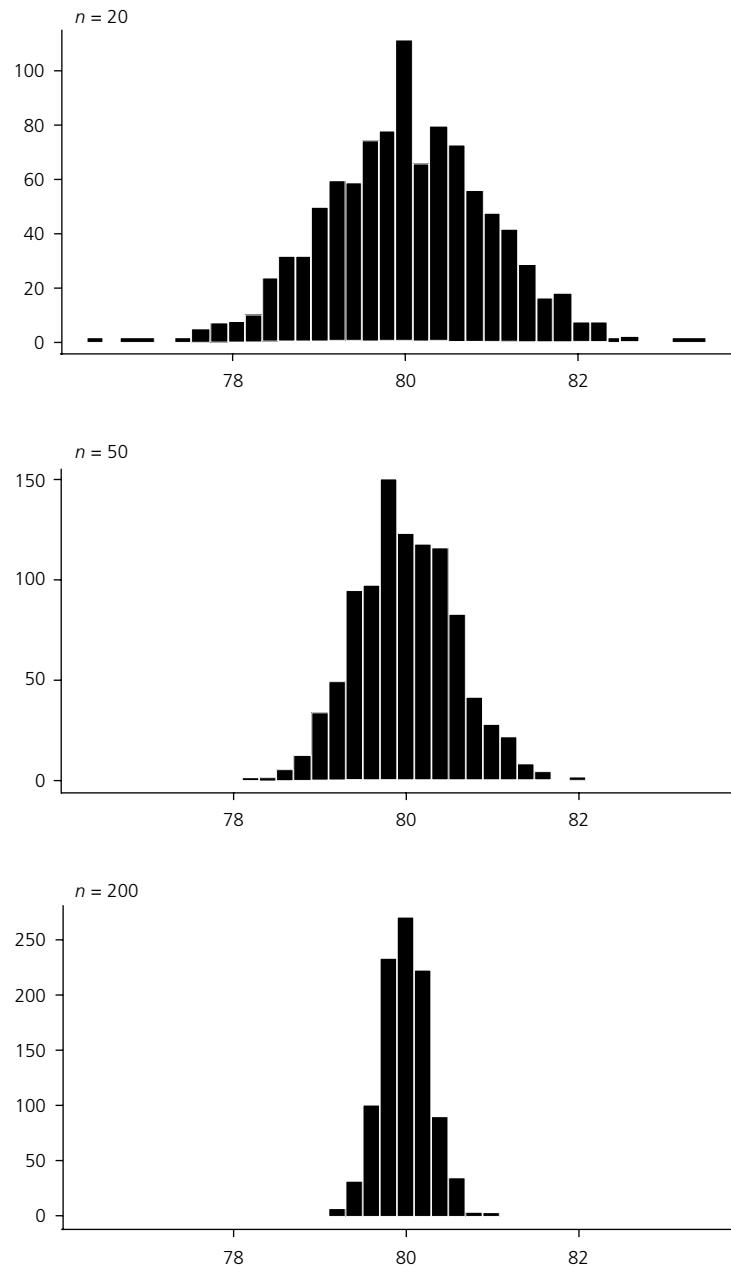
A small standard error tells us that we are in a reliable sampling situation where the repeat mean values are very likely to be closely bunched together; a large standard error tells us we are in an unreliable situation where the mean values vary considerably.

It is not possible at this stage to say precisely what we mean by small and large in this context; we need the concept of the confidence interval to be able to say more in this regard, and we will cover this topic in the next chapter. For the moment, just look upon the standard error as an informal measure of precision: high values mean low precision, and low values mean high precision. Further, if the standard error is small, it is likely that our estimate  $\bar{x}$  will be close to the true mean,  $\mu$ . This is because the  $\bar{x}$  values will be bunched together, as a consequence of a small standard error; and as they will vary symmetrically around the true mean, they will all be *close* to that true mean. If the standard error is large, however, there is no guarantee that we will be close to the true mean, as we do not have that close bunching of the  $\bar{x}$  values even though on average they will centre on the true mean.

Figure 2.7 shows histograms of  $\bar{x}$  values for sample sizes of 20, 50 and 200 from 100 simulations (samples) in each case. It is clear that for  $n = 20$ , there is considerable variation; there is no guarantee that the mean from a particular sample will be close to  $\mu$ . For  $n = 50$ , things are not quite so bad, although the sample mean could still be out at 82.0 or 78.2. For the sample size of 200, there is only a small amount of variability; over 250 of the 1000 mean values are within 0.1 units of the true mean. These histograms/distributions are referred to as *sampling distributions*. They are the distributions of  $\bar{x}$  from the sampling process. Remember, when you conduct a trial and get a mean value, it is just one realisation of such a sampling process. The standard error is the estimated standard deviation of the  $\bar{x}$  values in these histograms and measures their spread.

## 2.5 Standard errors more generally

The standard error concept can be extended in relation to any *statistic* (quantity) calculated from the data.



**Figure 2.7** Sampling distribution of the mean  $\bar{x}$ ; data from  $N(80, 16)$ , sample size  $n$

### 2.5.1 The standard error for the difference between two means

As a further example, imagine a placebo-controlled cholesterol-lowering trial. Generally, in such trials, patients in each of the treatment groups will receive lifestyle and dietary advice plus medication, either active or placebo, according to the randomisation scheme. Let  $\mu_1$  be the true mean reduction in total cholesterol in the active treatment group, and let  $\mu_2$  be the corresponding true mean reduction in the placebo group. So  $\mu_1$  is the mean reduction we would get if all patients in the population were given active treatment, and  $\mu_2$  is the mean we would get if all patients were given placebo. The lifestyle and dietary advice will, of itself, have a positive effect, and coupled with the *placebo effect*, we will most likely see a mean reduction in each of the two treatment groups. The issue is, are we seeing a larger reduction in the active group compared to the placebo group? With this in mind, the main interest lies not in the individual means but in their difference  $\mu_1 - \mu_2$ , the *treatment effect*.

Our best guess for the value of  $\mu_1 - \mu_2$  is the observed difference in the sample means  $\bar{x}_1 - \bar{x}_2$  from the trial.

Suppose that the value of  $\bar{x}_1 - \bar{x}_2$  turns out to be 1.4 mmol/l. We know full well that this will not be equal to the true difference in the means,  $\mu_1 - \mu_2$ . We also know that if we were to repeat the trial under identical circumstances with the same protocol, same investigators and so on, but of course with a different sample of patients, we would come up with a different value for  $\bar{x}_1 - \bar{x}_2$ .

We need to have some measure of precision and reliability, and this is provided by the standard error of  $\bar{x}_1 - \bar{x}_2$ . Again, we have a formula for this:

$$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Here  $n_1$  and  $n_2$  are the numbers of patients in each of the two treatment groups, and  $s_1$  and  $s_2$  are the standard deviations in each of those groups. This expression allows us to estimate the standard deviation of the  $\bar{x}_1 - \bar{x}_2$  values that we would get were we to repeat the trial.

#### Example 2.3 Standard error for the difference between two means

In a placebo-controlled trial in cholesterol lowering, Table 2.3 contains the data for the two treatment groups.

The standard error for the difference in the means,  $\bar{x}_1 - \bar{x}_2$ , is

$$\sqrt{\left(\frac{1}{24} + \frac{1}{23}\right) \times \frac{(24 - 1) \times 0.92 \times 0.92 + (23 - 1) \times 1.05 \times 1.05}{24 + 23 - 2}} = 0.29$$

**Table 2.3** Cholesterol-lowering data (artificial)

	<b>n</b>	<b>Mean (mmol/l)</b>	<b>sd (mmol/l)</b>
Active	24	2.7	0.92
Placebo	23	1.3	1.05

Small values of this standard error indicate high reliability; it is likely that the observed value  $\bar{x}_1 - \bar{x}_2$  for the treatment effect is close to the true treatment effect,  $\mu_1 - \mu_2$ . In contrast, a large value for the standard error tells us that  $\bar{x}_1 - \bar{x}_2$  is not a reliable estimate of  $\mu_1 - \mu_2$ . Again, we will not discuss specifically what is meant by *small* and *large* in this context, but we will come back to this in Chapter 3.

### 2.5.2 Standard errors for proportions

So far, we have considered standard errors associated with means and differences between means. When dealing with binary data and proportions, different formulas apply.

In Section 2.2.4, we let  $r$  denote a proportion in the sample and  $\theta$  the corresponding proportion in the population. For a single proportion  $r$ , the standard error formula is  $\sqrt{r(1-r)/n}$ , where  $n$  is the number of subjects in the sample.

For the difference between two proportions – for example, if we are looking at the difference  $r_1 - r_2$  between the response rate in the active group (group 1) and the response rate in the placebo group (group 2) – the standard error formula is  $\sqrt{r_1(1-r_1)/n_1 + r_2(1-r_2)/n_2}$ , where  $n_1$  and  $n_2$  are the numbers of subjects in groups 1 and 2, respectively.

### 2.5.3 The general setting

More generally, whatever statistic we are interested in, there is usually a formula that allows us to calculate its standard error. The formulas change, but their interpretation always remains the same; a small standard error is indicative of high precision and high reliability. Conversely, a large standard error means that the observed value of the statistic is an unreliable estimate of the true (population) value. It is also always the case that the standard error is an estimate of the standard deviation of the list of repeat values of the statistic that we would get were we to repeat the sampling process, a measure of the inherent sampling variability.

As discussed in the previous section, the standard error simply provides indirect information about reliability; it is not something we can use in any specific way, as yet, to tell us where the truth lies. We also have no way of saying what is large and what is small in standard error terms. However, in the next chapter, we will cover the concept of the confidence interval and see how this provides a methodology for using the standard error to enable us to make statements about where we think the true (population) value lies.

## CHAPTER 3

# Confidence intervals and $p$ -values

### 3.1 Confidence intervals for a single mean

#### 3.1.1 The 95% confidence interval

We have seen in the previous chapter that it is not possible to make a precise statement about the exact value of a population parameter based on sample data and that this is a consequence of the inherent variation in the sampling process. The *confidence interval* (CI) provides us with a compromise: rather than trying to pin down precisely the value of the mean  $\mu$  or the difference between two means;  $\mu_1 - \mu_2$ , for example; we give a range of values within which we are fairly certain that the true value lies.

We will first look at the way we calculate the CI for a single mean  $\mu$  and then talk about its interpretation. Later in this chapter, we will extend the methodology to deal with  $\mu_1 - \mu_2$  and other parameters of interest.

In the computer simulation in Chapter 2, the first sample ( $n = 50$ ) gave summary statistics (to two decimal places) as follows:

$$\bar{x} = 80.22 \text{ mmHg and } s = 4.33 \text{ mmHg}$$

The lower end of the CI, the *lower confidence limit*, is then given by

$$\bar{x} - 1.96 \frac{s}{\sqrt{n}} = 80.22 - \left( 1.96 \times \frac{4.33}{\sqrt{50}} \right) = 79.02$$

The upper end of the confidence interval, the *upper confidence limit*, is given by

$$\bar{x} + 1.96 \frac{s}{\sqrt{n}} = 80.22 + \left( 1.96 \times \frac{4.33}{\sqrt{50}} \right) = 81.42$$

The interval (79.02, 81.42) forms the *95% confidence interval*.

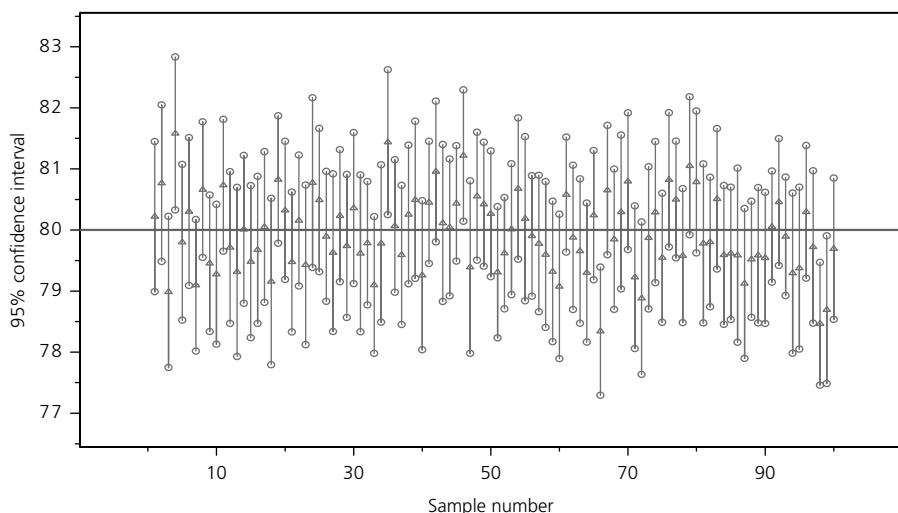
These data arose from a computer simulation where, of course, we know that the true mean  $\mu$  is 80 mmHg, so we can see that the method has worked in the sense that  $\mu$  is contained within the range 79.02 to 81.42.

The second sample in the computer simulation gave the following summary statistics:  $\bar{x} = 80.77$  mmHg and  $s = 4.50$  mmHg – and this results in the 95% CI as (79.52, 82.02). Again, we see that the interval has captured the true mean.

Now look at all 100 samples taken from the normal population with  $\mu = 80$  mmHg. Figure 3.1 shows the 95% CIs plotted for each of the 100 simulations. A horizontal line has been placed at 80 mmHg to allow the CIs to be judged in terms of capturing the true mean.

Most of the 95% CIs do contain the true mean of 80 mmHg, but not all. Sample number 4 gave a mean value  $\bar{x} = 81.58$  mmHg with a 95% CI (80.33, 82.83), which missed the true mean at the lower end. Similarly, samples 35, 46, 66, 98 and 99 gave CIs that do not contain  $\mu = 80$  mmHg. So, we have a method that seems to work most of the time, but not all the time. For this simulation as a whole, we have a 94% (94/100) success rate. If we were to extend the simulation and take many thousands of samples from this population, constructing 95% CIs each time, we would see a success rate of 95%; exactly 95% of those intervals would contain the true (population) mean value. This provides us with the interpretation of a 95% CI in practice: when we construct a 95% CI from data, we can be 95% certain that the true mean lies within the calculated range. Why? Because 95% of these intervals will indeed contain  $\mu$  in the long run. Of course, in any particular case, we do not know whether our CI is one of those 95% or whether we have been unlucky and gotten one of the 5% that do not contain the truth. In such cases, we will have been misled by the data.

Just as an aside, look back at the formula for the 95% CI. Where does the 1.96 come from? It comes from the normal distribution; 1.96 is the number of



**Figure 3.1** Computer simulation, 95% CIs,  $n = 50$ , population mean = 80 mmHg

standard deviations you need to move out to capture 95% of the values in the population. The reason we get the so-called 95% coverage for the CIs is directly linked to this property of the normal distribution.

### 3.1.2 Changing the confidence coefficient

We now have a procedure that allows us to make a statement about the value of  $\mu$  with 95% confidence, but we have to accept that such intervals will mislead us 5% of the time. You may feel that this is too risky and instead request a 99% CI, which will only mislead you 1% of the time. That's fine, but the formula will change, and instead of using 1.96 to give 95% coverage, we will need to use 2.576 to give us 99% coverage. The formula for the 99% CI is then

$$\bar{x} - 2.576 \frac{s}{\sqrt{n}} \text{ to } \bar{x} + 2.576 \frac{s}{\sqrt{n}}$$

For the first sample in the computer simulation, the 99% CI is (78.64, 81.80). This is a wider interval than the 95% interval; the more confidence we require, the more we have to hedge our bets. It is fairly standard to use 95% CIs, and this links with the conventional use of 0.05 (or 5%) for the cut-off for statistical significance. We will say more about this link in Section 9.8. Under some circumstances, we also use 90% CIs, and we will mention one such situation later. In multiple testing, it is also sometimes the case that we use *confidence coefficients* larger than 95%; again, we will discuss the circumstances where this might happen in a later chapter.

### 3.1.3 Changing the multiplying constant

The formula for the 95% CI (and also for the 99% CI) given previously is not quite correct. It is correct up to a point in that it will work for large sample sizes. For smaller sample sizes, we need to change the multiplying constant according to the values in Table 3.1.

The reason is again technical but relates to the uncertainty associated with the use of the sample standard deviation ( $s$ ) in place of the true population value ( $\sigma$ ) in the formula for the standard error. When  $\sigma$  is known, the multiplying constants given earlier apply precisely. When  $\sigma$  is not known (the usual case), we make the CI slightly wider to account for this uncertainty. When  $n$  is large, of course,  $s$  will be close to  $\sigma$ , so the earlier multiplying constants apply approximately.

Multiplying factors are given here for 90%, 95% and 99% CIs. Note that the constants 1.960 and 2.576, those used for 95% and 99% CIs previously, appear at the foot of the final two columns. The column on the left-hand side, labelled *degrees of freedom*, is closely linked to sample size. When calculating a CI for a mean, as in this section, we use the row corresponding to sample size ( $=n$ ) – 1, so degrees of freedom for a single sample =  $n$  – 1. Do not agonise over the term *degrees of freedom*; just think of getting the appropriate multiplying constant by

**Table 3.1** Multiplying constants for calculating confidence intervals

Degrees of freedom (df)	Confidence coefficient		
	90%	95%	99%
5	2.02	2.57	4.04
10	1.81	2.23	3.17
11	1.80	2.20	3.11
12	1.78	2.18	3.06
13	1.77	2.16	3.01
14	1.76	2.15	2.98
15	1.75	2.13	2.95
16	1.75	2.12	2.92
17	1.74	2.11	2.90
18	1.73	2.10	2.88
19	1.73	2.09	2.86
20	1.73	2.09	2.85
25	1.71	2.06	2.79
30	1.70	2.04	2.75
35	1.69	2.03	2.72
40	1.68	2.02	2.70
45	1.68	2.01	2.69
50	1.68	2.01	2.68
100	1.66	1.99	2.63
200	1.65	1.97	2.60
$\infty$	1.645	1.960	2.576

going into the row sample size –1. A more complete table can be found in many standard statistics textbooks. Alternatively, most statistics packages contain a function that will give the multiplying constants for any value of degrees of freedom.

So, if we were calculating a CI for a mean  $\mu$  from a sample of size 16, then we would look in row 15 for the multiplying constant and use 2.13 in place of 1.96 in the calculation of the 95% CI and 2.95 in place of 2.576 for the 99% CI.

### 3.1.4 The role of the standard error

Note the role played by the standard error in the formula for the CI. We have previously seen that the standard error of the mean provides an indirect measure of the precision with which we have estimated the value of the true mean. The CI has now translated the numerical value for the standard error into something useful in terms of being able to make a statement about where  $\mu$  lies. A large standard error will lead to a wide CI reflecting the imprecision and resulting limited information about the value of  $\mu$ . In contrast, a small standard error will produce a narrow CI, giving us a very definite statement about the value of  $\mu$ .

For sample sizes beyond about 30, the multiplying constant for the 95% CI is approximately 2. Sometimes, for reasonably large sample sizes, we do not

agonise over the value of the multiplying constant and simply use the value 2 as a good approximation. This gives us an approximate formula for the 95% CI as  $(\bar{x} - 2se, \bar{x} + 2se)$ .

Finally, returning to the formula for the standard error,  $s/\sqrt{n}$ , we can, at least in principle, see how we could make the standard error smaller, increase the sample size  $n$  and/or reduce the patient-to-patient variability. These actions will translate into narrower CIs.

**Example 3.1** Confidence interval for a single mean

In an asthma trial comparing two short-acting treatments, the hypothetical data in Table 3.2 were obtained for the increase in  $\text{FEV}_1$ .

**Table 3.2** Asthma data

Treatment	<i>n</i>	$\bar{x}$	<i>s</i>
A	18	54.6	14.6
B	21	48.8	12.9

95% and 99% CIs for  $\mu_1$  and  $\mu_2$ , the population mean increases in  $\text{FEV}_1$  in treatment groups A and B, are calculated as shown in Table 3.3.

**Table 3.3** Confidence intervals for asthma data

Treatment group	Multiplying constants (95%/99%)	$s/\sqrt{n}$	95% CI	99% CI
A	2.11/2.90	3.44	(47.3, 61.9)	(44.6, 64.6)
B	2.09/2.85	2.82	(42.9, 54.9)	(40.8, 56.8)

## 3.2 Confidence intervals for other parameters

### 3.2.1 Difference between two means

At the end of the previous chapter, we saw how to extend the idea of a standard error for a single mean to a standard error for the difference between two means. The extension of the CI is similarly straightforward. Consider the placebo-controlled trial in cholesterol lowering described in Example 2.3 in Chapter 2. We had an observed difference in the sample means  $\bar{x}_1 - \bar{x}_2$  of 1.4 mmol/l and a standard error of 0.29 mmol/l. The formula for the 95% CI for the difference between two means ( $\mu_1 - \mu_2$ ) is

$$\bar{x}_1 - \bar{x}_2 - (\text{constant} \times se) \text{ to } \bar{x}_1 - \bar{x}_2 + (\text{constant} \times se)$$

This expression is essentially the same as that for a single mean: statistic  $\pm$  (constant  $\times$  se). The rules for obtaining the multiplying constant, however, are slightly different. For the difference between two means, we use Table 3.1 as before, but now we go into that table at the row  $n_1 + n_2 - 2$ , where  $n_1$  and  $n_2$  are the sample sizes for treatment groups 1 and 2, respectively.

So, for the data in Example 2.3 ( $n_1 = 24$  and  $n_2 = 23$ ), the multiplying constant (from row 45) is 2.01 and the calculation of the 95% CI is as follows:

$$\text{Lower confidence limit} = 1.4 - 2.01 \times 0.29 = 0.8 \text{ mmol/l}$$

$$\text{Upper confidence limit} = 1.4 + 2.01 \times 0.29 = 2.0 \text{ mmol/l}$$

The interpretation of this interval is essentially as before; we can be 95% confident that the true difference in the (population) means,  $\mu_1 - \mu_2$ , is between 0.8 and 2.0. In other words, these data are telling us with 95% confidence that the mean reduction  $\mu_1$  in the active group is greater than the corresponding mean reduction  $\mu_2$  in the placebo group by between 0.8 mmol/l and 2.0 mmol/l.

### 3.2.2 Confidence interval for proportions

The previous sections in this chapter are applicable when we are dealing with means. As noted earlier, these parameters are relevant when we have continuous or score endpoints. With binary data, we will be looking to construct CIs for proportions plus differences between those proportions.

#### **Example 3.2** Trastuzumab in HER2-positive breast cancer

The data in Table 3.4 are taken from Piccart-Gebhart et al. (2005), who compared trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer with observation-only. The binary outcome here is one or more serious adverse events (SAEs) versus no SAEs during the one-year trial. The proportion in the observation-only group provides the background incidence of SAEs.

**Table 3.4** Trastuzumab data

	$\geq 1$ SAE	No SAEs	Total
Trastuzumab	117	1560	1677
Observation	81	1629	1710
Total	198	3189	3387

This display is termed a  $2 \times 2$  *contingency table*.

The incidence rates/proportions in the test treatment and control groups, respectively, are

$$r_1 = \frac{117}{1677} = 0.070 \quad r_2 = \frac{81}{1710} = 0.047$$

For Example 3.2, if we label the true SAE incidence proportions in the population as a whole as  $\theta_1$  (assuming all patients in the population received trastuzumab) and  $\theta_2$  (assuming all patients were only observed), then we will be interested in the CIs for the individual proportions  $\theta_1$  and  $\theta_2$  and also the difference in those proportions,  $\theta_1 - \theta_2$ .

In Section 2.5.2, we set down the formulas for the standard error for both individual proportions and the difference between two proportions. These lead naturally to expressions for the CI.

For example, for the trastuzumab group, the 95% CI for  $\theta_1$  is given by

$$0.070 \pm 1.96 \sqrt{\frac{0.070(1-0.070)}{1677}} = (0.058, 0.082)$$

The 95% CI for  $\theta_1 - \theta_2$ , the difference in the SAE rates, is given by

$$\begin{aligned} (\bar{r}_1 - \bar{r}_2) &\pm 1.96 \sqrt{\frac{\bar{r}_1(1-\bar{r}_1)}{n_1} + \frac{\bar{r}_2(1-\bar{r}_2)}{n_2}} \\ &= (0.070 - 0.047) \pm 1.96 \sqrt{\frac{0.070(1-0.070)}{1677} + \frac{0.047(1-0.047)}{1710}} \\ &= (0.007, 0.039) \end{aligned}$$

So, with 95% confidence, we can say that the absolute difference in SAE rates between trastuzumab and observation-only is between 0.7% and 3.9%.

Note that for binary endpoints and proportions, the multiplying constant is 1.96, the value used previously when we first introduced the CI idea. Again, this provides an approximation, but in this case, the approximation works well except in the case of very small sample sizes.

### 3.2.3 General case

In general, the calculation of the CI for any statistic, be it a single mean, the difference between two means, a median, a proportion, the difference between two proportions and so on, always has the same structure

$$\text{statistic} \pm (\text{constant} \times se)$$

where the *se* is the standard error for the statistic under consideration.

There are invariably rules for how to obtain the multiplying constant for a specific confidence coefficient, but as a good approximation (and provided the sample sizes are not too small), using the value 2 for the 95% CI and 2.6 for the 99% CI would get you very close.

This methodology applies whenever we are looking at statistics based on single-treatment groups or those relating to differences between treatment groups. When we are dealing with ratios, such as the odds ratio or the hazard

ratio, the methodology is changed slightly. We will cover these issues in a later chapter (see Chapter 4.5.5, for example, for CIs for the odds ratio).

### 3.2.4 Bootstrap confidence interval

The methods covered so far in Section 3.2 are applicable whenever we have a mathematical formula for the standard error. There are some complex situations, however, where a formula does not exist, and we need to use a technique known as *bootstrapping* to obtain first the *standard error* and then the CI.

Suppose that we have two treatment groups, A and B, including, respectively, 450 and 460 patients, and we are looking to obtain the standard error of the difference between the two medians. We do have a formula for calculating the standard error for the difference in the two medians, but for the purposes of this illustration, let's suppose we do not. Step 1 is to take a sample of patients at random (with replacement) of size 450 from group A. By *with replacement* here, we mean that once we have chosen the first patient to go into our sample at random, we choose the second patient at random from the complete group of 450 patients; in other words, we put that first patient back into the mix. So, in the sample of 450 patients, each patient can appear more than once and some patients will not appear at all. We repeat this for group B, giving us what we call a *bootstrap sample* that contains 450 group A patients and 460 group B patients. In step 2, we calculate the difference in the medians (median group A - median group B) from this bootstrap sample. In step 3, we repeat this sampling exercise many times, typically 1000 or so, giving 1000 bootstrap samples and 1000 values for the difference in the medians. The final step involves calculating the standard deviation of these 1000 median difference values using the method detailed in Section 2.2.3. This calculated standard deviation is the standard error for the difference in the medians, and our bootstrap CI is then simply the observed difference in the medians plus/minus the usual multiple of the standard error. For a 95% CI, this multiple will be 1.96, and for a 99% CI, this will be 2.576. This approach mimics the sampling process we went through in Section 2.4.

## 3.3 Hypothesis testing

In our clinical trials, we generally have some very simple questions:

- Does the drug work?
- Is treatment A better than treatment B?
- Is there a dose response?
- Are treatments A and B clinically equivalent?

To evaluate the truth or otherwise of these statements, we begin by formulating the questions of interest in terms of hypotheses. The simplest (and most common) situation is the comparison of two treatments: for example, in a

placebo-controlled trial, where we are trying to detect differences and demonstrate that the drug works.

Assume that we are dealing with a continuous endpoint, such as fall in diastolic blood pressure, and we are comparing means. If  $\mu_1$  and  $\mu_2$  denote the mean reductions in groups 1 and 2, respectively, then our basic question is as follows:

is  $\mu_1 = \mu_2$  or is  $\mu_1 \neq \mu_2$ ?

We formulate this question in terms of two competing hypotheses:

$H_0 : \mu_1 = \mu_2$ , termed the *null hypothesis*

and

$H_1 : \mu_1 \neq \mu_2$ , termed the *alternative hypothesis*

We base our conclusion regarding which of these two statements (hypotheses) we *prefer* on data, and the method that we use to make this choice is built around the *p*-value.

### 3.3.1 Interpreting the *p*-value

The '*p*' in *p*-value stands for probability, and as such, it lies between 0 and 1. I am sure we all know that if the *p*-value falls below 0.05, we declare statistical significance and conclude that the treatments are different: that is,  $\mu_1 \neq \mu_2$ . In contrast, if the *p*-value is above 0.05, then we talk in terms of non-significant differences. We will now explore just how this *p*-value is defined, and later, we will see the principles behind its calculation.

In the context of the comparison of an active treatment (A) with a placebo treatment (B) in lowering diastolic blood pressure, assume that we have the following summary statistics

$$\bar{x}_1 = 9.6 \text{ mmHg (active)}$$

$$\bar{x}_2 = 4.2 \text{ mmHg (placebo)}$$

with the difference  $\bar{x}_1 - \bar{x}_2 = 5.4 \text{ mmHg}$ .

Suppose that the *p*-value turns out to be 0.042. What does this *p*-value actually measure? We can see of course that it is  $< 0.05$ , so we have statistical significance, but what does the probability 0.042 refer to? What is it the probability of?

Usually, people give one of two responses to this question:

- Proposed definition 1: *There is a 4.2% probability that  $\mu_1 = \mu_2$* ,
- Proposed definition 2: *There is a 4.2% probability that the observed difference of 5.4 mmHg is due to chance*.

One of these definitions is correct, and one is incorrect. Which way round is it?

Well, the second definition is the correct one. The first definition is not only incorrect but also a common mistake that many people make. We will explore

in Section 9.10.1 why this definition causes so many problems and misunderstandings. For the moment, however, we will explore the correct definition in more detail. It is worthwhile expanding on the various components of the definition:

- There is a 4.2% probability that
- the observed difference *or a bigger difference in either direction (A better than B or B better than A)*
- is a chance finding *that has occurred with equal treatments ( $\mu_1 = \mu_2$ ), so when the null hypothesis is true.*

Commit this definition to memory: it is important!

To complete the logic, we consider this statement and argue as follows: there is only a 4.2% chance of seeing a difference as big as the one observed with equal treatments. This is a small probability, and it is telling us that these data are not at all likely to have occurred with equal treatments, and it is on this basis that we do not believe that the treatments are equal. We declare statistically significant differences between the treatment means.

In contrast, if the *p*-value had been, say, 0.65, then the definition says that there is a 65% probability of seeing a difference as big (or bigger) than the one observed, with equal treatments. Now, 65% is quite a high probability, and what we are seeing in this case is a difference that is entirely consistent with  $\mu_1 = \mu_2$ ; it is the kind of difference we would expect to see with equal treatments, and therefore, we have no reason to doubt the equality of the population means.

Another way of thinking about the *p*-value is as a measure of how consistent the difference is with equal treatments (or, equivalently, with the null hypothesis). A low *p*-value says that the difference is not consistent with equal treatments, while a high value says that the difference is consistent with equal treatments. The conventional cut-off between *low* and *high* is 0.05.

Many people ask at this stage: why 0.05? Well, in one sense it is an arbitrary choice – the cut-off could easily have been 0.04 or 0.075 – but 0.05 has become the agreed value, the convention. We will explore the implications of this choice later when we look at type I and type II errors.

The way that the hypotheses are set up is that we always structure  $H_1$  to be our *objective*.  $H_1$  represents the desirable outcome; we want to come out of the clinical trial concluding in favour of  $H_1$ . The *p*-value measures how consistent the data are with  $H_0$ ; if the *p*-value is small, the data are not consistent with  $H_0$ , and we declare statistical significance and decide in favour of  $H_1$ . In this way, we are essentially trying to disprove  $H_0$ . This is the *scientific method* with its roots in philosophical reasoning; the way science advances is by disproving things rather than by proving them. For example, proving that *all swans are white* is very difficult, but you only have to see one black swan to disprove that statement.

### 3.3.2 Calculating the *p*-value

We will start with a very simple situation to see how we calculate a *p*-value. Suppose we want to know whether a coin is a fair coin; by that, we mean that when we flip the coin, it has an equal chance of coming down heads (H) or tails (T).

Let  $\text{pr}(H)$  denote the probability of the coin coming down heads. We can then formulate null and alternative hypotheses as follows:

$$H_0 : \text{pr}(H) = \frac{1}{2} (\text{fair coin}) \quad H_1 : \text{pr}(H) \neq \frac{1}{2} (\text{unfair coin})$$

We now need some data on which to evaluate the hypotheses. Suppose we flip the coin 20 times and end up with 15 heads and 5 tails. Without thinking too much about probabilities and *p*-values, what would your intuition lead you to conclude? Would you say that the data provide evidence that the coin is not fair, or are the data consistent with the coin being fair?

We will now be a little more structured about this. Because this is such a simple situation, we can write down everything that could have happened in this experiment and fairly easily calculate the probabilities associated with each of those outcomes *under the assumption that the coin is fair*. The possible outcomes and probabilities are contained in Table 3.5.

**Table 3.5** Outcomes and probabilities for 20 flips of a fair coin (see the following discussion for the method of calculating the probabilities)

Heads (H)	Tails (T)	H – T	Probability (coin fair)
20	0	20	0.00000095
19	1	18	0.000019
18	2	16	0.00018
17	3	14	0.0011
16	4	12	0.0046
15	5	10	0.015
14	6	8	0.037
13	7	6	0.074
12	8	4	0.120
11	9	2	0.160
10	10	0	0.176
9	11	-2	0.160
8	12	-4	0.120
7	13	-6	0.074
6	14	-8	0.037
5	15	-10	0.015
4	16	-12	0.0046
3	17	-14	0.0011
2	18	-16	0.00018
1	19	-18	0.000019
0	20	-20	0.00000095

Note that we have included a column H – T; this is the number of heads minus the number of tails. This is done in order to link with what we do when we are comparing treatments where we use *differences* to measure treatment effects.

So, with a fair coin, getting 12 heads and 8 tails, for example, will happen on 0.120 (12%) of occasions. The most likely outcome with a fair coin, not surprisingly, is 10 heads and 10 tails, and this will happen 17.6% of the time. The extreme outcomes are not at all likely, but even 20 heads and 0 tails can still occur, and we will see this outcome 0.000095% of the time!

Our data were 15 heads and 5 tails, so how do we calculate the *p*-value? Well, remember the earlier definition and translate that into the current setting: *the probability of getting the observed difference or a bigger difference in either direction with a fair coin*. To get the *p*-value, we add up the probabilities (calculated when the null hypothesis is true; coin fair) associated with our *difference* (15 heads and 5 tails gives a difference of H – T = 10) or a bigger difference in either direction. This is given by

$$\begin{aligned} &= (0.00000095 + 0.000019 + 0.00018 + 0.0011 + 0.0046 + 0.015) \times 2 \\ &= 0.0417998 \text{ or } 0.0418 = p \end{aligned}$$

This means there is only a 4.18% probability of seeing the 15/5 split or a more extreme split (either way) with a fair coin. This probability is below the magical 5%, we have a statistically significant result, and the evidence suggests that the coin is not fair.

Had we seen a 14/6 split, however, the *p*-value would have increased to  $0.0417998 + 2 \times 0.037 = 0.1157998$ , a non-significant result; the 14/6 split is not sufficiently extreme for us to be able to reject the null hypothesis (according to the conventional cut-off at 5%). The 15/5 split ( $H - T = 10$  or  $-10$ ) therefore is the smallest split that just achieves statistical significance.

Now we will look at a simple example and show how to calculate the probabilities for a fair coin. Suppose we flip the coin just three times. The possible combinations are as follows:

	Probability
H H H	0.125
H H T	0.125
H T H	0.125
T H H	0.125
T T H	0.125
T H T	0.125
H T T	0.125
T T T	0.125

If the coin is fair, these are all equally likely: each of these outcomes has probability  $1/2 \times 1/2 \times 1/2 = (1/2)^3 = 0.125$  of occurring.

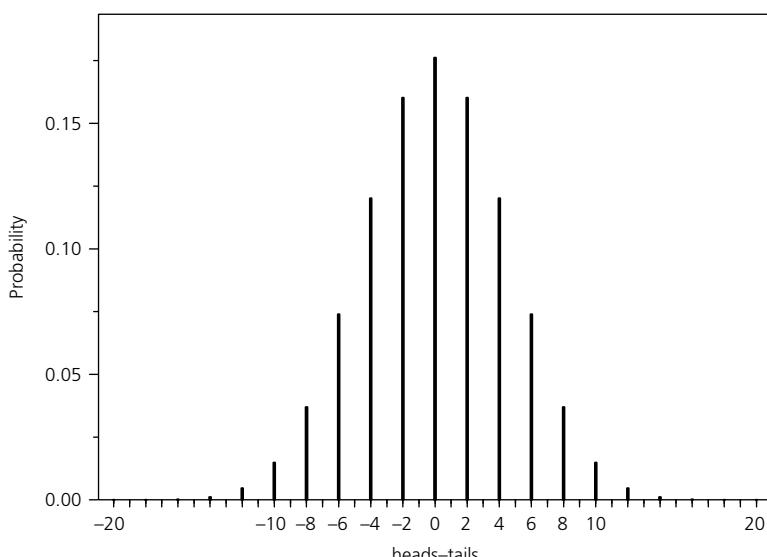
In terms of numbers of heads (H) and numbers of tails (T), there are just four possibilities, and we simply add up the probabilities corresponding to the individual combinations.

Heads (H)	Tails (T)	Probability (coin fair)
3	0	$0.125 = 1 \times (1/2)^3$
2	1	$0.375 = 3 \times (1/2)^3$
1	2	$0.375 = 3 \times (1/2)^3$
0	3	$0.125 = 1 \times (1/2)^3$

For 20 flips, we get the probabilities by multiplying  $(1/2)^{20}$  by the number of combinations that give rise to that particular outcome. For example, with 12 heads and 8 tails, this is  $20! \div (12! \times 8!)$  where  $n!$  denotes  $n \times n - 1 \times \dots \times 2 \times 1$ .

It is useful to look at this visually. Figure 3.2 plots each of the outcomes on the  $x$ -axis with the corresponding probabilities, calculated when the null hypothesis (*under the null hypothesis*) is true, on the  $y$ -axis. Note that the  $x$ -axis has been labelled according to *heads – tails* (H – T), the number of heads minus the number of tails. This identifies each outcome uniquely and allows us to express each data value as a difference. More generally, we will label this the *test statistic*; it is the statistic on which the  $p$ -value calculation is based. The graph and the associated table of probabilities are labelled the *null distribution (of the test statistic)*.

Using the plot, we calculate the  $p$ -value firstly by identifying the outcome we saw in our data and secondly by adding up all those probabilities associated with that outcome and more extreme outcomes (*bigger* differences in terms of the test statistic) in both directions (positive and negative). Returning to the H – T



**Figure 3.2** Null distribution for 20 flips of a fair coin

difference of 10 or  $-10$  (the 15/5 split), this outcome is at the boundary between  $p < 0.05$  and  $p > 0.05$  and just achieves statistical significance. We call this value for the test statistic (10) the *critical value*.

### 3.3.3 A common process

In the previous section, we calculated the  $p$ -value in a very simple situation. Nonetheless, in more complex situations, the steps required to calculate  $p$  are basically the same. Just briefly returning to the coin, all possible outcomes were expressed in terms of the difference  $H - T$ . This quantity is what we referred to in Section 1.8.1 as the signal. Large differences give strong signals – strong evidence that the coin is not fair – and small differences, in contrast, give weak signals.

We will develop the methodology for calculating  $p$ -values in relation to the comparison of two independent means for a parallel-group design. The resulting statistical test is known as the *unpaired* or *two-sample t-test*. The evidence for differences will depend on the strength of the signal together with the extent of the noise and the sample size.

The null and alternative hypotheses are specified as follows:

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

The signal is measured by the observed difference  $\bar{x}_1 - \bar{x}_2$ , and the noise is captured by the standard deviations  $s_1$  and  $s_2$  observed in treatment groups 1 and 2, respectively; and finally, we also need to factor in the sample sizes  $n_1$  and  $n_2$ .

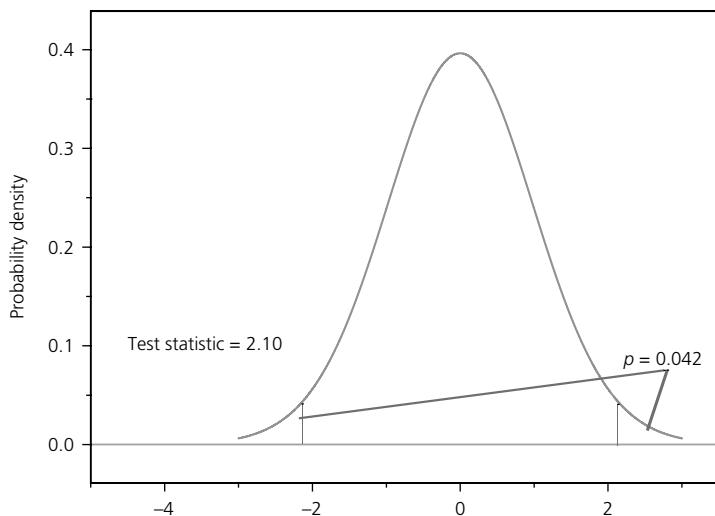
Consider the earlier example comparing two treatments for the reduction of blood pressure. The sample sizes in the two treatment groups were  $n_1 = 20$  and  $n_2 = 20$ , and the sample means were

$$\bar{x}_1 = 9.6 \text{ mmHg (active)}$$

$$\bar{x}_2 = 4.2 \text{ mmHg (placebo)}$$

The observed difference  $\bar{x}_1 - \bar{x}_2 = 5.4$  mmHg is the signal, and to calculate the  $p$ -value, we need to calculate the probability of seeing a difference between the treatment means at least as large as 5.4 (in either direction) when the true treatment means are the same. There is a mathematical trick that allows us to do this in a fairly straightforward way by considering the signal divided by the standard error. The reason for doing this is that when the null hypothesis is true ( $\mu_1 = \mu_2$ ), the ratio  $\left(\frac{\bar{x}_1 - \bar{x}_2}{se}\right)$  has a predictable behaviour, and we can calculate probabilities associated with its values. Remember, *randomness is entirely predictable*: when the null hypothesis is true, any differences seen between the sample means are just due to chance (random variation), and we can calculate the required probabilities under those circumstances.

Applying the formula, the standard error calculation (see Section 2.5.1) gives  $se = 2.57$ , and the numerical value of the difference between the means



**Figure 3.3** The t-distribution on 38 degrees of freedom

divided by the standard error ( $= 5.4/2.57$ ) is then 2.10. In this example, this is what we will call the *test statistic*, and it is going to be the statistic on which the test (*p*-value calculation) is based. This ratio is also known as the *z-score*, and we will use these two terms interchangeably at various points in our development.

Seeing a signal  $> 5.4$  mmHg is equivalent to seeing a signal/se ratio  $> 2.10$ , assuming that the noise and sample sizes are fixed, and it is this ratio that will lead to the *p*-value.

The probabilities associated with values of this signal/se ratio, under the assumption that the population means are the same, are given by a particular distribution: the t-distribution. Figure 3.3 displays these probabilities for the example we are considering. Note that we have labelled this the t-distribution on 38 degrees of freedom (df); we will say more about where the 38 comes from in the next chapter.

Probability density on the y-axis is the probability per unit on the x-axis. This makes the total area under the curve equal to 1 or 100%.

Using computer programs, we can add up all the probabilities associated with the observed value 2.10 of the test statistic and more extreme values in both directions (A better than B and B better than A), as seen in the identified areas under the t-distribution, to give the *p*-value, 0.042.

This calculation of the *p*-value comes from the probabilities associated with these signal/se ratios (z-scores), and this forms a common theme across many statistical test procedures. To reiterate, the signal/se ratio is also referred to as the test statistic. The distribution of the test statistic when the null hypothesis is true (equal treatments) is termed the *null distribution*.

In general, we can think of the  $p$ -value calculation as a series of five steps as follows:

- 1 Formulate null and alternative hypotheses. In all cases, the alternative hypothesis represents the *desirable* outcome. In a superiority trial, this means the null hypothesis is equality (or no effect/no change/no dependence), while the alternative hypothesis is inequality (there is an effect/a change/a dependence).
- 2 Calculate the value of the test statistic (usually = signal/se = z-score). The formula for the test statistic will be based on a standard approach determined by the endpoint type, the design of the trial (between or within patient) and the hypotheses of interest. Mathematics has provided us with optimum procedures for all the common (and not so common) situations, and we will see numerous examples in subsequent chapters.
- 3 Determine the null distribution of the chosen test statistic: that is, what are the probabilities associated with all the potential values of the test statistic when the null hypothesis is true? Again, mathematical theory (*randomness is entirely predictable*) has provided us with solutions to this, and all of these null distributions are known; we simply have to look them up.
- 4 Obtain the  $p$ -value by adding up the probabilities associated with the calculated value of the test statistic and more extreme values when the null hypothesis is true. This will correspond to adding up the probabilities associated with the observed value or more extreme values of the signal (treatment difference).
- 5 Draw conclusions. If  $p \leq 0.05$ , then declare statistical significance; if  $p > 0.05$ , then the differences are not statistically significant.

There is another way to think about the test statistic in terms of *evidence* for true differences. If the signal is strong (large differences) and the standard error is small, indicating that we have a reliable estimate of the true difference, then the ratio of the signal and the standard error (the test statistic) will be numerically large, either in a positive or in a negative direction depending on the direction of the treatment effect. When this happens, we will have a value of the test statistic that is well away from zero and in what we call the *tails* of the null distribution. Figure 3.3 indicates a situation where that has happened, and this is precisely the setting where we end up with small, statistically significant  $p$ -values. The ratio of the signal to the standard error or z-score can therefore be thought of as a measure of evidence for true differences, where large values allow us to conclude in favour of such differences.

Finally, what we are talking about here links back to our discussion on the signal-to-noise ratio, sample size and evidence in Chapter 1. We said then that a large value of the signal-to-noise ratio points towards a treatment difference, whereas a small value of the signal-to-noise ratio does not. We also added that a large sample size tells us this ratio is reliable, while a small sample size tells us this the ratio is unreliable. The sample size combines with the signal-to-noise ratio to produce the signal/se ratio to give us our measure of evidence. In

particular, for the difference between two means with an equal number ( $n$ ) of patients per group and an assumed equal standard deviation in each of the treatment groups, it is easy to show that the signal/se ratio is equal to the square root of  $n/2$  multiplied by the signal-to-noise ratio, as mentioned at the end of Section 1.8.3.

### 3.3.4 The language of statistical significance

There is a fair amount of language that we wrap around this process. We talk in terms of a *test of significance*. If  $p \leq 0.05$ , we declare statistical significance and *reject the null hypothesis at the 5% level*. We call 5% the *significance level* – it is the level at which we declare statistical significance. If  $p > 0.05$ , then we say that we have a non-significant difference and we are *unable to reject the null hypothesis at the 5% level*.

Our conventional cut-off for statistical significance is 5%, but we also use other levels, notably 1% and 0.1%. If  $p \leq 0.01$ , then the evidence is even stronger that there are differences, and we have highly significant differences. If  $p \leq 0.001$ , then the evidence is even stronger still and we have very highly significant differences.

There is often quite a bit of discussion when we see  $0.05 < p \leq 0.10$ : *almost significant, a trend towards significance, approaching significance* and other imaginative phrases! I have some sympathy with such comments. One thing we do have to remember is that the *p*-value scale is a continuum, and to have a strict cut-off at 0.05 is in a sense unrealistic. There really is little difference, from a strength of evidence point of view, between  $p = 0.048$  and  $p = 0.053$ , yet one gives statistical significance, and one does not. Unfortunately, many practitioners (including regulators) have a strict demarcation at 0.05. In one sense, this is understandable; having a strict cut-off at 0.05 removes any ambiguity.

### 3.3.5 One-sided and two-sided tests

The *p*-value calculation detailed in the previous section gives what we call a *two-sided* or a *two-tailed test* since we calculate  $p$  by taking into account values of the test statistic equal to, or more extreme, than that observed, in both directions. So, for example, with the coin, we look for movement away from *coin fair* both in terms of *heads more likely than tails* and *tails more likely than heads*.

In part, this is because of the way we set up the hypotheses; in our earlier discussion, we asked, ‘is the coin fair?’ or ‘is the coin not fair?’ We could have asked a different question – ‘are heads equally/less likely than tails?’ or ‘are heads more likely than tails?’ – in which case we could have been justified in calculating the *p*-value only in the direction corresponding to *heads more likely than tails*. This would have given us a *one-sided* (or a *one-tailed*) *p*-value. Under these circumstances, had we seen 17 tails and 3 heads, this would not have led to a significant *p*-value, and we would have discounted that outcome as of no interest; it is not in the direction we are looking for.

Clearly, one-sided  $p$ -values are of interest to sponsors: firstly, they are smaller and more likely to give a positive result; and, secondly, many sponsors would argue that they are only interested in departures from the null hypothesis in one particular direction – the one that favours their drug. While this may be an argument a sponsor might use, regulators (and the scientific community more generally) may not think this way. Regulators are interested in differences both ways and insist that generally,  $p$ -values are two-sided. It must be said, though, that in many cases, regulators are also comfortable with one-sided  $p$ -values; but when considering one-sided  $p$ -values, they also state that the significance level that should be used is 0.025 rather than 0.05. Now, because most situations are symmetric, two-sided  $p$  is usually equal to  $2 \times$  one-sided  $p$ , so it actually makes no difference operationally whether we use one-sided or two-sided  $p$ -values in terms of detecting a positive outcome for the experimental treatment!

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'It is important to clarify whether one- or two-sided tests of statistical significance will be used, and in particular to justify prospectively the use of one-sided tests . . . The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings'.*

We will discuss type I (and type II) errors in detail in Section 8.1, but for the moment, for *type I error* read *significance level*.

## CHAPTER 4

# Tests for simple treatment comparisons

### 4.1 The unpaired t-test

In Section 3.3.3, we introduced the general structure for a significance test with the comparison of two means in a parallel-group trial. This resulted in a procedure that goes under the general heading of the *two-sample* (or *unpaired*) *t-test*. This test was developed for continuous endpoints, although it is applicable more widely and is frequently used for score endpoints.

The test was developed almost 100 years ago by William Sealy Gosset. Gosset was a chemist by training and was employed by the Guinness brewery, initially in Dublin, Ireland, but subsequently in London. He became interested in statistics and in the application of statistics to the improvement of quality within the brewing process. Gosset's work was based on a combination of mathematics and empirical experience (trial and error), but the procedures he came up with have certainly stood the test of time; the unpaired t-test is undoubtedly the most commonly used (although not always appropriately) statistical test of them all.

The calculation of the *p*-value in the example in Section 3.3.3 consisted of adding up the probabilities, associated with values of the signal/se ratio greater than the observed value of 2.10, given by the t-distribution. It turns out that these probabilities depend upon the number of patients included in the trial. There are an infinite number of t-distributions ( $t_1, t_2, t_3, \dots$ ), and the one we choose is indexed based on calculating the total sample size (both groups combined) and subtracting two. We will see that this t-distribution is used in other settings where the rule for choosing the index is different, but the rule for the unpaired t-test is  $\text{index} = n_1 + n_2 - 2$ . This quantity, as noted in Section 3.3.3, is called the *degrees of freedom* (df).

There is a connection between what we are seeing here and the calculation of the confidence interval (CI) in Chapter 3. Recall Table 3.1 in Section 3.1.3, *Changing the multiplying constant*. It turns out that *p*-values and CIs are linked, and we will explore this further in a later chapter. The multiplying constants for  $\text{df} = 38$  are 2.02 for 95% confidence and 2.71 for 99% confidence. If we were to look at the  $t_{38}$  distribution, we would see that  $\pm 2.02$  cuts off the outer 5% probability, while  $\pm 2.71$  cuts off the outer 1% probability.

Having calculated the  $p$ -value, we can also calculate the 95% CI for the difference  $\mu_1 - \mu_2$  to give us information about the magnitude of the treatment effect. For the data in the example in Section 3.3.3, this CI is given by

$$(5.4 \pm 2.02 \times 2.57) = (0.2, 10.6)$$

So, with 95% confidence, we can say that the true treatment difference/effect ( $\mu_1 - \mu_2$ ) is somewhere in the interval 0.2 mmHg to 10.6 mmHg.

## 4.2 The paired t-test

The *paired t-test*, also known as the *one-sample t-test*, was also developed by Gosset. This test is primarily used for the analysis of endpoints arising from within-patient designs, although we also see it applied when comparing a baseline value with a final value within the same treatment group.

Consider a two-period, two-treatment crossover trial in asthma comparing an active treatment (A) and a placebo treatment (B) in which the peak expiratory flow (PEF) (l/min) data in Table 4.1 were obtained, in terms of the value at the end of each period.

Patients 1–16 received treatment A followed by treatment B, while patients 17–32 received treatment B first followed by treatment A.

The final column has calculated the A – B differences, and as we shall see, the paired t-test works entirely on the column of differences. Again, we will follow through several steps for the calculation of the  $p$ -value for the A versus B comparison:

**1** Let  $\mu$  be the population mean value for the column of differences. The null and alternative hypotheses are expressed in terms of this quantity:

$$H_0 : \mu = 0 \quad H_1 : \mu \neq 0$$

A non-zero value for  $\mu$  reflects treatment differences; a positive value tells us that the active treatment is effective.

**Table 4.1** Data from a crossover trial in asthma (hypothetical)

Patient	A	B	Difference (A – B)
1	395	362	33
2	404	385	19
3	382	386	-4
.			
.			
.			
32	398	344	54

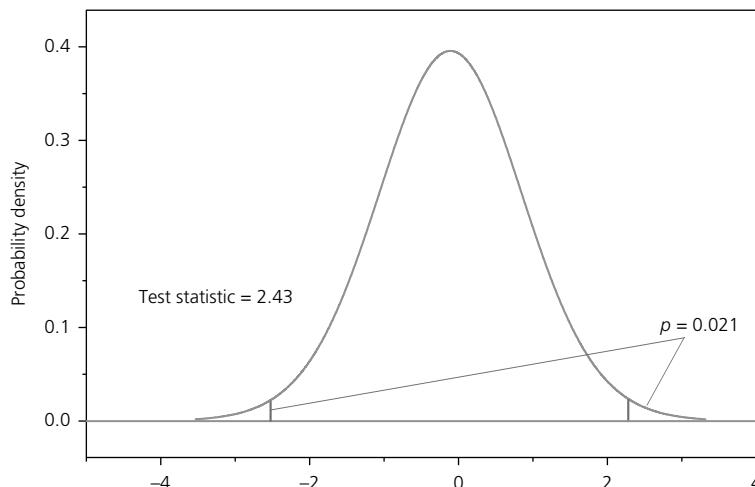
- 2 Again, the test will be based on the signal/se ratio. In this case, the signal is the observed mean  $\bar{x}$  of the column of differences, and se is the standard error associated with that mean. For these data,

$$\bar{x} = 28.41 \text{ l/min} \quad se(\text{of } \bar{x}) = 11.71 \text{ l/min}$$

The se here is obtained from the standard deviation of the differences divided by the square root of 32, the number of differences in the final column.

The test statistic = signal/se =  $28.4/11.7 = 2.43$  captures the evidence for treatment differences. Larger values, either positive or negative, are an indication of treatment differences.

- 3 The probabilities associated with the values that this signal/se ratio can take when the treatments are the same ( $\mu = 0$ ) are again given by the t shape; in this case, the appropriate t-distribution is  $t_{31}$ , the t-distribution on 31 df. Why  $t_{31}$ ? The appropriate t-distribution is indexed by the number of patients or number of differences – 1.
- 4 Our computer programs now calculate the *p*-value, the probability associated with getting a value for the signal/se ratio at least as large as 2.43, in either direction, when the null hypothesis is true. This value turns out to be 0.021 (see Figure 4.1). This value also reflects, assuming that the se is fixed, the probability of seeing a mean difference at least as big as 28.4 l/min by chance (with equal treatments). This signal is sufficiently strong for us to conclude in favour of a real treatment effect.
- 5 The *p*-value is  $< 0.05$ , giving statistical significance at the 5% level, and we conclude, based on the evidence, that treatment A (active) is more efficacious than treatment B (placebo); the active treatment works.



**Figure 4.1** The t-distribution on 31 df

This test is based entirely on the column of differences; once the column of differences is calculated, the original data are no longer used. An alternative approach might have been to simply calculate the mean value on A and the mean value on B and compare the two using the unpaired t-test. Would this have worked? In fact, no: using the unpaired t-test in this way would have been incorrect; the unpaired t-test is used to compare means across two independent samples. The paired t-test uses the data in the most efficient way; forming the column of differences links the observations on the same patient and effectively uses each patient as their own control. Calculating the A and B means separately and taking their difference would have given the same signal as the mean of the column of differences, but the se would have been different. It would not reflect the patient-to-patient variability in the A – B differences but would be based on the patient-to-patient (within-group) variability in the value of the endpoint itself.

Does it matter which way round the differences are calculated, A – B or B – A? No: as long as we are consistent across all of the patients, it does not matter and will simply have the effect of changing the sign of the test statistic – the two-sided *p*-value will remain unchanged.

As with the unpaired t-test, it would be useful to calculate a CI for the treatment effect,  $\mu$ . This is given by

$$\begin{aligned} (\bar{x} \pm 2.04 \times se) &= (28.4 \pm 2.04 \times 11.7) \\ &= (4.5, 52.3) \end{aligned}$$

Here, 2.04 is the appropriate multiplication constant for 95% confidence with 32 patients. So, we can be 95% confident that the treatment difference (active treatment effect),  $\mu$ , is between 4.5 l/min and 52.3 l/min.

Before we move on, it should be pointed out that the paired t-test provides a valid analysis for continuous endpoints arising from a crossover trial only when the trial is balanced: that is, when the number of patients following the A/B sequence is the same as the number of patients following the B/A sequence. When this is not the case, the analysis needs to be modified slightly. This is because in many settings, there will also be a *period effect*: irrespective of treatment, patients may respond differently in period I compared to period II. This could be caused, for example, by the temporal nature of the underlying disease, by external conditions that may be changing over time, or through a learning effect. If a period effect is present, but the trial is balanced, then there will be equal numbers of A and B patients giving responses in period I and in period II, and under fairly general conditions, the period effect will cancel out in the paired t-test comparison of the treatments. When balance is not present, this effect will not cancel out, and there will be bias. See Senn (2002), Section 3.6, for details of how to deal with this. We have also assumed in this development of the crossover trial that there are no carry-over effects. Carry-over effects are present when the treatment received in period 1 influences the outcome in period 2. For example, if one of the treatments being evaluated (say, A) has an effect on the underlying disease

condition, this effect will carry over into period 2 and influence the outcome of treatment B. The outcome of B will then not be a pure treatment B effect but in the analysis will be considered as such. See Senn (2002) for further discussion on this point.

### 4.3 Interpreting the t-tests

The following example illustrates several issues and problems associated with the interpretation of  $p$ -values arising out of the t-tests. The setting is a very common one where a variable is measured at baseline and then subsequently at the end of the treatment period and the analysis focuses on the relative effects of the two treatments.

**Example 4.1** Comparison of two active treatments for the treatment of major depressive disorder in a randomised control trial

The primary endpoint in this trial was the 17-point Hamilton Depression Scale (HAMD-17), and the data presented in Table 4.2 correspond to the mean ( $se$ ).

**Table 4.2** Data on HAMD-17 (hypothetical)

	Baseline	Final (week 8)	Change from baseline
Active A ( $n = 36$ )	27.4 (1.18)	15.3 (0.92)	12.1 (0.95)
Active B ( $n = 35$ )	26.8 (1.22)	11.8 (1.32)	15.0 (1.17)

There are a number of comparisons that we can undertake:

- 1 Unpaired t-test comparing treatment means at baseline,  $p = 0.73$
- 2 Unpaired t-test comparing treatment means at week 8,  $p = 0.030$
- 3 Unpaired t-test comparing the mean change from baseline in the active A group with the mean change from baseline in the active B group,  $p = 0.055$
- 4 Paired t-test of baseline with week 8 in the active A group,  $p \ll 0.001$  (this means the  $p$ -value is very much less than 0.001)
- 5 Paired t-test of baseline with week 8 in the active B group,  $p \ll 0.001$

Let us consider each of these tests in turn and their interpretation:

- 1 Test 1 tells us that the treatment means are comparable at baseline, which is what we would expect to see given that this is a randomised trial. Of course, chance differences can sometimes occur. Indeed, in a randomised trial, we would expect to see  $p \leq 0.05$  for such a baseline comparison 5% of the time. See Section 6.9 for a further discussion on this point.
- 2 This test compares the treatment groups at week 8. The  $p$ -value suggests a treatment difference, but does this test necessarily provide an analysis of the data that uses all the information? Is there a possibility that we are being misled? Note that even though the earlier comparison of baseline means gave a non-significant  $p$ -value, the mean in active A group at baseline is slightly higher than the mean in the B group. Could this have contributed to the observed difference at week 8?

**3** Test 3 for the comparison of the mean change from baseline between the groups is non-significant. It would appear that the difference seen in test 2 is, in part, caused by the differences already seen at baseline. Looking at change from baseline has accounted for the minor baseline imbalances and, in general, is the basis for a more appropriate and sensitive analysis than simply looking at the week 8 means. It would be inappropriate to place too much emphasis on the fact that in test 3, the *p*-value is technically non-significant; recall that 0.05 is a somewhat arbitrary cut-off with regard to what we define as *statistical significance*. It is test 3, however, that provides the most appropriate information for evaluating the relative effects of the two treatments.

In Chapter 5, on adjusting the analysis, we will say quite a bit about additional improvements to this kind of analysis that increase sensitivity further and also avoid the problem of regression towards the mean. For the moment, though, test 3 is the best way to compare the treatments.

**4** Test 4 has given a very impressive *p*-value, but what is the correct interpretation of this test? The fall in HAMD score from 27.4 to 15.3 surely indicates that active A is an effective treatment! Well, in point of fact, it does not. The fall seen in this group could indeed have been caused by the medication, but equally, it could have been caused, for example, by the ancillary counselling that all patients receive or as a result of the placebo effect (the psychological impact of being in the trial and receiving *treatment*), and we have no way of knowing which of these factors is causing the effect and in what combination. The only way of identifying whether active drug A is efficacious is to have a parallel placebo group and undertake test 3; this would isolate the effect due to the specific medication from the other factors that could be causing the fall.

**5** Test 5 should be interpreted in exactly the same way as test 4. The fall is impressive, but is it due to the active medication? We don't know, and in the absence of a placebo group, we will never know.

Suppose that test 4 had given  $p = 0.07$  and test 5 had given  $p = 0.02$ . Would that therefore mean that active B is a better treatment than active A? No: to evaluate the relative effect of two treatments, we have to compare them! Directly, that is – not indirectly through the test 4 and test 5 comparisons back to baseline.

## 4.4 The chi-square test for binary endpoints

### 4.4.1 Pearson chi-square

The previous sections have dealt with the t-tests, methods applicable to continuous endpoints. We will now consider tests for binary endpoints, where the outcome at the subject level is a simple dichotomy: success/failure. In a between-patient, parallel-group trial, our goal here is to compare two proportions.

In Section 3.2.2, we presented data from a clinical trial comparing trastuzumab to observation only after adjuvant chemotherapy in HER2-positive breast cancer. Incidence proportions in the test treatment and control groups were, respectively, 7.0% and 4.7%.

The proportion of patients suffering serious adverse events (SAEs) in the trastuzumab group (7.0%) is clearly greater than the corresponding proportion

in the observation-only group (4.7%), but is this difference (signal) strong enough for us to conclude that there are real differences, or could this be a difference that is compatible with chance?

The chi-square test for comparing two proportions was developed by Karl Pearson around 1900 and predates the development of the t-tests. The steps involved in the *Pearson chi-square test* can be set down as follows:

- 1 The null and alternative hypotheses relate to the two true proportions,  $\theta_1$  and  $\theta_2$ :

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$

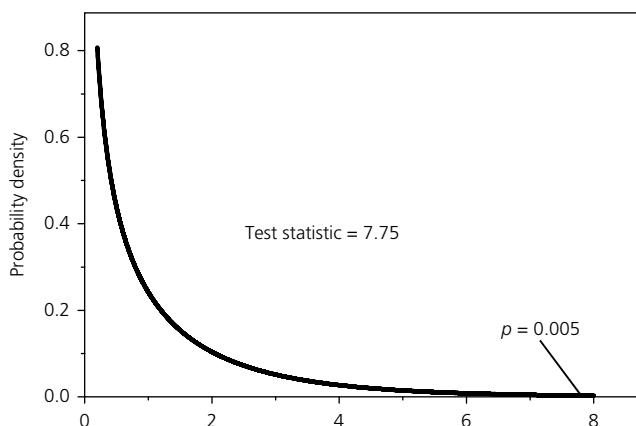
- 2 In forming a test statistic, Pearson argued in the following way. In the data, we see a total of 198 patients suffering SAEs. Had there been no differences between the groups in terms of the incidence of SAEs, then we would have seen equal proportions of patients suffering SAEs in the two groups. This would have meant seeing  $198 \times (1677/3387) = 98$  patients with SAEs in the trastuzumab group and  $198 \times (1710/3387) = 100$  in the observation-only group. Similarly, we should have seen 1579 patients not suffering SAEs in the trastuzumab group and 1610 such patients in the observation-only group. We term these values the *expected frequencies* and denote them by  $E$ ; the *observed frequencies* are denoted by  $O$ . These observed and expected frequencies are set down in Table 4.3, where the entries in the  $2 \times 2$  contingency table are  $O(E)$ .

We now need a measure of how far we are from *equal treatments*. Clearly, if the  $E$ s (what we should have seen with equal treatments) are close to the  $O$ s (what we have actually seen), then we have little evidence of a real difference in the incidence proportions between the groups. However, the further the  $O$ s are from the  $E$ s, the more we believe the true SAE proportions are different. The test statistic is formed by looking at each of the four cells of the table and firstly calculating  $(O - E)$ . Some of these values will be positive and some negative: for example, +19 in the trastuzumab  $\geq 1$  SAE cell and -19 in the observation-only  $\geq 1$  SAE cell. We then square these values (this gets rid of the sign), divide by the corresponding  $E$  (we will say more on this later) and add up the resulting quantities across the four cells, as follows:

$$\begin{aligned} & \sum \frac{(O-E)^2}{E} \\ &= \frac{(117-98)^2}{98} + \frac{(1560-1579)^2}{1579} + \frac{(81-100)^2}{100} + \frac{(1629-1610)^2}{1610} \\ &= 7.75 \end{aligned}$$

**Table 4.3** Observed and expected  $O(E)$  frequencies for the trastuzumab data

	$\geq 1$ SAE	No SAEs	Total
Trastuzumab	$O_1(E_1) = 117(98)$	$O_2(E_2) = 1560(1579)$	$n_1 = 1677$
Observation	$O_3(E_3) = 81(100)$	$O_4(E_4) = 1629(1610)$	$n_2 = 1710$
Total	198	3189	3387



**Figure 4.2** Chi-square distribution on one df

This value captures the evidence in support of treatment differences. If what we have seen ( $O_s$ ) is close to what we should have seen with *equal* treatments ( $E_s$ ), this statistic will have a value close to zero. However, if the  $O_s$  and the  $E_s$  are well separated, this statistic will have a large positive value; the more the  $O_s$  and the  $E_s$  disagree, the larger it will be. For the moment, this test statistic is not in the form of a signal/se ratio, but we will see later that it can also be formulated in that way.

- 3 Pearson calculated the probabilities associated with values of this test statistic when the treatments are the same, to produce the null distribution. This distribution is called the *chi-square distribution on one df*, denoted  $\chi_1^2$ , and is displayed in Figure 4.2. Note that values close to zero have the highest probability. Values close to zero for the test statistic will only result when the  $O_s$  and the  $E_s$  agree closely, whereas large values for the test statistic are unlikely when the treatments are truly the same.
- 4 To obtain the  $p$ -value, we now need to add up the probabilities associated with values of the test statistic at least as large as the value observed, 7.75 in our data. As can be seen from Figure 4.2, this value is well out to the right of the distribution and gives  $p = 0.005$ .
- 5 The  $p$ -value is  $\leq 0.01$ , and so we have a highly significant result: a highly significant difference between the treatment groups in terms of SAE proportions. Trastuzumab is associated with a significant increase in the proportion of patients suffering SAEs compared to observation only.

Several aspects of this  $p$ -value calculation deserve mention:

- The calculation of the test statistic involves division by  $E$ . This essentially weights the evidence from the different cells so that a cell with a smaller expected frequency gets more weight and a cell with a larger expected frequency is down-weighted. This makes sense since an  $O - E$  difference of 19 in 100 is a much more relevant difference than a difference of 19 in 1600, and the test statistic is more influenced by the former than the latter.

- The  $(O - E)^2$  values are all equal to 361, so algebraically, the test statistic could be written as

$$(O - E)^2 \left[ \frac{1}{E_1} + \frac{1}{E_2} + \frac{1}{E_3} + \frac{1}{E_4} \right] = 361 \times \left[ \frac{1}{98} + \frac{1}{1579} + \frac{1}{100} + \frac{1}{1610} \right]$$

- The null distribution for the t-test depended on the number of subjects in the trial. For the chi-square test comparing two proportions, providing the sample size is reasonably large, this is not the case; the null distribution is always  $\chi^2$ . As a consequence, we become very familiar with  $\chi^2$ . The critical value for 5% significance is 3.841, while 6.635 cuts off the outer 1% probability and 10.83 the outer 0.1%.

#### 4.4.2 The link to a ratio of the signal to the standard error

The formulation of the chi-square test procedure in the previous section, using observed and expected frequencies, is the standard way in which this test is developed and presented in most textbooks. It can be shown, however, that this procedure is akin to a development following the earlier signal/se ratio approach.

In comparing two proportions, the signal is provided by the observed difference  $r_1 - r_2$  in the proportions. The standard error for that difference is given by the expression (see Section 2.5.2)  $\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}$ . The probabilities associated with the resulting signal/se ratio (the test statistic) when the true proportions  $\theta_1$  and  $\theta_2$  are equal are provided by a special case of the normal distribution,  $N(0, 1)$ : the normal distribution with mean zero and standard deviation 1. For the trastuzumab example, the signal takes the value 0.023, and the value of the standard error is 0.0081. So to obtain the  $p$ -value based on this null distribution, we compare the value of our test statistic ( $= 0.023/0.0081 = 2.84$ ) with the  $N(0, 1)$  distribution giving a  $p$ -value of 0.0046. The Pearson chi-square test had earlier given  $p = 0.0054$ , a very similar value.

In general, it can be shown that the approach following the signal/se ratio method is mathematically very similar to the standard formulation of the chi-square test using observed and expected frequencies, and in practice, they will invariably give very similar results. Altman (1991) (Section 10.7.4) provides more detail on this connection.

### 4.5 Measures of treatment benefit

The chi-square test has given a  $p$ -value, and this provides the evidence in relation to the existence of a treatment difference. Through the CI for the difference in the proportions of patients with SAEs calculated earlier in Section 3.2.2, we have some idea of the extent of the treatment effect in absolute terms. There are,

however, other measures of treatment benefit/harm in common usage for binary endpoints. Each of these measures – odds ratio (OR), relative risk (RR), relative risk reduction (RRR), absolute risk reduction (ARR) and number needed to treat (NNT) – is a way of expressing the treatment benefit. They each have their advantages and disadvantages, and I think it is fair to say that none of them is universally accepted as the *single best approach*. We will define each in turn and provide an interpretation and critique of their use. For further details and discussion, see Grieve (2003).

#### 4.5.1 Odds ratio

To understand the OR, you first of all need to understand odds. For the data of Example 3.3, consider each of the treatment groups separately.

For the trastuzumab group, the odds of a patient suffering one or more SAEs are  $117/1560 = 0.075$ ; for every patient free from SAEs, there are 0.075 patients suffering one or more SAEs. For the observation-only group, the odds of a patient suffering one or more SAEs are  $81/1629 = 0.050$ . In this group, for every patient not suffering SAEs, there are 0.050 patients who do suffer one or more SAEs.

The *odds ratio* is then the ratio of the odds of a patient suffering one or more SAEs:

$$\text{OR} = 0.075 / 0.050 = 1.51$$

An OR of one, or close to one, tells us that the treatments are the same (or at least similar). An OR greater than one tells us that you are worse off in the test treatment group and vice versa, but over and above that, the interpretation is not straightforward. It is the definition itself that provides this interpretation; the value 1.51 for the OR indicates that the odds of suffering at least one SAE in the trastuzumab group is 1.51 times the odds of suffering at least one SAE in the observation-only group or, alternatively, a 51% increase in the odds of suffering at least one SAE. What makes this difficult to interpret is that these are statements on the *odds scale*, and the odds scale is something that we find difficult to work with.

Usually, the odds relating to the test treatment group go on the top when calculating the ratio (the numerator), while the odds for the control group go on the bottom (the denominator). However, there is no real convention regarding whether we calculate the odds in favour of success or the odds in favour of failure. Had we chosen to calculate the odds in favour of no SAE, the OR would have been  $\frac{1/0.075}{1/0.050}$ , which has the value 0.66 ( $= 1/1.51$ ); so take care that when you see an OR presented, you are clear how the calculation has been organised.

#### 4.5.2 Relative risk

The RR is defined again as a ratio, this time in relation to the risks calculated for the two treatments. For the trastuzumab group, the *risk* is the proportion of patients suffering one or more SAEs and takes the value  $117/1677 = 0.070$ ,

while for the observation-only group, this is  $81/1710 = 0.047$ . The *relative risk* (sometimes called the *risk ratio*) is then the ratio of these risks:

$$RR = 0.070 / 0.047 = 1.47$$

A RR of one, or close to one, is again indicative of similar treatments. A RR above one, as here, is saying that the risk in the test treatment group is higher than the risk in the control group. The interpretation beyond that is simpler than for the OR. The RR of 1.47 tells us that the risk in the trastuzumab group is 47% higher than the risk in the observation-only group.

There are also conventions with RR. As with the OR, we usually put the risk for the test treatment group as the numerator and the risk for the control group as the denominator. But now, because we are calculating risk, there should be no confusion regarding what we view as the event; we tend to calculate relative risk and not relative benefit.

### 4.5.3 Relative and absolute risk reduction

Consider the data presented in Table 4.4 relating to the binary outcome died/survived in a parallel-group trial.

The RR for death is  $0.20/0.35 = 0.57$ .

When the RR is less than one, as in this case, we often also calculate the reduction in the RR as relative risk reduction (RRR) =  $1 - RR$  or the absolute risk reduction (ARR) as  $r_2 - r_1$ .

In the example, RRR = 0.43: there is a 43% relative reduction in the risk (of death) in the active group compared to control. And ARR =  $0.35 - 0.20 = 0.15$ : there is a 15% absolute reduction in the risk of death in the active group compared to control.

We use RRR and ARR where the intervention is having a benefit in reducing the risk. In the earlier example involving trastuzumab and the incidence of patients suffering one or more SAEs, the active treatment was associated with an increase in risk. In this case, we speak in terms of a RR increase of 0.47 ( $= 1.47 - 1$ ), a 47% increase in the risk of suffering one or more SAEs. Similarly, there is an absolute increase in the risk of suffering an SAE of 0.023 ( $0.070 - 0.047$ ).

**Table 4.4** Active/placebo comparison: Binary outcome survival (hypothetical)

	Died	Survived	Total
Active	20	80	100
Placebo	35	65	100
Total	55	145	200

#### 4.5.4 Number needed to treat

In the example in Table 4.4, 80% of patients in the active group survived compared to 65% in the placebo group. So out of 100 patients, we would expect to see, on average, an additional 15% ( $80\% - 65\%$ ) surviving in the active group. The *number needed to treat* (NNT) is then  $100/15$  or 6.7. We need to treat on average an additional 6.7 patients with the active treatment to save one additional life.

A convenient formula for NNT is

$$\text{NNT} = \frac{1}{(0.80 - 0.65)}$$

The denominator here is the difference in the survival proportions. Note also that  $\text{NNT} = 1/\text{ARR}$ .

We usually round this up to the nearest integer, so  $\text{NNT} = 7$  in our example. We need to treat seven patients with the active medication to see one extra patient survive compared to placebo.

There may be some situations where the test treatment is, in fact, harmful relative to the control treatment in terms of a particular endpoint. In these circumstances, it does not make sense to talk about NNT, and we refer instead to *number needed to harm (NNH)*. So, for the data in Example 3.3, the SAE proportion in the trastuzumab group was 6.98 % compared to 4.74 % in the observation-only group, and the NNH is equal to  $1/(0.0698 - 0.0474)$ , which rounds up to 45.

#### 4.5.5 Confidence intervals

We saw earlier (Section 3.2.2) how to calculate CIs for the difference in the true proportions of patients with the event of interest. We will now look at methods for calculating a CI for the OR.

Calculating CIs for ratios is trickier than calculating CIs for differences. We saw in Chapter 3 that, in general, the formula for the CI is

$$\text{statistic} \pm (\text{constant} \times se)$$

With a ratio, it is not possible to obtain a standard error formula directly; however, it is possible to obtain standard errors for log ratios. (Taking logs converts a ratio into a difference with  $\log A/B = \log A - \log B$ ). So, first of all, we calculate CIs on the log scale. It does not make any difference what base we use for the logs, but by convention, we usually use natural logarithms, denoted  $\ln$ .

The standard error for the  $\ln$  of the OR is given by

$$\sqrt{\frac{1}{O_1} + \frac{1}{O_2} + \frac{1}{O_3} + \frac{1}{O_4}}$$

where the  $O$ s are the respective observed frequencies in the  $2 \times 2$  contingency table (see Table 4.3). In the trastuzumab example, this is given by

$$\sqrt{\frac{1}{117} + \frac{1}{1560} + \frac{1}{81} + \frac{1}{1629}} = 0.149$$

The 95% CI for the  $\ln$  of the OR is then

$$\ln 1.51 \pm (1.96 \times 0.149) = (0.120, 0.704)$$

Finally, we convert this back onto the OR scale by taking antilogs of the ends of this interval to give a 95% CI for the OR as (1.13, 2.02). We can be 95% confident that the true OR lies within this range.

In a similar way, we can calculate a CI for the RR. The method is the same as for the OR but with a different formula for the standard error. The standard error for the log of the RR is given by

$$\sqrt{\frac{1}{O_1} + \frac{1}{O_3} - \frac{1}{n_1} - \frac{1}{n_2}}$$

In the trastuzumab example, this is given by

$$\sqrt{\frac{1}{117} + \frac{1}{81} - \frac{1}{1677} - \frac{1}{1710}} = 0.140$$

The 95% CI for the  $\ln$  of the RR is then

$$\ln 1.47 \pm (1.96 \times 0.140) = (0.110, 0.660)$$

Converting this back onto the RR scale gives a 95% CI for the RR as (1.12, 1.93), and we can be 95% confident that the RR lies within this range.

Previously, when we calculated a CI – for example, for a difference in proportions or a difference in means – the CI was symmetric around the estimated difference; in other words, the estimated difference sat squarely in the middle of the interval, and the ends of the interval were obtained by adding and subtracting the same amount ( $2 \times se$ ). When we calculate a CI for the OR (or the RR), the interval is symmetric only on the log scale. Once we convert back to the OR scale by taking antilogs, that symmetry is lost. This is not a problem, but it is something that you will notice. It is a property of all standard CIs calculated for ratios.

CIs for NNT are a little more complicated; see Grieve (2003) and Altman (1998) for further details.

#### 4.5.6 Interpretation

In large trials and with events that are rare, the OR and RR give very similar values. We can see this in the trastuzumab example, where the OR was 1.51 and the RR was 1.47. In smaller trials and with more common events, however, this will not be the case. Comparable values for the OR and the RR arise more frequently in cohort studies where the sample sizes are generally large and the events being investigated are often rare, and these measures tend to be used interchangeably. As a result, there seems to be some confusion as to the distinction, and it is my experience that the OR and RR are occasionally labelled incorrectly in clinical research papers, so take care.

It is possible to convert from an OR to an RR (and vice versa). The conversion formula is

$$RR = \frac{OR}{1 - r_c + (r_c \times OR)}$$

where  $r_c$  is the absolute risk in the control group. When calculating an OR or, alternatively, taking a value for an OR from a publication, the value for the absolute risk in the control group may not be available directly. However, it may be the case that a value for this risk is available from an alternative source, and an approximate value for the RR can then be calculated. Converting from an OR to an RR can be useful since, as discussed earlier, the OR can be difficult to interpret. The RR is much easier in this regard.

A question that is sometimes asked is, why do we use the ORs when they are such difficult quantities to interpret? Well, there are essentially two reasons. Firstly, in case-control studies in epidemiology, it is not possible to calculate a RR; it is only possible to calculate an OR. But it is usually the case in such studies that we are dealing with rare events, and the RR and the OR are again numerically close together, so the OR can be interpreted as if it were a RR. I will say more about this issue in Section 17.8.2. Secondly, we have developed several mathematical techniques, such as those relating to meta-analysis (see Chapter 18) and logistic regression (see Section 6.6.1), that revolve around the OR, and consequently, it tends to dominate our way of thinking about binary endpoints. In summary, although not ideal, the OR is central to the way we analyse and report binary endpoints. However, as pointed out by Grimes and Schulz (2008), ‘For most clinicians, odds ratios will remain . . . well, odd’.

It is also worth mentioning that all the measures – difference in event proportions, OR, RR, RRR, ARR and NNT – expressed in isolation have limitations. What we are trying to do with such quantities is to use a single measure to summarise the data. All information is contained in the two event proportions  $r_1$  and  $r_2$ , and attempting to summarise two numbers by a single number is inevitably going to lead to problems in particular cases. Beware of those limitations and revert to  $r_1$  and  $r_2$ , if need be, to tell the full story.

## 4.6 Fisher's exact test

The Pearson chi-square test is what we refer to as a *large sample test*; this means that provided the sample sizes are fairly large, it works well. Unfortunately, when the sample sizes in the treatment groups are not large, there can be problems. Under these circumstances, we have an alternative test, *Fisher's exact test*.

The way this works is as follows. Consider Table 4.5.

Given there are only seven successes in total (and 41 failures), we can easily write down everything that could have happened (recall the way we looked at flipping a coin in Section 3.3.2) and calculate the probabilities associated with each of these outcomes when there really are no differences between the treatment success proportions (Table 4.6).

We observed the 6/1 split in terms of successes across the two treatment groups in our data, and we can calculate the two-sided  $p$ -value by adding up the probabilities associated with those outcomes – which are as extreme, or more extreme, than what we have observed – when the null hypothesis is true (equal treatments). This gives  $p = 0.097$  ( $= (0.0439 + 0.0047) \times 2$ ). The corresponding chi-square test applied (inappropriately) to these data would have given  $p = 0.041$ , and the conclusion would have been slightly different.

The rule of thumb for the use of Fisher's exact test is based on the expected frequencies ( $E$ ) in the  $2 \times 2$  contingency table; each of these should be at least five for the chi-square test to be applicable. In our example, the expected frequencies in each of the cells corresponding to *success* are 3.5, supporting the use of Fisher's exact test in this case.

**Table 4.5** Data for Fisher's exact test

	Success	Failure	Total
Group A	6	18	24
Group B	1	23	24
Total	7	41	48

**Table 4.6** Probabilities for Fisher's exact test under the null hypothesis

Successes on A	Successes on B	Probability
7	0	0.0047
6	1	0.0439
5	2	0.1593
4	3	0.2921
3	4	0.2921
2	5	0.1593
1	6	0.0439
0	7	0.0047

In fact, Fisher's exact test could be used under all circumstances for the calculation of the  $p$ -value, even when the sample sizes are not small. Historically, however, we tend not to do this; Fisher's test requires some fairly hefty combinatorial calculations in large samples to get the null probabilities, and in the past, this was difficult. For larger sample sizes,  $p$ -values calculated using either the chi-square test or Fisher's exact test will be similar, so we tend to reserve the use of Fisher's exact test for only those cases where it is needed and use the chi-square test outside of that.

## 4.7 Tests for categorical and ordered categorical endpoints

### 4.7.1 Categorical endpoints

The Pearson chi-square test extends in a straightforward way when there are more than two outcome categories.

Consider four outcome categories labelled A, B, C and D and the comparison of two treatments in terms of the distribution across these categories. Taking the example of categorical data from Chapter 1, we might have

A = death from cancer causes

B = death from cardiovascular causes

C = death from other causes

D = survival

Consider the hypothetical data displayed in Table 4.7.

The chi-square test proceeds, as before, by calculating expected frequencies. These are given in Table 4.8.

**Table 4.7** Observed frequencies ( $O$ )

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>Total</b>
Group 1	15	13	20	52	100
Group 2	17	20	23	40	100
Total	32	33	43	92	200

**Table 4.8** Expected frequencies ( $E$ )

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>Total</b>
Group 1	16	16.5	21.5	46	100
Group 2	16	16.5	21.5	46	100
Total	32	33	43	92	200

As before, we then compute

$$\sum \frac{(O - E)^2}{E}$$

with the sum being over all eight cells.

The resultant test statistic is then compared to the chi-square distribution but this time on three df, written  $\chi_3^2$ . As we mentioned earlier, the particular chi-square shape that we use is not determined by the number of patients; rather, it depends upon the size of the contingency table. In this example, we have a  $2 \times 4$  table (four outcome categories), and with two treatment groups, the df for the chi-square distribution are equal to the number of categories – 1.

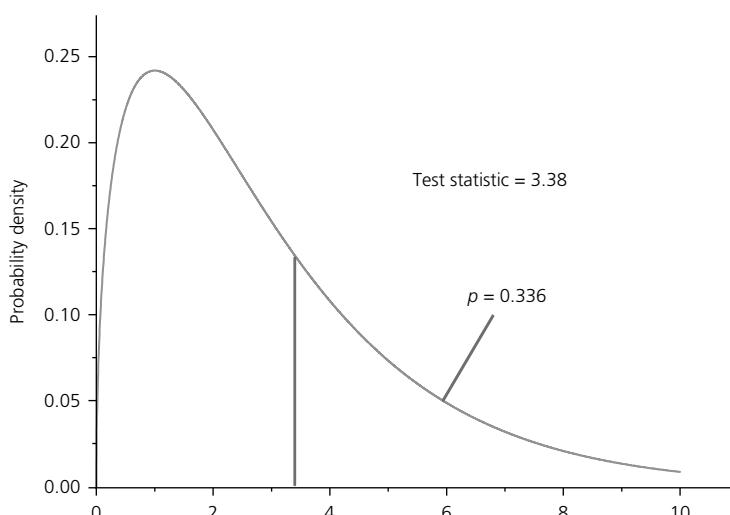
In the example, the test statistic value is 3.38, and Figure 4.3 illustrates the calculation of the  $p$ -value, which for these data turns out to be  $p = 0.336$ . This is a non-significant result.

Although this test provides a valid comparison of the treatment groups in relation to the outcome categories, it is not of any great value in providing a useful conclusion from a clinical perspective. The procedure provides a test of the null hypothesis

$$H_0 : \theta_{1A} = \theta_{2A} \text{ and } \theta_{1B} = \theta_{2B} \text{ and } \theta_{1C} = \theta_{2C} \text{ and } \theta_{1D} = \theta_{2D}$$

where the suffices label both the treatment group and the outcome category against what we call the *general alternative hypothesis*:

$$H_1 : \text{the opposite of } H_0$$



**Figure 4.3** Chi-square distribution on three df

But suppose we see a significant  $p$ -value; what does it mean? Well, it simply means there are some differences somewhere across the categories, but nothing more specific than that – and that is not particularly useful.

A further issue is that we very rarely see strictly categorical outcomes in practice in our clinical trials; it is much more common to have an ordering of the categories, giving us an ordered categorical outcome, and we will deal with this in the next section. Finally, the problem we discussed in the previous section regarding small sample sizes applies here, and Fisher's exact test should be used when this is the case. The rule of thumb again is that all the expected frequencies should be at least five for the chi-square test to be valid. Computer programs tend to give both  $p$ -values (chi-square test and Fisher's exact test) automatically, so it is no great problem to pick the appropriate one depending on sample size and the rule of thumb.

### 4.7.2 Ordered categorical (ordinal) endpoints

Clinical trials in many different therapeutic settings frequently use an endpoint consisting of ordered categories.

The Pearson chi-square test is not appropriate for ordered categorical endpoints as it does not take any account of the ordering of the outcome categories. The appropriate test is the *Mantel–Haenszel (MH) chi-square test* (Mantel and Haenszel, 1959). This test takes account of the ordering by scoring the ordered categories (e.g. improved = 1, no change = 2, worse = 3) and comparing the average score in one treatment group with the average score in the second treatment group. It is not quite as simple as this, but that is the general idea! The scores are chosen by the observed data patterns in the two groups combined.

The formula for the test statistic is somewhat complex, but again, this statistic provides the combined evidence in favour of treatment differences. When Mantel and Haenszel developed this procedure, they calculated that when the treatments are identical, the probabilities associated with the test statistic follow a  $\chi^2_1$  distribution. This is irrespective of the number of outcome categories, and the test is sometimes referred to as the *chi-square one degree of freedom test for trend*.

**Example 4.2** Flutamide plus leuprolide compared to leuprolide alone in the treatment of prostate cancer

The data in Table 4.9 are taken from a randomised controlled add-on trial (Crawford et al., 1989) comparing leuprolide + placebo (L + P) with leuprolide + flutamide (L + F) in prostate cancer. The endpoint here is improvement in pain at week 4.

**Table 4.9** Flutamide data

	Improved	No change	Worse	Total
L + P	50	180	33	263
L + F	73	174	20	267
Total	123	354	53	530

**Table 4.10** Flutamide data as percentages

	Improved	No change	Worse	Total
L + P	19%	68%	13%	263
L + F	27%	65%	7%	267

In this example, comparison of the test statistic value with  $\chi^2$  gives  $p = 0.006$ , a highly significant result.

A significant  $p$ -value coming out of this test is indicative of a shift or trend in one direction across these categories for one treatment compared to the other. It is illustrative to look at the percentages in the various categories as shown in Table 4.10.

In the L + F treatment group, there has been a shift towards the improvement end of the scale compared to the L + P treatment group, and it is this trend that the MH chi-square test has picked up. The Pearson chi-square test does not look for such trends and would have simply compared the 19% with the 27%, the 68% with the 65% and the 13% with the 7% in an overall, average way. Had the ordering been ignored and the Pearson chi-square test applied (incorrectly), then the  $p$ -value would have been 0.023. Although this is still statistically significant, indicating treatment differences, the  $p$ -value is somewhat different to the correct  $p$ -value of 0.006, and in many cases, such differences could lead to incorrect conclusions.

As with binary and categorical endpoints, is there an issue with small sample sizes? Well, no, there is not. The MH test is a different kind of chi-square test and is not built around expected frequencies. As a consequence, it is not affected by small expected frequencies and can be used in all cases for ordered categorical endpoints. There are some pathological cases where it will break down, but these should not concern us in practical settings.

### 4.7.3 Measures of treatment benefit

Measures such as the difference in the proportions of patients with events, OR, RR, RRR, ARR and NNT do not easily translate into the categorical data context. If we wanted to construct such measures in these cases, we would collapse the outcome categories to two, the binary case, and proceed as before. In the categorical example covered earlier, this could involve collapsing categories A, B and C to produce a binary outcome death/survival. As a further example, if the categorical outcome was *Main Reason for Discontinuation*, which was classified as Adverse Event, Withdrawal of Consent, Protocol Violation or Other, there might be interest in expressing an OR in relation to Adverse Event, in which case we would collapse the other three categories and proceed as in the binary case.

For ordered categorical endpoints, we again collapse adjacent outcomes: for example, by looking at *improved* versus *not improved* (= no change or worse). Another approach is to work with the average OR across the range of outcomes. With three outcome categories this involves forming two  $2 \times 2$  contingency

tables by firstly collapsing *improved* and *no change* and secondly collapsing *no change* and *worse*. In each of these two separate tables, we calculate the OR; the average OR is then obtained as an *average* of the two. In our example, the two ORs are 1.77 and 1.60, so the average OR will be somewhere in the middle of these two values. The averaging process is a little complex, and we will not go into details here. The quoted OR is then the odds in favour of a better outcome on average, with ‘better outcome’ considered as either *improved* or a combination of *improved* and *no change*.

## 4.8 Count endpoints

We have said little so far about count endpoints and how to measure treatment effects in relation to those kinds of endpoints. Most commonly, counts occur as *recurrent events* – for example, epileptic seizures over a 6-month period, moderate/severe asthma attacks over a 12-month period or number of incontinence episodes recorded in a 3-day diary – and the purpose of treatment is to reduce the rate at which those events occur. One possibility in theory would be to observe each subject over the same time period and perhaps compare the mean number of events per patient between the treatment groups. In practice, however, it is unlikely that all patients will be observed for the same period due to dropouts, withdrawals, missing data, etc. or in association with an interim analysis where, inevitably, the observation period will not be the same across all patients. A more appropriate methodology for analysis is based on evaluating the number of events through the *negative binomial model* with group comparisons based on a ratio of the event rates per unit of time: the *rate ratio*. As this analysis revolves around modelling, we will delay further discussion until Chapter 6.

## 4.9 Extensions for multiple treatment groups

In this section, we will discuss the extension of the t-tests for continuous endpoints and the chi-square tests for binary, categorical and ordered categorical endpoints to deal with more than two treatment arms.

### 4.9.1 Continuous endpoints

In this setting, a technique termed *one-way analysis of variance (one-way ANOVA)* gives an overall *p*-value for the simultaneous comparison of all treatments. Suppose, for example, that we have four treatment groups with means  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$  and  $\mu_4$ . This procedure gives a *p*-value for the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

against the general alternative hypothesis:

$H_1$  : the opposite of  $H_0$ , that there are differences somewhere

A significant  $p$ -value from this test would cause us to reject the null hypothesis, but the conclusion from this only tells us that there are some differences somewhere; at least two of the  $\mu$ 's are different. At that point, we would want to look to identify where those differences lie, and this would lead us to pairwise comparisons of the treatment groups and reverting to a series of unpaired t-tests. It could be argued that the question posed by the one-way ANOVA technique is of little value and that it is more relevant to start directly with a set of structured questions relating to the comparisons of pairs of treatments.

For example, let us suppose that we have three treatment groups: test treatment ( $\mu_1$ ), active control ( $\mu_2$ ) and placebo ( $\mu_3$ ). In a superiority setting, there could be two questions of interest:

**1** Does the test treatment work?

$$H_0 : \mu_1 = \mu_3 \quad H_1 : \mu_1 \neq \mu_3$$

**2** Is the test treatment better than the control treatment?

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Both questions are answered by the unpaired t-test.

However, one small advantage of one-way ANOVA is that it uses all data from all of the treatment groups to give us a measure of noise, so even when we are comparing the test treatment with the control treatment, we are using information on the patient-to-patient variation in the placebo group to help estimate the noise; we work with a pooled estimate of the standard deviation. There are ways of adapting the unpaired t-test to incorporate this broader base of information, but even then, the gains are small except in very small trials where information regarding the noise is at a premium. See Julious (2004) for further details.

In summary, there is not much to be gained in using one-way ANOVA with multiple treatment groups. A simpler analysis structuring the appropriate pairwise comparisons will more directly answer the questions of interest. One final word of caution, though: undertaking multiple comparisons in this way raises another problem, that of multiplicity. For the time being, we will put that issue to one side; we will return to it in Chapter 10.

#### **4.9.2 Binary, categorical and ordered categorical endpoints**

As with continuous endpoints, the most relevant questions for binary, categorical and ordered categorical endpoints will invariably relate to pairwise comparisons of treatments. We mentioned earlier for continuous endpoints that one minor advantage was the ability to use information from all of the treatment

groups to estimate patient-to-patient variation assuming a common standard deviation across the treatment groups. With binary, categorical and ordered categorical endpoints, however, even this small advantage does not apply; there is no standard deviation of this kind with these endpoint types. The recommendation again, therefore, is to focus on the chi-square procedures developed earlier in this chapter for pairwise treatment comparisons.

### 4.9.3 Dose-ranging studies

The discussion so far in this section has assumed that the treatment groups are unordered. But in some situations, these multiple treatment groups correspond to placebo and then increasing dose levels of a drug. It could still be in these circumstances that we are looking to compare each dose level with placebo to identify, for example, the minimum effective dose – and again we are back to the pairwise comparisons.

However, there will be some circumstances where we are just interested in trends; if we increase the dose, does the mean response increase?

For continuous endpoints, there is a procedure within the one-way ANOVA methodology that focuses on this; we would be looking for a trend across the treatment groups.

For binary, categorical and ordered categorical endpoints, there is also an approach that is a further form of the MH chi-square test. You will recall that the MH test is used for ordered categorical responses comparing two treatments. Well, this procedure generalises to allow ordering across the treatment groups in addition for each of the binary, categorical and ordered categorical endpoint types. More details can be found in Stokes, Davis and Koch (2000).

### 4.9.4 Further discussion

For the remainder of this book, as we investigate further designs and methods of analysis, we will focus our developments on comparing two treatment groups. When we have more than two treatment groups, our questions are usually in relation to pairwise comparisons in any case, as discussed earlier, and these can be handled directly by reducing to those specific evaluations. For binary, categorical and ordered categorical endpoints, this is precisely the approach. For continuous endpoints, there are some advantages in efficiency in using a combined estimate of the standard deviation from the complete experiment, and this is what is usually done.

However, specific mention will be made in multiple treatment group settings where issues arise that require considerations outside of these.

## CHAPTER 5

# Adjusting the analysis

### 5.1 Objectives for adjusted analysis

The main reason we might want to adjust the analysis is to account for imbalances in baseline factors. We have already discussed in Chapter 1 the importance of stratifying the randomisation to avoid baseline imbalances and confounding between treatment group and factors at baseline that are key determinants of outcome. However, even if the randomisation is stratified, there may still be minor imbalances that could influence our measures of treatment difference. Also, we cannot stratify for everything, and in addition, factors may come to light while the trial is ongoing that are predictors of outcome that clearly were not included in any stratified randomisation. There is also a more technical statistical point. From a theoretical perspective, adjusting the analysis to take account of the stratification is required to preserve the statistical properties of *p*-values and confidence intervals (CIs) in a strict sense.

The methodology we will present in this chapter also provides a framework for evaluating the homogeneity of treatment effect in subgroups defined by the baseline factors. This is important in terms of underpinning our ability to generalise the results of the study to the population as a whole, something the regulators refer to as *generalisability*.

### 5.2 Comparing treatments for continuous endpoints

Consider the following hypothetical situation comparing an active treatment + dietary advice with placebo + dietary advice in cholesterol lowering, where the primary endpoint is the reduction (baseline minus final) in total cholesterol (mmol/l). Summary statistics are provided in Table 5.1, and these have been broken down by age group.

Two features of these data are worth noting at this stage. Firstly, there are imbalances in the distribution of age across the two treatment groups. The patients in the control group tend to be younger; the mean age in the active

**Table 5.1** Sample sizes and mean values for reduction in total cholesterol (mmol/l) in a (hypothetical) trial comparing an active treatment + dietary advice with placebo + dietary advice. Data Set 1

Age group	Placebo	Active
<50	n = 28, mean = 0.36	n = 21, mean = 0.68
≥50, <60	n = 17, mean = 0.17	n = 16, mean = 0.39
≥60	n = 33, mean = 0.10	n = 43, mean = 0.36

**Table 5.2** Treatment differences by age category. Entries are sample sizes and mean values. Data Set 1

Age group	Placebo (P)	Active (A)	Difference (A – P) in means
<50	n = 28, 0.36	n = 21, 0.68	n = 49, 0.32
≥50, <60	n = 17, 0.17	n = 16, 0.39	n = 33, 0.22
≥60	n = 33, 0.10	n = 43, 0.36	n = 76, 0.27

group was 57.8 years compared to 55.6 years in the control group. Secondly, irrespective of treatment group, the younger patients on average performed better than the older patients, and the mean reduction in total cholesterol was greatest in both treatment groups in the patients aged <50 and smallest in the patients aged ≥60. Ignoring age category, the overall mean reduction in the active group was 0.45 mmol/l ( $n = 80$ ), while the overall mean reduction in the placebo group was 0.21 mmol/l ( $n = 78$ ), so that, to three decimal places, the treatment difference was 0.246 mmol/l. This results in a  $p$ -value from the unpaired t-test of 0.072, not quite statistically significant at the 5% level. However, given the imbalance in age across the two treatment groups, and also noticing that older patients didn't do quite as well, we need to recognise that the active group has been penalised by having more than its fair share of older patients. Had we not had this imbalance, we might well have seen statistically significant differences in favour of the active treatment.

Let's think about how we might adjust the analysis to correct for those baseline imbalances. The first step in this adjustment approach is to calculate the treatment difference within each age category. These differences are given in the final column in Table 5.2.

Having calculated the differences between the active mean and the placebo mean, we now average these three values, 0.32, 0.22 and 0.27, to give us our overall measure of treatment difference. This is the *adjusted treatment difference* and for these data takes the value 0.274. It is not the simple average of these three values but a weighted average, weighted according to the sample size in each of the age categories overall. We do this because averages based on larger sample sizes are more reliable and more precise than those based on smaller sample sizes and so are given more weight. Don't worry too much about this

weighting; it is a technical issue relating to optimising the statistical properties of the procedure. Had there been equal numbers of patients overall in the three age categories, there would not have been any weighting, and we would have taken just the straight numerical average, so think of it in those terms.

Note that the adjusted treatment difference of 0.274 is larger than the overall difference calculated earlier of 0.246, which was based on calculating the difference at the group level. The process of calculating treatment differences for each age category and then averaging those differences is not influenced by the imbalances between the treatment groups in terms of age. For example, the difference of 0.32 seen for the patients in the <50 group has simply been calculated from the individual treatment means of 0.68 and 0.36 and is not in any way affected by there being fewer patients <50 in the active group compared to the placebo group. The adjusted difference is then a better measure of the true treatment difference; it compares like with like in terms of two groups balanced according to the age distribution of the patients. The adjusted difference of 0.274 is now the signal, while the noise is a weighted combination of the individual standard deviations in each of the six cells of the table. The resulting *p*-value is 0.045, which now interestingly indicates statistical significance at the 5% level.

Adjusting for baseline factors in this way provides a fair view of the treatment difference, something that would be lost if we simply compared the overall means in an unadjusted analysis. One common question is, if the treatment groups are perfectly balanced, isn't this adjusted analysis a waste of time? The answer to that question is no – it is not a waste of time. Adjusting the analysis never harms you. If the treatment groups are perfectly balanced for baseline factors, then adjusting the analysis will make no difference, and the *p*-value for the adjusted analysis will pretty much agree with the *p*-value comparing the overall treatment means. If the treatment groups are not perfectly balanced, then the adjusted analysis will correct for those imbalances. Of course, at the planning stage, you do not know whether or not you are going to see imbalances, so specifying the adjusted analysis is an insurance policy just in case.

This method of analysis is referred to as *two-way analysis of variance (ANOVA)* or alternatively as a *stratified analysis*. The focus is to compare the treatment groups while recognising potential treatment group differences in the distribution of baseline factors. To enable this to happen, we allow the true treatment means  $\mu_A$  and  $\mu_B$  for groups A and B to be different for the different age groups, as seen in Table 5.3.

**Table 5.3** True means according to age group

Age group	Treatment A	Treatment B	Difference
<50	$\mu_{A1}$	$\mu_{B1}$	$\mu_{A1} - \mu_{B1}$
≥50, <60	$\mu_{A2}$	$\mu_{B2}$	$\mu_{A2} - \mu_{B2}$
≥60	$\mu_{A3}$	$\mu_{B3}$	$\mu_{A3} - \mu_{B3}$

The null hypothesis we evaluate is then

$$H_0 : \mu_{A1} = \mu_{B1} \text{ and } \mu_{A2} = \mu_{B2} \text{ and } \mu_{A3} = \mu_{B3}$$

against the general alternative hypothesis  $H_1$  that there are differences somewhere.

This null hypothesis says that the treatment means are the same within each age group but not necessarily across age groups. Two-way ANOVA gives us a  $p$ -value for the adjusted treatment difference that evaluates this null hypothesis. If that  $p$ -value is significant ( $p \leq 0.05$ ), then we reject the null hypothesis, and there is evidence supporting a treatment difference.

This analysis implicitly assumes that the treatment effect is consistent across the various age categories. For the data being considered in this section, this seems a reasonable assumption, but we will return to a more formal evaluation of this assumption in a later section. The weighted average of the treatment differences provides the best estimate of the overall treatment effect. In our example, this was 0.274 mmol/l, and we can construct CIs around this value to allow an interpretation of the size of the true treatment difference.

There is a further advantage for this kind of analysis: it improves *power*, our ability to detect treatment differences if they truly exist. This is true irrespective of whether there are imbalances at baseline. This is because, generally, adjusting the analysis in this way reduces the noise. When we undertake an unadjusted analysis, the standard deviation is calculated as the *average* of the two standard deviations for the complete groups of patients in treatment groups A and B. With the adjusted analysis as discussed earlier, the standard deviation is calculated as the *average* of the six standard deviations in the six cells of the table, one for each treatment and age group combination. The latter will tend to be smaller than the former since in general there will be greater homogeneity within each treatment and age category combination than there would be taking each of the treatment groups as a whole.

In the next chapter, we will talk about adjusting the analysis for several factors simultaneously. We often call the factors we wish to adjust for *covariates*, and there is considerable regulatory guidance on adjustment for covariates. Adjusting in this way is frequently part of the primary analysis for the key endpoints, and as with all analyses, details should be set down in the protocol and the statistical analysis plan.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'In some instances an adjustment for the influence of covariates or for subgroup effects is an integral part of the planned analysis and hence should be set out in the protocol'.*

In the CHMP ‘Guideline on adjustment for baseline covariates in clinical trials’, there is the clear recommendation to use adjusted analyses to account for baseline imbalances.

**CHMP (2015): ‘Guideline on adjustment for baseline covariates in clinical trials’**

*‘When there is some imbalance between treatment groups in a baseline covariate that is solely due to chance then adjusted treatment effects may account for this observed imbalance when unadjusted analyses do not’.*

One final point concerns the *requirement* to adjust for factors that were used as stratification factors at the randomisation stage.

**CHMP (2015): ‘Guideline on adjustment for baseline covariates in clinical trials’**

*‘The primary analysis should reflect the restriction on the randomisation implied by the stratification. For this reason, stratification variables, if not solely used for administrative reasons, should usually be included as covariates or stratification variables in the primary analysis regardless of their prognostic value’.*

### 5.3 Least squares means

In our example, the two overall (raw) treatment means were 0.45 in the active group and 0.21 in the placebo group. But we recognise, because of the baseline age imbalances, that these means are not directly comparable. In conjunction with adjusting the treatment comparison, it seems natural therefore to adjust the individual treatment means to also account for the baseline imbalances. The total sample size is 158, and had the groups been perfectly balanced for age, we

would have seen a proportion  $0.31\left(=\frac{49}{158}\right)$  in the <50 category in each of the treatment groups, a proportion  $0.21\left(=\frac{33}{158}\right)$  in each of the treatment groups in the middle age category and finally a proportion  $0.48\left(=\frac{76}{158}\right)$  in each of the treatment groups in the  $\geq 60$  category. We can then reconstruct both the active and placebo means that would have been seen had we had this perfect balance. These reconstructed (adjusted) means are obtained as weighted combinations of the observed means across the three age categories as follows:

$$\text{Active adjusted mean} = 0.31 \times 0.68 + 0.21 \times 0.39 + 0.48 \times 0.36 = 0.47$$

$$\text{Placebo adjusted mean} = 0.31 \times 0.36 + 0.21 \times 0.17 + 0.48 \times 0.10 = 0.20$$

As we would expect, these adjusted means are slightly further apart than the original raw means, which were 0.45 for the active group and 0.21 for the placebo group. Had the groups been perfectly balanced, we would have observed a bigger treatment difference. This adjustment of the individual treatment means fits exactly with our adjusted method for comparing the treatments in that the difference  $0.47 - 0.20 = 0.27$ , which was the adjusted treatment difference we calculated previously; this is a mathematical connection. For reasons that we will explain in the next chapter, we usually refer to these adjusted means as *least squares (LS) means*.

## 5.4 Evaluating the homogeneity of the treatment effect

### 5.4.1 Treatment-by-factor interactions

The adjusted framework is based on calculating treatment differences within subgroups defined by the factor being adjusted for and then averaging those differences. As we have said, there is an implicit assumption within that framework that these differences are similar across the various levels of the factor. Evaluating whether this assumption is appropriate is an important next step as this could impact the generalisability of the trial results. If the treatment effect is not homogeneous, then we will talk in terms of having *treatment-by-factor* or *treatment-by-covariate* interactions. The ICH E9 guideline is quite clear on the need to evaluate the homogeneity of the treatment effect.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The treatment effect itself may also vary with subgroup or covariate – for example, the effect may decrease with age or may be larger in a particular diagnostic category of subjects. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis'.*

We will firstly discuss a significance test for the treatment-by-covariate interaction and then talk about graphical methods.

Consider again the (hypothetical) setting evaluating a treatment for lowering cholesterol with new data and new summary statistics as set out in Table 5.4.

**Table 5.4** Treatment differences by age category. Entries are sample sizes and mean values. Data Set 2

Age group	Placebo (P)	Active (A)	Difference (A – P) in means
<50	$n = 28, 0.36$	$n = 21, 0.68$	$n = 49, 0.32$
$\geq 50, <60$	$n = 17, 0.17$	$n = 16, 0.29$	$n = 33, 0.12$
$\geq 60$	$n = 33, 0.10$	$n = 43, 0.16$	$n = 76, 0.06$

In these data, the active treatment is performing on average better than placebo, but now the extent of the difference is not consistent. There is a large difference among the <50 age category, a smaller difference in the middle age category and an even smaller difference in the ≥60 age category.

To assess this inconsistency, two-way ANOVA additionally provides a *p*-value for the hypothesis

$$H_0 : \mu_{A1} - \mu_{B1} = \mu_{A2} - \mu_{B2} = \mu_{A3} - \mu_{B3}$$

against the general alternative hypothesis that the treatment differences are not all equal.

This null hypothesis says that the treatment difference/effect is consistent. If the *p*-value from this test is significant, then the data support heterogeneity of treatment effect according to age, and we have a significant treatment-by-age (sometimes written as treatment × age) interaction.

Power and sample size calculations (see Chapter 9 on this topic) focus on testing the main effect of treatment and not on the evaluation of the treatment-by-age interaction. As a consequence, this test for interaction will have low power and only pick up marked heterogeneity. To counter this to an extent, we sometimes evaluate the significance of the interaction test at a less strict significance level than is usual, so, for example, using  $p \leq 0.10$  as a guide to reject the null hypothesis of homogeneity rather than when  $p \leq 0.05$ . In discussing the treatment-by-covariate significance test, ICH E9 states:

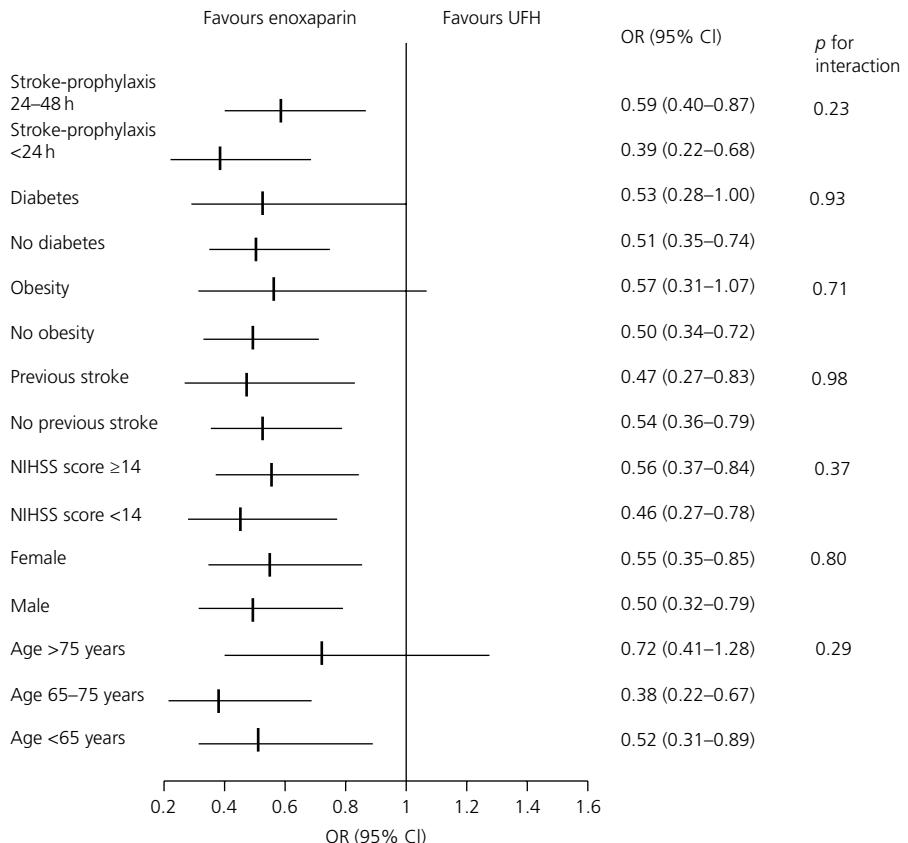
**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'When using such a statistical significance test, it is important to recognise that this generally has low power in a trial designed to detect the main effect of treatment'.*

Therefore, it is important not to rely overly on statistical significance when evaluating interactions and to use clinical judgment about what constitutes an important differential treatment effect numerically. Indeed, the CHMP *Points to Consider* paper, published several years after ICH E9, takes a rather extreme view and cautions against the use of formal significance testing for interactions altogether. Other commentators, however, do view these interaction tests as being of some value.

**CHMP (2015): 'Guideline on adjustment for baseline covariates'**

*'Tests for interactions often lack statistical power and the absence of statistical evidence of an interaction is not evidence that there is no clinically relevant interaction. Conversely, an interaction cannot be considered as relevant on the sole basis of a significant test for interaction. Assessment of interaction terms based on statistical significance tests is therefore of little value'.*



**Figure 5.1** Forest plot for risk for venous thromboembolism in patients with acute ischaemic stroke by patient characteristics for enoxaparin and unfractionated heparin. NIHSS, National Institutes of Health Stroke Scale; OR, odds ratio; UFH, unfractionated. Source: Sherman DG, Albers GW, Bladin C, Fieschi C, Gabbai AA, Kase CS et al. (2007). The efficacy and safety of enoxaparin versus unfractionated heparin for the prevention of venous thromboembolism after acute ischaemic stroke (PREVAIL Study): an open-label randomized comparison. Lancet, 369, 1347–1355. Reproduced with permission from Elsevier.

The ICH guideline also mentions the use of graphical methods, and we will discuss these further in conjunction with subgroup evaluation in Section 10.8. Figure 5.1, taken from Sherman et al. (2007), is the kind of plot they are looking for. This *forest plot* displays treatment differences together with 95% CIs and also p-values for interaction.

#### 5.4.2 Quantitative and qualitative interactions

ICH E9 makes a distinction between quantitative and qualitative interactions. A *quantitative interaction* refers to the situation where the treatment difference is consistently in one direction (e.g. treatment A is always better than treatment B) but there are differences in terms of magnitude. A *qualitative interaction* is where

the treatment difference is in a different direction for some level or levels of the factor (e.g. treatment A is better than treatment B for patients  $<50$ , but treatment B is better than treatment A for patients  $\geq 50$ ). All the interactions seen in Figure 5.1 are quantitative interactions. Had the OR for patients aged  $>75$  years, for example, been  $>1$ , favouring UFH, then we would have had a qualitative interaction for that factor.

If heterogeneity of treatment effect is found, this could possibly undermine the generalisability of the results. For example, with a qualitative interaction, one treatment is performing better on average in one or more subgroups but worse in other subgroups, and it will be difficult to draw a general conclusion of ‘treatment A is a better treatment than treatment B’. Even a quantitative interaction will sometimes give problems in terms of estimating with confidence the magnitude of the treatment effect. We will further discuss the interpretation of these forest plots in Section 10.8 in relation to subgroup testing and multiplicity.

## 5.5 Methods for binary and ordered categorical endpoints

The *Cochran–Mantel–Haenszel (CMH) tests* are a collection of procedures that extend the simple chi-squared tests introduced in Chapter 4 to allow adjustment for baseline factors when dealing with binary, categorical and ordered categorical endpoints. Landis, Heyman and Koch (1978) provide further details.

When adjusting for a baseline factor with binary endpoints, we will have a series of  $2 \times 2$  tables, one for each level of the factor. For categorical and ordered categorical endpoints with  $c$  categories, we will have a series of  $2 \times c$  tables. The CMH test in the first instance provides a single  $p$ -value for the treatment difference.

In terms of summary statistics for binary and ordered categorical endpoints, we usually work with the OR, and when we adjust for a baseline factor, these ORs are averaged over the different levels of the factor to produce a so-called *common OR*. It is also possible to obtain, in a similar way, *average* values for both the reduction in event rates and for relative risk (RR) if these are needed.

The same issues arise with the heterogeneity of the treatment effect as in the preceding text with continuous endpoints, and indeed, the ICH and CHMP comments detailed there are relevant for all endpoint types. For binary endpoints, there is a significance test, the Breslow–Day test (Breslow and Day, 1994), which provides a  $p$ -value for the homogeneity of the treatment effect across levels of the factor. Again, graphical methods are also available and follow the approach seen earlier, plotting estimated ORs for each baseline factor and the subgroups defined by levels of those factors together with their corresponding 95% CIs. Indeed, the example displayed in Figure 5.1 is based on a binary outcome and ORs.

## 5.6 Multi-centre trials

### 5.6.1 Adjusting for centre

As indicated in the ICH E9 guideline, there are two reasons why we conduct multi-centre trials:

- To recruit sufficient numbers of patients within an appropriate timeframe
- To enable the evaluation of the consistency of the treatment effect across a range of centres to provide a basis for generalisability

The first issue is of practical importance; there is probably no other way the required numbers of patients could be recruited. The second issue is very much in line with our discussions earlier in this chapter in relation to the evaluation of the homogeneity of treatment effect according to levels of baseline factors. A multi-centre structure enables us to look at treatment differences in different centres or clusters/groups of centres to assess whether what we are seeing is a consistent effect. Without this consistency, it would be difficult to draw conclusions about the value of the treatment across a broad patient population.

In many cases, where the trial has been set up to recruit from a small number of large centres, the randomisation will be stratified by centre. In this case, the analysis should be adjusted for centre. Following on from that, we can investigate the homogeneity of treatment effect across the different centres as we did for levels of baseline covariates in Section 5.4. Alternatively, some trials may be conducted across a large number of small centres – for example, GP studies. It may be in these cases that the randomisation has not been stratified by centre but rather by groupings of centres defined at the design stage (e.g. by geographical region). Under these circumstances, the groupings form pseudo-centres, and these would be adjusted for in the statistical analysis, with homogeneity being judged according to those same groupings

### 5.6.2 Significant treatment-by-centre interactions

Suppose that in a particular setting there is some evidence of treatment-by-centre interactions. A simple interpretation of the data is then not straightforward.

In the section on multi-centre trials, the ICH E9 guideline makes the following point:

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'If heterogeneity of treatment is found, this should be interpreted with care and vigorous attempts should be made to find an explanation in terms of other features of trial management or subject characteristics . . . In the absence of an explanation, heterogeneity of treatment effect as evidenced, for example, by marked quantitative interactions implies that alternative estimates of the treatment effect may be required, giving different weights to the centres, in order to substantiate the*

*robustness of the estimates of treatment effect. It is even more important to understand the basis of any heterogeneity characterised by marked qualitative interactions, and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted'.*

Work is therefore needed to find an explanation when interactions are seen. That explanation may come from looking, for example, at differential compliance within the centres or different characteristics of the patient sample recruited at the different centres. Inevitably, a treatment-by-centre interaction is simply a surrogate for some *hidden* explanation. It must also be said that the explanation may be *chance!* In a trial with a reasonable number of centres, where treatment A really is a better treatment than treatment B, seeing one treatment reversal (a centre in which treatment B is numerically superior to treatment A) would not be unlikely. Senn (2007, Section 14.2.6) investigates this and shows, under some fairly general conditions, that the probability of seeing at least one reversal goes above 50% with six or more centres, so be cautious with over-interpretation with regard to such interactions.

### **5.6.3 Combining centres**

The ideal situation in many multi-centre trials is to have a small number of large centres (or pre-defined pseudo-centres). This gives the necessary consistency and control yet still allows the evaluation of homogeneity. In practice, however, we do not always end up in this situation, and combining centres at the data analysis stage will sometimes need to be considered since, from a statistical perspective, adjusting for small centres in the analysis is problematic and can lead to unreliable estimates of treatment effect.

#### **CHMP (2015): 'Guideline on adjustment for baseline covariates'**

*'Adjusting for many small centres might be possible but raises analytical problems for which there is no best solution. Analyses either ignoring centres used in the randomisation or adjusting for a large number of small centres might lead to unreliable estimates of the treatment effect and P-values that may be either too large or too small'.*

There are no fixed rules for these combinations, but several points should be noted:

- Combining centres just because they are small or combining centres to produce centres of similar size has no scientific justification.

#### **CHMP (2015) Guideline on adjustment for baseline covariates in clinical trials**

*'Furthermore, pooling small centres to form one centre of size comparable to that of other centres has little or no scientific justification'.*

- Combinations should be based on similarity (by region, by country, by type, etc.). For example, a trial in depression may be run across many centres with at most 10 patients being recruited at each centre; some of the centres will be GP centres, while others will be specialist psychiatric facilities. In this case, combining by centre type (GP or specialist psychiatrist facility) would make sense and would allow the homogeneity of treatment effect between GP centres and specialist psychiatric centres to be investigated.
- Ideally, rules for combining centres should be detailed in the statistical analysis plan.
- Any final decisions regarding combinations should be made at the blind review stage prior to revealing treatment group allocations.

We will discuss the decision-making process for the statistical analysis plan and the blind review in Section 21.3.4.

## CHAPTER 6

# Regression and analysis of covariance

### 6.1 Adjusting for baseline factors

We saw in the previous chapter how to adjust for single factors in treatment comparisons using two-way ANOVA for continuous and score endpoints and the CMH test for binary, categorical and ordinal data. Generally, it can be advantageous to adjust for several factors simultaneously. To a certain extent, this could be done using the methods of Chapter 5 by using combinations of levels of the different factors to define the strata. For example, suppose we wish to adjust for both age ( $<50$  versus  $\geq 50$ ) and sex. We could do this by defining four strata as follows:

Males,  $<50$  years

Males,  $\geq 50$  years

Females,  $<50$  years

Females,  $\geq 50$  years

This can work perfectly well, although it is not easy to use such an analysis as a basis for investigating treatment-by-factor interactions. For example, if the treatment effect was different for males and females but was not influenced by age, this would be difficult to identify within this structure.

Further, as the number of baseline factors increases, this approach becomes a little unwieldy. The method of analysis of covariance (ANCOVA) is a more general methodology that can deal with the increase in complexity caused by wanting to adjust simultaneously for several factors, and that can also give improved ways of exploring interactions. We will develop this methodology later in the chapter, but as a lead into that, we will firstly discuss regression.

### 6.2 Simple linear regression

Regression provides a collection of methods that allow the investigation of dependence, how an outcome variable depends upon something that we measure at baseline CMH.

As an example, suppose in an oncology study we wish to explore whether time to disease recurrence from entry (months) into the study depends upon the size of the primary tumour measured at baseline (diameter in cm). The scatter plot in Figure 6.1 represents (artificial) data on 20 subjects.

A visual inspection of the plot would suggest that there is some dependence, but in many cases this will not be quite so clear-cut. We explore the dependence from a statistical point of view by fitting a straight line to the data.

The equation of a straight line is

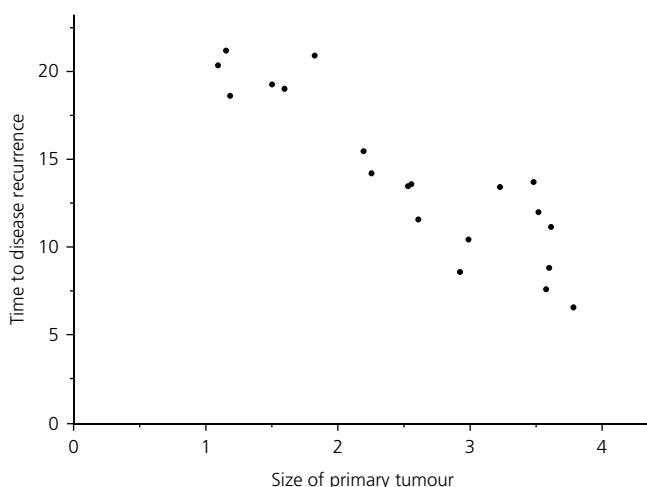
$$y = a + b x$$

where  $a$  is the intercept (the value of  $y$  where the line crosses the  $y$ -axis) and  $b$  is the slope (the amount by which  $y$  increases when  $x$  increases by one unit).

The value of  $b$  is of the greatest importance. If  $b$  is positive, there is a positive dependence; as  $x$  increases, so does  $y$ . If  $b$  is negative, there is a negative dependence; as  $x$  increases,  $y$  decreases. Finally, if  $b = 0$ , there is no dependence; as  $x$  increases, nothing happens to  $y$ .

The method we use to fit the straight line so that it describes the data in the best possible way is called *least squares*. This involves measuring the vertical distance of each point from a line placed on the plot, as shown in Figure 6.2, squaring each of those distances (this, among other things, gets rid of the sign) and choosing the line that makes the average of these squared distances as small as possible. In our example, this *least squares regression line* has the equation

$$y = 25.5 - 4.48x$$



**Figure 6.1** Scatter plot for dependence of time to disease recurrence on size of primary tumour

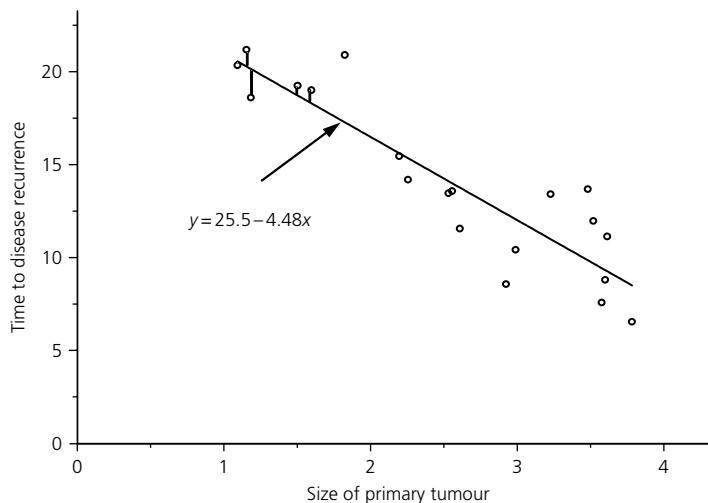


Figure 6.2 Least-squares regression line

The value of the slope is  $-4.48$ , and this estimates the average increase in time to disease recurrence as the tumour size at baseline increases by 1 cm. The primary question of interest here is, ‘does time to disease recurrence depend upon tumour size at baseline?’ To address this question, we, as usual, formulate null and alternative hypotheses

$$H_0 : b = 0 \quad H_1 : b \neq 0$$

and construct an appropriate test. This involves the signal, which is the estimate of  $b$  from the data, and a measure for the noise, which is equal to the standard deviation of the vertical distances of the points from the fitted line. The standard error (se) of the estimate of  $b$  involves this standard deviation, and there is a formula for its calculation. Finally, the test statistic is structured as before, the signal is divided by the associated standard error, and the null distribution is  $t_{n-2}$ , where  $n$  is the number of subjects. A significant  $p$ -value ( $p \leq 0.05$ ) from this test tells us that  $b$  is significantly different from zero, indicating dependence. A non-significant  $p$ -value tells us that there is insufficient evidence to conclude dependence. For these data,  $p \ll 0.001$ , a very highly significant dependence of time to disease recurrence on the size of the primary tumour at baseline. The slope of  $-4.48$  indicates that on average, for each 1 cm increase in the diameter of the primary tumour, there is a 4.48-month decrease in the time to disease recurrence. Finally, it is straightforward to construct a confidence interval around the estimated slope.

This technique of *simple linear regression* therefore provides a way to evaluate whether a particular baseline variable is predictive of outcome. We use the term *dependent variable* as a label for the  $y$  variable and *independent variable* as a label for the  $x$  variable. We will extend these ideas in the next section to evaluate several baseline variables/factors simultaneously.

### 6.3 Multiple regression

In Section 6.2, we saw how to study the dependence of an outcome variable on a variable measured at baseline. It could well be that there are several baseline variables that predict outcome, and in this section, we will see how to incorporate these variables simultaneously through a methodology termed *multiple (linear) regression*.

Taking up the example from the previous section, it may be that time to disease progression depends potentially not just on size of primary tumour but also on age and sex, and we would like to explore the nature of that dependence. Clearly, size of primary tumour and age are both numerical, while sex is not; we incorporate qualitative variables (termed *factors*) of this kind by using so-called *indicator variables*. Generally, these take the values zero and one according to the *value* of the variable. It also does not matter which way round they are coded. Switching the codes would simply result in the coefficient of that variable in the equation changing sign:

Let  $x_1$  = size of primary tumour

$x_2$  = age

$x_3$  = 0 male

1 female

The extension of simple linear regression to deal with multiple baseline variables is somewhat difficult to visualise, but algebraically it is just a matter of adding terms to the equation:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3$$

Now,  $b_1$  measures the effect of size on time to disease recurrence, while the coefficients  $b_2$  and  $b_3$  measure the effects of age and sex, respectively. More specifically,  $b_1$  and  $b_2$  are the changes in the average time to disease recurrence as size and age each increase by one unit, respectively, while  $b_3$  measures the sex effect, the average time to disease recurrence for females minus that for males. Each of these quantities, which we can estimate from the data again using the method of least squares, represents the contribution of each of the variables separately in the presence of the other variables.

The questions of interest revolve around the values of the  $b$  coefficients. Do age and size of primary tumour predict time to disease recurrence? Is there a sex effect? We address these questions by formulating hypotheses:

$$H_{01} : b_1 = 0 \quad H_{11} : b_1 \neq 0$$

$$H_{02} : b_2 = 0 \quad H_{12} : b_2 \neq 0$$

$$H_{03} : b_3 = 0 \quad H_{13} : b_3 \neq 0$$

Each of these is evaluated by dividing the estimate of the corresponding  $b$  value by its standard error and comparing it to the  $t_{n-4}$  distribution (the degrees of freedom for the appropriate t shape is the number of subjects minus [1 + the number of  $x$  variables in the model]). Note that multiple regression with just a single variable reduces to simple linear regression.

Suppose that the fitted equation turns out to be

$$y = 23.8 - 3.51 x_1 - 0.174 x_2 + 0.47 x_3$$

Therefore, for a 1 cm increase in the tumour size, time to disease recurrence reduces by on average an estimated 3.51 months; for each additional year in age, there is on average an estimated 0.174-month reduction in time to disease recurrence; and finally, the time to disease recurrence is estimated to be slightly higher for females by on average 0.47 months.

The  $p$ -values associated with each term in this model were

$$H_{01} : b_1 = 0 \quad p = 0.007$$

$$H_{02} : b_2 = 0 \quad p = 0.02$$

$$H_{03} : b_3 = 0 \quad p = 0.48$$

This suggests that size of primary tumour and age are predictors of time to disease recurrence, while there is no evidence that sex has an impact.

Note that this approach is not the same as conducting three linear regression analyses on the baseline variables separately. In fact, such an approach could give a confused picture if the baseline variables being considered were correlated. For example, suppose that age and size of primary tumour are correlated, with older patients tending to present with larger tumours. Also, suppose size of primary tumour is the driver in terms of time to disease recurrence and that age has no effect additional to that. The separate linear regressions would indicate that both size of primary tumour and age predict outcome; age would be identified as a predictor of outcome only because of its correlation with tumour size. Multiple regression, however, would give the correct interpretation. Size of primary tumour would be seen as a predictor of outcome, but once that effect is accounted for, age would add nothing to this prediction.

There is sometimes discussion about whether to use categories for a continuous variable – for example, having a binary age variable taking the value 0 for patients aged  $<50$  years and taking the value 1 for patients aged  $\geq 50$  – or whether to include the continuous variable itself in the modelling. Categorisation will generally waste information and is not recommended; it is much better to retain the variable in its continuous form. See Royston, Altman and Sauerbrei (2006) for a detailed discussion on this point. Similar comments will apply later in this chapter when looking at covariates in ANCOVA.

With a large number of potential baseline variables, it may be of interest to select those variables that are impacting on outcome, and methods (*stepwise*

*regression*) are available for doing this. Using this methodology, the unimportant variables are eliminated, leaving a final equation containing just the important predictive factors. This methodology is often used to construct *prognostic indices*.

Note that what we are doing here for both linear regression and multiple regression is fitting a (*statistical*) *model* to the data and using that fitted model to draw inferences about the predictors of outcome.

## 6.4 Logistic regression for binary endpoints

Multiple regression as presented so far is for continuous and score outcome variables  $y$ . For binary, categorical and ordered categorical outcomes, the corresponding technique is called *logistic regression*. Suppose that in our earlier example we defined success to be *disease-free for five years*; we might be interested in identifying those variables/factors at baseline that were predictive of the probability of success.

Define  $y$  now to take the value one for a treatment success and zero for a treatment failure. For mathematical reasons, rather than modelling  $y$  as we did for continuous and score outcome variables, we now model the probability that  $y = 1$ , written  $\text{pr}(y = 1)$ .

This probability, by definition, will lie between zero and one, so to avoid numerical problems, we do not model  $\text{pr}(y = 1)$  directly but a transformation of  $\text{pr}(y = 1)$ , the so-called *logit* or logistic transform:

$$\ln \left\{ \frac{\text{pr}(y = 1)}{1 - \text{pr}(y = 1)} \right\} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Computer packages can fit these models, provide estimates of the values of the  $b$  coefficients together with standard errors and give  $p$ -values associated with the hypothesis tests of interest. The null hypotheses will be exactly as  $H_{01}$ ,  $H_{02}$  and  $H_{03}$  in Section 6.3.

Methods of stepwise regression are also available for the identification of a subset of the baseline variables/factors that are predictive of outcome. A good example of the application of logistic regression identifying the clinical signs that predict severe illness in neonates in countries of low or middle income is provided by The Young Infants Clinical Signs Study Group (2008). The logistic regression model extends both to categorical endpoints using the *polychotomous logistic model* and to ordered categorical endpoints using the *ordinal logistic model*. For an example of the former, see Marshall and Chisholm (1985) for an application in diagnosis.

### 6.4.1 Negative binomial regression for count endpoints

We briefly mention in Section 4.8 the use of the negative binomial model for the analysis of count endpoints. This model can be used to evaluate the dependence of a count outcome on baseline covariates/factors. Suppose that a particular

subject is observed for  $T$  units of time. Typically, a unit of time may be one year in the case of number of moderate/severe exacerbations in asthma, or one day in the case of incontinence episodes. Note that the period of observation could differ across the subjects. The model is built around the mean number of events per unit time (*event rate*),  $\mu$ , with

$$\ln \mu = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ln T$$

The term  $\ln T$  is called the *offset term*. If all subjects were to be observed for the same length of time, this term could be dropped from the model. Its inclusion in the model allows for the period of observation to be different for different subjects. Observing subjects for different periods of time is not usually done by design but will be the case, for example, if some subjects drop out or if the model is being used for an interim analysis.

## 6.5 Analysis of covariance for continuous outcomes

### 6.5.1 Main effect of treatment

We will return to the example where we have just a single baseline variable, size of primary tumour, predicting the outcome time to disease recurrence; but now, in addition, we have randomised the patients to one of two treatment groups: test treatment and placebo.

Figure 6.3 displays a possible pattern for the results. Note firstly that as before, there appears to be a dependence of time to disease recurrence on the size of the

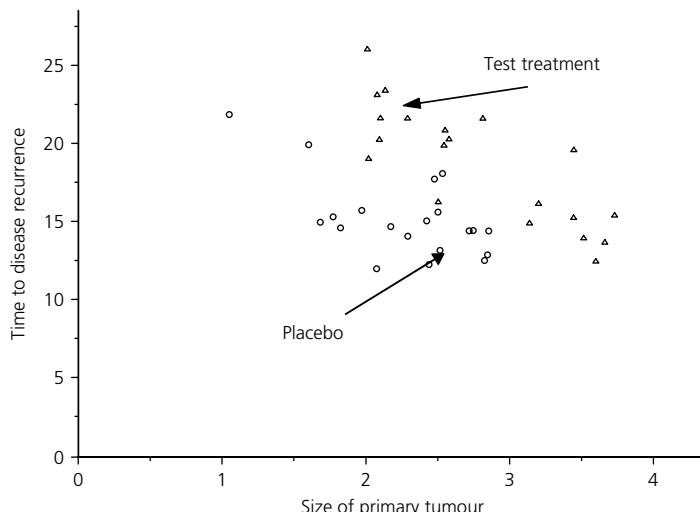


Figure 6.3 Scatter plot for two treatment groups

primary tumour at baseline. In addition, it also seems that the patients receiving the test treatment have longer times to disease recurrence compared to those receiving the control treatment, irrespective of the size of the primary tumour.

A formal comparison of the two treatments could be based on the unpaired t-test, comparing the mean time to disease recurrence in the test treatment group with the mean time to disease recurrence in the control group. While this is a valid test, it may not be particularly sensitive. The separation between the two groups is clear, but if we now simply read off the times to disease recurrence on the  $y$ -axis, we will see considerable overlap between the treatment groups; we will have lost some sensitivity by ignoring the size of the primary tumour variable.

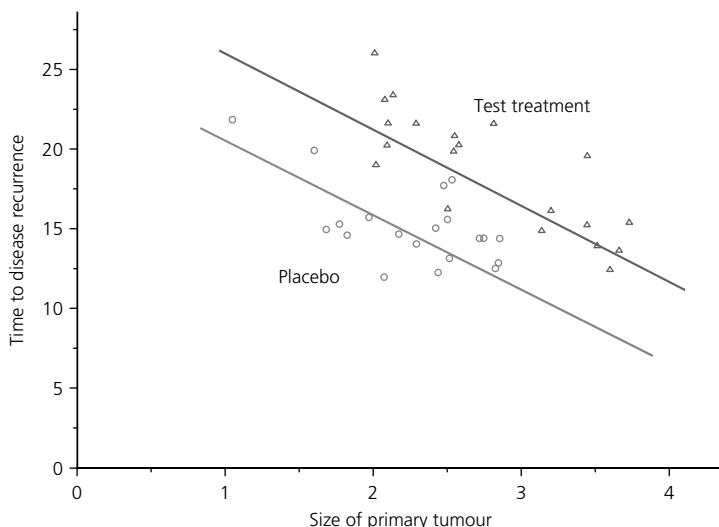
Consider an alternative approach, fitting simple linear regression lines to the data from these two groups of patients. The equations of these lines can be written as

$$y = a_1 + bx \quad \text{test treatment}$$

$$y = a_2 + bx \quad \text{placebo}$$

Figure 6.4 shows these lines fitted to the data. Note that we have constrained the slopes of these lines to be the same; we will return to this point later. The intercepts,  $a_1$  and  $a_2$ , are the points where the lines cross the  $y$ -axis.

Had the treatments been equally effective, the points for patients in the placebo group would not have been, in general, below the points for patients in the test treatment group; the lines would have been coincidental with  $a_1 = a_2$ . Indeed,



**Figure 6.4** Scatter plot and fitted lines for two treatment groups

the larger the treatment difference, the bigger the difference between the two intercepts,  $a_1$  and  $a_2$ . Our main interest is to compare the treatments, and within this framework, we compare the values of  $a_1$  and  $a_2$  through the null hypothesis  $H_0: a_1 = a_2$  and the alternative hypothesis  $H_1: a_1 \neq a_2$ . The signal is provided by the estimate of  $a_1 - a_2$ , and there is an associated standard error for that estimate; we compare the signal/se ratio to  $t_{n-3}$  to give the  $p$ -value.

The quantity  $a_1 - a_2$  is the vertical distance between the lines and represents the (adjusted) difference in the mean time to disease recurrence in the test treatment group minus the mean time to disease recurrence in the control group, the treatment effect. It is also straightforward to obtain a confidence interval around this *adjusted treatment effect* to capture the true difference.

This technique is called *analysis of covariance*, and size of the primary tumour at baseline is the covariate. Taking account of the covariate here has led to a much more powerful analysis than that provided by the simple unpaired t-test. Of course, the main reason we see an improvement in sensitivity is that the covariate is such a strong predictor of outcome. These improvements will not be quite so great with weaker predictors.

It is also possible to include more than one covariate in the analysis in cases where several are thought to be influential for the outcome by simply adding terms to the earlier equations as with multiple regression. That is,

$$y = a_1 + b_1x_1 + b_2x_2 + b_3x_3 \quad \text{test treatment}$$

$$y = a_2 + b_1x_1 + b_2x_2 + b_3x_3 \quad \text{placebo}$$

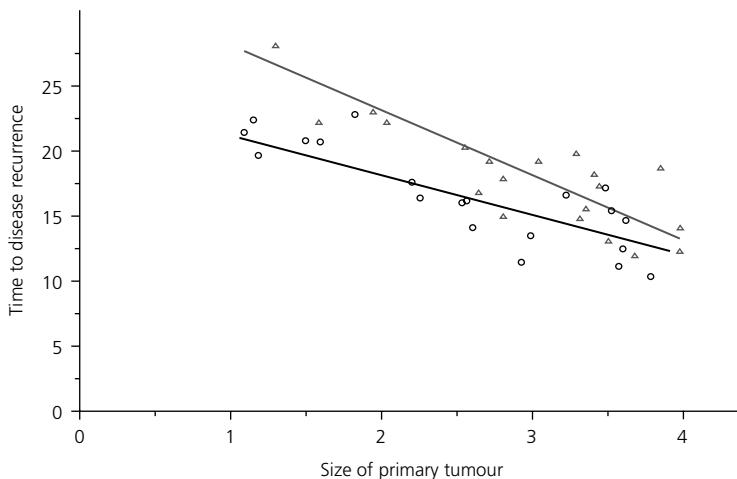
Again, we test the hypothesis  $H_0: a_1 = a_2$ . The estimate of  $a_1 - a_2$  is used as the signal, and the ratio of this estimate divided by its standard error (the test statistic) is compared to  $t_{n-q}$ , where  $n$  is the number of subjects and  $q$  is the number of covariates +2 to give the  $p$ -value.

We call these equations (or models) *main effects models*. In the next subsection, we will be adding treatment-by-covariate interaction terms to the main effects (of treatment and the covariates).

### 6.5.2 Treatment-by-covariate interactions

Returning to the case with a single covariate, we have assumed that the two lines are parallel. This may not be the case. Figure 6.5 shows a situation where it would not be appropriate to assume parallel lines. Here, patients presenting with small tumours do much better in the test treatment group compared to the placebo group, but there are virtually no differences between the treatments for those patients presenting with large tumours.

We have previously discussed treatment-by-factor interactions in Chapter 5, where the treatment effect is seen to be different for the different levels of the factor. More generally, we can consider treatment-by-covariate interactions, where the treatment effect depends on the value of the covariate. We can



**Figure 6.5** Scatter plot showing treatment by covariate interaction

investigate this by considering a model where we allow the lines to have different slopes as well as different intercepts:

$$y = a_1 + b_1 x \quad \text{test treatment}$$

$$y = a_2 + b_2 x \quad \text{placebo}$$

This now gives us the opportunity to assess whether there are any interactions by fitting these two lines to the data and formulating a test of the hypothesis  $H_0: b_1 = b_2$  against the alternative  $H_1: b_1 \neq b_2$ . A significant  $p$ -value – and as with treatment-by-factor interactions, we would be looking at  $p \leq 0.10$  as a guide for statistical significance – would indicate the presence of an interaction. In this case, talking in terms of the *treatment effect* makes little sense as a consistent treatment effect is not present. A non-significant  $p$ -value would suggest that it is safe to assume that the treatment effect is constant across the levels of the covariate, there is no evidence to the contrary, and the model with a common slope  $b$  provides an adequate description of the data.

If a significant treatment-by-covariate interaction is found, it could be useful to divide the patients into subgroups in terms of the size of the primary tumour – say, small, medium and large – and look at the treatment difference within those subgroups to try to better understand the nature of the interaction.

In the presence of several covariates, there will be a series of  $b$  coefficients, two for each covariate, as follows:

$$y = a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 \quad \text{test treatment}$$

$$y = a_2 + b_{21}x_1 + b_{22}x_2 + b_{23}x_3 \quad \text{placebo}$$

Assessing the treatment-by-covariate interactions is then based on comparing the  $bs$ :  $b_{11}$  with  $b_{21}$ ,  $b_{12}$  with  $b_{22}$  and  $b_{13}$  with  $b_{23}$  in separate hypothesis tests.

### 6.5.3 A single model

These models can be written in a more precise form by defining a binary indicator to denote treatment. Let  $z = 0$  for patients randomised to the placebo group, and let  $z = 1$  for patients randomised to the test treatment group. The model with a single covariate and assuming a common slope can then be written as

$$y = a + cz + bx$$

Now, when  $z = 0$  (placebo group),  $y = a + bx$ ; and when  $z = 1$  (test treatment group),  $y = (a + c) + bx$ .

The  $b$  in this model is as previously, but now  $a = a_1$  and  $c = a_2 - a_1$ . We refer to this in mathematics as a re-parameterisation; don't be put off by it! The hypothesis  $H_0: a_1 = a_2$  is replaced by the hypothesis  $H_1: c = 0$ . None of this changes the analysis in any sense; it is just a more convenient way to write down the model and will be useful later when we bring together the ideas of ANOVA and ANCOVA.

Although conceptually it is useful to think of fitting straight lines to each of the treatment groups separately, this is not how it is done in practice. We simply fit this single equation to the data. This also allows us to use the information on the noise from the two groups combined to obtain standard errors. Finally, we build in the interaction terms by adding on to this common equation a *cross-product term*,  $z \times x$ , and using another re-parameterisation:

$$y = a + cz + bx + dzx$$

So when  $z = 0$  (placebo group),  $y = a + bx$ ; and when  $z = 1$  (test treatment group),  $y = (a + c) + (b + d)x$ .

$b_1$  in the previous model in Section 6.5.2 is now  $b$  in this common model, while  $d = b_2 - b_1$ . Assessing the presence of a treatment-by-covariate interaction (common slope) is then done through the hypothesis  $d = 0$ .

For several covariates, we simply introduce a cross-product term for each covariate with corresponding coefficients  $d_1$ ,  $d_2$  and  $d_3$  to investigate interactions. The presence of treatment-by-covariate interactions can then be investigated through these coefficients.

In general, the focus of an analysis to estimate treatment effects is the *main effects model*, which, as we have seen, provides treatment effects adjusted for baseline imbalances. Interaction terms are investigated as an add-on to explore the homogeneity of the treatment effect and whether there is evidence of a differential effect according to the values of one or more covariates.

### 6.5.4 Connection with adjusted analyses

The ANCOVA main effects model is a form of adjusted analysis; we are providing an adjusted treatment effect in the presence of covariates. This is very much like the adjusted analysis we presented in the previous chapter. For the single covariate example, had we defined strata according to the size of the primary tumour –

say, small, medium and large – and then undertaken two-way ANOVA to compare the treatments, we would have got very similar results to those seen here through ANCOVA. This applies to the *p*-values for the assessment of treatment difference, the estimated (adjusted) treatment difference and the associated confidence intervals. The *p*-values for studying the homogeneity of the treatment effect by adding in the interaction term will give similar results to those using the methods of Section 5.4.1 looking for treatment-by-factor interactions.

### 6.5.5 Advantages of ANCOVA

ANCOVA offers several advantages over simple two-treatment-group comparisons:

- Improvements in efficiency (smaller standard errors, narrower confidence intervals, increased power).
- Correction for baseline imbalances. Randomisation will, on average, produce groups that are comparable in terms of baseline characteristics. It is inevitable, however, that small differences will still exist, and if these are differences in important prognostic factors, they could have an impact on the treatment comparisons. By chance, there will also be occasions when a substantial imbalance exists. In fact, in Figures 6.3 and 6.4, we have such imbalances, and these imbalances could cause bias in our evaluation of a treatment effect. Here, purely by chance, we have ended up with a predominance of patients with large tumours in the test treatment group and a predominance of patients with small and medium tumours in the control treatment group. A simple unpaired t-test would possibly fail to detect a treatment difference as a result of this baseline imbalance. ANCOVA helps correct for those baseline imbalances; once the two regression lines are estimated, ANCOVA *ignores* the data (and therefore the imbalance) and simply works with the distance between the lines for the treatment effect.
- Allows assessment of prognostic factors. Fitting the ANCOVA model provides coefficients for the covariates, and although this is not the primary focus of the analysis, these coefficients and associated confidence intervals provide information on the effect of the baseline covariates on outcome. This information can be useful for trial planning, for example choosing factors on which to stratify the randomisation for the future.
- Provides a convenient framework for the evaluation of treatment-by-covariate interactions; in rare cases, such interactions are anticipated, but in most cases, such analyses are exploratory.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'In most parts, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall.'*

The adjusted analyses discussed in Chapter 5 also share some of these advantages and provide improvements in efficiency, can also account for baseline imbalances, and allow the evaluation of the homogeneity of the treatment effect. On this final point, however, and as discussed in Section 6.1, simple adjusted analyses looking at combinations of factors are less able to identify the nature of those interactions. With ANCOVA, it is possible to say which specific covariates are causing such interactions. A further point to note here, and as mentioned earlier, is that simple adjusted analyses become more difficult as the number of covariates increases.

Should treatment-by-covariate interactions be found, either through a test of homogeneity in an adjusted analysis or through ANCOVA, analysis usually proceeds by looking at treatment differences within subgroups. Forest plots (Section 10.8) of treatment effects with associated confidence intervals within these subgroups are useful in this regard.

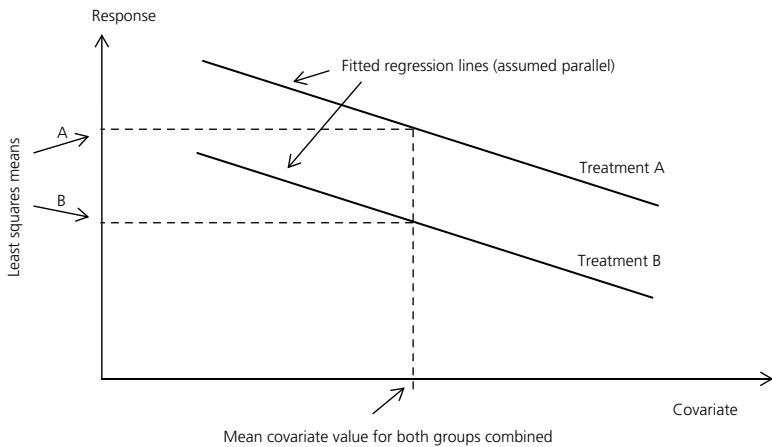
One disadvantage of ANCOVA is that the modelling does involve several assumptions, and if those assumptions are not valid, then the approach could mislead. For example, it is assumed (usually) that the covariates affect outcome in a linear way; there is invariably too little information in the data to be able to assess this assumption in any effective way. In contrast, with an adjusted analysis, assumptions about the way in which covariates affect outcome are not made, and in that sense, it can be seen as a more general approach. In some regulatory circles, simple adjusted analyses are preferred to ANCOVA for these reasons.

### 6.5.6 Least squares means

The method we use to fit ANCOVA models is again least squares, measuring the vertical distances of the points from the regression lines and choosing values for the parameters in the model that make the average of the resulting squared distances as small as possible.

We discussed in the previous section how ANCOVA can correct for baseline imbalances in the covariate; suppose again that those imbalances were present. The mean values for time to disease recurrence calculated from the data, and the numerical difference between those means, would give us a misleading impression of the true treatment benefit; we somehow need to correct those mean values for the baseline imbalance in tumour size. A straightforward way to do this would be to take the average tumour size for the two treatment groups combined and then use the fitted regression equations with the common slope to predict the time to disease recurrence for this *average* patient in each of the two treatment groups. Figure 6.6 shows this construction diagrammatically.

We refer to these *fitted* values as the *least squares means*; the mathematics here is very much in line with what we did in Section 5.3 when we were looking at simple methods for adjusting the analysis. They provide a better summary of the average efficacy of the two treatments in an absolute sense, but maybe more importantly, their difference provides a better estimate of the relative efficacy



**Figure 6.6** Calculating least squares means

of the two treatments. The difference between these least squares means is also the vertical distance between the two fitted lines and is numerically equal to  $a_1 - a_2$ , the difference in the intercepts. We discussed the use of this adjusted treatment difference in Section 6.5.1 and argued that it provided an appropriate measure of the true treatment difference.

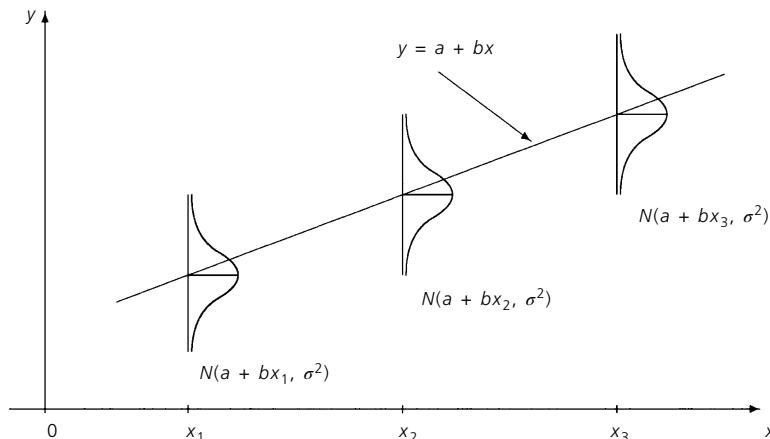
### 6.5.7 Random element

The equations that we have been fitting to data provide *statistical models* that allow the prediction of an outcome based on baseline characteristics of the subject and the treatment group to which the subject has been randomised. These models predict mean outcomes: what is happening on average. Of course, not all subjects with the same baseline characteristics in the same treatment group will respond with outcomes that are the same, and a further key element of these models is a *random element* that allows subjects to have individual values that vary around the mean value. The usual assumption we make regarding this random element is that it is a normal distribution of values with a certain variance (*residual variance*) that can be estimated from the data. Figure 6.7 displays this final aspect for the linear regression model. This same idea applies to the more complex models; there is always a random element that accompanies the model equations.

## 6.6 Other endpoint types

### 6.6.1 Binary endpoints and extensions

Analysis of covariance for a binary endpoint is based on logistic regression. With continuous endpoints, although we developed the concepts by looking at separate lines for the different treatment groups, we ended up writing down a single



**Figure 6.7** Normal distribution assumption for the distribution around the mean predicted by the model.

equation. This will be the start point with our discussion on binary, categorical and ordered categorical endpoints and logistic regression. We will focus this discussion on the most common setting, the binary case; extensions for ordered categorical endpoints are straightforward.

Again, let  $z = 0$  for patients in the control group and  $z = 1$  for patients in the test treatment group and assume that we have several covariates, say,  $x_1$ ,  $x_2$  and  $x_3$ . The main effects model looks at the dependence of the probability that the outcome  $y = 1$ , where, as previously,  $y = 1$  denotes treatment success and  $y = 0$  denotes treatment failure, on treatment and the covariates:

$$\ln \left\{ \frac{\text{pr}(y=1)}{[1-\text{pr}(y=1)]} \right\} = a + cz + b_1x_1 + b_2x_2 + b_3x_3$$

The coefficient  $c$  measures the impact that treatment has on  $\text{pr}(y=1)$ . If  $c = 0$ , then  $\text{pr}(y=1)$ , the probability of treatment success, is unaffected by which treatment group the patient is in; there is no treatment effect. Having fitted this model to the data and, in particular, obtained an estimate of  $c$  and its standard error, we can test the hypothesis  $H_0: c = 0$  in the usual way through the signal/se ratio.

The quantity  $c$  is very closely related to the odds ratio (OR);  $c$  is the log of the OR, adjusted for the covariates. The antilog of  $c$  (given by  $e^c$ ) gives the adjusted OR. Confidence intervals in relation to this OR can be constructed initially by obtaining a confidence interval for  $c$  itself and then taking the antilog of the lower and upper confidence limits for  $c$ .

**Example 6.1** Effect of betamethasone on incidence of neonatal respiratory distress

This randomised trial (Stutchfield et al., 2005) investigated the effect of betamethasone on the incidence of neonatal respiratory distress after elective caesarean section. In the ITT analysis, of the 492 women randomised to the active treatment, 11 babies were subsequently admitted to the special baby unit with respiratory distress compared to 24 babies out of 495 women randomised to the control group.

The OR for the binary outcome (baby admitted to the special baby unit for respiratory distress) is then 11/481 divided by 24/471, giving a value 0.449. The chi-square test comparing the treatments gave  $p = 0.027$ .

A logistic regression analysis was undertaken with

$$\begin{aligned} z &= 0 \text{ if mother randomised to control} \\ z &= 1 \text{ if mother randomised to betamethasone} \end{aligned}$$

The analysis was adjusted for gestational age and involved two indicator variables; the standard gestational age was 39 weeks and

$$\begin{aligned} x_1 &= 1 \text{ if gest. age} = 37 \text{ weeks} \\ &\quad 0 \text{ if otherwise} \\ x_2 &= 1 \text{ if gest. age} = 38 \text{ weeks} \\ &\quad 0 \text{ if otherwise} \end{aligned}$$

The coefficient of the treatment indicator  $z$  was  $-0.840$ , giving an (adjusted) OR for the treatment effect of  $0.432 (e^{-0.840} = 0.432)$ . The coefficient  $b_1$  of  $x_1$  was  $2.139$ , and the coefficient  $b_2$  of  $x_2$  was  $1.472$ . So, a value of  $x_1 = 1$  (gestational age = 37 weeks) gives an increase of  $2.139$  on the log odds scale or, equivalently, an increase of  $8.5 (e^{2.139} = 8.5)$  on the odds of being admitted to the special baby unit compared to the standard 39 weeks' gestational age. For 38 weeks' gestational age, the increase in the odds of being admitted is  $4.4 (e^{1.472} = 4.4)$  compared to a gestational age of 39 weeks.

We can also investigate the presence of treatment-by-covariate interactions by adding cross-product terms:

$$\ln \left\{ \frac{\text{pr}(y=1)}{[1 - \text{pr}(y=1)]} \right\} = a + cz + b_1x_1 + b_2x_2 + d_1zx_1 + d_2zx_2$$

Questions relating to those interaction terms are addressed through the  $d$  coefficients as before for continuous/score endpoints. In this example, looking for treatment-by-covariate interactions would be asking whether the treatment benefit, in terms of a reduction in the likelihood of the baby suffering respiratory distress, was the same for babies delivered at 37, 38 and 39 weeks.

Logistic regression offers similar advantages as ANCOVA for continuous/score endpoints: correcting for baseline imbalances, allowing the evaluation of the effects of the covariates, and providing a convenient framework for the

identification of treatment-by-covariate interactions. When we discussed models for continuous and score endpoints earlier in this chapter, we introduced the idea of a random element to the model. This does not work in quite the same way for binary and categorical endpoints where the random element is, in a certain sense, built into the logistic model as it stands. A subject assigned to a particular treatment group will have a probability of success/response provided by the model, which will depend additionally on baseline characteristics. Let's suppose that for a particular subject assigned to the active group, this is equal to 0.75. This tells us that 75% of those kinds of subjects in that treatment group will record a response, while 25% will record non-response, and there is the random element. It is like tossing a coin that has a 75% probability of coming down heads (success/response) and a 25% probability of coming down tails (failure/non-response).

### 6.6.2 Count endpoints

The negative binomial model for the analysis of a count endpoint is easily extended to incorporate covariates, as we did with the logistic model for binary outcomes. The model is expressed in terms of  $\ln \mu$ , where  $\mu$  is the mean event rate.

The right-hand side of the equations will include the treatment indicator, covariates, and possibly treatment  $\times$  covariate interactions if these are being formally investigated, plus the addition of an offset term  $\ln T$ , where  $T$  is the period of observation for each subject.

In the main effects model,  $e^{\epsilon}$  is the ratio of the adjusted event rates (adjusted rate ratio), adjusting for imbalances across the treatment groups in the baseline covariates.

**Example 6.2** Solifenacin and mirabegron in patients with overactive bladder syndrome (OAB)

Martina et al. (2014) report on a modelling exercise for the negative binomial model applied to the data from two sets of studies in OAB, from the clinical development of solifenacin and mirabegron. Patients in the four solifenacin studies were randomised to receive once-daily doses of 5 mg solifenacin (two studies), 10 mg solifenacin (four studies) or placebo (four studies) or 2 mg tolterodine (two studies) twice daily for 12 weeks. Patients in the three mirabegron studies were randomized to receive placebo (three studies), 25 mg mirabegron (one study), 50 mg mirabegron (three studies) or 100 mg mirabegron (two studies) or tolterodine (one study) once daily for 12 weeks. The endpoint of interest was the number of incontinence episodes recorded in a three-day diary at the end of the treatment period, and an offset term was included to account for patients who had missing diary days ( $T = 1, 2$  or  $3$  depending on the number of available diary days). Covariates included in the models were sex, age, study and the mean number of incontinence episodes per day at baseline. Four sets of models were fitted to the data based on solifenacin (5 mg) versus placebo, solifenacin (10 mg) versus placebo, mirabegron (25 mg) versus placebo and mirabegron (50 mg) versus placebo. Table 6.1 provides the rate ratios and 95% confidence intervals for the treatment comparisons of interest.

**Table 6.1** Negative binomial models for solifenacin and mirabegron treatment effects

Development programme	Treatment	n	Adjusted rate ratio (95% CI)	p-value
Solifenacin	Placebo	781		
	5 mg	314	0.57 (0.46, 0.72)	<0.001
	10 mg	778	0.59 (0.50, 0.69)	0.001
Mirabegron	Placebo	878		
	25 mg	254	0.70 (0.55, 0.90)	0.004
	50 mg	862	0.74 (0.64, 0.85)	0.001

CI: confidence interval

The adjusted rate ratios were 0.57 and 0.59 for solifenacin 5 mg and 10 mg versus placebo, respectively. The adjusted rate ratios were 0.70 and 0.74 for mirabegron 25 mg and 50 mg versus placebo. So, for example, there was a 43% reduction in the rate of incontinence episodes for solifenacin (5 mg) versus placebo with a 95% CI (28% to 54%). All treatment differences versus placebo were statistically significant.

## 6.7 Mixed models

In the analyses and modelling that we have considered so far, each subject has provided just one observation on outcome. Sometimes, for a variety of different reasons, we are interested in including several observations on outcome for each subject. In a 12-week placebo-controlled trial in Parkinson's disease, for example, where the primary outcome is the amount of daily time spent in the so-called ON state without troublesome dyskinesia (good-quality ON time), suppose measurements on the outcome variable are recorded in patient diaries at baseline, 2 weeks, 4 weeks, 8 weeks and 12 weeks. The models we have discussed so far can be extended to include such *repeated measures* in a single model. These models go under the heading of *mixed models for repeated measures (MMRM)* and model the dependence of outcome at each of these timepoints, expressed as change from baseline, on treatment and other factors.

These models can be used to compare treatments at specific times during the treatment period, with primary interest at 12 weeks. The term *mixed model* comes from the fact that two sources of variation need to be accounted for: *between-subject variation* (subject-to-subject variation) and *within-subject variation* (visit-to-visit variation for the same subject). In the example given, there would be several terms in the model:

- A series of three binary indicators for visit (weeks 2, 4, 8 and 12) to allow outcome at each visit to be different.
- Baseline amount of good quality ON time to account for any regression towards the mean (see the definition and discussion in Section 6.8).

- Stratification factors, and possibly additional covariates thought to be predictive of outcome.
- Treatment indicator ( $z = 0/1$ ) and treatment-by-visit interaction terms to allow the treatment effect to be different for each visit.

The primary focus for this analysis is the treatment effect at week 12, and MMRM will be the basis of the primary comparison in the statistical analysis. Treatment comparisons at other time points could also be of interest, and this model would allow such comparisons to be undertaken.

This modelling framework has several other advantages, and we will see in Chapter 7 how the model is used to account for missing data under certain assumptions about the nature of those missing data. In the current example, suppose that a subject in the trial provides baseline, week 2 and week 4 values for the amount of good-quality ON time but withdraws from the study following the week 4 visit and thus does not provide outcomes at 8 weeks and 12 weeks. The data for this subject up to and including week 4 are used in estimating the coefficients in the model, the noise, and contribute to the treatment comparisons.

In general, we use the MMRM only when we have continuous and score endpoints. There are extensions of the logistic regression models that apply to binary and ordered categorical outcomes, but repeated measures for such endpoint types are rarely encountered in practice. Mixed models of this kind are not relevant for a time-to-event endpoint due to the nature of the endpoint.

## 6.8 Regulatory aspects of the use of covariates

ICH E9 and the CHMP (2015) ‘Guideline on adjustment for baseline covariates in clinical trials’ make several useful points on baseline covariates in addition to those already set down in Chapter 5:

- Pre-planning. It is important that the covariates to be included in the modelling are decided in advance.

### **CHMP (2015): ‘Guideline on adjustment for baseline covariates in clinical trials’**

*‘Covariates to be included in the analysis must be pre-specified in the protocol. When a confirmatory (typically phase III) trial starts, the important covariates should have already been identified through previous trials and other available evidence. However, if the state of knowledge changes between the writing of the protocol and the completion of the study it may be appropriate to re-consider and update the description of the analysis in the statistical analysis plan prior to any unblinding’.*

If new knowledge becomes available regarding important covariates after completion of the statistical analysis plan, the recommendation is to *modify* the plan at the blind review stage.

- Baseline imbalances.

**CHMP (2015): 'Points to Consider on Adjustment for Baseline Covariates'**

*'A pronounced baseline imbalance is not expected a priori in a randomised trial: if the randomisation process has worked correctly, any observed imbalance is likely to be a random phenomenon. Therefore, if a baseline imbalance is observed this should not be considered an appropriate reason to include this baseline measure as a covariate in the primary analysis. In case the baseline imbalance is for a prognostic factor, sensitivity analyses including the baseline measure as a covariate should be performed in order to assess the robustness of the primary analysis'.*

In Section 6.8, we will say a little more on this point.

- Covariates affected by treatment allocation. Variables measured after randomisation (e.g. compliance, duration of treatment) should not be used as covariates in a model for evaluation of the treatment effect as these may be influenced by the treatment received. A similar issue concerns *late baselines*: that is, covariate measures that are based on data captured after randomisation. The term *time-dependent covariate* is sometimes used in relation to these considerations.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'It is not advisable to adjust the main analyses for covariates measured after randomisation because they may be affected by the treatments.'*

**CHMP (2015): 'Guideline on adjustment for baseline covariates in clinical trials'**

*'A covariate that may be affected by the allocated treatment (for example, a covariate measured after randomisation such as duration of treatment, level of compliance or use of rescue medication) should not be included in the primary analysis of a confirmatory trial. When a covariate is affected by the treatment either through direct causation or through association with another factor, the adjustment may hide or exaggerate the treatment effect. It therefore makes the treatment effect difficult to interpret.'*

- It is good practice for continuous and score outcome variables recording change from baseline to adjust the analysis or include the baseline value of the outcome variable in the analysis as a covariate. When this is done, including the outcome variable itself or the change from baseline in that variable makes no difference mathematically to the analysis; the same *p*-values and estimates of treatment effect will be obtained, so the choice is one of interpretability. If the baseline value is not included in this way, there could be problems with regression towards the mean.

*Regression towards the mean* is a phenomenon that frequently occurs with data in a wide variety of situations. For example, in a study in migraine an entry criterion may be that the patient has suffered a certain number of migraine headaches in the previous month. Migraine headaches often occur in clusters and almost irrespective of treatment, they will likely improve because of the cyclical nature of the condition, and the current interventions being considered are applied at a point in time when patients are having more major problems. If change from baseline was used as the variable to measure the effectiveness of a treatment, the mean change from baseline in each of the two groups would undoubtedly overestimate the absolute benefit of treatment in each of those treatment groups; part of those improvements would be due to *regression towards the mean*. Further, in a randomised comparison, if one treatment group, by chance, contained patients with poorer baseline values, then comparing the mean change from baseline in one group with the mean change from baseline in the other group could give a biased conclusion. Of course, randomisation should protect against this, but in some cases, imbalances will be present, and including the baseline value as a covariate in ANCOVA or adjusting through ANOVA should correct for this bias.

**CHMP (2015): 'Guideline on adjustment for baseline covariates in clinical trials'**

*'When the analysis is based on a continuous outcome there is commonly the choice of whether to use the raw outcome variable or the change from baseline as the primary endpoint. Whichever of these endpoints is chosen, the baseline value should be included as a covariate in the primary analysis. The use of change from baseline with adjustment for baseline is generally more precise than change of baseline without adjustment. Note that when the baseline is included as a covariate in a standard linear model, the estimated treatment effects are identical for both 'change from baseline' (on an additive scale) and the 'raw outcome' analysis. Consequently, if the appropriate adjustment is done, then the choice of endpoint becomes solely an issue of interpretability'.*

- How many covariates? It is usually not appropriate to include lots of covariates in an analysis.

**CHMP (2015): 'Guideline on adjustment for baseline covariates in clinical trials'**

*'No more than a few covariates should be included in the primary analysis. Even though methods of adjustment, such as analysis of covariance, can theoretically adjust for a large number of covariates it is safer to pre-specify a simple model. Results based on such a model are more likely to be numerically stable, the assumptions underpinning the statistical model are easier to validate and generalisability of the results may be improved'.*

Remember, however, that variables used to stratify the randomisation should be included. It also is not usually appropriate to select covariates within ANCOVA models using stepwise or, indeed, any other variable selection techniques. The main purpose of the analysis is to compare the treatment groups, not to select covariates.

**CHMP (2015): '*Guideline on adjustment for baseline covariates in clinical trials*'**

*'Methods that retrospectively select covariates by choosing those that are most strongly associated with the primary outcome (often called 'variable selection methods') should be avoided in confirmatory clinical trials. The clinical and statistical relevance of a covariate should be assessed and justified from a source other than the current dataset'.*

Finally, including covariates that are highly correlated with each other adds little to the analysis and should be avoided. Clinical knowledge of such correlations should help to prevent this from happening.

## 6.9 Baseline testing

It is generally accepted among statisticians that baseline testing – that is, producing *p*-values for comparisons between the treatment groups at baseline – is of little value in randomised trials. If randomisation has been performed correctly, then 5% of significance test comparisons at baseline will give statistically significant results; any imbalances seen at baseline must be due to chance. The only value to such testing is to evaluate whether the randomisation has been performed correctly, for example, in the detection of fraud. Altman (1991), Section 15.4, provides an extensive discussion on the issue of baseline testing.

It is nonetheless appropriate to produce baseline tables of summary statistics for each of the treatment groups. These should be looked at from a clinical perspective and imbalances in variables that are potentially predictive of outcome noted. Good practice hopefully will have ensured that the randomisation has been stratified for important baseline prognostic factors and/or the important prognostic factors have been included in some kind of adjusted analysis, for example, ANCOVA. If this is not the case, then sensitivity analyses should be undertaken through ANOVA or ANCOVA to make sure that those imbalances are not the sole cause of an observed positive (or negative) treatment difference.

## 6.10 Correlation and regression

We have used the term *correlation* in this chapter loosely to denote a relationship between two variables. A *correlation coefficient* can be calculated that measures the strength of that relationship. Correlation coefficients lie between -1 and +1.

A correlation coefficient of +1 represents perfect positive correlation where one variable is directly a linear (straight-line) function of the other variable, with both increasing together. A correlation of -1 represents a perfect negative correlation where again, one variable is a direct linear function of the other, but now as one increases, the other decreases. A correlation coefficient of zero tells us that the two variables are not related. In our clinical trial applications, we are usually more interested in dependence – when a baseline variable changes, what impact does that have on the outcome variable? – and regression addresses this dependence question. Correlation coefficients are more appropriate when we are looking for the connection between two outcome variables or between two baseline variables. To assess the presence of a relationship between two variables, we can test the null hypothesis that the correlation coefficient ( $\rho$ ) is equal to zero ( $H_0: \rho = 0$ ) against the alternative hypothesis that the correlation coefficient is non-zero ( $H_1: \rho \neq 0$ ).

Two correlation coefficients are in common use. If both variables can be assumed to be normally distributed, we use the *Pearson correlation coefficient*, while if one or both variables is non-normal, the *Spearman (rank) correlation coefficient* is used. One issue to be aware of is that the p-value associated with the test of  $\rho = 0$  contains no information regarding the magnitude of the correlation and therefore its clinical relevance. A statistically significant correlation does not necessarily mean that the two variables are closely related; it is simply telling us that there is evidence that a relationship between the two exists: in other words, the two variables are not completely independent.

There is a mathematical connection between the Pearson correlation coefficient and the slope of the linear regression line. If  $s_x$  denotes the standard deviation of the independent variable, while  $s_y$  denotes the standard deviation of the dependent variable, the correlation coefficient is equal to  $bs_x/s_y$ , where  $b$  is the calculated value of the regression slope.

We will return to some further discussion on correlation in Chapter 20.

## CHAPTER 7

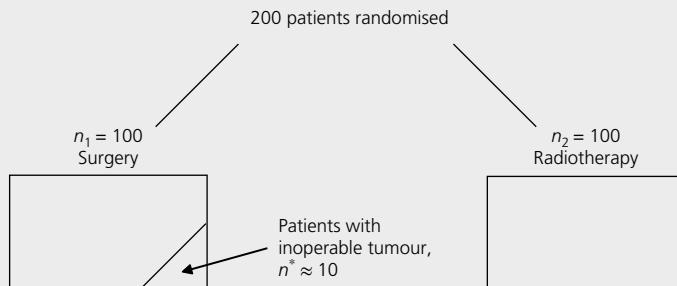
# Intention-to-treat, analysis sets and missing data

### 7.1 The principle of intention-to-treat

When we analyse data, there are inevitably questions that arise regarding how to deal with patients who withdraw, patients who have violated the protocol, patients who have taken a banned concomitant medication and so on. The principle of intention-to-treat (ITT) helps guide our actions. We will firstly explain the principle through two examples and then discuss various aspects of its interpretation before addressing the application of the principle in practice.

#### Example 7.1 Surgery compared to radiotherapy in operable lung cancer

Consider a (hypothetical) trial comparing surgery and radiotherapy in the treatment of operable lung cancer. Assume that a total of 200 patients were randomised to one of these two groups, as shown in Figure 7.1.



**Figure 7.1** Randomised trial comparing surgery and radiotherapy

Unfortunately, some of the patients randomised to the surgery group were found on the operating table not to have operable tumours. The intended operative procedure could not be undertaken for these patients, and they were simply closed up and given the radiotherapy regimen defined for the radiotherapy group. Assume that there were 10 such patients. The primary endpoint is survival time, and all 200 patients were followed up until death. At the analysis stage, we have to decide how to deal with the 10 patients who were assigned to

surgery but did not receive the intended surgery. Several possibilities exist for the analysis, and we will consider the following three:

*Option 1:* Compare the mean survival time of the 90 who received the intended surgery with the mean survival time of the 100 who were assigned to radiotherapy.

Remember that all 200 patients provide data on the primary endpoint, so this analysis ignores the data on 10 patients, those who were assigned to surgery but did not receive the intended operative procedure. Perhaps we can think of ways of including those data. Options 2 and 3 include these patients.

*Option 2:* Compare the mean survival time of the 90 patients who received the intended surgery with the mean survival time of the 100 + 10 patients who ended up getting radiotherapy.

The argument here is that the 10 patients who *switched* from surgery to radiotherapy are most likely to behave like the 100 patients initially assigned to the radiotherapy group.

*Option 3:* Compare the groups according to the randomisation scheme: that is, the mean survival time of the 100 patients initially assigned to surgery with the mean survival time of the 100 patients initially assigned to radiotherapy.

These three options differ merely in the way the 10 patients who were assigned to surgery but did not receive the intended surgery are handled. Option 1 ignores them, option 2 puts them in the radiotherapy *group*, while option 3 leaves them in the surgery *group*.

There are two things to note immediately. Firstly, the 10 patients in the surgery group who do not have operable tumours are likely to have very poor prognoses. These are the patients who have advanced tumours, maybe multiple tumours, that cannot easily be excised. Secondly, randomisation will have given us balanced groups so, equally, there will be approximately 10 patients in the radiotherapy group who also do not have operable tumours and similarly have a very poor prognosis, but these patients remain unseen and unidentified.

Now, consider the implications arising out of each of these options.

*Option 1:* This option removes 10 patients with very poor prognoses from the surgery group but leaves a similar group of patients with very poor prognoses in the radiotherapy group. This clearly introduces bias as it is not comparing like with like. The radiotherapy group will, because of the omission of 10 very specific poor prognosis patients in the surgery group, be on average a better prognosis group than the surgery group to which they are being compared.

*Option 2:* This option is twice as bad! The 10 patients with very poor prognoses from the surgery group have been transferred to the radiotherapy group to give a total of 20 patients with very poor prognoses in that group. The bias in the resulting analysis is likely to be even bigger than the bias under option 1.

*Option 3:* This is the only valid option. It is the only option that compares groups that are alike in terms of the mix of patients. This is the ITT option, which compares the groups as randomised.

A number of issues arise out of these considerations. While option 3 makes sense from a statistical perspective, does it make sense from a clinical standpoint? Remember that the purpose of this trial is to compare surgery and radiotherapy in operable lung cancer, yet we are including 10 patients in the surgery group who did not receive the planned surgery. Well,

in fact, option 3 is not comparing surgery with radiotherapy; it is comparing two treatment strategies. Strategy 1 is to give the patient surgery; however, if it is found that the tumour is not operable, then close the patient up and give them radiotherapy. On the other hand, strategy 2 is to give radiotherapy.

Although you may agree that this provides a comparison of these two treatment strategies, you may say that this is not of interest to you clinically, and you are looking for a comparison of pure surgery with pure radiotherapy. On that basis, you may therefore prefer to go for option 1, which seems to provide exactly what you want, as it gives a comparison of pure surgery with pure radiotherapy! This view, however, would be both naive and incorrect. Option 1 does not provide a *valid* comparison of pure surgery and pure radiotherapy because the groups are not alike; it is subject to bias because the radiotherapy group in this comparison is on average a better prognosis group than the surgery group.

A question that sometimes arises here is, can't we just remove 10 patients from the radiotherapy group and then compare the 90 who received surgery with the resulting radiotherapy group? No: although this would equalise the numbers in the two groups being compared, it would not necessarily equalise the mix of patients in the two groups. Remember that the 10 patients who did not get surgery are not just any 10 patients; they are a selected group of patients with very poor prognoses. As an alternative, could we remove the 10 patients in the radiotherapy group who turn out to have the worst survival experience? The answer again is no; this would not necessarily make the resulting comparison valid as it is based on a very strong assumption, which could never be verified, that the 10 patients who did not get the intended surgery are indeed the 10 worst patients from that group. Unfortunately, the comparison of pure surgery and pure radiotherapy is not possible in a straightforward way in this trial based on a simple selection of patient groups. The only question that can be answered relates to the evaluation of the two treatment strategies mentioned earlier.

### **Example 7.2** Clofibrate in the reduction of mortality after myocardial infarction

This is a well-known placebo-controlled trial that evaluated clofibrate in terms of reducing mortality in patients suffering a myocardial infarct and was reported by the Coronary Drug Project Research Group (1980).

The groups were compared overall, and the five-year death rate among the 1103 patients randomised to clofibrate was 20.0% compared to a five-year death rate among the 2789 placebo patients of 20.9%. These differences were not statistically significant, with  $p = 0.55$ .

The trialists then investigated the impact of compliance on these results. Patients were defined as good compliers if they took at least 80% of the prescribed dose during the treatment period. Poor compliers were patients who took less than 80%. In the clofibrate group, the good compliers were seen to have only a 15.0% five-year death rate, while the poor compliers had a 24.6% five-year death rate, a clear difference both clinically and statistically, with  $p = 0.001$ . So, the active medication does work; it is simply a matter of taking the medication. Patients who take the medication do well, while those who fail to take the medication do not.

However, this same comparison was then undertaken among the placebo patients. The good compliers on placebo had only a 15.1% five-year death rate, while the poor compliers

had a 28.3% five-year death rate with  $p = 0.0000000000000047$ ! These placebo tablets are remarkable!

A moment's thought will suffice to realise that the conclusions we are drawing from these analyses are nonsense. The bottom line is that the active medication has no effect, and the initial overall comparison tells us that. However, compliance is linked to other things that are potentially having an effect, such as giving up smoking, starting to take regular exercise, modifying one's diet to reduce the amount of sugar consumed and so on. The patients who do these things are the ones who do everything that their doctors tell them to do, including taking the medication! So, taking the medication is correlated with some other things that are beneficial and causing an apparent *treatment effect* related to compliance.

These two examples should give a clear indication of the dangers of compromising the randomisation at the analysis stage. Even small departures in terms of excluding patients from the analysis could have a major impact on the validity of the conclusions.

The *principle of ITT* tells us to compare the patients according to the treatments to which they were randomised. Randomisation gives us comparable groups; removing patients at the analysis stage destroys the randomisation and introduces bias. Randomisation also underpins the validity of the statistical comparisons. If we depart from the randomisation scheme, the statistical properties of our tests are compromised.

The 1998 FDA guideline on antimicrobial drugs captured the issues well.

#### **FDA (1998): 'Developing Antimicrobial Drugs – General Considerations for Clinical Trials'**

*'The intent-to-treat principle suggests that eligible, randomised patients should be evaluated with respect to outcome based on original treatment assignment regardless of modifications to treatment occurring after randomisation. The statistical analysis seeks to establish if the particular assignment received is predictive of outcome, and the study can be interpreted as a strategy trial where the initial assignment is only the beginning of the treatment strategy. However, many researchers seek to glean results from the clinical trial that would have been observed if all patients had been able to remain on their initial assignment. This leads to analysis of subsets that exclude patients with imperfect compliance or follow-up data. However, the validity of these analyses rests on the assumption that the two treatment groups, after excluding such patients, differ only by the treatment received. This assumption could be violated in many subtle ways. For example, differential toxicity related to severity of illness could lead to selection bias. Similarly, the subjects unable to comply with medication may be those most at risk of a negative outcome and their exclusion may bias the treatment comparison.'*

## 7.2 The practice of intention-to-treat

### 7.2.1 Full analysis set

The previous section clearly indicates the need to conform to the principle of ITT in relation to the randomisation to ensure that the statistical comparison of the treatment groups remains valid. In practice, compliance with this principle is sometimes difficult, and the regulators, recognising these difficulties, allow a compromise. This involves the definition in particular trials of the *full analysis set*, which gets us as close as we possibly can get to the ITT ideal.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The intention-to-treat principle implies that the primary analysis should include all randomised subjects. Compliance with this principle would necessitate complete follow-up of all randomised subjects for study outcomes. In practice this ideal may be difficult to achieve, for reasons to be described. In this document the term "full analysis set" is used to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomised subjects.'*

The regulators are telling us, therefore, to get as close as possible, and they go on in the ICH E9 guideline to outline circumstances where it will usually be acceptable to omit subjects without causing bias.

These potential exclusions are:

- Subjects who violate the inclusion/exclusion criteria
- Subjects who fail to take at least one dose of study medication
- Subjects who do not provide any post-baseline data

These omissions will not cause bias only under some circumstances. In particular, subjects in each of the treatment groups should receive equal scrutiny for protocol violations, and all such violators should be excluded, in relation to the first point. For the second and third points, the fact that patients do not take study medication or do not provide any post-baseline data should be unrelated to the treatments to which such subjects were assigned. Any potential bias arising from these exclusions should be fully investigated.

The term *full analysis set* (FAS) was introduced to separate the practice of ITT from the principle, but practitioners still frequently use the term *ITT population* when referring to this analysis set. The term *modified ITT population* (*mITT*) is also in common use within companies and also by regulators in some settings where exclusions from strict ITT are considered. Different therapeutic situations will invariably give rise to different considerations on how best to define the most appropriate analysis sets. Increasingly, we are seeing very specific regulatory guidance in this regard within therapeutic-specific guidelines. In the treatment of bacterial infections, for example, the CHMP (2012) guideline recommends;

*'In studies with antibacterial agents that have a clinical primary endpoint it is suggested that the all treated population and the clinically-evaluatable population should be viewed as co-primary. In studies with a microbiological primary endpoint it is suggested that the co-primary analysis populations should be all-treated with a pathogen and microbiologically-evaluatable'.*

In superiority trials, the full analysis set or something akin to that is invariably the basis for the primary analysis. The regulatory preference for this stems in part because the full analysis set also tends to give a conservative view of the treatment difference, as a result of including in the analysis subjects who have not conformed entirely with the protocol. The regulators can be assured that if the analysis based on this analysis set gives a statistically significant result, then the treatment being evaluated is effective. This preference, however, only applies when considering superiority trials. In equivalence and non-inferiority, this analysis set tends to be anti-conservative. This issue will be discussed later in Chapter 12 on equivalence and non-inferiority testing.

### **7.2.2 Per-protocol set**

The *per-protocol set (PPS)* is described as follows by the regulators.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The "per-protocol" set of subjects, sometimes described as the "valid cases", the "efficacy" sample or the "evaluable subjects" sample, defines a subset of the subjects in the full analysis set who are more compliant with the protocol. . .'*

The definition of a PPS of subjects in principle allows us to get closer to the scientific question by including only those patients who comply with the protocol to a defined extent. The PPS, like the full analysis set, must be pre-specified in the protocol and then defined at the patient level at the blind review, following database lock but before breaking the blind. It must be noted, however, that the PPS is subject to bias as it does not comply with randomisation. In our discussions in Chapter 8 on estimands, we will see that analyses based on the PPS have fallen out of favour in recognition of their inherent bias, and the scientific and clinical questions previously addressed by looking at this analysis set are dealt with more appropriately through considerations within the estimand framework.

### **7.2.3 Further aspects of ITT**

The focus so far in this chapter has been on the set of subjects to be included in the analysis. A further aspect of the principle of ITT concerns the data that are to be included in the analysis, with the emphasis on inclusion of all available data irrespective of the compliance of the subject to the protocol. The following text is taken from the glossary to the ICH E9 guideline.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The principle asserts that the effect of a treatment policy can be assessed by evaluating on the basis of the intention to treat a subject (i.e. the planned regimen) rather than the actual treatment given. It has the consequence that subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment.'*

In summary, therefore, the principle of ITT tells us to include in the analysis all subjects in the groups to which they were randomised and all data collected on those subjects.

For a per-protocol analysis, it may well be more appropriate to not only exclude some subjects from the analysis but also to exclude some data from those subjects who remain in the analysis. For example, data following the taking of rescue medication, following switching to an alternative treatment, or after simply stopping the assigned medication, perhaps due to side effects, could be excluded from an analysis based on the PPS. Again, the thinking here is to try to look more at the pure effects of treatment.

It has been good statistical practice to evaluate the sensitivity of the conclusions to different choices of the analysis sets.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'In general, it is advantageous to demonstrate a lack of sensitivity of the principle trial results to alternative choices of the set of subjects analysed. In confirmatory trials it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per-protocol analysis, so that any differences between them can be the subject of explicit discussion and interpretation'.*

*'When the full analysis set and the per-protocol set lead to essentially the same conclusions, confidence in the trial results is increased. . . '*

As mentioned earlier, however, since the introduction of the estimand framework, much less emphasis has been placed on analyses associated with the PPS and considerations with regard to which data to include and which data to exclude are more precisely addressed through that framework.

## 7.3 Missing data

### 7.3.1 Introduction

The discussion in the previous section regarding the practical application of the principle of ITT does not give the full picture. While this principle plus consideration of the PPS may clearly define the sets of subjects to be analysed and the data

on those subjects that are to be the basis of those analyses, we still have to decide how to deal with the missing data caused by withdrawal from the study or loss to follow-up for example.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfill all the requirements of the protocol concerning the collection and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, nonetheless, provided the methods of dealing with missing values are sensible, and particularly if those methods are pre-defined in the protocol.'*

It is worth remembering that there is no single, perfect way of handling missing data. The only truly effective way of dealing with missing data is not to have any in the first place! In addition, we need to consider the sensitivity of our trial results to the methods employed to handle missing data. This is particularly true if there are a large amount of such data.

There are a number of alternative simple approaches to deal with missing data in common practice. Among these so-called *single imputation* methods are the following:

- Complete cases analysis
- Last observation carried forward (LOCF)
- Baseline observation carried forward (BOCF)
- Success/failure classification
- Worst-case/best-case imputation

In recent years, we have seen the development of a collection of more sophisticated methods that go under the general heading of *multiple imputation*. These are based on certain assumptions about the mechanisms that cause the data to be missing. We will discuss this framework in a later section.

### **7.3.2 Complete cases analysis**

One very simplistic way of handling missing data is to remove those patients with missing data from the analysis in a *complete cases analysis* or *completer analysis*. By definition, this is a per-protocol analysis that omits all patients who do not provide a measure on the primary endpoint and is of course subject to bias. Such an analysis could be acceptable in an exploratory setting where we may be looking to get some idea of the treatment effect if every subject were to follow the protocol perfectly, but it would in general not be acceptable in a confirmatory setting as a primary analysis.

### 7.3.3 Last observation carried forward (LOCF)

This analysis takes the final observation for each patient and uses it as that patient's endpoint in the analysis. For example, in a 12-month trial in acute schizophrenia, a patient who withdraws at month 7 due to side effects will have their month 7 value included in the month 12 analysis of the data.

In one sense, this approach has clinical appeal. The final value provided by the patient who withdrew at month 7 is a valid measure of how successful we have been in treating this patient with the assigned treatment and so should be part of the overall evaluation of the treatment. In some circumstances, however, this argument breaks down. For example, if there is an underlying worsening trend in disease severity, then patients who withdraw early will tend to provide better outcomes than those who withdraw later in the treatment period. If one treatment has more early drop-outs than the other, possibly because of side effects, there will be bias caused by the use of this method. Multiple sclerosis and Alzheimer's disease are settings where this could apply. The opposite will of course be true in cases where the underlying trend is one of improvement; depression could be one such setting.

### 7.3.4 Baseline observation carried forward (BOCF)

This approach takes the baseline value and carries it forward for those subjects with a missing outcome value irrespective of any data that have been collected during the study. The thinking here is that once treatment is removed, any benefits of the treatment are lost, and the subject returns to their condition at the point of randomisation. If the variable being used as the basis for analysis is change from baseline, the endpoint will take the value zero. This could be an appropriate approach, for example, where the treatments under study are simply being evaluated for symptom relief and there is no underlying trend in the disease condition.

### 7.3.5 Success/failure classification

One particularly simple way of dealing with missing data is to use a binary outcome as the endpoint, with drop-out being classified as treatment failure. Success (or response) could be defined in terms of a certain improvement in an outcome variable from baseline. However, reducing the outcome to a dichotomy in this way will lead to a loss of power, and this loss of power needs to be considered when calculating sample size at the planning stage. This kind of analysis is often referred to as a *responder analysis*.

A success/failure approach is particularly effective if the endpoint is already binary: for example, cured/not cured in a trial of an anti-infective or responder/non-responder in an oncology study, with patients with missing data, perhaps due to withdrawal of consent or loss to follow up, considered as failures/non-responders.

### 7.3.6 Worst-case/best-case classification

This method gives those subjects who withdraw for positive reasons the best possible outcome value for the endpoint and those who withdraw for negative reasons the worst value. This may seem a little extreme, and a lesser position would be to look at the distribution of the endpoint for the completers and use, say, the upper quartile (the value that cuts off the best 25% of values) for those subjects withdrawing for positive reasons and the lower quartile (the value that cuts off the worst 25% of values) for those subjects who withdraw for negative reasons.

Alternative applications of these rules would be to use the best-case value for subjects in the control group and the worst-case value for subjects in the test treatment group. Certainly, if the treatment comparison is preserved under such a harsh scheme, we can be confident of a true benefit for the test treatment!

#### **Example 7.3** Bosentan therapy in pulmonary arterial hypertension

This was a placebo-controlled trial (Rubin et al., 2002) using change from baseline to week 16 in exercise capacity (six minute walking distance, 6MWD) as the primary endpoint and change from baseline to week 16 in the Borg dyspnea index and the WHO functional class as key secondary endpoints. Missing data at week 16 were handled as follows:

*'For patients who discontinued the study medication because of clinical worsening, the values recorded at the time of discontinuation were used; patients for whom no value was recorded (including patients who died) were assigned the worst possible value (0 m). For all other patients without a week 16 assessment, the last six-minute walking distance, score on the Borg dyspnea index, and WHO functional class were used as week 16 values.'*

So, for patients without week 16 values, LOCF was used unless the patient withdrew because of clinical worsening, in which case either the data at the time of withdrawal or a worst-case imputed value (zero metres for the walk test, Borg index = 0 or WHO functional class = IV) was used if there were no data at withdrawal.

### 7.3.7 Sensitivity

In all cases, and particularly where the extent of missing data is substantial, several analyses will usually be undertaken to assess the sensitivity of the conclusions to the method used to handle missing data. If the conclusions are consistent across these different analyses, we are in a good position. However, if our conclusions are seen to change or to depend heavily on the method used for dealing with missing data, the validity of those conclusions will be drawn into question.

Our discussion so far regarding sensitivity has focused on using several different approaches to both the definition of the analysis sets and the handling of missing data. One of the main goals of a trial is to estimate the magnitude of the

treatment benefit, and these sensitivity evaluations will give a series of estimates. For estimation purposes – and in particular, we are thinking in terms of point estimates and confidence intervals – we should choose a method of imputation that makes sense clinically rather than go for extremes that, while providing a conservative view of the treatment effect for the purposes of evaluating statistical significance, do not give a sensible, realistic estimate of the clinical benefit.

The whole issue of sensitivity analyses and how these can be best structured will be revisited in conjunction with the use of estimands in Chapter 8.

### 7.3.8 Avoidance of missing data

The CHMP guideline on missing data (CHMP, 2010) includes several key points on the avoidance of missing data. As we can see from the earlier discussion, missing data causes problems, and we should avoid it wherever possible.

#### ***CHMP (2010): 'Points to Consider on Missing Data in Confirmatory Clinical Trials'***

*'Several major difficulties arise as a result of the presence of missing values and these are aggravated as the number of missing values increases. Thus, it is extremely important to avoid the presence of unobserved measurements as much as possible, by favouring designs that minimise this problem, as well as strengthening data collection regardless of the patient's adhesion to the protocol and encouraging the retrieval of data after the patient's drop-out. Continued collection of data after the patient's cessation of study treatment is strongly encouraged, in particular data on clinical outcome. In some circumstances, in particular where this type of "retrieved drop-out" information represents the progression of the patient without (or before) impact of further therapeutic intervention, these data give the best approximation to the Full Analysis Set and would generally be seen as a sound basis for the primary analysis.'*

Allowing dose reductions or drug *holidays* will possibly keep patients in a trial and avoid them dropping out, in addition to providing a model closer to what will happen in practice in real life. Continued follow-up for patients who withdraw from medication is essential for application of ITT and will in any case give much more flexibility when it comes to analysing the data once the trial is complete. There are of course ethical issues associated with continuing follow-up for patients who wish to withdraw from treatment. In some countries, it is disallowed by law, and that of itself may rule out the potential for doing this; alternatively, we can restrict recruitment to countries where it is feasible. If continued follow-up following withdrawal from treatment is to happen, the informed consent form will need to be structured to take account of this. It should be the expectation that data continue to be collected for subjects following withdrawal from treatment, and both subjects and

investigators should be aware of this at the informed consent stage. Investigators sometimes need to be educated on the importance of continued data collection in such cases.

### 7.3.9 Classification of missing data

Many of the methods we have discussed for dealing with missing data have involved some form of (*single*) *imputation*. LOCF, BOCF, success/failure classification and worst-case/best-case classification are all examples where the missing data values have been replaced by certain assumed values. In this section, we will briefly discuss more sophisticated forms of imputation in conjunction with various assumptions regarding the nature of the missing data. We will also return to these issues in the next chapter on estimands.

A theoretical framework has been developed that is based on a classification according to the nature of the underlying *missingness* mechanism:

- An observation that is missing purely by chance with no connection with the condition of the patient at that point in time is said to be *missing completely at random (MCAR)*. This form of missingness occurs, for example, if a subject misses a visit because of some administrative mistake in booking the appointment or when a blood sample has been contaminated during transit or in storage. Omitting patients with missing data from the analysis under these circumstances will not cause bias.
- An observation is said to be *missing at random (MAR)* if that observation can be predicted from data from other similar patients who do not have missing observations. Omitting patients with missing data under these circumstances will cause bias. However, those missing data values can be  or *imputed* based on a model fitted to those patients with complete data, and the characteristics of the patient(s) with missing data, to remove that bias.
- When the observation is neither MCAR nor MAR, it is said to be *missing not at random (MNAR)*. When an observation is MNAR, it cannot be predicted directly from data from other similar patients who do not have missing observations. Under these circumstances, omitting patients with missing data will again cause bias; not only that, there is no easy way to correct for this bias through prediction/modelling, as is the case with MAR.

Note that in the definitions of MCAR and MAR, when we talk about the *observed data for that patient*, we are thinking in terms of baseline characteristics and observations taken at earlier visits. Further details at a technical level on this framework are provided by Carpenter and Kenward (2007), while Sterne et al. (2009) give a basic overview.

The difficulty with applying this thinking under most circumstances is that it is not possible to know definitively which situation we are in. Nonetheless, we can make assumptions about the missingness mechanism and undertake sensitivity analyses based on varying those assumptions.

If the data truly are MCAR, then we can simply undertake an analysis based on those patients without missing data to get an unbiased view of the treatment differences. This is the only situation where analysing the data on completers gives an unbiased view of the treatment effect. In practice, however, we are rarely in this situation; and when we are, the assumption will usually only apply to at most a handful of patients.

If the data are MAR, various strategies are available to deal with those missing data:

- For either continuous or score endpoints where observations on the variable that is the basis for the endpoint are available at various visits through follow-up, the usual approach is to analyse based on an MMRM model fitted to the non-missing data. This model implicitly assumes MAR and utilises only those data that are available across the group of patients being evaluated.
- Alternatives involve imputing the missing observations through multiple imputation, and we will discuss the methodology associated with this approach in the next section.

### **7.3.10 Multiple imputation**

Consider initially continuous or score endpoints. *Multiple imputation* (MI) proceeds in a series of steps:

- 1 Fit a regression model (*imputation model*) to the data from those patients who have complete data. This is a multiple regression model with the outcome variable as the dependent variable and the baseline characteristics as independent variables. The variance associated with the random element (Section 6.5.7) is also predicted from the data on patients with complete data.
- 2 Predict the mean outcome for each of the patients with missing outcomes based on their baseline characteristics, simulating from the normal distribution with the observed variance to add the random element and give a unique set of outcomes for the patients with missing values. Estimate the treatment effect from the set of patients with complete data plus those patients with estimated outcomes from the imputation model.
- 3 Repeat this process, say 20 times, getting a new estimate of the treatment effect each time. Average those 20 values to give the overall estimate of the treatment effect. Use *Rubin's Rules* (Carpenter and Kenward (2007)) to obtain a standard error for the estimated treatment effect and allow calculation of a *p*-value for treatment differences and the associated confidence interval.

As can be seen from the method used to predict the missing outcomes, we are assuming that those outcomes are missing at random, and this allows us to undertake the prediction based on patients with complete data.

When we have binary outcomes, we follow a similar process for multiple imputation but now with the imputation model based on logistic regression. The logistic regression model fitted to data on completers allows us to predict the

probability of success/response/event based on baseline characteristics. As mentioned in the previous chapter in our discussion of logistic regression, the fitted model contains a random element through the predicted probability, and we simulate an actual outcome from that prediction for a patient with a missing outcome. To be more specific, suppose the predicted response probability is 0.8. We can imagine placing 10 tickets in a hat, 8 marked ‘response’ and 2 marked ‘non-response’, and choosing one at random to give our ‘outcome’ with the correct background probability of response. We do this for each patient with a missing outcome and augment those patients without missing outcomes to produce a complete set of data. We then analyse those data and calculate a treatment difference, perhaps as an odds ratio, and repeat 20 times, averaging the odds ratios over those 20 repeats to give our final estimate of the treatment effect. Rubin’s Rules give us a standard error that provides a *p*-value and confidence interval. A straightforward extension allows us to deal with an ordered categorical outcome.

A start point for statistical analysis in many cases will be to assume that data are missing at random. The usual strategy is to then use the MMRM model for statistical analysis for continuous or score endpoints when the outcome is measured at various visits through time but to use multiple imputation if we are dealing with a single outcome measured at one point in time. For binary endpoints, such as ‘cure’ or ‘response’, we will be using MI.

The missing-at-random assumption is only an assumption, and a key step in any statistical analysis is to relax that assumption to judge the robustness of the results to departures from MAR. There are many different MNAR assumptions that can be utilised. Two common approaches are *jump to reference* and the *delta method*. For the purposes of this discussion, assume that the patient has withdrawn from the study due to a serious adverse event. It is standard practice and the basis for a conservative analysis to continue assuming MAR for patients in the control group but use MNAR assumptions for patients in the experimental group to penalise those outcomes appropriately.

Jump to reference assumes that a patient in the experimental group behaves exactly like a control group patient following withdrawal. MI can be used to predict the outcome for a patient in the control group with the same characteristics as the patient in the experimental group who has the missing outcome and this value used as the imputed value for that patient. In this method, it is assumed that any benefit the experimental group patient has accrued prior to withdrawal has been lost. A less extreme assumption, termed *copy reference*, assumes that the benefit accrued by the experimental group patient up to the point of withdrawal is retained. Finally, the delta method uses standard MI under the MAR assumption for predicting the outcome for the experimental group withdrawal but then discounts this by an amount  $\delta$ , which could, for example, be a proportion of the treatment difference in outcomes. See Kenward (2015) for further discussion on these and other techniques.

## 7.4 Intention-to-treat and time-to-event data

To illustrate the kinds of arguments and considerations that are needed in relation to ITT, the discussion in this section will consider a set of applications where problems can arise. In Chapter 13, we will cover methods for the analysis of time-to-event endpoints or so-called *survival data*; but for the moment, I would like to focus on endpoints within these areas that do not use the time point at which randomisation occurs as the start point for the time-to-event measure. Examples include the time from rash healing to complete cessation of pain in herpes zoster, the time from six weeks after start of treatment to first seizure in epilepsy, the time from eight weeks to relapse among responders at week 8 in severe depression and finally the duration of response in oncology.

In each of these situations, there is clinical interest in looking at these endpoints. From a statistical point of view, however, each endpoint gives an analysis that is in violation of the principle of ITT and can result in bias as a consequence. We will look at each of the settings in turn.

In the case of a randomised trial in herpes zoster, patients have the potential to cease pain prior to rash healing, and these patients would not enter the analysis of time to cessation of pain from rash healing. Invariably, the likelihood that pain will cease early in this way will depend upon the treatment received. Therefore, the sets of patients in each of the two treatment groups entering the analysis of time to cessation of pain from rash healing will not necessarily be alike. This selection phenomenon will result in a violation of ITT, and the resultant analysis will be biased. See Kay (1995) for further discussion on this point. There have been some attempts (see, e.g. Arani et al., 2001) to justify such an analysis using complex statistical modelling, but this approach has been shown (Kay (2006)) to be flawed, and the problem of violation of ITT remains. In herpes zoster, the pain that remains following rash healing is known as post-herpetic neuralgia (PHN), and there is strong interest in evaluating the relative effects of treatments on PHN. Looking at time from rash healing to cessation of pain is an attempt to focus on this. Unfortunately, it is not possible – at least, using this simple approach, even in a randomised trial – to analyse this endpoint in isolation in an unbiased way. The best way to identify the relative effect of the two treatments on PHN is to compare the proportion of all patients in the two treatment groups still with pain at specified points through time according to the randomised groups. Those patients who cease pain prior to rash healing by definition do not suffer PHN, and including those patients in an analysis is needed to get a fair treatment comparison in relation to that aspect of the condition.

In newly diagnosed epilepsy, it is not uncommon to use the time from six weeks (or sometimes three months) following the start of treatment to the first seizure as the primary – or certainly important secondary – endpoint. Again, however, there are problems with selection effects and ITT. Brodie et al. (1995) evaluate lamotrigine compared to carbamazepine in a randomised trial in

patients with newly diagnosed epilepsy. In a secondary analysis of time (from randomisation) to withdrawal, by six weeks, approximately 18% of the patients have withdrawn in the lamotrigine group, while approximately 27% of patients have withdrawn in the carbamazepine group over this period. The analysis of time from six weeks to first seizure excludes these patients. This is a long way from being a randomised comparison and is potentially subject to bias. Even if the withdrawal rates had been the same, the potential for bias would remain. It is not so much that the numbers of patients withdrawing are different; it is that the comparability in terms of the mix of patients has been compromised, and the differential withdrawal rates just make things worse. From a clinical point of view, excluding the first six weeks of treatment does make sense as it is recognised that it takes some time to stabilise the dose, but again, unfortunately, the endpoint that apparently captures this – time from six weeks to first seizure – cannot be evaluated in a straightforward, unbiased way. An alternative and appropriate approach to look at the effectiveness of treatment following dose stabilisation has been suggested by Brodie and Whitehead (2006). These authors (using three months as the stabilisation period rather than six weeks) consider time from randomisation to withdrawal, whenever this occurs, or to first seizure from three months onwards. This endpoint combines both tolerability (withdrawal) and efficacy (first seizure) but does not penalise a treatment in terms of seizures during the stabilisation phase (the first three months). From a pragmatic perspective and from the patients' point of view, this alternative endpoint makes a lot of sense; the important issue is longer-term stabilisation, free of seizures, and this endpoint captures that.

In severe depression, many trials are designed to investigate treatment relapse in patients who have responded following treatment. Response could be defined, for example, by a reduction in the score on the 17-point Hamilton Depression Scale (HAMD-17) to below 15, with relapse defined as an increase to 16 or above. Typically, response is assessed following eight weeks of treatment, and the endpoint of interest in evaluating relapse is the time from eight weeks to relapse. Patients who have not responded by week 8 are usually withdrawn for lack of efficacy. These extension studies in responders are not randomised comparisons, and the analysis is based solely on those patients who are responders at eight weeks. Storosum et al. (2001) recognise the potential for bias in this analysis. In common with the previous two settings, there is a violation of the principle of ITT. An alternative analysis that looks at time to treatment failure with treatment failure defined as withdrawn from treatment for lack of efficacy up to week 8 or  $\text{HAMD-17} \geq 16$  beyond week 8 would maintain all patients in the treatment comparison. This comparison takes account of possible differential effects of treatment up to and including week 8 in terms of achieving a response, and beyond week 8 is looking at the proportion of patients whose response is maintained.

Finally, it is quite common in oncology studies to look at the duration of response among responders as a way of judging the quality of a response to treatment. Having a longer duration of response is much more desirable than having a response that is short-lived. Clearly, a patient needs to respond before they can have a measurement on duration of response, and it is often the case, as would be expected, that an effective treatment produces more responders. Comparing treatment groups for duration of response invariably looks at two groups of patients who are not alike in their characteristics, and calculation of *p*-values for those comparisons is not recommended. However, plotting Kaplan-Meier curves can be helpful in understanding how the quality of the response is playing out across the groups.

In general, care should be taken with time-to-event endpoints that do not use the point of randomisation as the start point as there is always the potential for patient selection to take place between the point of randomisation and when the clock starts ticking for the proposed endpoint. We will see in the next chapter, when we discuss estimands, possible alternative approaches that do take into account these intermediate events and address questions relating to associated endpoints in a structured way.

## 7.5 General questions and considerations

One question that is frequently asked is, what do you do with patients who were given the wrong treatment by mistake? It must be said that this does not happen very frequently, but when it does, it is necessary to dig a little and try to find out why this has happened. If it is an isolated case and clearly an administrative error, it seems most reasonable to include that patient in the group according to treatment received. Note that strict ITT would retain this patient in the group to which they were randomised, and such an analysis may be imposed by regulators. However, if it is not an isolated case – maybe there are several such mistakes in the same centre – this draws into question the validity of what is happening at that centre, and one starts to think in terms of poor quality, misconduct or indeed fraud; has the investigator correctly followed the randomisation scheme? In such situations, there may ultimately be consideration of removing all data from that centre from the analysis.

The considerations so far in this chapter have been on the evaluation of efficacy. For safety, we usually define the *safety set* as the set of subjects who receive at least one dose of study medication. Usually, the safety set coincides with the full analysis set, but not always. There may well be a patient who started on medication but withdrew immediately because of a side effect. This patient is unlikely to have provided post-baseline efficacy data and so could be excluded from the full analysis set but would still be included in the safety set.

In crossover trials, considerations of analysis sets and missing data are somewhat different. In these trials, each subject provides a response on each of the treatments. The analysis of such data focuses on the treatment difference within each subject. When a subject drops out during the second period and therefore fails to give a response for the treatment given in period 2, then it is not possible to calculate a treatment difference, and so this patient would not be included in the analysis. So, in crossover trials, we are usually forced to exclude the drop-outs. Does this compromise the validity of the treatment comparison? In terms of bias, the answer is not usually, since exclusions of this kind will deplete each of the treatment *groups* equally as the same subject is being omitted from each of them, although the potential for bias should always be considered. In terms of extrapolating the conclusions from the trial to the general population, however, there could be problems. If certain kinds of patients are being omitted from the analysis essentially because they are prone to side effects from one of the two treatments being compared, the trial population may not be representative of the population defined by the inclusion/exclusion criteria.

In phase I studies with healthy volunteers, these aspects are unlikely to be an issue, and it is common practice to *replace* drop-outs with other subjects to achieve the required sample size.

A key aspect of the definition of analysis sets and the way that missing data are to be handled is pre-specification. Usually, these points will be covered in the protocol or, if not, in the statistical analysis plan. If methods are not pre-specified, there will be problems, as the way that these issues are dealt with could then be data driven, or at least there may be suspicion of that. This is, of course, not unique to analysis sets and missing data but is true more generally in relation to the main methods of statistical analysis.

To conclude this discussion, it is worth covering just a few misconceptions:

- Does having equal numbers of subjects in the treatment groups at the statistical analysis stage protect against bias?

*This corresponds to similar drop-out rates across the treatment groups. The answer to the question is no! It is the mix of patients that is the basis of a valid comparison, not the numbers of patients. It is almost inevitable that if two treatments are truly different, then different kinds of subjects will drop out of the two groups. For example, in a placebo-controlled trial, those withdrawing from the active treatment group could well be withdrawing for side effects, while the drop-outs in the placebo group could be withdrawing because of lack of effect.*

- Does basing the sample size calculation on the PPS and then increasing the sample size to allow for drop-outs ensure that the PPS will not be subject to bias?

*No! It has often made sense to power for the PPS and then factor upwards to allow for drop-outs, as this has ensured that there is enough power for the full analysis set, provided that any extra patient-to-patient variation in the full analysis set does not*

*counterbalance the increase in sample size. However, the analysis based on the PPS is still subject to bias. See Section 9.5.2 for a further aspect of this discussion.*

- Does pre-specifying in the protocol that the analysis based on the PPS will be the primary analysis protect against bias?

*As mentioned elsewhere, it is good scientific practice to pre-specify the main methods of statistical analysis in the protocol; but just because something is specified in the protocol does not mean it is correct. So again, the answer is no.*

The focus of the next chapter is estimands, and considerations in association with this concept will build on many aspects considered in this chapter.

# CHAPTER 8

## Estimands

### 8.1 ICH E9 (R1)

In 2019, an addendum to the ICH E9 ‘Statistical Principles for Clinical Trials’ guideline was published that dealt with the topic of estimands (ICH E9 (R1)). The concept of an estimand is not new within the statistical community, but within the regulatory environment it signals a major shift in the way we view the clinical questions that are posed and how we estimate treatment effects that connect with those clinical questions.

***ICH E9(R1) (2019): ‘Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials’***

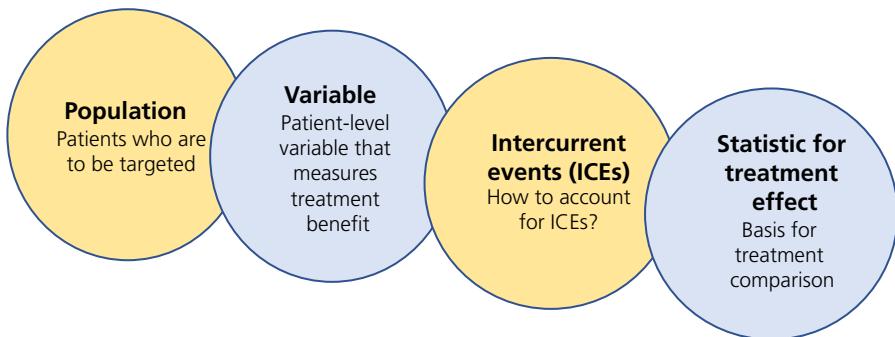
*‘This addendum presents a structured framework to link trial objectives to a suitable trial design and tools for estimation and hypothesis testing. This framework introduces the concept of an estimand, translating the trial objective into a precise definition of the treatment effect that is to be estimated’.*

One key element of this framework concerns missing data and how these are dealt with to allow alignment with the clinical question(s). A further aspect is a shift away from consideration of the randomised population of subjects as the only population for which valid inferences are possible. The framework also allows a more structured consideration of sensitivity analyses and their role.

We will begin this chapter by setting up the structure for the construction of an estimand and move on to develop the various elements of that structure in detail.

### 8.2 Attributes of an estimand

An *estimand* consists of four attributes: population, variable, intercurrent event(s) and statistic for treatment effect, as displayed in Figure 8.1. Much of what is contained within these attributes is not new, but setting out these elements this way enables each component to be considered in turn so that they link appropriately with the clinical question. A further component that is not



**Figure 8.1** Components of an estimand

displayed in Figure 8.1 but also requires consideration concerns which treatments/interventions are being compared. For example, in a placebo-controlled setting evaluating drug A on top of standard of care, changes in the standard of care in light of a patient's response to the treatment regimen could be considered either part of, or outside of, the treatments being compared. Which of these two situations we are in will determine certain aspects of how we define our estimands.

### 8.2.1 Population

This is the population of subjects that forms the basis for the clinical question. In many cases, this is self-evident and defined according to the inclusion/exclusion criteria in the protocol. Alternatively, it could be a predefined subgroup of the population defined according to certain baseline characteristics. However, the estimand framework also allows the possibility, through the principle stratification strategy, to consider a subset of subjects within the population who satisfy certain conditions relating to events that occur post-baseline: for example, those who would comply with treatment or who would be able to tolerate the treatment. This kind of subsetting will clearly cause problems in terms of intention-to-treat (ITT) and the comparison of randomised groups, but these problems can be overcome to a degree if we are prepared to make some assumptions and correct for the lack of alignment with the randomisation. This aspect will be discussed further in Section 8.3.4.

### 8.2.2 Variable

This is the subject/patient level variable that is to be the focus of the clinical question: for example, the change from baseline in total cholesterol, time to death or number of epileptic seizures in a 12-month period. Note that, in line with the way we defined an *endpoint* in Section 1.12, this is defined at the subject/patient level and not at the group level.

### 8.2.3 Intercurrent event (ICE)

Explicit consideration of intercurrent events is the aspect of estimands that is new. These are events that occur post-randomisation that can lead to missing data or data being collected following the ICE that are not relevant to the clinical question of interest. Examples of ICEs include taking rescue medication, withdrawal from treatment due to an adverse event (AE), death and so on. In consideration of the ICE, we need to think through how the data (if any) following the ICE are to be incorporated (if at all) into the definition and evaluation of the estimand. Also, if the data following the ICE are either missing or thought not to be relevant, we need to use methods that will deal with those missing/non-relevant data. Coming back to the earlier discussion relating to the treatments/interventions being compared, if changes in background therapy are considered as part of the treatments being compared, these should not be considered as ICEs. However, if those changes are not considered as part of the treatments being compared, they should be viewed as possible ICEs.

### 8.2.4 Statistic for treatment effect

This is the summary statistic to be used to measure the treatment effect or treatment difference. Examples here would include the difference between the mean change from baseline for the experimental group and the mean change from baseline in the control group, the odds ratio, or the hazard ratio.

To address each clinical question in any particular case, these four attributes need to be brought together to define the estimand that addresses that question.

## 8.3 Estimand strategies

### 8.3.1 Five strategies

The ICH E9 (R1) guideline proposes five estimand strategies that can capture all the usual clinical questions of interest in a clinical trial. We will cover each of these in turn.

We will discuss these strategies in the context of a placebo-controlled trial in diabetes. The approaches discussed in relation to this example are broadly in line with the proposals set down in the CHMP (2018) guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus. See also Aroda et al. (2019) for further discussion of this example.

The population of interest in the example is the population of patients with type 2 diabetes, where the primary endpoint is based on HbA1c over the 26-week (6-month) treatment period. The two intercurrent events of interest are taking rescue medication allowed by protocol and withdrawal from treatment due to AEs. The summary statistic for treatment difference depends on the precise definition of the primary endpoint/estimand but can be based either on a continuous measure, change from baseline in HbA1c, or on a responder endpoint defined in terms of achieving HbA1c < 7%.

### 8.3.2 Treatment policy, composite and hypothetical strategies

The *treatment policy strategy* essentially ignores the ICE and uses all data irrespective of the occurrence of the ICE. This aligns very much with our earlier considerations with regard to ITT: all patients, all data.

The *composite strategy* builds the ICE into the variable itself. Often, but not always (see Example 8.1), a composite strategy involves defining a binary endpoint (success/failure) according to a certain level of improvement, with a patient who suffers the ICE counted as a ‘treatment failure’ irrespective of the data collected after the occurrence of the ICE. This approach is in line with what was previously discussed in Section 7.3.5, where we introduced the idea of a simple success/failure classification.

The *hypothetical strategy* attempts to address a question that concerns the pure treatment effect: the treatment effect had the ICEs not occurred. It is important that any hypothetical strategy is realistic from a clinical perspective. For the diabetes example, it could be entirely reasonable to consider what would have happened had rescue treatment not been available but somewhat unrealistic to consider what would have happened had the patient not suffered the AE. The toxicity of the treatment is a feature of the treatment and cannot be divorced from it.

**Example 8.1** Randomised placebo-controlled trial in treatment of COVID-19 infection

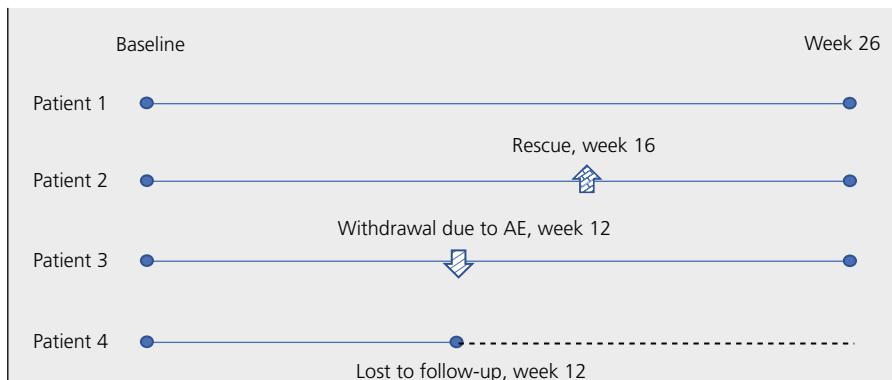
Beigel et al. (2020) report on the use of intravenous Remdesivir in the treatment of adults hospitalised with COVID-19 with evidence of lower respiratory tract infection. The primary variable of interest was the time from entry into the intensive care unit (ICU) to recovery within the first 29 days, where recovery was defined as points 1, 2 or 3 on the WHO 8-point clinical status scale (1 = not hospitalised with no limitation of activities; 2 = not hospitalised, with limitation of activities, home oxygen requirement, or both; through to 8 = death).

The intercurrent event in this case was death, and a censored observation equal to 29 days was assigned to patients who died within the 29-day period (see Chapter 13 for a discussion of the analysis of time-to-event endpoints and censoring). This is giving the worst possible outcome to those patients who die in this early period of time and is an example of a composite strategy where the ICE is directly built into the variable.

**Example 8.2** Randomised placebo-controlled trial in Type 2 diabetes

The ICEs of interest in this example are (i) taking rescue medication and (ii) withdrawing from treatment due to an AE. Consider four patients in the 26-week trial, as shown in Figure 8.2:

- Patient 1 completes the trial without taking rescue medication or suffering an AE.
- Patient 2 takes rescue medication from week 16 onwards.



**Figure 8.2** Diabetes trial, four hypothetical patients

- Patient 3 suffers an AE at week 12 and withdraws from the medication (but remains in the trial).
- Patient 4 is lost to follow-up at week 12.

We will consider two possible estimands. Estimand 1 is based around using a hypothetical strategy for taking rescue medication but a treatment policy strategy for a patient who withdraws from medication due to an AE. This estimand is as recommended by the CHMP 2018 guideline on diabetes. Estimand 2 uses a composite strategy for the ICEs, both considered failure of the treatment. Treatment success/failure outside of that is defined as an absolute HbA1c < 7.0%/≥ 7.0% at week 26. Table 8.1 sets out the main elements of these two estimands.

The treatment policy strategy for withdrawal from treatment due to an AE for estimand 1 uses the observed data on HbA1c at week 26. The hypothetical strategy for rescue medication for estimand 1 will ignore the data on HbA1c following that ICE and impute values based on a missing not at random (MNAR) assumption. In this context, we can assume that if rescue medication had not been available, the patient would have behaved like a control group patient receiving placebo (see Section 7.3.10) since receiving rescue medication is signalling that the assigned study treatment is not effective in maintaining glycaemic control.

The final element in relation to the analysis of these data is how we would deal with missing data for patients who are lost to follow-up: for example, patient 4 in the example. Multiple imputation can provide predicted outcomes for change in baseline in HbA1c under the assumption of MAR for such patients. For estimand 2, the predicted values at week 26 would be classified according to HbA1c <7% or ≥7% to provide a complete data set for analysis.

Estimand 1 answers the following clinical question: 'What is the treatment effect in the targeted population of patients with type 2 diabetes had rescue medication not been available?'

Estimand 2 answers the following clinical question: 'What is the treatment effect in the targeted population of patients with type 2 diabetes where taking rescue medication and discontinuation of trial product are considered as failure of the treatment?'

**Table 8.1** Estimand definitions for the diabetes example

<b>Estimand 1</b>	Population	Patients with type 2 diabetes
	Variable	Change from baseline in HbA1c to week 26
	ICE	Taking rescue medication
	Strategy	Hypothetical
	Data	Impute data on HbA1c following ICE
	ICE	Withdrawal from treatment due to AE
	Strategy	Treatment policy
	Data	Use observed data on HbA1c following ICE
	Summary statistic for treatment effect	Difference in LS mean change from baseline in HbA1c
	Method of statistical analysis	Analysis of covariance with HbA1c at baseline as a covariate
<b>Estimand 2</b>	Population	Patients with type 2 diabetes
	Variable	Success = HbA1c < 7.0% at week 26
	ICE	Taking rescue medication
	Strategy	Composite
	Data	Patient considered treatment failure
	ICE	Withdrawal from treatment due to AE
	Strategy	Composite
	Data	Patient considered treatment failure
	Summary statistic for treatment effect	Odds ratio for treatment success
	Method of statistical analysis	Logistic regression with HbA1c at baseline as a covariate

### 8.3.3 While on treatment

In this strategy, the outcomes are measured only up to the ICE of interest. For example, the focus could be the incidence of AEs while the patient is receiving treatment or up to 14 days following discontinuation of treatment. In this case, discontinuation of treatment (or discontinuation of treatment + 14 days) is the ICE. In a study involving a terminal illness where the ICE is death, interest may be around measures of quality of life prior to death.

### 8.3.4 Principal stratification

There are many potential applications of this strategy. Much of the theory is outside the scope of this book, but we will provide a flavour of the approach by considering one example. We are again considering the diabetes example, and the ICE in this case is ‘taking rescue medication’. Assume that we are interested in the treatment effect in the stratum of patients who require rescue medication on placebo (drug B) but do not require rescue on drug A (experimental). This is the stratum of patients for whom the experimental drug is effective and required to provide glycaemic control.

When receiving either drug A or drug B, some patients will require rescue and others will not. Consider patients divided into four strata: those who would not require rescue, whichever treatment they were assigned to; those who would require rescue on drug A but not on drug B; those who would require rescue on drug B but not on drug A; and those who would require rescue whichever treatment group they were assigned to. If we label these groups of patients respectively as  $S_{00}$ ,  $S_{01}$ ,  $S_{10}$  and  $S_{11}$ , the structure is as in Figure 8.3.

Assume further that the proportions of patients in each of these strata are, respectively,  $p_{00}$ ,  $p_{01}$ ,  $p_{10}$  and  $p_{11}$ . Of course, the observed values for these quantities are not known in practice since patients are only ever observed in one of the treatment groups. These hypothetical proportions rely on what are called *counterfactual* arguments: what would have happened had a particular patient who was randomised to drug A been alternatively randomised to drug B, and vice versa. However, we do know the proportions of patients in our study who (i) require/do not require rescue medication on drug A ( $p_A$ ,  $1 - p_A$ ) and (ii) require/do not require rescue medication on drug B ( $p_B$ ,  $1 - p_B$ ). Each of these proportions estimates the sum of pairs of the proportions  $p_{00}$ ,  $p_{01}$ ,  $p_{10}$  and  $p_{11}$  and, in particular,

$$p_A = p_{01} + p_{11}, 1 - p_A = p_{00} + p_{10}, p_B = p_{10} + p_{11}, 1 - p_B = p_{00} - p_{01}$$

Now comes the first key assumption that enables us to obtain the counterfactual proportions from the observed proportions: that a patient who would not require rescue on placebo (drug B) would also not require rescue on the experimental drug (drug A), so that  $p_{01} = 0$ . Under this assumption, the remaining proportions in the table can be calculated as follows:

$$p_{00} = 1 - p_B, p_{10} = p_B - p_A, p_{11} = p_A$$

Let's turn our attention to the endpoint we are evaluating and consider an endpoint measured on a continuous scale, such as change from baseline in HbA1c to week 26. We need to make two additional assumptions, and these now involve the outcomes themselves:

**Assumption 1:** The true treatment difference among the stratum of patients who require rescue medication on both drug A and drug B (placebo) is zero.

		Drug A	
		No rescue	Rescue
Drug B	No rescue	$S_{00}$ , $p_{00}$	$S_{01}$ , $p_{01}$
	Rescue	$S_{10}$ , $p_{10}$	$S_{11}$ , $p_{11}$

$S_{00}$ ,  $S_{01}$ ,  $S_{10}$ ,  $S_{11}$  are labels to identify the cells.  $p_{00}$ ,  $p_{01}$ ,  $p_{10}$ ,  $p_{11}$  are the population proportions in those cells.

**Figure 8.3** Patient strata

Assumption 2: The true treatment difference among the stratum of patients who do not require rescue medication on both drug A and drug B is also zero.

These assumptions are based on the notion that if rescue is required on drug, the drug is not working and providing acceptable glycaemic control (Assumption 1) and is not required on placebo, placebo is providing adequate glycaemic control (Assumption 2).

Now let  $Y_A - Y_B$  represent the true treatment difference. The overall treatment difference is made up of the four differences, one for each cell in the table, weighted by the proportions of patients in each of those cells:

$$Y_A - Y_B = (Y_A - Y_B)_{00} \times p_{00} + (Y_A - Y_B)_{01} \times p_{01} + (Y_A - Y_B)_{10} \times p_{10} + (Y_A - Y_B)_{11} \times p_{11}$$

Several terms in this equation take the value zero because of the assumptions we have made. In particular,  $p_{01}$ ,  $(Y_A - Y_B)_{00}$  and  $(Y_A - Y_B)_{11}$  are all equal to zero. It then follows that

$$Y_A - Y_B = (Y_A - Y_B)_{10} \times p_{10}$$

$(Y_A - Y_B)_{10}$  is precisely the quantity we want to estimate: the treatment difference in the stratum of patients for whom the experimental drug is effective and required to provide glycaemic control. If  $d$  is the estimated difference  $Y_A - Y_B$  based on a suitable model applied to the ITT population,  $(Y_A - Y_B)_{10} = d/p_{10}$ . Standard errors for this estimate of treatment effect in our principal stratum can be calculated, and a 95% confidence interval together with the  $p$ -value to assess statistical significance then follow. See Mallinckrodt et al. (2020), Chapter 24, for further details on these and similar considerations in other settings.

## 8.4 Sensitivity and supplementary analyses

### 8.4.1 Main estimator

The objectives of the clinical trial will lead to one or more estimands that will be the basis of the primary and key secondary analyses. Each estimand will have an associated *main estimator*, which will provide a *main estimate* of the treatment effect.

#### **ICH E9(R1) (2019): 'Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials'**

*'An estimand for the effect of treatment relative to a control will be estimated by comparing the outcomes in a group of subjects on the treatment to those in a similar group of subjects on the control. For a given estimand, an aligned method of analysis, or estimator, should be implemented that is able to provide an estimate on which reliable interpretation can be based'.*

### 8.4.2 Sensitivity analyses

Various assumptions will have been made regarding missing data and replacing observed data that is deemed to be not relevant for the clinical question being addressed by that estimand. Further assumptions for the statistical model (normality of residuals, covariates relevant for inclusion in the model and so on) will also have been made. These assumptions should be the most plausible given the setting and the requirements imposed by the clinical question. The main estimate together with the associated *p*-value and confidence interval will lead to conclusions regarding the clinical question. Those conclusions, however, should then be assessed for robustness by changing the methods that deal with features of the data such as missingness and outliers.

***ICH E9(R1) (2019): 'Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials'***

*'Inferences based on a particular estimand should be robust to limitations in the data and deviations from the assumptions used in the statistical model for the main estimator. This robustness is evaluated through a sensitivity analysis. Sensitivity analysis should be planned for the main estimators of all estimands that will be important for regulatory decision making and labelling in the product information'.*

The estimand may, for example, be based on a hypothetical strategy that has used a copy reference approach (see Section 7.3.10) to impute outcomes for patients in the active treatment group following the ICE, taking rescue medication. The main estimator will likely be expressed as the difference in least squares (LS) mean changes from baseline based on an analysis of covariance. Sensitivity estimator 1 may consider an alternative assumption regarding the imputation method using, for example, a delta method (see Section 7.3.10), where the delta is varied to see how sensitive the main estimate is to changes of this kind. If there are large numbers of patients who have taken rescue medication, the impact of changing these assumptions could be substantial. Sensitivity estimator 2 may alternatively look at adding to the statistical model covariates that are known to be predictive of outcome, to assess the sensitivity of the main estimate to imbalances in baseline covariates.

Prior to the introduction of the estimand framework within the regulatory setting, it was common practice in a superiority trial to base the primary analysis on the full analysis set (FAS) and a sensitivity analysis on the per-protocol set (PPS). Indeed, the ICH E9 guideline (Section 5.2.2) explicitly states that this should be done. However, it was recognised that analysis based on the PPS is subject to bias because of lack of alignment with the principle of ITT. Analysis based on PPS has fallen out of favour and has been replaced with a more structured view of what constitutes a sensitivity analysis in the context of estimands.

**ICH E9(R1) (2019): 'Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials'**

'Estimands might be constructed, with aligned method of analysis, that better address the objective usually associated with the analysis of the PPS. If so, analysis of the PPS might not add additional insights'.

### **8.4.3 Supplementary analyses**

Supplementary analyses can be undertaken in addition to sensitivity analyses to target possibly different estimands or target the main estimand but using a different analytic technique.

**ICH E9(R1) (2019): 'Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials'**

'Supplementary Analysis:

*A general description for analyses that are conducted in addition to the main and sensitivity analysis with the intent to provide additional insights into the understanding of the treatment effect'.*

The distinction between a sensitivity analysis and a supplementary analysis is best illustrated through examples.

In our earlier discussion of the trial in diabetes, we considered two separate estimands. Estimand 1 was based on the change from baseline in HbA1c, while estimand 2 looked at the binary outcome defined in terms of HbA1c at week 26  $</\geq 7.0\%$ . Estimand 1 would generally be considered as providing the most clinically relevant measure of treatment benefit. Estimand 2 can be considered as a supplementary analysis and provides a responder/non-responder view of the treatment effect.

Removing outliers and repeating the ANCOVA analysis, would constitute a sensitivity analysis. Replacing the ANCOVA model with a non-parametric approach where the treatment difference is expressed as the difference between two medians would constitute a supplementary sensitivity analysis; it is a different analytic approach for estimating the treatment effect.

In any setting, sensitivity analyses will always be needed since every statistical method of analysis involves assumptions, and the robustness of results and conclusions given changes in those assumptions needs to be evaluated. However, supplementary analyses are not always needed: for example, where the estimand is clearly structured in the most reliable way to answer the clinical question with little to be gained from consideration of a separate estimand or a different analytical method for analysis.

## CHAPTER 9

# Power, sample size and clinical relevance

### 9.1 Type I and type II errors

Unfortunately, the statistical test procedures we use are not perfect, and from time to time we will be fooled by the data and draw incorrect conclusions. For example, we know that 17 heads and 3 tails can (and will) occur with 20 flips of a fair coin (the probability from Section 3.3.2 is 0.0011); however, that outcome would give a statistically significant  $p$ -value, and we would conclude incorrectly that the coin was not fair. Conversely, we could construct a coin that was biased 60%/40% in favour of heads and in 20 flips; see, for example, 13 heads and 7 tails. That outcome would lead to a non-significant  $p$ -value ( $p = 0.224$ ) for the null hypothesis that the coin was fair, and we would fail to pick up the bias. These two potential mistakes are termed type I and type II errors.

To explain in a little more detail, consider a parallel-group trial in which we are comparing two treatment means using the unpaired t-test. The null hypothesis  $H_0: \mu_1 = \mu_2$  that the treatment means are equal is either true or not true; God knows, we don't! We mere mortals have to make do with data, and on the basis of data, we will see either a significant  $p$ -value ( $p \leq 0.05$ ) or a non-significant  $p$ -value ( $p = \text{NS}^*$ ). The various possibilities are contained in Table 9.1.

Suppose the truth is that  $\mu_1 = \mu_2$ ; the treatment means are the same. We would hope that the data would give a non-significant  $p$ -value and our conclusion would be correct; we are unable to conclude that differences exist. Unfortunately, that does not always occur, however; on some occasions, we will be hoodwinked by the data and get  $p \leq 0.05$ . On that basis, we will declare statistical significance and draw the conclusion that the treatment means are different. This mistake is called the *type I error*. It is the *false positive*, sometimes referred to as the  $\alpha$  error.

Conversely, suppose that in reality,  $\mu_1 \neq \mu_2$ ; the treatment means are truly different. In this case, we would hope that  $p \leq 0.05$ , in which case our conclusion

---

\* $p = \text{NS}$  is shorthand to say that  $p$  is not statistically significant at the 5% level. Its use in reporting trial results is not recommended; exact  $p$ -values should be used but is used here for convenience.

**Table 9.1** Type I and type II errors

	$H_0$ true, $\mu_1 = \mu_2$	$H_0$ not true, $\mu_1 \neq \mu_2$
Data gives $p = \text{NS}$ (cannot conclude $\mu_1 \neq \mu_2$ )	✓	✗
Data gives $p \leq 0.05$ (conclude $\mu_1 \neq \mu_2$ )	✗	✓

will be the correct one: the evidence supports treatment differences. Again, this will not always happen, and there will be occasions when, under these circumstances, we get  $p = \text{NS}^*$ , a non-significant  $p$ -value. On this basis, we will say that we do not have enough evidence to conclude differences. This second potential mistake is called the *type II error*. This is the *false negative* or the  $\beta$  error: the treatment means really are different, but we have failed to pick that up!

There is a well-known theorem in statistics, called the Neyman–Pearson Lemma, which shows that for a given sample size, it is simply not possible to eliminate the potential for these two mistakes; we must always trade them off against each other. Usually, the type I error is fixed at 0.05 (5%). This is because we use 5% as the significance level, the cut-off between statistical significance ( $p \leq 0.05$ ) and non-significance ( $p > 0.05$ ). The null distribution tells us precisely what will happen when the null hypothesis is true; we *will* get extreme values in the tails of that distribution, even when  $\mu_1 = \mu_2$ , some of the time. However, when we do see a value in the extreme outer 5%, we declare significant differences; and, by definition, this will occur 5% of the time when  $H_0$  is true.

The type II error is a little more difficult to pin down. It is related to another quantity called *power*. If type II error is 10%, then power is 90%; *power* is 100 minus type II error. Type II error is missing a real difference – power is capturing a real difference; if there is a 10% chance of missing the bus, there is a 90% chance of catching the bus, and they are opposites in this sense! We control type II error by controlling power; for example, we may design our trial to have 80% power, in which case the type II error is controlled at 20%.

## 9.2 Power

As seen in the previous section, power measures our ability to detect treatment differences. A convenient mathematical way of thinking about power is

$$\text{power} = \text{probability } (p \leq 0.05)$$

When we say that a trial has 80% power to detect a certain level of effect, for example, 4 mmHg, what we mean is that if the true difference really is 4 mmHg and we conduct the trial, there is an 80% chance of coming out of the trial with a significant  $p$ -value and declaring differences. In other words, if we were to run this same trial 10 times, then on 8 of those occasions, on average, we would get

a statistically significant result at the 5% level, and on 2 occasions, on average, we would get a non-significant result.

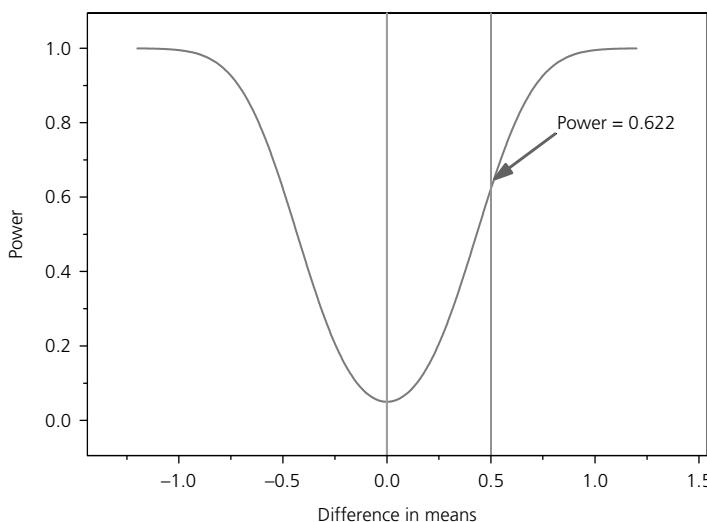
We can calculate power in advance of running the trial by speculating about what may happen. Assume that in a parallel-group cholesterol-lowering study comparing a test treatment with placebo, there are 50 patients per group. The unpaired t-test will be used to compare the mean reduction in total cholesterol between the groups at the conventional two-sided significance level of 0.05. Assume also that the standard deviation for the reduction in total cholesterol is 1.09 mmol/l. For various values for the treatment difference, the calculated power is given in Table 9.2.

So, for example, if the true difference between the treatment means was 0.50 mmol/l, this trial would have a 62.2% chance of coming out with a significant  $p$ -value ( $p \leq 0.05$ ). Similarly, if the true difference were 0.75 mmol/l, the chance of getting a statistically significant result would be 92.6%.

Figure 9.1 plots the values for power against the true difference in the treatment means. Certain patterns emerge. Power increases with the magnitude of the treatment difference, large differences give high values for power, and the

**Table 9.2** Power for various treatment differences,  
 $n = 50$  per group

Treatment difference, $\mu_1 - \mu_2$	Power
0.25	0.206
0.50	0.622
0.75	0.926
1.00	0.995



**Figure 9.1** Power curve for  $n = 50$  per group

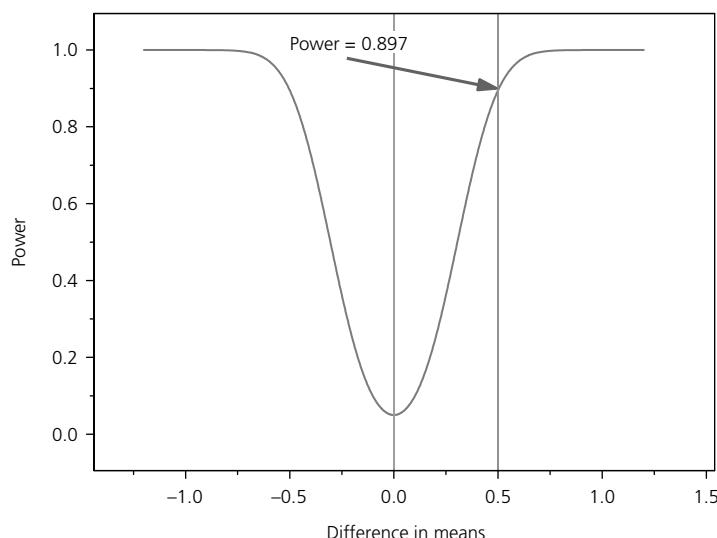
value for power approaches one as the treatment difference increases in either a negative or a positive direction. The implication here is that large differences are easy to detect, and small differences are more difficult to detect. The power curve is symmetric about zero, and this is because our test is a two-sided test; a difference of +1 mmol/l has just the same power as a difference of -1 mmol/l.

Suppose now that the trial in the example were a trial in which a difference of 0.5 mmol/l was viewed as an important difference. Maybe this reflects the clinical relevance of such a difference, or perhaps from a commercial standpoint; it would be a worthwhile difference to have for a new drug in the marketplace. Under these circumstances, only having 62.2% power to detect such a difference would be unacceptable; this corresponds to a 37.8% type II error, an almost 40% chance of failing to declare significant differences if 0.5 mmol/l were the true treatment effect. Well, there is only one thing you can do, and that is to increase the sample size. The recalculated values for power are given in Table 9.3, with a doubling of the sample size to 100 patients per group.

The power to detect a difference of 0.5 mmol/l is now 89.7%, a substantial improvement on 62.2%. Figure 9.2 shows the power curve for a sample size of

**Table 9.3** Power for  $n = 100$  per group

Treatment difference, $\mu_1 - \mu_2$	Power
0.25	0.365
0.50	0.897
0.75	0.998
1.00	1.000



**Figure 9.2** Power curve for  $n = 100$  per group

100 patients per group, and as can be seen, the values for power have increased across all the potential values for the true treatment difference. These arguments form the basis of the sample size calculation; we think in terms of what level of effect it is important to detect, either from a clinical, regulatory, or commercial perspective, and choose the sample size to give high power for detecting such an effect. In our example, if we had said that we require 80% power to detect a difference of 0.50 mmol/l, then a sample size of 76 per group would have given us exactly that. For 90% power, we would need 101 patients per group.

Before moving on to discuss sample size calculations in more detail, it is worth noticing that the power curve does not come down to zero at a difference of 0.0, and the curve crosses the  $y$ -axis at the significance level, 0.05. Recall that power can be thought of as the probability ( $p \leq 0.05$ ). Even when the treatments are identical (difference = 0.0), there is still a 0.05 chance of getting a significant  $p$ -value, and this is the type I error and the reason why the power curve cuts at this value. This issue tells us what happens if we want to change the significance level, for example, from 0.05 to 0.01. (We sometimes do this when dealing with multiplicity, and we will look in more detail at this issue in Chapter 10.) Reducing the significance level will pull down the power curve so that it crosses at 0.01, and the effect of this will be to reduce all the power values. Even when there are true treatment differences, achieving  $p \leq 0.01$  is much more difficult than achieving  $p \leq 0.05$ , so the power comes down. In practice, of course, we may need to consider increasing the sample size to compensate for this reduction in the significance level to recover the required power.

### 9.3 Calculating sample size

Once the requirements of a trial have been specified, then calculating sample size is straightforward, and formulas exist for all the commonly occurring situations.

In all cases, we need to specify the required values of the type I error and the power. Usually, we will be setting the type I error at 5%, and the recommended minimum value for power is 80%, although for important trials, 80% is not enough, and 90% at least is recommended.

The remaining quantities that need to be considered when calculating sample size depend upon the statistical test to be used:

- For the unpaired t-test, we need to specify the standard deviation,  $\sigma$ , for the primary endpoint and the level of effect,  $d$ , we are looking to detect with, say, 90% power.

There is usually an implicit assumption in this calculation that the standard deviations are the same in each of the treatment groups. In general, this assumption is a reasonable one to make as the effect of treatment will usually be to change

the mean without affecting the subject-to-subject variability. We will say a little more in a later section about dealing, at the analysis stage, with situations where this is not the case. The sample size calculation, however, is also easily modified, if needed, to allow unequal standard deviations.

- For the paired t-test, the standard deviation of the within-subject differences for the primary endpoint needs to be specified, and again, the level of effect to be detected.
- For the  $\chi^2$  test, we need to know the success/event rate in the control group and as usual some measure of the treatment difference we are looking to detect.

We commonly refer to the level of effect to be detected as the *clinically relevant difference (crd)*; what level of effect is an important effect from a clinical standpoint. Note also that crd stands for *commercially relevant difference*; it could well be that the decision is based on commercial considerations. Finally, crd stands for *cynically relevant difference*! It does happen from time to time that a statistician is asked to ‘do a sample size calculation, oh, and by the way, we want 200 patients!’ The issue here of course is budget and feasibility, and the question really is, what level of effect are we able to detect with a sample size of 200?

The standard deviation referred to earlier sometimes provides the biggest challenge. The information for this will come from previous data; for that same endpoint, from a similar population/sample of subjects, same treatment

#### **Example 9.1** Unpaired t-test

In a placebo-controlled hypertension trial, the primary endpoint is the fall in diastolic blood pressure. It is required to detect a clinically relevant difference of 8 mmHg in a 5% level test. Historical data suggests that  $\sigma = 10$  mmHg. Table 9.4 provides sample sizes for various levels of power and differences around 8 mmHg; the sample sizes are per group.

For 90% power, 33 patients per group are required to detect a difference of 8 mmHg. Smaller differences are more difficult to detect, and 59 patients per group are needed to have 90% power to detect a difference of 6 mmHg. Lowering the power from 90% to 80% reduces the sample size requirement by just over 25%.

**Table 9.4** Sample sizes per group

crd	Power		
	80%	85%	90%
6 mmHg	44	50	59
8 mmHg	24	29	33
10 mmHg	16	18	22

**Example 9.2**  $\chi^2$  test

In a parallel-group, placebo-controlled trial in acute stroke, the primary endpoint is success on the Barthel index at month 3. Previous data suggest that the success rate on placebo will be 35%, and it is required to detect an improvement in the active treatment group to 50%. How many patients are needed for 90% power in a 5% level test?

For 90% power, 227 patients per group are needed. For 80% power, the sample size reduces to 170 patients per group. If the success rate in the placebo group, however, were to be 40% and not 35%, then the sample size requirements per group would increase to 519 for 90% power and to 388 for 80% power to detect an improvement up to 50% in the active group.

duration and so on. Similar comments apply for the success/event rate in the control group for a binary endpoint. We should try and match as closely as possible the conditions of the historical data to those pertaining to the trial being planned.

Machin et al. (2011) provide extensive tables in relation to sample size calculations and include in their book formulas and many examples. In addition, there are several software packages specifically designed to perform power and sample size calculations, namely nQuery ([www.statsol.ie](http://www.statsol.ie)) and PASS ([www.ncss.com](http://www.ncss.com)). For those with access to SAS®, O'Brien and Castelloe (2010) give details on the use of that package for sample size calculations.

It is generally true that sample size calculations are undertaken based on simple test procedures, such as the unpaired t-test and the  $\chi^2$  test. In dealing with both continuous and binary endpoints, it is likely that the primary analysis will ultimately be based on adjusting for important baseline prognostic factors. Usually, such analyses will give higher power than the simple alternatives. These more complex methods of analysis, however, are not usually taken into account in the sample size calculation for two reasons. Firstly, it would be very complicated to do so and would involve specifying the precise nature of the dependence of the endpoint on the factors to be adjusted for and knowledge regarding how those baseline factors will be distributed within the target population. Secondly, using the simple approach is a conservative approach as, in general, the more complex methods of analysis that we end up using will lead to an increase in power.

There is often an allowance built into the sample size calculation for drop-outs. Suppose for example that the sample size calculation suggests we require 180 patients for 80% power but also that we expect 10% of patients to drop out. Increasing the required sample size to 200 will 'allow for the drop-outs', and removing 10%, that is 20 patients from those randomised, will still leave the 180 needed for 80% power. This adjustment is oversimplistic. The focus for

the primary analysis will be all patients randomised or something close to that through the full analysis set, and it is not as simple as ‘removing’ the 20 patients who withdraw from the analysis. Also, considerations of estimands will require us to think through how those 20 patients are going to be included in the analysis. For example, with a binary outcome (responder/non-responder), will we adopt the composite strategy and consider the drop-outs as non-responders? Or will we use a hypothetical strategy estimand where the outcome for the drop-outs will be predicted through some multiple imputation methodology? The estimand approach we adopt will determine what will happen to the drop-outs and needs to be built into the assumptions we will be making for the responder rates, for example, rather than adopting a simple ‘factoring up for the drop-outs’. Assumptions regarding the event rate in the control group in the case of a binary endpoint, and the standard deviation for a continuous endpoint, will usually be based on historical data. It is important when extracting those historical data that the analysis set on which they have been based, and the methods for handling missing data, are taken into account to align with the methods of analysis that are to be used in the current trial. For example, the variability in a continuous outcome measure across a full analysis set may well be larger than it is in a per-protocol type analysis.

Finally, note that in our considerations, we have worked with groups of equal size. It is straightforward to adapt the calculations for unequal randomisation schemes, and the computer packages mentioned earlier can deal with these. Altman (1991) (Section 15.3) provides a simple method for adapting the standard sample size calculation to unequal group sizes as follows. If  $N$  is the calculated sample size based on an equal randomisation, and  $k$  represents the ratio of the number of patients in one treatment group compared to the other treatment group, then the required number of patients for a  $k$  to 1 randomisation is

$$N' = N \frac{(1+k)^2}{4k}$$

So, for example, if a 2 to 1 randomisation is required and 200 patients would have been needed for a 1 to 1 allocation, then the revised sample size is

$$N' = 200 \times \frac{9}{8} = 225$$

This is a modest increase. In general, a 2 to 1 randomisation will lead to a 12.5% increase in sample size compared to 1 to 1; a 3 to 1 randomisation would lead to a 33.3% increase.

## 9.4 Impact of changing the parameters

### 9.4.1 Standard deviation

It is interesting to see the impact of a change in the standard deviation on the required sample size. Consider the example from the previous section where we were looking to detect a treatment effect of 8 mmHg with a standard deviation of 10 mmHg. For 90% power, the total sample size requirement was 66 patients. If the standard deviation was not 10 mmHg but 20 mmHg, then the required sample size would be 264. A doubling of the standard deviation has led to a fourfold increase in the sample size. The formula for sample size contains not the standard deviation by itself but the variance (=standard deviation squared), and this is what drives this increase. Even a modest increase in the standard deviation, say, from 10 mmHg to 12 mmHg, would require 96 patients in total compared to 66.

There are several implications of this sensitivity of sample size on the standard deviation:

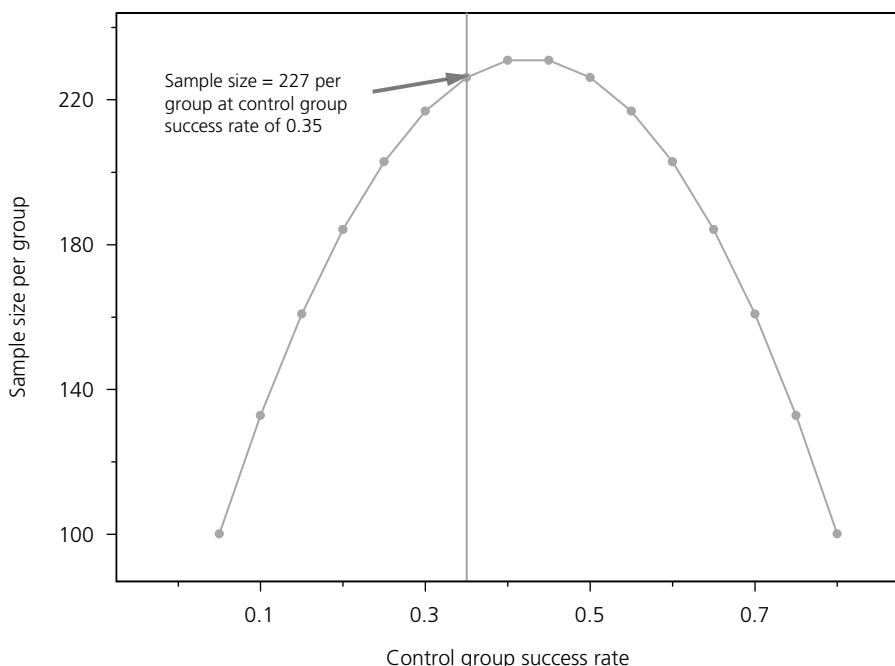
- Good information is needed for the standard deviation,  $\sigma$ ; if you get it slightly wrong, you could be severely underpowered. Be realistic and, if anything, conservative.
- Work hard to control the subject-to-subject variability, which not only depends on in-built subject differences but also on extraneous variability caused by an inconsistent measurement technique, data recording and sloppy methodology. Tightening up on these things will over time bring  $\sigma$  down and help to keep sample sizes lower than they would otherwise be.

### 9.4.2 Event rate in the control group

Again, referring to an example in the previous section where the event was success on the Barthel index at month 3, we had an event rate in the control group of 35%, and we were looking to detect an improvement of 15% in absolute terms to 50%. A sample size of 227 per group gave 90% power. Figure 9.3 illustrates how this sample size depends upon the success rate in the control group; note that we are looking in each case for an absolute 15% improvement.

The curve is symmetric around a rate of 0.425 (halfway between 0.35 and 0.50) since we are undertaking two-tailed tests, and changing the labels for success and failure will simply repackage the same calculation. For example, comparing 30% to 45% produces the same sample size as comparing 70 ( $= 100 - 30$ )% to 55 ( $= 100 - 45$ )%. The sample sizes are much reduced as the success rates move either down towards 0% or up towards 100%. In those regions, of course, the relative changes in either the success rate or the failure rate are large, and it is this that impacts the calculation.

One point worth making here is that information on event rates in the control group will inevitably come from studies that took place some time ago. As general patient care is improving in many circumstances over time, these



**Figure 9.3** Sample size for detecting absolute improvement of 0.15 in success rates

historical rates may not be reflective of what will happen in the future for the trial that is being planned. Such historical rates, when we are looking at treatment success, for example, could well be underestimates for the current trial. This could lead to the situation where the actual difference in the rates between the control and experimental groups is less than what was assumed in the sample size calculation and a study that is consequently underpowered.

#### 9.4.3 Clinically relevant difference

For a continuous endpoint, the sample size is inversely proportional to the square of the clinically relevant difference. If the crd is reduced by a factor of two, then the sample size is increased by a factor of four; if the crd is increased by a factor of two, then the sample size is reduced by a factor of four. In our earlier example, the sample size requirement to detect a difference of 8 mmHg was 33 patients per group. To detect a difference of 4 mmHg, we require 132 patients per group.

For a binary endpoint, this same relationship between the crd, in terms of the absolute difference in success rates, and the sample size is approximately true. In the example, we were looking to detect an improvement in the success rate from 35% to 50%, an absolute difference of 15%, and we needed a sample size of 227 patients per group. If we were to halve that difference and look for an improvement from 35% to 42.5%, then the sample size requirement would be 885 per group, an increase in the sample size by a factor of 3.9.

## 9.5 Regulatory aspects

### 9.5.1 Power $\geq 80\%$

The recommendation for at least 80% power comes from the ICH E9 guideline:

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed. . . The probability of type II error is conventionally set at 10% to 20%; it is in the sponsor's interest to keep this figure as low as feasible especially in the case of trials that are difficult or impossible to repeat'.*

The guideline stresses that it is *in the sponsor's interest* to have power as high as possible. Too often, researchers see the power calculation as merely something for the protocol to satisfy Ethics Committees (ECs), Institutional Review Boards (IRBs) and regulators and go for the *minimum* requirement. Also, it is tempting to choose ambitious values for the key parameters, such as the standard deviation of the primary endpoint, the event rate in the control group or the crd to produce a sample size that is comfortable from a budgeting or practical point of view, only to be disappointed once the data appear. Be realistic in the choice of these quantities and recognise that 80% power is 20% type II error, a one in five chance of failing to achieve statistical significance even if everything runs perfectly and in line with assumptions.

### 9.5.2 Sample size adjustment

It will sometimes be the case that there are gaps in our knowledge, and it will not be possible to give values for the standard deviation or for the event rate in the control group with any degree of confidence. In these circumstances, it is possible to revisit the sample size calculation once the trial is underway.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'In long term trials there will usually be an opportunity to check the assumptions which underlay the original design and sample size calculations. This may be particularly important if the trial specifications have been made on preliminary and/or uncertain information. An interim check conducted on the blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions. . . '.*

Note that this calculation must be undertaken on the blinded data – in other words, using data on the treatment groups combined. Even evaluating the data

with the two groups separated using arbitrary labels A and B (partial unblinding) would not be acceptable. If a comparison based on either partial or complete unblinding were to be made, this would introduce bias in the final analysis. We will say much more about this in two later sections dealing with interim analysis (Section 8.5.3) and adaptive designs (Section 16.1.2).

Using all data as a single group to calculate the standard deviation in the continuous case, however, will give an overestimate of the within-group standard deviation, particularly if the treatment differences are large. It must be accepted therefore that this could lead to some overpowering, although in practice, experience suggests that this is minor. For a binary endpoint, we will end up with a combined event rate, and this must be unpicked to enable a sample size recalculation to be undertaken. For example, if the original calculation is based upon detecting an increase in the success rate from 35% to 50%, then we are expecting an overall, average, success rate of 42.5%. If at the interim check this overall rate turns out to be 45%, then we could be looking for an increase from 37.5% to 42.5% if we want to retain the ability of the trial to have enough power to detect an absolute 15% improvement.

## 9.6 Reporting the sample size calculation

A detailed statement of the basis of the sample size calculation should be included in the protocol, in the Clinical Study Report (CSR) report and in publications. This statement should contain the following:

- Significance level to be used. This will usually be 5% and relate to a two-sided test or <5% depending on issues of multiplicity.
- Required power ( $\geq 80\%$ ).
- The primary endpoint on which the calculation is based, together with the statistical test procedure used as the basis for the sample size calculation.
- Estimates of the basic quantities needed for the calculation, such as the standard deviation or the event rate in the control group and the sources of those estimates.
- The clinically/commercially relevant difference (crd). If the expected difference is larger than this, then it could be worth considering powering for the expected effect, and the sample size will then, be lower.

The CONSORT statement (Schulz et al., 2010) sets down standards for the reporting of clinical trials, and their recommendations in relation to the sample size calculation are in line with these points. There may, of course, be cases, especially in the early exploratory phase, where the sample size has been chosen on purely practical or feasibility grounds. This is perfectly acceptable in that context, and the sample size section in the protocol should clearly state that this is the case.

**Example 9.3** Xamoterol in severe heart failure

Below is the sample size statement from the Xamoterol in Severe Heart Failure Study Group (1990):

*'It was estimated that 228 patients would have to complete the study to give a 90% chance of detecting a 30-second difference in exercise duration between placebo and xamoterol at the 5% level of significance. The aim was therefore to recruit at least 255 patients to allow for withdrawals. A blinded re-evaluation of the variance of the exercise data after the first 63 patients had completed the study and a higher drop-out rate (15%) than expected (10%) caused the steering committee (in agreement with the safety committee) to revise the recruitment figure to at least 450'.*

Consider each of the elements in the calculation and reporting of that calculation in turn:

- The primary endpoint on which the calculation is based in the exercise duration.
- The required power was set at 90%, a type II error of 10%.
- The type I error was set at 5%. Note in general the alternative phrases for the type I error: significance level,  $\alpha$  error and false-positive rate.
- Which statistical test was to be used for the comparison of the treatment groups in terms of the primary endpoint, do you think? This is a comparison between two independent groups in a parallel-group trial, and the primary endpoint is continuous, so the sample size calculation will undoubtedly have been based on the two-sample t-test (although this is not specified).
- The trial has been powered in terms of the completers. Recruitment was set at 255 patients to allow for withdrawals. Is this appropriate? Note the earlier comments on this issue.
- There were two reasons for increasing the sample size: a larger than expected standard deviation (variance) for the primary endpoint and a higher drop-out rate (15% compared to 10%).

Most of the elements are contained within the sample size section according to the requirements set down in the CONSORT statement; the only omissions seem to be the specification of the statistical test on which the sample size calculation was based, the assumed standard deviation of the primary endpoint and the basis for that assumption.

## 9.7 Post hoc power

Suppose a study is powered at 90% to detect a reduction in day 28 mortality from 35% (placebo) to 27% (experimental) with a two-sided significance level of 5%. The sample size requirement is 701 subjects per group. The trial is targeting an 8% absolute reduction in day 28 mortality. Suppose now that the observed death rates are 35 and 30% in the placebo and experimental groups, respectively: a 5% reduction in the mortality rate, somewhat lower than the targeted reduction. Will these data lead to a statistically significant result? The answer is yes, they will! In fact, a reduction of 4.9% or greater will yield  $p \leq 0.05$ . An

observed reduction of 8% will give  $p = 0.00094$ . These outcomes may be surprising, and there is a common misunderstanding that if you fail to achieve an observed treatment difference in line with the difference that was targeted in the sample size calculation, the result will be non-significant.

The power calculation tells us that if the true rates are 35 and 27%, there is a 9 out of 10 chance of achieving statistical significance, and this takes account of the fact that the observed values of the two mortality rates and in particular, their difference, will be distributed around the true values. Also be aware that the assumed rates are just that – assumed!

Considerations of power and sample size are for planning purposes at the design stage. Once the data become available, power becomes somewhat irrelevant, and post hoc power calculations are of little value. If a trial powered at, say, 90% fails to achieve statistical significance, there are two possible reasons:

- 1 The true treatment difference is very much lower than that targeted at the planning stage.
- 2 The true treatment difference is equal to that targeted (or even higher), but you have been unlucky, and there is a type II error. Remember, there is a 10% potential for this with 90% power.

Of course, you have no way of knowing which of these two situations you are in. All you have are the data – and speculating retrospectively about whether you were powered appropriately or not is largely fruitless. The following is a quote from Zhang et al. (2019):

*'Power analysis is an indispensable component of planning clinical research studies. However, when used to indicate power for outcomes already observed, it is not only conceptually flawed but also analytically misleading'.*

## 9.8 Link between *p*-values and confidence intervals

In Chapter 3, we developed the concepts of both the confidence interval (CI) and the *p*-value. At that stage, these ideas were kept separate. There is a close link between the two, and in this section, we will discuss that link.

Consider an application of the unpaired t-test with

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Note that the null hypothesis could be rewritten as  $H_0: \mu_1 - \mu_2 = 0$ .

The 95% CI ( $a, b$ ) for the difference in the treatment means,  $\mu_1 - \mu_2$ , provides a range of plausible values for the true treatment difference. With 95% confidence, we can say that  $\mu_1 - \mu_2$  lies somewhere within the range from  $a$  to  $b$ .

Consider the results presented in Table 9.5 for two separate trials. In trial 1, the *p*-value provides evidence that the treatment means are different. The CI also supports treatment differences, with the magnitude of that difference lying

**Table 9.5** *p*-values and confidence intervals for two trials (hypothetical)

	<b><i>p</i>-value</b>	<b>95% CI for <math>\mu_1 - \mu_2</math></b>
Trial 1	$p \leq 0.05$	(2.1 mmHg, 5.7 mmHg)
Trial 2	$p = \text{NS}$	(−1.3 mmHg, 4.6 mmHg)

between 2.1 and 5.7 mmHg with 95% confidence. So at least informally in this case, the *p*-value and the CI are telling us similar things; the treatment means are different. In trial 2, according to the *p*-value, there is not enough evidence to declare differences, while the 95% CI tells us that  $\mu_1$  could be bigger than  $\mu_2$  by as much as 4.6 units but also that  $\mu_2$  could be bigger than  $\mu_1$  by as much as 1.3 units and everything in between is also possible, so certainly 0 is a plausible value for  $\mu_1 - \mu_2$ . In this case, again, the *p*-value and CI are saying similar things, and neither can discount the equality of the treatment means.

The link between the *p*-value and the CI, however, is not just operating at this informal level; it is much stronger, and there is a mathematical connection between the two, as follows:

- If  $p < 0.05$ , the 95% CI for  $\mu_1 - \mu_2$  will exclude zero (and vice versa).
- If  $p > 0.05$  (NS), the 95% CI for  $\mu_1 - \mu_2$  will include zero (and vice versa).

Note that if *p* is exactly equal to 0.05, one end of the CI will be equal to zero; this is the boundary between the two conditions above.

One element that makes the link work is the correspondence between the significance level (5%) and the *confidence coefficient* (95%). If we were to use 1% as the cut-off for statistical significance, then the same link would apply but now with the 99% CI.

There is a misunderstanding regarding a similar potential link between the *p*-value and the CIs for the individual means. A significant *p*-value does not necessarily correspond to non-overlapping CIs for the individual means. See Julious (2004) and Cumming (2009) for further discussion on this issue.

This link also applies to the *p*-value from the paired t-test, and the CI for  $\mu$ , the mean difference between the treatments, and in addition extends to adjusted analyses including ANOVA and ANCOVA and similarly for regression. For example, if the test for the slope  $b$  of the regression line gives a statistically significant *p*-value (at the 5% level), then the 95% CI for the slope will not contain zero, and vice versa.

When dealing with binary data a similar link applies with the CI for the odds ratio and the *p*-value for the  $\chi^2$  test, with one important difference; it is the value one (and not zero) that is excluded or included from the CI when *p* is either significant or non-significant, respectively. Recall that for the odds ratio, it is the value one that corresponds to equal treatments. The link for binary data is not exact in the strict mathematical sense, but in practice, this correspondence can be assumed to apply pretty much all the time except on the boundary of 0.05 for the *p*-value,

where from time to time one end of the CI may not quite fall on the appropriate side of one. Similar comments to these also apply to the relative risk, the risk ratio for count endpoints, and the hazard ratio for time to event endpoints.

## 9.9 Confidence intervals for clinical importance

Example 9.4 presents some hypothetical data from four trials in hypertension.

- Trial 1 has given statistical significance and has detected something of clinical importance.
- Trial 2 has also given statistical significance, but the difference detected is clinically unimportant.

In comparing the results from trials 1 and 2, the *p*-value does not tell the whole story. In terms of statistical significance, they are indistinguishable, but the first trial has demonstrated a clinically important difference, while trial 2 has detected something that is clinically irrelevant.

- Trial 3 has given non-significance statistically, and inspecting the CI tells us that there is nothing in terms of clinical importance either. With 95% confidence, the benefit of the active treatment is at most 3.1 mmHg.
- Trial 4 is different, however. We do not have statistical significance, but the CI suggests that there could still be something of clinical importance with potential differences of 5 mmHg, 10 mmHg and even 14 mmHg. This is classically the trial that is too small with low power to detect even large differences.

Again, the *p*-value is not giving the whole story. There is clearly nothing of clinical importance in trial 3, but in trial 4, there could be something worthwhile – it is just that the trial is too small.

**Example 9.4** A series of trials in hypertension (hypothetical)

In a collection of four placebo-controlled trials in hypertension, a difference of 4 mmHg in terms of mean fall in diastolic bp is to be considered of clinical importance; anything less is unimportant. The results are given in Table 9.6, where  $\mu_1$  and  $\mu_2$  are the mean reductions in diastolic bp in the active and placebo groups, respectively.

**Table 9.6** *p*-values and confidence intervals for four trials

	<i>p</i> -value	95% CI for $\mu_1 - \mu_2$
Trial 1	$p \leq 0.05$	(3.4 mmHg, 12.8 mmHg)
Trial 2	$p \leq 0.05$	(1.2 mmHg, 2.9 mmHg)
Trial 3	$p = \text{NS}$	(−3.5 mmHg, 3.1 mmHg)
Trial 4	$p = \text{NS}$	(−2.6 mmHg, 14.3 mmHg)

Note the mathematical connection again, with the first two trials giving significant *p*-values and the second two trials giving non-significant *p*-values.

It should be clear from the development in this example that statistical significance and clinical importance are somewhat different things. The *p*-value tells us nothing about clinical importance. Just because we have statistical significance, it does not mean, necessarily, that we have detected a clinically important effect. Vice versa, a non-significant result statistically does not necessarily indicate the absence of something of clinical importance. The most appropriate way to provide information on clinical benefit is by presenting observed treatment differences together with CIs.

Gardner and Altman (1989) capture the essence of this argument:

*'Presenting p-values alone can lead to them being given more merit than they deserve. In particular, there is a tendency to equate statistical significance with medical importance or biological relevance. But small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small'.*

The regulators are not only interested in statistical significance but also in clinical importance. This allows them, and others, to appropriately balance benefit and risk. Therefore, it is good practice to present both *p*-values and CIs, and indeed, this is a requirement within a regulatory submission. Most journals nowadays also require results to be presented in the form of CIs in addition to *p*-values.

## 9.10 Misinterpretation of the *p*-value

### 9.10.1 Conclusions of similarity

In Section 3.3.1, we defined the *p*-value and briefly mentioned a common incorrect definition. We will return now to discuss why this leads to considerable misinterpretation. In the example of Section 3.3.1, we had observed a treatment difference of 5.4 mmHg with a *p*-value of 0.042 (4.2%), and the proposed incorrect definition was that 'there is a 4.2% probability that  $\mu_1 = \mu_2$ '.

The problem with this definition is the misinterpretation when the *p*-value is large. As an extreme case, suppose that we ran a trial in hypertension with two patients per group, and suppose that even though, in truth, the true treatment means were very different, the patients in the active group had blood pressure reductions of 7 mmHg and 5 mmHg, respectively, while in the placebo group, the reductions were 2 mmHg and 10 mmHg. These data give a mean reduction in the active group of 6 mmHg and a mean reduction in the placebo group of 6 mmHg. In a two-sample t-test, the resulting *p*-value would be one ( $\text{signal} = \bar{x}_1 - \bar{x}_2 = 0$ ). A *p*-value of 1 or 100% corresponds to certainty and, taking the aforementioned definition of the *p*-value, tells us, on the basis of the observed data, that it is certain that the true treatment means are identical! I hope we would all agree that this conclusion based on two patients per group would be entirely inappropriate.

This example, of course, is purely hypothetical, but in practice, we do see large  $p$ -values, say, of the order of 0.70 or 0.80, that have come from situations where, in truth, the treatments could be very different, but we have ended up with a large  $p$ -value merely as a result of a small sample size or a large amount of patient-to-patient variation (or both), and as a consequence, we have a large amount of noise, a small signal-to-noise (and signal-to-standard error) ratio and a large  $p$ -value. A  $p$ -value of this order of magnitude, under this (incorrect) definition, is giving a probability of something close to certainty that the treatment means are identical.

It is not uncommon to see a conclusion that treatments are the same (or similar) simply on the back of a large  $p$ -value; this is not necessarily the correct conclusion. Presentation of the 95% CI will provide a statement about the possible magnitude of the treatment difference. This can be inspected, and only then can a conclusion of similarity be made if this interval is seen to exclude clinically important differences. We will return to a more formal approach to this in Chapter 12, where we discuss equivalence and non-inferiority.

### 9.10.2 The problem with 0.05

A further aspect of the  $p$ -value that causes some problems of interpretation is the cut-off for significance at 0.05. This issue was briefly raised in Section 3.3.4, where it was pointed out that 0.05 is a completely arbitrary cut-off for statistical significance and that  $p$ -values close to 0.05, but sitting on opposite sides of 0.05, should not really lead to different conclusions.

Too often, a  $p$ -value  $\leq 0.05$  is seen as definitive proof that the treatments are different, while a  $p$ -value above 0.05 is seen as proof that they are the same or similar. The  $p$ -value is a measure of the compatibility of the data with equal treatments; the smaller the  $p$ -value, the stronger the evidence against the null hypothesis. The  $p$ -value is a measure of evidence in relation to the null hypothesis; treating  $p \leq 0.05$ ,  $p > 0.05$  in a binary way as proof that there is, or is not, a difference is a gross oversimplification, and we must never lose sight of that.

## 9.11 Single pivotal trial and 0.05

The conventional level for two-sided statistical significance for efficacy is 0.05, but there are some circumstances where a more stringent level is appropriate: in particular, where a regulatory submission is to be based on a single pivotal trial. A discussion of the two-pivotal-trial rule and under what conditions sponsors may be allowed to deviate from that requirement is included in CPMP (2001) 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study'.

The repeatability of findings gives strong scientific support for the existence of a true treatment effect and is essentially the reason the regulators in general like to see two pivotal trials that are positive ( $p \leq 0.05$  in favour of the experimental

treatment) – this clearly constitutes convincing evidence. In conjunction with this, the two-trial rule provides an opportunity to investigate the treatment effect in different settings, and a demonstration of an effect in both trials adds support to the robustness of that positive finding. The policy of running two separate trials with the same protocol by simply dividing up the centres is not consistent with this thinking and should be avoided.

The two-trial rule – for example, in a placebo-controlled setting – effectively translates into a very stringent requirement for statistical significance. In a single trial, the conventional two-sided type I error rate is 0.05. It follows that to obtain a positive result from such a trial, we need the effect to be statistically significant and in favour of the active treatment. The type I error associated with this false-positive result is 0.025 (which is 1 in 40). In two trials, therefore, obtaining two false-positive results carries a combined false-positive rate of  $0.025 \times 0.025 = 0.000625$  (which is 1 in 1600). In other words, if the active treatment were to be truly ineffective, we would see two positive trials by chance on only 1 in 1600 occasions.

In therapeutic settings where there are practical reasons why two trials cannot be easily undertaken or where there is a major unmet public health need, it may be possible for a claim to be based on a single pivotal trial. The regulatory authorities do allow this, but only under certain conditions.

***CPMP (2001): 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study'***

*'In cases where the confirmatory evidence is provided by one pivotal study only, this study will have to be exceptionally compelling, and in the regulatory evaluation special attention will be paid to:*

- *The internal validity. There should be no indications of a potential bias*
- *The external validity. The study population should be suitable for extrapolation to the population to be treated*
- *Clinical relevance. The estimated size of the treatment benefit must be large enough to be clinically valuable*
- *The degree of statistical significance. Statistical evidence considerably stronger than  $p < 0.05$  is usually required*
- *Data quality*
- *Internal consistency. Similar effects demonstrated in different pre-specified subpopulations. All important endpoints showing similar findings*
- *Centre effects. None of the study centres should dominate the overall result, neither in terms of number of subjects nor in terms of magnitude of effect*
- *The plausibility of the hypothesis tested'*

Statistical evidence stronger than  $p < 0.05$  is open to interpretation, but certainly one-sided  $p \leq 0.000625$  would be a lower bound on this. In practice, the precise

value would depend on the therapeutic setting and the unmet need, although  $p \leq 0.01$  would seem a fair target in many situations. The only exceptions are in orphan indications, where the rules tend to be relaxed somewhat. It is difficult to make general statements as things depend so much on the specific situation. With a single pivotal trial in an orphan setting, however, it is unlikely that this requirement for the  $p$ -value to be considerably below 0.05 would apply.

It is also worth commenting on the *internal consistency* issue. In general, the regulators are looking for demonstration of a robust treatment effect. Within the context of the intended label, there needs to be clear evidence that the treatment is effective in all sub-populations – age groups, disease severity, race, sex (if appropriate) and so on – and this is what is meant by internal consistency. The two-trial rule gives an opportunity to evaluate the treatment across different environments: for example, different hospital types and different geographies. A single trial will only provide a similar level of assurance if it recruits across the broad range of settings, consistent with the label, followed by a thorough demonstration of homogeneity of treatment effects across those settings.

# CHAPTER 10

## Multiple testing

### 10.1 Inflation of the type I error

#### 10.1.1 False positives

Whenever we undertake a statistical test in a situation where the two treatments being compared are the same (e.g. in terms of equal means or an odds ratio of 1), there is a 5% probability of getting a statistically significant result purely by chance; this is the type I error. If we were to conduct several tests in this same setting, the probability of seeing one or more significant *p*-values purely by chance would start to mount up. For example, if we conduct five tests on independent sets of data – say, on five distinct subgroups – then the probability of getting at least one false-positive result is 22.6%.<sup>\*</sup> For 50 tests, this probability becomes 92.3%, virtual certainty. This should come as no surprise; the 1 in 20 probability of a false positive on each occasion will result in something coming up purely by chance. Certainly, with 50 tests under these circumstances, the most surprising thing would be if you did not see a false positive at least once.

The problem with this so-called *multiplicity* or *multiple testing* arises when we draw a positive confirmatory conclusion on the basis of a positive result that has been *generated* simply because we have undertaken lots of comparisons and *cherry-picked* the statistically significant differences that are in our favour. Inflation of the type I error rate in this way is of great concern to the regulatory authorities and the clinical community more generally; regulators do not want to be registering, and the clinical community does not want to be adopting, treatments that do not work. It is therefore necessary to control this inflation. Most of this chapter is concerned with ways in which the potential problem can be controlled, but firstly, we will look to an example to further illustrate the issues.

#### 10.1.2 A simulated trial

Lee et al. (1980) report on lessons to be learned in multiple testing by analysing data from a simulated randomised clinical trial. These authors took data from 1073 consecutive, medically treated coronary artery disease patients from the

---

\*The probability of no significant results in five tests is  $0.95 \times 0.95 \times 0.95 \times 0.95 \times 0.95 = 0.774$ , so the probability of one or more significant results is  $1 - 0.774 = 0.226$ .

Duke University data bank and split these patients at random into two groups. They then labelled these groups 1 and 2 and proceeded to analyse these data as if the two groups had in fact received different treatments. These patients had all been treated medically, their data on baseline factors and outcome (survival time) was already complete, and the treatment 1 and treatment 2 groups were totally spurious – they did not exist; this was just a random split of a database.

The overall comparison of the two treatment groups in terms of the primary outcome, survival time, not surprisingly gave a non-significant  $p$ -value ( $p > 0.05$ ). The authors then proceeded to investigate the data in subgroups. The two recognised key prognostic factors are the number of diseased vessels (1, 2 or 3) and left ventricular contraction pattern at baseline (LVCP – normal or abnormal). This defined six subgroups, and comparing treatment 1 with treatment 2 in each of these subgroups again yielded non-significant  $p$ -values. However, the  $p$ -value in the subgroup of patients with three diseased vessels and abnormal LVCP gave a  $p$ -value between 0.10 and 0.15, perhaps indicating that there might be something of interest (a *trend!*) in that subgroup. They then analysed the data in this subgroup further by subdividing by a third prognostic factor, history of congestive heart failure (CHF, yes or no). The comparison in the subgroup (number of diseased vessels = 3, abnormal LVCP and no history of CHF) gave a statistically significant difference between treatment 1 and treatment 2 with  $p \leq 0.01$ . Note that this final comparison came after a non-significant result overall and then looking in eight initial subgroups. In a setting where the treatments are truly identical, a statistically significant result will eventually appear.

Clearly, this difference must be a false positive; the separate treatment group 1 and treatment group 2 are totally artificial – they do not exist. Lee et al. have used this simulated study to illustrate the false conclusions that potentially can arise through multiple testing. It is not that uncommon, unfortunately, to see trialists go off on a *fishing trip* in this post hoc way, looking for treatment differences; and the strategy that they undertook to *discover* this statistically significant  $p$ -value is maybe not untypical of what could happen in a real situation. This is very dangerous and can easily *uncover* effects that are not real.

## 10.2 How does multiplicity arise?

There are numerous settings that can result in multiplicity:

- Multiple endpoints
  - Multiple pairwise comparisons in multi-arm trials
  - Comparing treatments within many subgroups
  - Interim analyses
  - Using different statistical tests on the same data
  - Using different analysis sets or different algorithms for missing data
- This list is not exhaustive, but these represent the main areas of concern.

We will explore each of these in turn, but before doing so, it is worth making some preliminary points. Firstly, not all multiple testing is a bad thing. For example, it is good practice to evaluate several different algorithms for missing data (the final bullet point) to gauge the robustness of the results to that choice. It can also be of value to look at treatment differences in various subgroups to assess the homogeneity of the overall finding across the complete population. The problem arises when the results of these comparisons are *cherry-picked*, with only those analyses that have given significant results then being used to make a confirmatory claim and those giving non-significant results just ignored or pushed to the background. Secondly, if this process of cherry-picking is to be in any sense allowed, there will be a price to pay in terms of reducing the level at which statistical significance can be declared. We will say more about specific methods for making this reduction later, but basically, the idea is to divide up the 5% allowable false-positive rate across the numerous tests that will be the basis of any confirmatory claims. For example, if five tests make up the confirmatory analysis and a claim is going to be made on any of these comparisons that yield a statistically significant result, then the level at which statistical significance can be declared will be reduced from 5% to 1%; the argument here is that five lots of 1% make up 5% so the overall type I error rate remains controlled at 5%. For 10 tests, the *adjusted significance level* (sometimes denoted by  $\alpha'$ ) would be 0.5%. This is the simplest form of adjustment and is known as the *Bonferroni correction*.

The basic rule throughout is that the false positive rate across the complete trial should be controlled at 5%. This overall rate is known as the *family-wise error rate* (FWER).

### 10.3 Regulatory and scientific view

The regulatory position on multiplicity is well expressed in ICH E9.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'When multiplicity is present, the usual frequentist approach to the analysis of clinical trial data may necessitate an adjustment to the type I error. Multiplicity may arise for, example, from multiple primary variables, multiple comparisons of treatments, repeated evaluation over time and/or interim analyses. Methods to avoid or reduce multiplicity are sometimes preferable when available, such as the identification of the key primary variable (multiple variables), the choice of a critical treatment contrast (multiple comparisons), the use of a summary measure such as "area under the curve" (repeated measures). In confirmatory analyses, any aspects of multiplicity which remain after steps of this kind have been taken should be identified in the*

*protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan'.*

Note that these recommendations relate to confirmatory claims and statements. For exploratory investigations, there are no restrictions in this multiplicity sense. Any findings arising from such analyses alone, however, cannot be viewed as confirmatory.

These comments are directed primarily at efficacy and are not generally applied to routine safety comparisons unless a specific safety claim (e.g. drug A reduces the incidence of neutropenia compared to drug B) is to be made. With the routine evaluation of safety, if *p*-values are being used as a flag for potential concerns, we tend to be conservative and not worry about inflating the type I error. It is missing a real safety concern, the type II error, which troubles us more.

Specific guidelines on multiplicity have been produced both in the US and in Europe – FDA (2017) ‘Multiple Endpoints in Clinical Trials. (Draft) Guidance for Industry’ and CHMP (2017) ‘Guideline on multiplicity issues in clinical trials’ – and we will mention these further in this chapter.

Concerns about multiplicity have been recognised within the regulatory community for many years, but there has been noticeably less concern outside until recently. Many journal editors now recognise, however, that multiplicity – the uncontrolled use of *p*-values and the misinterpretation of those *p*-values among the general clinical community – continues to cause problems in relation to the overinterpretation of studies and false claims. This is a quotation from an editorial from the *New England Journal of Medicine* (Harrington et al., 2019) regarding new guidelines for the reporting of *p*-values in the journal:

*‘The new guidelines discuss many aspects of the reporting of studies in the Journal, including a requirement to replace *P* values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity’.*

The general message is that claims of statistical significance will not be allowed unless there is appropriate control of multiple testing and inflation of the type I error rate. Increasingly, leading journals are adopting this same position.

## 10.4 Methods for adjustment

### 10.4.1 Bonferroni correction

The Bonferroni method of adjustment was mentioned earlier in this chapter as a method of preserving the overall 5% type I error rate. In general, if there are *m* confirmatory comparisons with claims to be made on whichever of these is statistically significant, the Bonferroni correction requires that each comparison

be evaluated at level  $\alpha' = \alpha/m$ . In a strict statistical sense, this is the correct adjustment only for tests based on independent sets of data. For example, if there are four non-overlapping (independent) subgroups of patients – males aged under 65, males aged 65 or over, females aged under 65 and females aged 65 or over – then an adjustment that uses the 0.0125 level of significance for each of the subgroups will have an overall type I error rate of 5%. In most cases, however, when we use this adjustment, the tests that make up the set of comparisons will not be independent in this sense. With multiple primary endpoints, there will undoubtedly be correlation between those endpoints. Where this is the case, the Bonferroni correction provides a conservative procedure; in other words, the effective overall type I error rate will be at most 5%. As an extreme example of this conservativeness, suppose that two primary endpoints were perfectly correlated, so that if one is statistically significant by chance then so will the other one. The Bonferroni adjustment would require each of the endpoints to be evaluated at the 2.5% level of significance; but because of the perfect correlation, the overall type I error rate would also be 2.5% – considerably less than the 5% requirement.

The considerations so far are based on the presumption that the type I error rate is divided equally across the comparisons. This does not always make sense, and indeed, it is not a requirement that it be done in this way. For example, with two comparisons, there would be nothing to prevent having a 4% type I error rate for one of the comparisons and a 1% type I error rate for the other comparison, provided this was clearly set down in the protocol. With interim analyses, for example, it is advantageous to divide the error rate unequally.

#### 10.4.2 Holm correction

The Bonferroni correction is the simplest form of correction but often not the most efficient. A more efficient correction that protects the overall type I error is due to Holm (1979). This approach, sometimes also referred to as *Bonferroni-Holm*, involves ordering the  $p$ -values from smallest to largest over the  $m$  endpoints to be considered within the confirmatory structure of the trial. If the smallest  $p$ -value is  $\leq 0.05/m$ , that endpoint can be declared statistically significant. The procedure then looks at the second-smallest and compares it to  $0.05/(m - 1)$  for statistical significance, and so on, through to the largest  $p$ -value compared to 0.05. Statistical testing stops as soon as a non-significant result is found at one of these steps. For example, suppose there are three endpoints. The smallest  $p$ -value across those three endpoints is compared to 0.017. If statistical significance is not achieved for this first endpoint, testing stops. If statistical significance is achieved, however, a difference is declared for that endpoint, and the second-smallest  $p$ -value is compared to 0.025. Again, if statistical significance is not achieved in this case, testing stops. If statistical significance is achieved, a difference is declared for that endpoint, and finally, the largest  $p$ -value is compared to 0.05. Table 10.1 provides examples of how the procedure works in

**Table 10.1** Holm and Hochberg corrections

	Case 1			Case 2			Case 3		
	p-value	Holm	Hochberg	p-value	Holm	Hochberg	p-value	Holm	Hochberg
Endpoint 1	0.014	✓	✓	0.014	✓	✓	0.019		✓
Endpoint 2	0.020	✓	✓	0.031			0.020		✓
Endpoint 3	0.045	✓	✓	0.056			0.056		

practice. In case 1, statistical significance can be declared for all three endpoints. In case 2, statistical significance can be declared for endpoint 1 but not for endpoints 2 and 3, since the second-smallest *p*-value is not  $\leq 0.025$ . In case 3, statistical significance cannot be declared for any of the endpoints since the smallest *p*-value is not  $\leq 0.017$ . Note that had the Bonferroni correction been used, only endpoint 1 would have been declared statistically significant in case 1. In cases 2 and 3, the use of the Bonferroni correction would have led to the same conclusions as the Holm correction.

### 10.4.3 Hochberg correction

This approach involves ordering the *p*-values in the opposite way to the Holm correction, from largest down to smallest, and moving through these *p*-values starting with the largest. If the largest *p*-value is  $\leq 0.05$ , then all endpoints can be declared statistically significant. If the largest *p*-value is  $> 0.05$ , then non-significance is declared for that endpoint, but now we are allowed to move to the second-largest *p*-value. If the *p*-value for that endpoint is  $\leq 0.05/2 = 0.025$ , then that endpoint and all endpoints lower in the order can be declared statistically significant. However, if that endpoint gives a *p*-value  $> 0.025$ , non-significance is declared for endpoint number 2 in the ordering, and we move to the endpoint with the third-largest *p*-value. This is then compared to  $0.05/3 = 0.017$ , and so on down through the list, with the endpoint in position *m* being compared to  $0.05/m$ . Looking at Table 10.1, the order in which the endpoints would be considered under the Hochberg procedure would be from endpoint 3 up through to endpoint 1. In case 1, all endpoints are statistically significant since the largest *p*-value is  $\leq 0.05$ . In case 2, since the *p*-value for endpoint 3 is  $> 0.05$  and the *p*-value for endpoint 2 is  $> 0.025$ , neither of these two endpoints is statistically significant. The *p*-value for endpoint 3, however, is  $\leq 0.017$ , so statistical significance can be declared for that endpoint. In case 3, endpoint 3 is not statistically significant under the Hochberg scheme, but endpoint 2 and endpoint 3 are statistically significant since the *p*-value for endpoint 2 is  $\leq 0.025$ . See Table 10.1 for the results from the Hochberg procedure and a comparison of Holm and Hochberg.

In general, of these three procedures, Hochberg is the most efficient and provides the most power to detect treatment differences, followed by the Holm procedure, with the Bonferroni procedure the least efficient. The Bonferroni and

Holm procedures are known to control the FWER under all circumstances. However, this is not true for the Hochberg procedure, where in some (albeit exceptional) cases, the FWER is not controlled. The FDA (2017) draft guideline ‘Multiple Endpoints in Clinical Trials: Guidance for Industry’ provides more detail on when the method does, and does not, provide adequate control of the type I error, but given this uncertainty goes on to say, *‘Therefore, beyond the aforementioned cases where the Hochberg procedure is known to be valid, its use is generally not recommended for the primary comparisons of confirmatory clinical trials unless it can be shown that adequate control of Type I error rate is provided’*.

One question might be, given that the Holm procedure always outperforms Bonferroni, why would we ever use Bonferroni? Well, while this is true in the simple settings outlined here, things are not quite so straightforward on some occasions when these procedures are combined with hierarchical testing (see Section 10.5.3) or when the analyses of the endpoints are separated in time. The fact that the order in which the endpoints are looked at is determined only once the data are in hand for both the Holm and Hochberg procedures creates some practical difficulties, and reverting to Bonferroni can provide a much clearer pathway through the required statistical testing in certain cases.

#### 10.4.4 Interim analyses

Interim analyses arise when we want to look at the data as they accumulate, with the possibility of stopping the trial at the interim stage if the data suggest overwhelming efficacy of the test treatment compared to control. If we introduce, say, two interim looks in addition to the final analysis at the end of the trial, then we have an overall testing strategy that consists of three tests, and some account of the multiplicity is required. There has been a considerable amount of theory developed in this area, and the resulting procedures not only preserve the 5% type error rate but also do not pay as big a price as Bonferroni. Remember that Bonferroni adjusted significance levels are only strictly correct when based on independent tests. In the context of interim analysis, the data sets that are being analysed are not independent and overlap in a very structured way. With a sample size of 600 and three looks, the first interim analysis after 200 patients provides precisely half of the data on which the second interim analysis, based on 400 patients, is to be undertaken, while these 400 patients provide two-thirds of the data on which the final analysis is to be conducted. This induced correlation between the three analyses allows a less stringent significance level to be used compared to the setting with three independent tests.

Pocock (1977) developed a procedure that divides the type I error rate of 5% equally across the various analyses. In the example earlier, with two interim looks and a final analysis, Bonferroni would suggest using an adjusted significance level of 0.017 (= 0.05/3). The Pocock method, however, gives us the correct adjusted significance level as 0.022, which exactly preserves the overall 5% type I error rate.

While this equal division of the type I error may work for some settings, it is more likely that we would want, firstly, to keep back most of the 5% for the final and most important analysis and, secondly, only stop a trial early in the case of overwhelming evidence for efficacy. The methods of O'Brien and Fleming (1979) divide up the type I error rate unequally, with very stringent levels at the early interims, becoming less stringent at subsequent analyses and leaving most of the 5% over for the final analysis. In the case of two interim looks and a final analysis, equally spaced in terms of patient numbers, the adjusted significance levels are 0.00052, 0.014 and 0.045; adjusted significance levels that are very stringent early on, with most of the 0.05 left over for the final analysis.

It is also possible to stop trials for reasons other than overwhelming efficacy: for example, for futility, where at an interim stage it is clear that if the trial were to continue, it would have little chance of giving a positive, statistically significant, result. We will say more about futility and interim analyses in general in Chapter 14, where we will look at the practical application of these methods.

## 10.5 Avoiding adjustment

### 10.5.1 Co-primary endpoints

As mentioned in the previous section, multiplicity can lead to adjustment of the significance level. There are, however, some situations when adjustment is not needed, although these situations tend to have restrictions in other ways. We will focus this discussion on endpoints that are to be included within the confirmatory structure of the trial and in subsequent sections use similar arguments to deal with other aspects of multiple testing.

As ICH E9 points out, 'There should generally be only one primary variable', and when this is the case, there is clearly no need for adjustment. However, there may well be good scientific and commercial reasons for including more than one primary variable: for example, to cover the different potential effects of the new treatment. Indeed, in some therapeutic settings, the regulators require us to demonstrate effects in terms of two or more endpoints. For example, in COPD, we look for effects both in terms of lung function and symptoms.

**CHMP (2012): '*Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (COPD)*'**

*'Measurement of lung function parameters alone is considered to be insufficient in the assessment of therapeutic effect. If lung function is selected as a primary endpoint (FEV1 would be the parameter of choice), additional evidence of efficacy must be demonstrated through the use of a co-primary endpoint, which should either be a symptom-based endpoint or a patient-related endpoint'.*

Under such circumstances, no adjustment to the significance level is required; statistical significance needs to be shown for both endpoints. These are referred to in the regulatory setting as *co-primary endpoints*.

#### **CHMP (2017) 'Guideline on multiplicity issues in clinical trials'**

*'If more than one primary endpoint is used to define study success, this success could be defined by a positive outcome in all endpoints or it may be considered sufficient, if one out of a number of endpoints has a positive outcome. Whereas in the first definition the primary endpoints are designated as co-primary endpoints, the latter case is different and would require appropriate adjustment for multiplicity.'*

Note that for the 'one out of several giving a positive outcome' comment, we would think in terms of a Bonferroni, Holm or Hochberg adjustment to the significance level.

Requiring significance on two co-primary endpoints will impact power. If the power attached to each endpoint were 90%, for example, then the combined power (the probability of getting statistical significance on both endpoints) could be as low as 81% ( $81\% = 90\% \times 90\%$ ). Increasing the sample size to give power of at least 95% for each individual endpoint will ensure a combined power of at least 90% ( $90.025\% = 95\% \times 95\%$ ).

#### **10.5.2 Composite endpoints**

Another way of avoiding adjustment is to combine the multiple measurements into a single *composite endpoint*. Examples would be disease-free survival in oncology, where the variable is the time to disease recurrence or death, whichever occurs first, or a composite of major cardiovascular events (MACE, often defined as nonfatal stroke, nonfatal myocardial infarction [MI], and cardiovascular death), a binary outcome in a cardiovascular setting. Note that this particular version of the MACE endpoint is commonly referred to as the three-point MACE. This approach does not require adjustment of the significance level; we are back to having a single primary endpoint. Such endpoints are used extensively in various settings where there is simply not enough power to look at the components individually.

There are some additional requirements, however, when using composite endpoints. A large positive effect in one of the components could potentially be masking a negative effect in a different component, and this would be unacceptable. Ideally, all component endpoints should be giving differences in the same direction; indeed, it should be expected that treatment will affect all components in a similar way. At the very least, none of the clinically important components should be giving differences in the wrong direction, and the data should be presented for those separate components to confirm this.

In some applications, especially in cardiovascular disease, the separate components (or some combinations of those components) may be secondary endpoints. If specific claims are to be made for any of the components, however, these should be organised within the confirmatory structure of the testing procedures: for example, using hierarchical testing (see Section 10.5.3). The CHMP cover these points in their guidance on multiplicity.

**CHMP (2017) '*Guideline on multiplicity issues in clinical trials*'**

*'It is recommended to analyse in addition the single components and clinically relevant groups of components separately, to provide supportive information. There is, however, no need for an adjustment for multiplicity provided significance of the primary endpoint is achieved. If claims are to be based on (subgroups of) components, this needs to be pre-specified and embedded in a valid confirmatory analysis strategy.'*.

When regulators speak here about analysing the separate components, they do not necessarily mean to produce  $p$ -values but are merely referring to presentation of the individual event rates, for example. There is an acceptance that the power for the individual components may make the  $p$ -values somewhat irrelevant.

### 10.5.3 Hierarchical testing

It may be possible with several primary endpoints to rank these in a hierarchy in terms of their clinical importance and within this structure adopt a testing strategy that avoids having to adjust the significance level. This ranking, which of course should be pre-specified in the protocol, determines the order in which the statistical testing is done. No adjustment to the significance level is required, but claims cannot be made beyond the first non-significant (NS) result in the hierarchy. Consider the setting with three primary endpoints, ranked according to their clinical relevance (Table 10.2). In case 1, claims can be made on endpoints 1 and 2. In case 2, a claim can be made on endpoint 1 only, because endpoint 2 is non-significant and we are then not allowed to make a claim for endpoint 3, even though in this case that endpoint may give  $p \leq 0.05$ . In case 3, no claims can be made. Any  $p$ -values  $\leq 0.05$  lower in the hierarchy than a non-significant endpoint cannot provide the basis for a confirmatory statement/

**Table 10.2** Hierarchical testing

	Case 1	Case 2	Case 3
Endpoint 1	$p \leq 0.05$	$p \leq 0.05$	$p = \text{NS}$
Endpoint 2	$p \leq 0.05$	$p = \text{NS}$	
Endpoint 3	$p = \text{NS}$		

claim; those findings would only be viewed as supportive/exploratory. This method is also referred to as the *fixed-sequence method*.

The CPMP (2002) 'Points to Consider on Multiplicity Issues in Clinical Trials' specifically mentions some examples of the hierarchical strategy:

*'Typical examples are: (i) acute effects in depressive disorders followed by prevention of progression, (ii) reduction of mortality in acute myocardial infarction followed by prevention of other serious events'.*

Clearly, it is very important that we get the hierarchy correct. Generally, this would be determined by the clinical relevance of the endpoints, although under some circumstances it could be determined, in part, by the likelihood of seeing statistical significance with the easier *hits* towards the top of the hierarchy. These ideas can be considered as a way of dealing with secondary endpoints, which might be considered for inclusion in a claim. In many cases, secondary endpoints are simply primary endpoints lower down in the hierarchy.

**Example 10.1** Ticagrelor versus clopidogrel in patients with acute coronary syndromes (the PLATO study) (Adapted from Wallentin et al., 2009)

PLATO was a randomised double-blind trial comparing ticagrelor and clopidogrel in the treatment of patients with acute coronary syndromes for the prevention of cardiovascular events. The primary endpoint was the time to first occurrence of the composite of death from cardiovascular causes, MI or stroke. The principle secondary consideration was the same endpoint but in the subgroup of patients for whom invasive management was planned at randomisation. Additional secondary endpoints (considered for the complete patient population) were organised in a hierarchy below the primary and key secondary endpoints in the following order:

- The composite of death from any cause, MI or stroke
- The composite of death from vascular causes, MI, stroke, severe recurrent cardiac ischaemia, recurrent cardiac ischaemia, transient ischemic attack or other arterial thrombotic events
- MI alone
- Death from cardiovascular causes alone
- Stroke alone
- Death from any cause

The results for these endpoints are given in Figure 10.1. The primary endpoint gave statistical significance with  $p < 0.001$ . The  $p$ -value for the principal secondary endpoint, the primary in a subgroup, gave  $p = 0.003$ . The next four secondary endpoints in the hierarchy all gave statistical significance at the 5% level. The endpoint stroke, however, gave a non-statistically significant result with  $p = 0.22$ , so a claim cannot be made for that specific endpoint or for any endpoints, irrespective of their statistical significance, below stroke in the hierarchy. Interestingly, the endpoint below stroke is all-cause mortality with  $p < 0.001$ . In a formal sense, a confirmatory conclusion for all-cause mortality in terms of its statistical significance cannot be made since doing so would violate the strict control of the type I error.

End point	Ticagrelor group	Clopidogrel group	Hazard ratio for ticagrelor group (95% CI)	p-value†
Primary end point: death from vascular causes, MI or stroke : no./total no. (%)	864/9333 (9.8)	1014/9291 (11.7)	0.84 (0.77–0.92)	<0.001‡
Secondary end points : no./total no. (%)				
Death from any cause, MI, or stroke	901/9333 (10.2)	1065/9291 (12.3)	0.84(0.77–0.92)	<0.001‡
Death from vascular causes, MI stroke, severe recurrent ischaemia, recurrent ischaemia, TIA or other arterial thrombotic event	1290/9333 (14.6)	1456/9291 (16.7)	0.88 (0.81–0.95)	<0.001‡
MI	504/9333 (5.8)	593/9291 (6.9)	0.84 (0.75–0.95)	0.005‡
Death from vascular causes	353/9333 (4.0)	442/9291(5.1)	0.79 (0.69–0.91)	0.001‡
Stroke	125/9333 (1.5)	106/9291 (1.3)	1.17 (0.91–1.52)	0.22
Ischaemic	96/9333 (1.1)	91/9291 (1.1)		0.74
Haemorrhagic	23/9333 (0.2)	13/9291(0.1)		0.10
Unknown	10/9333 (0.1)	2/9291(0.02)		0.04
Other events : no./total no.(%)				
Death from any cause	399/9333 (4.5)	506/9291 (5.9)	0.78 (0.69–0.89)	<0.001
Death from causes other than vascular causes	46/9333 (0.5)	64/9291 (0.8)	0.71 (0.49–1.04)	0.08
Severe recurrent ischaemia	302/9333 (3.5)	345/9291 (4.0)	0.87 (0.74–1.01)	0.08
Recurrent ischaemia	500/9333 (5.8)	536/9291 (6.2)	0.93 (0.82–1.05)	0.22
TIA	18/9333 (0.2)	23/9291 (0.3)	0.78 (0.42–1.44)	0.42
Other arterial thrombotic event	19/9333 (0.2)	31/9291 (0.4)	0.61 (0.34–1.08)	0.09
Death from vascular causes, MI, stroke : no./total no.(%)				
Invasive treatment planned§	569/6732 (8.9)	668/6676 (10.6)	0.84 (0.75–0.94)	0.003‡
Event rate, days 1–30	443/9333 (4.8)	502/9291 (5.4)	0.88 (0.77–1.00)	0.045
Event rate, days 31–361¶	413/8763 (5.3)	510/8688 (6.6)	0.80 (0.70–0.91)	<0.001
Stent thrombosis : no. of patients who received a stent/ total no. (%)				
Definite	71/5640 (1.3)	106/5649 (1.9)	0.67 (0.50–0.91)	0.009
Probable or definite	118/5640 (2.2)	158/5649 (2.9)	0.75 (0.59–0.95)	0.02
Possible, probable or definite	155/5640 (2.9)	202/5649 (3.8)	0.77 (0.62–0.95)	0.01

\* The percentages are Kaplan–Meier estimates of the rate of the end point at 12 months. Patients could have had more than one type of end point. Death from vascular causes included fatal bleeding. Only traumatic fatal bleeding was excluded from the category of death from vascular causes. MI denotes myocardial infarction and TIA transient ischaemic attack.

† p values were calculated by means of Cox regression analysis.

‡ Statistical significance was confirmed in the hierarchical testing sequence applied to the secondary composite efficacy end points.

§ A plan for invasive or non-invasive (medical) management was declared before randomisation.

¶ Patients with any primary event during the first 30 days were excluded.

**Figure 10.1** Efficacy endpoints in the PLATO study. Source: Adapted from Wallentin et al., 2009

There is also the possibility of mixing hierarchical considerations with alpha adjustment. For example, in the case of a single primary endpoint and two secondary endpoints of equal importance to each other, the primary endpoint would be evaluated at  $\alpha = 0.05$ , while each of the secondary endpoints would use  $\alpha = 0.025$ , for example based on a Bonferroni correction. Claims could only be considered for the secondary endpoints if the primary endpoint gave  $p \leq 0.05$ , but then additional claims could be made on whichever of the secondary endpoints gives  $p \leq 0.025$ .

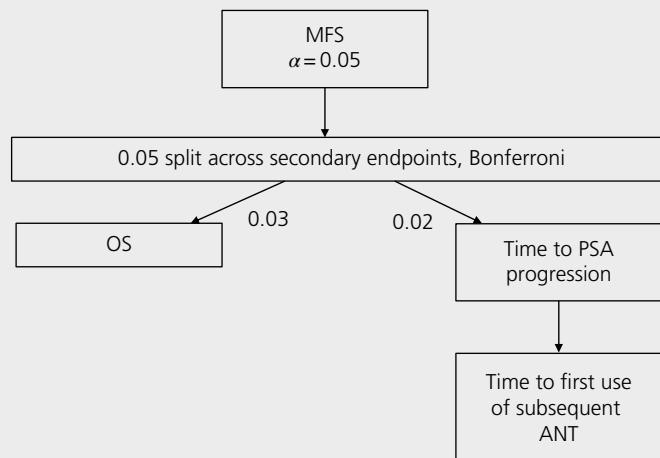
## 10.6 Fallback procedure

The fallback procedure is a methodology that firstly combines a Bonferroni splitting of the type I error with hierarchical testing but secondly allows unused alpha to be reused later in the testing sequence. It is best explained through an example.

**Example 10.2** Enzalutamide in nonmetastatic, castration-resistant prostate cancer (the PROSPER study) (Adapted from Hussain et al., 2018)

Hussain et al. (2018) report on a double-blind trial of Enzalutamide in men with nonmetastatic, castration-resistant prostate cancer. The primary efficacy endpoint was metastasis-free survival (MFS), defined as the time from randomisation to radiographic progression or death from any cause. Secondary efficacy endpoints were time to PSA progression, time to the first use of a subsequent antineoplastic therapy and overall survival (OS).

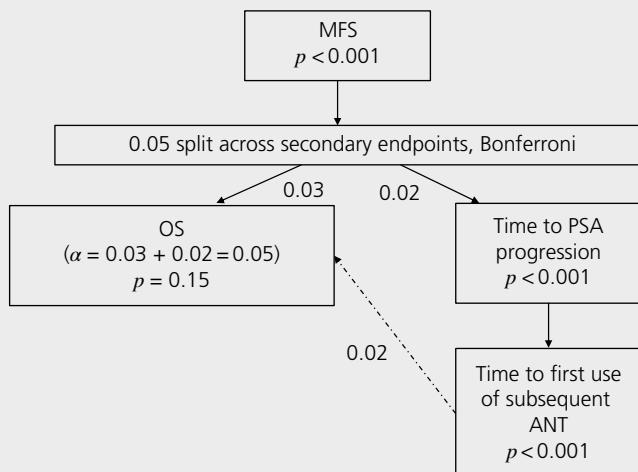
Figure 10.2 provides the pre-planned testing structure for the primary and key secondary endpoints. MFS was initially tested at the 5% two-sided level; following this, there was a Bonferroni split of the 0.05, with 0.03 being initially assigned to OS and 0.02 to the remaining two secondary endpoints, time to PSA progression and time to the first use of a subsequent antineoplastic therapy (ANT). The final two secondary endpoints were organised in a mini hierarchy in the order shown. The testing sequence was pre-specified to be MFS followed by time to PSA progression, time to the first use of a subsequent ANT and finally OS. If the primary endpoint was statistically significant at the 5% level and then both time to PSA progression and time to the first use of a subsequent ANT were found to be statistically significant at the 2% level, the 0.02 alpha level was to be added back into the 0.03 level already assigned to OS to give an overall level for OS of 0.05. The argument here is that had there been a third endpoint in the mini hierarchy, then 0.02 would have been available for a third endpoint, but instead of using 0.02 this way, it is available for use with OS. Had either of the  $p$ -values for time to PSA progression and/or time to the first use of a subsequent ANT



**Figure 10.2** Pre-planned testing structure for enzalutamide clinical trial in prostate cancer

been  $>0.02$ , this would not have been possible, and the test for OS would be undertaken at the initially assigned 3% level.

At the analysis stage, MFS was statistically significant with  $p < 0.001$  (Figure 10.3) Both time to PSA progression and time to the first use of a subsequent ANT were statistically significant when judged at  $\alpha = 0.02$ , with  $p < 0.001$  in each case. This makes 0.02 available to be added back into the 0.03 already assigned to OS to give a significance level of 0.05 for that final secondary endpoint. However, the  $p$ -value for OS was non-significant, with  $p = 0.15$ .



**Figure 10.3** Efficacy results for enzalutamide clinical trial in prostate cancer

## 10.7 Multiple comparisons of treatments

In the case of multiple treatment groups, it is important to recognise the objectives of the trial. For example, in a three-arm trial with test treatment, active comparator and placebo, the primary objective may well be to demonstrate the effectiveness of the test treatment by showing the superiority of the test treatment over placebo, and this will be the basis of the claim, while a secondary objective may be to demonstrate the non-inferiority, or perhaps superiority, of the test treatment compared to the active control. This secondary objective may, for example, be driven by market positioning. In this case, we have a hierarchy with the primary objective based on a test undertaken at the 5% level of significance, with the test treatment versus active control comparison as a second level in the hierarchy, and again, this would be conducted with  $\alpha = 0.05$ . Of course, this second comparison cannot be undertaken if the primary objective is not achieved; this makes sense because it would have little value in this scenario if we were unable to demonstrate that the test treatment works by comparing with placebo.

As a second example, consider a trial with four treatment arms: placebo and low, medium, and high doses of drug A. If we wanted to come out of this trial with a confirmatory statement concerning the effectiveness of drug A at a particular dose level, then one strategy would be to undertake three tests, each dose level against placebo, and make a claim based on whichever of these is statistically significant. An adjustment would be required, and Bonferroni would give an adjusted significance level of 0.017. There is an alternative to the Bonferroni procedure in the setting where several experimental groups are to be compared to a control group, known as Dunnet's test (Dunnet [1955]), which has some advantages over Bonferroni in terms of power. Alternatively, it may be worth considering a hierarchy in the order high dose versus placebo, medium dose versus placebo and low dose versus placebo, with no adjustment of the 5% significance level. The constraint here, of course, is that you can only make claims down to the first non-significant result. This strategy would get you to the *minimum effective dose* provided that things are well behaved, and there is an underlying monotonic dose-response relationship (the higher the dose, the bigger the effect).

## 10.8 Subgroup testing

Subgroup testing through a post hoc evaluation of treatment effects within those subgroups cannot in general be used to recover a *failed* study. This is another form of multiplicity, searching for positive, statistically significant effects in one or more subgroups.

### **CHMP (2019) 'Guideline on the investigation of subgroups in confirmatory clinical trials'**

*'The key problem of exploring subgroups is closely related to issues with multiple testing. Multiple factors are available on which subgroups can be identified and opportunities to select how the subgroup should be constructed (e.g. with different categorisations of a continuous factor) both introduce multiplicity and analysis of these subgroups may lead to contradictory conclusions simply due to the play of chance'.*

If a claim is to be considered for a specific subgroup, this needs to form part of the pre-planned confirmatory strategy.

Usually, evaluation of treatment effects within subgroups is undertaken to assess the homogeneity of the treatment effect. It is common to display treatment differences according to subgroups defined by baseline factors, in the form of point estimates of the treatment effect and 95% confidence intervals, possibly together with *p*-values for treatment-by-covariate interactions (see Sections 5.4.1 and 6.5.2) displayed in a Forest plot. The *p*-value for interaction compares the treatment effect in one subgroup (for example males) with the treatment effect in the opposite subgroup (females).

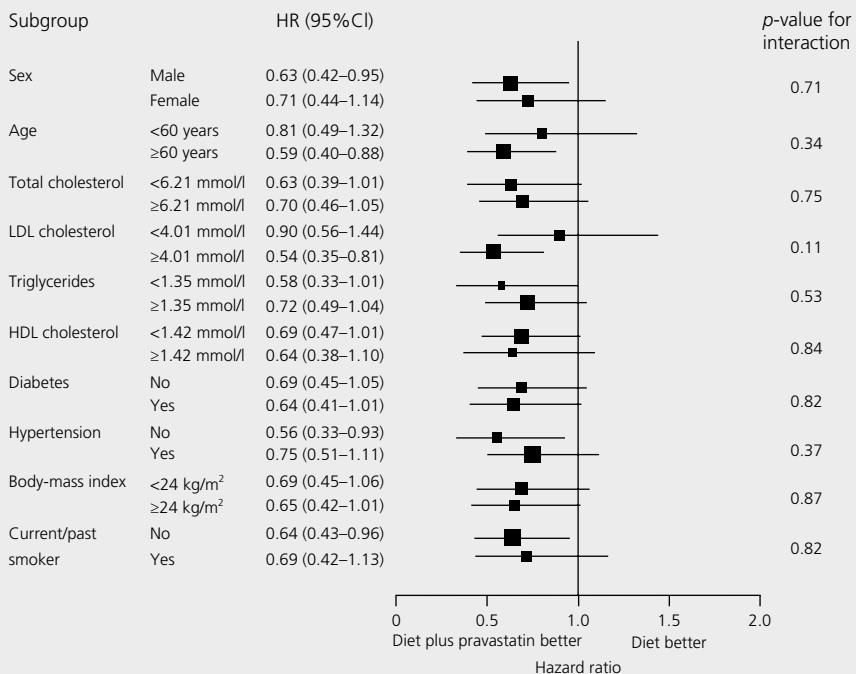
**CHMP (2019) 'Guideline on the investigation of subgroups in confirmatory clinical trials'**

'A Forest plot that includes all relevant subgroups and shows consistency in the direction and magnitude of the treatment effect is generally accepted as adding validity to the overall conclusion that the outcome of the trial applies to the studied patient population. For continuous variables, plots should be presented to characterise how the estimated effect of treatment changes over the range of the factor.'

Recall that when assessing interactions, we generally use a significance level of 0.10 rather than 0.05 due to a lack of power. Example 10.3 looks at a trial evaluating pravastatin in preventing cardiovascular disease.

**Example 10.3** Pravastatin in preventing cardiovascular disease

Figure 10.4 is taken from Nakamura et al. (2006), who reported a large placebo-controlled randomised trial evaluating the effect of pravastatin in preventing cardiovascular disease.



**Figure 10.4** Assessing the homogeneity of treatment effect. Source: Nakamura H, Arakawa K, Itakura H, et al., for the MEGA Study Group (2006) 'Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA study): a prospective randomised controlled trial'. *The Lancet*, **368**, 1155–1163. Reproduced by permission of Elsevier

The overall treatment effect was positive, with a hazard ratio of 0.67, 95% CI (0.49, 0.91),  $p = 0.01$ . We will cover hazard ratios and their use in survival analysis in Chapter 13; for the moment, simply note that, like the odds ratio and relative risk, a value of one corresponds to equal treatments. The homogeneity of the treatment effect can be assessed by looking at the  $p$ -value for the treatment-by-covariate interaction and calculating the hazard ratio together with the 95% CI separately in various subgroups defined by baseline factors of interest, as seen in Figure 10.4. Note that in this case, all of the point estimates in subgroups are within the 95% confidence interval for the overall treatment effect, and this pattern strongly supports the homogeneity of treatment effect.

In Figure 10.4, all interactions give  $p$ -values above 0.10, and based on these results, there is no evidence of treatment-by-covariate interactions for the baseline factors considered. This homogeneity of treatment effect can also be seen visually by inspecting the forest plot and observing similar values for the hazard ratios in the various subgroups and 95% confidence intervals between those subgroups that are almost completely overlapping. The hazard ratio in all cases is  $<1$ , consistent with the overall result. The  $p$ -value for the treatment-by-LDL cholesterol interaction is close to statistical significance when compared with the 0.10 cut-off, and one might argue some support for a possible differential treatment effect. However, we must recognise that we have undertaken 10 tests for interaction, and the multiplicity element of this inevitably means that apparent interactions will come up purely by chance. With a significance level of 10%, 1 in 10  $p$ -values will be significant by chance. There are two important aspects that the forest plot as displayed in this example does not show. Firstly, the plot should include the overall hazard ratio together with the associated 95% confidence interval to allow assessment of the overlap between what is seen in subgroups compared to what is seen in the complete population. Secondly, the numbers of patients and number of events for each subgroup for each treatment group should be provided in a column in the plot. Small subgroups – and for a time-to-event endpoint, subgroups with fewer events – tend to give wider confidence intervals and potentially treatment effects that are less precise.

Two settings cause some confusion and merit further discussion.

### **Setting 1: The overall result is statistically significant, but there are non-significant effects in some subgroups.**

Is it appropriate to discount subgroups showing non-significant effects in isolation since clear evidence for efficacy in those subgroups is absent? The short answer to that question is no, this is not appropriate. The trial is not powered to detect effects in subgroups that will have sample sizes that are considerably lower than the overall sample size. It is almost inevitable that several subgroups will show non-significant effects with confidence intervals that cross the ‘no treatment effect’ line. In Example 10.3, more than half of the subgroups have CIs that

cross 1, but there should be no concerns about homogeneity of treatment effect in this trial, as discussed earlier, and no considerations regarding discounting subgroups. The only concern in this kind of setting would be for a subgroup where the point estimate is either close to the no-effect line or in the direction favouring the control group. Nonetheless, these patterns can appear purely by chance, so we should be cautious and not jump to conclusions without careful additional considerations. The ‘interaction’ should be plausible pharmacologically/clinically and ideally should have some supportive evidence external to the trial.

**Setting 2: The overall result is statistically non-significant, but there is good evidence for an effect in one or more subgroups.**

This is the setting we mentioned briefly in the opening sentences of this section. The key problem here is multiplicity, and the potential for false positive conclusions has been explained through the simulated trial example in Section 10.1.2. In exceptional circumstances, regulators may be willing to give this finding further consideration, but this would be rare.

***CHMP (2019) ‘Guideline on the investigation of subgroups in confirmatory clinical trials’***

*‘Assessment scenario 3: The clinical data presented fail to establish statistically persuasive evidence in the primary analysis population but there is interest in identifying a subgroup where a relevant treatment effect and compelling evidence of a favourable risk-benefit profile can be assessed.*

*This relates to the use of a subgroup to rescue a trial that has formally failed, such that the primary objective of the trial could not be demonstrated (usually classified as  $p > 5\%$ , two-sided). From a formal statistical point of view, no further confirmatory conclusions are possible in a clinical trial where the primary null hypothesis cannot be rejected.*

*One or more additional trials should usually be conducted. In rare instances there may a basis for pursuing regulatory approval without conducting additional studies. This may be the case for a clinical setting where trials are not feasible to repeat or situations where trials are of considerable size (like in cardiovascular diseases) and even subpopulations may bear considerable amounts of randomised evidence that can be assessed for decision making about efficacy and a positive risk-benefit’.*

## **10.9 Other aspects of multiplicity**

### **10.9.1 Using different statistical tests**

Using several different statistical methods – for example, an unpaired t-test, an analysis adjusted for centre effects, ANCOVA adjusting for centre and including baseline risk as a factor, etc. – and possibly choosing whichever method produces the smallest  $p$ -value are other forms of multiplicity and are inappropriate. There is often a temptation to ‘torture the data until it confesses’.

It is standard practice to pre-specify in the protocol, or certainly in the statistical analysis plan, the statistical method to be used for analysis for each of the endpoints within the confirmatory part of the trial. This avoids the potential for bias at the analysis stage, which could arise if a method were chosen, for example, which maximised the apparent treatment difference. Changing the method of analysis following unblinding of the study in an unplanned way, even if there seems to be sound statistical reasons for doing so, is problematic. Such a switch could be supported if there was a clear algorithm contained within the statistical analysis plan that specified the rules for the switch. An example of this would be as follows:

*'The treatment means will be compared using the unpaired t-test. If, however, the group standard deviations are significantly different according to the F-test, then the comparison of the means will be based on Welch's form of the unpaired t-test'.*

The blind review does offer an opportunity to make some final changes to the planned statistical methods, and this opportunity should not be missed – but remember that this is based on blinded data.

### **10.9.2 Different analysis sets and methods for missing data**

In a superiority trial, the primary analysis is based on the full analysis set. The per-protocol set has conventionally been used as the basis for a supportive secondary analysis. However, as we argued in Chapter 8, the role of the per-protocol set is unclear and should no longer be the focus of a sensitivity analysis. The form of the analysis does depend on the methods to be used to account for missing data, and these should be clearly pre-specified. As mentioned in Section 8.4.2, it is good practice to explore the robustness of the conclusions to the methods used for handling missing data. These analyses again will be supportive (or not) of the main conclusions, and no multiplicity aspects arise.

In equivalence and non-inferiority trials (see Chapter 12), the full analysis and per-protocol sets have conventionally been given equal status and considered as co-primary on the basis that in this setting, the FAS is not conservative. However, such considerations have changed based on the introduction of estimands, and this point will be considered further in Chapter 12.

### **10.9.3 Pre-planning**

We have mentioned on several occasions that methods of statistical analysis should be pre-planned. This gives the regulators and others confidence that we are not adapting the methods of analysis or choosing new methods of analysis that somehow maximise the observed treatment benefit. This clearly would be inappropriate. In a sense, not having a pre-planned analysis and having the flexibility to look at several different methods of analysis once the data are in hand constitutes a further form of multiplicity. Any methods that are chosen having seen the data would have little chance of being the basis for confirmatory claims.

As is pointed out in ICH E9, '*Only results from analyses envisaged in the protocol (including amendments) can be regarded as confirmatory*'.

In a situation where there is a need to change key aspects of the statistical methods following finalisation of the protocol, and once the trial is underway, it is always advisable to issue a protocol amendment rather than just document this change in the statistical analysis plan. Such a change, for example, could be motivated by some new knowledge regarding what might be the most efficient method of analysis. However, making such changes in an unblinded study could still lead to concerns if there was suspicion that they were data-driven.

#### **10.9.4 Nominal significance**

In this chapter, we have covered a range of methodologies that control the FWER across the trial at 5%. In terms of reporting, we have seen that journal editors in many cases are unwilling to allow the presentation of *p*-values for endpoints that have not been appropriately controlled for multiplicity. Nonetheless, from time to time, you will see *p*-values reported that are, for example,  $\leq 0.05$  but below a non-significant endpoint in a hierarchical testing sequence or  $\leq 0.05$  when there has been a Bonferroni split of the alpha and formal statistical significance has not been achieved.

In these cases, the term *nominally significant* is sometimes used. These endpoints have given a *p*-value that is below the conventional threshold for statistical significance in isolation but do not satisfy the multiplicity requirements that have been imposed. These *p*-values do not constitute formal statistical significance and are purely exploratory and considered as hypothesis-generating for possible confirmation based on existing independent data or in the future.

## CHAPTER 11

# Non-parametric and related methods

### 11.1 Assumptions underlying the t-tests and their extensions

The t-tests and their extensions (ANOVA, ANCOVA and regression) all make assumptions about the distribution of the data in the background populations. If these assumptions are not appropriate, then, strictly speaking, the *p*-values coming out of those tests together with the associated confidence intervals (CIs) are not correct.

The assumptions are essentially of two kinds: *homogeneity of variance* and *normality*. Consider, to begin with, the unpaired t-test. This test assumes firstly that the two population distributions from which the data are drawn have the same standard deviation (homogeneity of variance) and secondly that they have the normal distribution shape. For the extensions – ANOVA, ANCOVA and regression – both homogeneity of variance and normality assumptions underpin the methods: these relate to the distribution of the individual subject values around the equation for the mean, the so-called *residuals*. Figure 6.7 in Section 6.5.7 shows the homogeneity of variance and normality assumptions. We will focus primarily on the simple settings in exploring the issues associated with these assumptions and in presenting other methods that are available if these assumptions do not hold.

The previous procedures make one additional assumption, and that is independence: the way a particular subject responds is not linked to the way another subject responds. This assumption is unlikely to be violated in a randomised clinical trial, and we will not discuss the issue further here.

### 11.2 Homogeneity of variance

In this section, we will focus the development on the unpaired t-test. The constant variance assumption can be assessed by undertaking the F-test relating to the hypotheses:

$$H_0 : \sigma_1 = \sigma_2 \quad H_1 : \sigma_1 \neq \sigma_2$$

Here,  $\sigma_1$  and  $\sigma_2$  are the true standard deviations within treatment groups 1 and 2, respectively. A significant  $p$ -value from this test would indicate that constant variance cannot be assumed and therefore that the  $p$ -value coming out of the unpaired t-test is not correct. Should this happen, we would need to use an alternative form of the unpaired t-test that allows for non-constant variance. This form of the unpaired t-test is known as *Welch's approximation*. It involves a slightly different formula for the standard error (se) of the difference  $\bar{x}_1 - \bar{x}_2$  and a different calculation of the degrees of freedom for the t-distribution on which the  $p$ -value calculation is based. These details need not concern us here; suffice it to say that the issue of non-constant variance is straightforward to deal with. In addition, our experience in clinical trials tells us that non-constant variance does not seem to occur particularly often in practice, and when it does, it tends to be associated additionally with violation of the normality assumption. Conveniently, taking care of the normality assumption by transforming the data (see Section 11.4) often also takes care of non-constant variance.

### 11.3 The assumption of normality

While constancy of variance does not seem to be too much of a concern in our clinical trials, it is not uncommon for the assumption of normality to be violated, which is more concerning. Laboratory variables sometimes do not display the normal distribution shape; and in pharmacokinetics, several of the quantities that we routinely calculate, such as AUC and  $C_{\max}$ , frequently have distributions like that shown in Figure 11.1. We talk in terms of the data being *positively skewed*. We

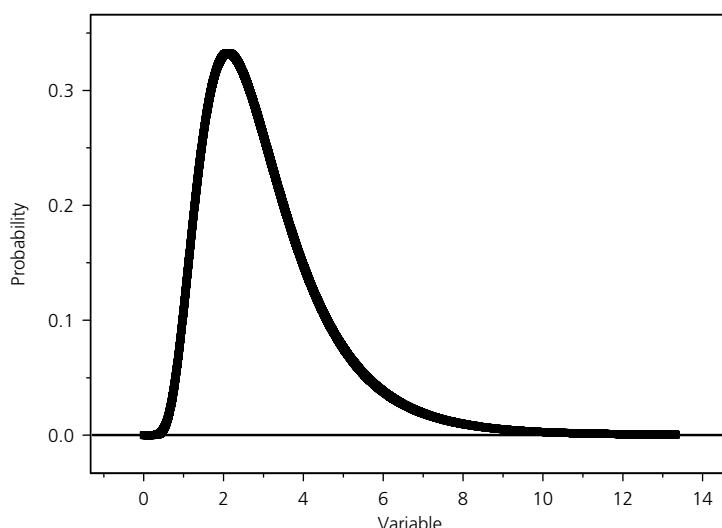
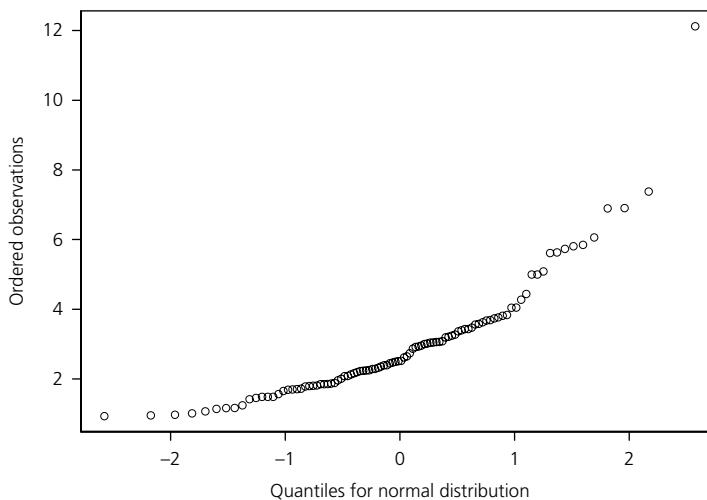


Figure 11.1 Positively skewed (non-normal) distribution



**Figure 11.2** Quantile–quantile plot to assess normality

often see these positively skewed distributions because there is a physical boundary at zero; it is not possible to observe a negative value, although it is possible to see larger values for some patients that are far from the bulk of the data.

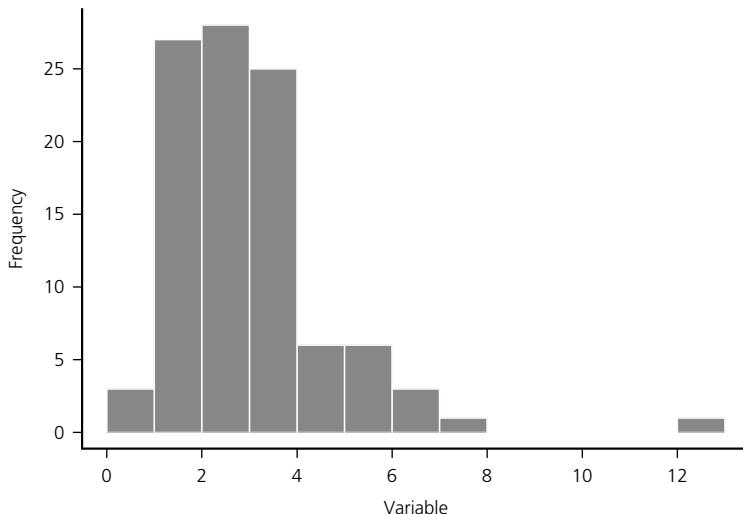
Checking the assumption of normality can be undertaken in one of two ways. Firstly, we have graphical methods, such as a *quantile–quantile plot* (also known as a *normal probability plot*), where normal data displays itself as a straight line. Departures from a straight-line plot are indicative of non-normality. Figure 11.2 is a quantile–quantile plot to assess the normality of 100 observations simulated from the distribution displayed in Figure 11.1, while Figure 11.3 shows the histogram for these same data. The quantile–quantile plot clearly does not conform well to a straight line, and the histogram reflects the positive skewness of these data.

This visual approach based on inspecting the normal probability plot may seem somewhat crude. Most of the test procedures, however, such as the unpaired t-test, are what we call *robust* against departures from normality. In other words, the *p*-values resulting from these tests remain approximately correct, unless we depart substantially from normality, particularly with large sample sizes. The normal probability plot is sensitive enough to detect such substantial departures.

Secondly, we have a statistical test, the *Shapiro–Wilks test*, that gives a *p*-value for the following setting:

$$H_0 : \text{normal} \quad H_1 : \text{non-normal}$$

A significant *p*-value provides evidence that the data are not normally distributed and leads to the rejection of  $H_0$ ; a non-significant *p*-value tells us that there is no evidence for non-normality, and in practice it will be safe to assume that the data are at least approximately normally distributed. See Altman (1991) Section 7.5.3 for further details of this test and some examples.



**Figure 11.3** Histogram for simulated data

Our discussions here suggest that we look for normality in each treatment group separately. In practice, this is not quite what we do; we look at the two groups combined and evaluate the normality of the residuals: that is, the distance between each subject observation and the mean value for that subjects' group. This calculation allows the residuals to be considered as a single group for the purpose of evaluating the assumption of normality. Similarly, this approach is used to deal with ANOVA where there are several *groups* (e.g. treatment group A and treatment group B observations in each of several centres, for example) and also with more complex structures that form the basis of ANCOVA and regression. For example, in regression, we assess the normality of the *residuals*: the vertical differences between each observation and the corresponding value on the fitted line.

## 11.4 Non-normality and transformations

We will concentrate primarily in this section on the parallel-group case with two treatments where for normally distributed data we would be undertaking the unpaired t-test. If the data are clearly non-normal, then the first approach in analysing these data would be to consider transforming to recover normality. As we mentioned in the previous section, it is not uncommon to have positively skewed data with values that cannot be negative. A transformation that often successfully transforms the data to be normal is the log transformation. It does not matter to which base we take these logs; the usual choices would be to base

**Table 11.1** The log transformation

<i>x</i>	$\log_{10}x$
0	$-\infty$
1	0
10	1
100	2
1000	3

e (natural logarithms) or base 10, each of which is equal to a constant multiple of the other. Table 11.1 shows the effect of taking logs of various values to base 10.

The effect of the log transformation on the values 1, 10, 100 and 1000 is to effectively bring them closer together. On the original scale, these numbers get progressively further apart, whereas on the log scale, they become equally spaced. Also, the log transformation gives a negative value for values on the original scale between zero and one. The log transformation *brings in* the large positive values and *throws out* the values below one to become negative, and this has the effect of making the positively skewed distribution more symmetric. If this transformation is successful in recovering normality, we simply analyse the data on the log scale using the unpaired t-test. The resulting *p*-value provides a valid comparison of the two treatments in terms of the means, albeit on the log scale. Similarly, we can calculate 95% CIs for the difference in the means on the log scale.

While the *p*-value allows us to judge statistical significance, the clinical relevance of the finding is difficult to evaluate from the calculated CI because this is now on the log scale. It is usual to *back-transform* the lower and upper confidence limits, together with the difference in the means on the log scale, to give us something on the original data scale that is more readily interpretable. The back-transform for the log transformation is the antilog.

Mean values are usually calculated as *arithmetic means*. However, there is another kind of mean value: the *geometric mean*. The *arithmetic mean* is  $(x_1 + x_2 + \dots + x_n)/n$ , while the geometric mean is defined as  $\sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$ : the *n*th root of all the data values multiplied together. When the antilog is applied to the difference in the means on the log scale, the result is numerically equal to the ratio of the geometric means for the original data. This happens in pharmacokinetics, where it is standard practice to log-transform  $C_{\max}$  and AUC before analysis; it is the ratio of geometric means together with a CI for that ratio that is usually quoted. More generally, we often quote geometric means when we use the log transformation in data analysis.

In other settings, such as ANOVA, ANCOVA and regression, log-transforming the outcome variable is always worth trying, where the outcome variable is a strictly positive quantity, as an initial attempt to recover normality, should the residuals indicate a violation of the normality assumption.

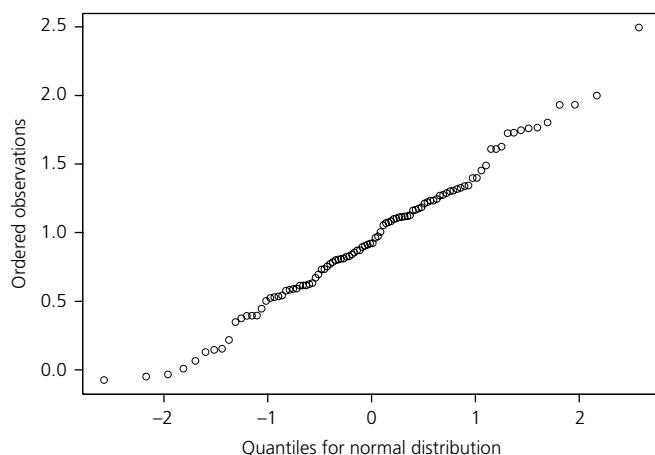
Finally, log-transforming positively skewed data often kills two pigeons with one pebble, recovering both normality and constant variance. With skewed data, the group with the larger outcome values tends to have more spread, and the log transformation generally brings the spread of the data in each group more into line.

The log transformation is by far the most common transformation, but several other transformations are used from time to time to recover normality and constancy of variance. The *square root transformation*,  $\sqrt{x}$ , is sometimes used with a count endpoint, while the *logit transformation*,  $\ln\{x/(1-x)\}$ , can be used where the patient provides a measure that is a proportion, such as the proportion of questions out of 20 answered correctly in a memory test. Note that  $\ln$  is the symbol that denotes natural logarithms (log to base  $e$ ). One slight problem with the logit transformation is that it is not defined when the value of  $x$  is either zero or one. To cope with this in practice, we sometimes add 1/2 to  $x$  and  $(1-x)$  as a *fudge factor* before taking the log of the ratio.

Figure 11.4 is the quantile–quantile plot for the log-transformed data from Figure 11.3, while Figure 11.5 is the histogram of these same log-transformed data. The quantile–quantile plot is approximately linear, indicating that the log transformation has recovered normality for these data, and the histogram conforms more closely to the normal distribution shape. An assumption of normality would now be entirely reasonable on this transformed scale.

If we assume that the log of the endpoint follows a normal distribution, we say that the distribution of the original endpoint is *log-normal*.

Several other distributions are used for certain specialised settings on which statistical tests can be based. The *beta distribution* is often a valid assumption for an endpoint recorded as a proportion. The *gamma distribution* can sometimes be used for *positively skewed data* where the bulk of the data is towards the lower end



**Figure 11.4** Quantile–quantile plot for log-transformed data

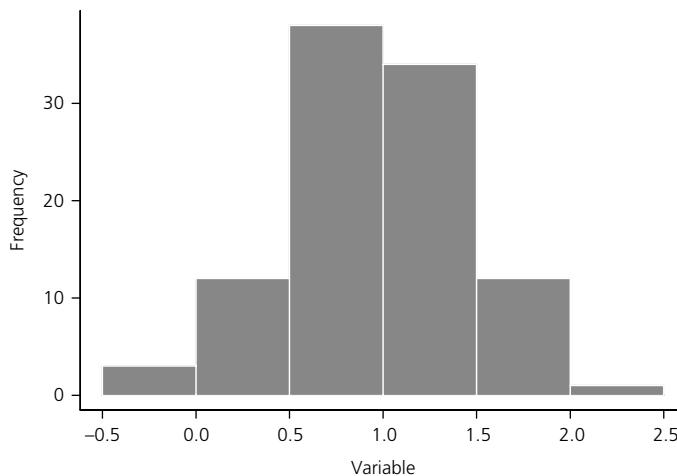


Figure 11.5 Histogram for log-transformed data

of the scale, and there is a long right tail to the distribution. Statistical tests can be developed for both the beta and gamma distributions as an alternative to taking logs and assuming normality. Finally, the *Poisson distribution* is often suited to situations where the endpoint under consideration is a count. These methods are all based on assuming particular parametric functions for the endpoint distributions and are referred to in general as *parametric methods*. Methods that make no assumptions or very weak assumptions about the data distributions are known as *non-parametric methods*.

## 11.5 Non-parametric tests

### 11.5.1 The Mann–Whitney U-test

The hypothetical data in Table 11.2 are simulated from two distinct non-normal distributions. The population from which the observations in group A are taken has a mean of 3, while the population from which the B group observations are taken has a mean of 5.

The Mann–Whitney U-test is equivalent to an alternative test called the *Wilcoxon rank sum test*. These tests were developed independently but subsequently shown to be mathematically the same. We will develop the test using the Wilcoxon rank sum methodology.

The first step in undertaking this test is to order (or *rank*) the observations in both groups combined, from smallest to largest. This ranking is seen in Table 11.2. There are 31 observations in total, and the average of the numbers 1 through to 31 is 16 ( $16 = (1 + 31)/2$ ). If there were no differences on average between the two populations, then the average rank in each of the two groups would be close

**Table 11.2** Hypothetical non-normal data for two groups

Group A ( $n = 15$ )	Rank	Group B ( $n = 16$ )	Rank
1.45	1	3.54	18
2.55	11	3.04	15
4.35	21	5.33	24
1.93	5	3.16	17
3.95	19	3.10	16
7.90	29	2.69	12
1.78	3	2.95	14
4.60	22	5.89	26
1.84	4	5.17	23
2.54	10	2.13	7
5.47	25	4.30	20
2.29	8	8.48	30
2.91	13	12.88	31
1.46	2	6.43	27
2.30	9	2.01	6
		7.03	28

to 16. The *signal* for the Wilcoxon rank sum test is the difference between the observed average rank in one of the two groups minus the expected average rank (equal to 16) under the assumption that the treatments on average are the same. Choose either group to work with. The average of the ranks attached to the 15 observations in our chosen group (group A) is equal to 12.13 (= 182/15 since the total of the ranks in this group is 182). This average is below 16, indicating that in group A, the observations look to be smaller than the observations in group B; and the value of the signal is  $-3.87 (= 12.13 - 16)$ . The standard error attached to this signal is given by

$$se = \sqrt{\frac{n_2(n_1 + n_2 + 1)}{12n_1}} = \sqrt{\frac{16 \times (15 + 16 + 1)}{12 \times 15}} = 1.69$$

where  $n_1$  and  $n_2$  are the sample sizes in groups A (our chosen group) and B, respectively. The signal-to-se ratio for these data is then  $-2.29$ . We obtain the *p*-value by comparing the value of this test statistic with what we would expect to see if the treatments were the same. This null distribution is a special form of the normal distribution called the *standard normal*: the normal distribution with mean zero and standard deviation equal to one. The two-sided *p*-value turns out to be 0.022, a statistically significant result indicating treatment differences. Upon inspecting the data, it is clear that group A, on average, contains smaller observations than group B.

The methodology assumes that the observations are distinct, so a unique ranking can be defined. In many cases, however, this will not be true, and we will see tied values in the data. In these situations, we assign average ranks to

the tied values. For example, had the observations 2.54 and 2.55 both been equal to 2.55, we would have attached the rank 10.5 (the average of the ranks 10 and 11 corresponding to these two observations) to both values; the ranking would then be 8, 9, 10.5, 10.5, 12, 13 and so on. With three values all equal, say, in positions 14, 15 and 16, the average rank 15 is attached to each of the observation. Provided the number of tied values is not too large, the same formulas as earlier can be used to calculate the value of the test statistic. For more frequent ties, the formula for the standard error needs to be modified. More details can be found in van Belle et al. (2004), Section 8.6.3. Example 11.1 provides an application of the Mann–Whitney U-test in a setting where the data are heavily tied.

**Example 11.1** Natalizumab in the treatment of relapsing multiple sclerosis

Miller et al. (2003) report a trial comparing two dose levels of natalizumab (3 mg/kg and 6 mg/kg) with placebo. The primary endpoint was the number of new brain lesions during the six-month treatment period. Table 11.3 presents these data.

The distribution of the number of new lesions (a count endpoint) was clearly non-normal within each treatment group. There is a peak at zero in each group, with progressively fewer patients as the number of lesions increases. A log transformation would not work here because of the zero values for the endpoint. The authors used the Mann–Whitney U-test to compare each natalizumab dose group with placebo obtaining  $p < 0.001$  in each case. Each dose level is significantly better than placebo in reducing the number of new enhancing lesions.

**Table 11.3** Number of new enhancing lesions

	Placebo (n = 71)	Natalizumab (3 mg, n = 68)	Natalizumab (6 mg, n = 74)
No lesions	23 (32%)	51 (75%)	48 (65%)
1–3 lesions	18 (25%)	14 (21%)	20 (27%)
4–6 lesions	13 (18%)	1 (1%)	5 (7%)
7–9 lesions	0	0	0
10–12 lesions	3 (4%)	1 (1%)	0
>12 lesions	14 (20%)	1 (1%)	1 (1%)

### 11.5.2 The Wilcoxon signed rank test

This test is the non-parametric equivalent of the paired t-test. The paired t-test makes a single assumption that the population of differences for each patient follows the normal shape. If this assumption is violated, the  $p$ -value from the paired t-test is not correct – although, as with the unpaired t-test, the paired t-test is robust against modest departures from normality.

The test is again based on a ranking procedure. Under the assumption that the treatments being compared, A and B, are in truth on average the same, the number of positive A – B differences should be approximately equal to the number of negative A – B differences.

For example, suppose there are 12 patients; under the assumption of equal treatments we should see approximately six positive A – B differences and six negative A – B differences. Further, the magnitude of the positive and negative differences should be similar. Having calculated the A – B differences, the first step is to assign ranks to all patients according to the magnitude of those differences (ignoring the sign). Secondly, we add up the ranks attached to those differences that were positive. The average rank of the positive differences should be equal to the average rank of the negative differences, and both should be equal to 6.5 (the average of the numbers 1 to 12) under the null hypothesis of no treatment differences. The signal for the test statistic can be calculated from the observed average rank for the positive differences minus the expected average rank for those positive differences. The standard error associated with this signal is then calculated, and we compare the signal-to-se ratio with the standard normal distribution to give us the *p*-value.

Had we chosen to calculate the differences B – A, the signal would equal the signal based on the A – B differences, but with the opposite sign, and the two-sided *p*-value would be completely unchanged.

### 11.5.3 General comments

Non-parametric tests, as seen in the two procedures outlined earlier in Section 11.5, are based on some form of data ranking. Once the data are ranked, the test is based entirely on those ranks; the original data play no further part. Therefore, the behaviour of ranks determines the properties of these tests, and it is this element that gives them their robustness. Whatever the original data look like, once the rank transformation is performed, the patterns in the data become predictable under the null hypothesis.

It may seem strange that the normal distribution plays a part in the *p*-value calculations in Sections 11.5.1 and 11.5.2. The appearance of this distribution is in no sense related to the underlying distribution of the data. For the Mann–Whitney U-test, for example, it relates to the behaviour of the average of the ranks within each of the individual groups under the null hypothesis that, on average, the treatments are the same, where the ranks in those groups of sizes  $n_1$  and  $n_2$  are simply a random split of the numbers 1 through to  $n_1 + n_2$ .

In terms of summary statistics, means are less relevant because of the inevitable skewness of the original data (otherwise, we would not be using a non-parametric test). This skewness frequently produces extremes, which then tend to dominate the calculation of the mean. Medians are usually a better, more stable description of *average*.

Extending non-parametric tests to more complex settings, such as regression, ANOVA and ANCOVA, is not straightforward (although there are some simple extensions), but this aspect of these methods limits their usefulness. The van Elteren test (van Elteren, 1960) is a stratified form of the Mann–Whitney U-test and gives the possibility to undertake simple adjusted analyses.

## 11.6 Advantages and disadvantages of non-parametric methods

It is fair to say that statisticians disagree somewhat regarding the value of non-parametric methods. Some statisticians view them very favourably, while others are reluctant to use them unless there is no other alternative.

Clearly, the main advantage of a non-parametric method is that it makes no assumptions about the underlying distribution of the data. In contrast, the corresponding parametric methods make specific assumptions: for example, that the data are normally distributed. Does this matter? Well, as mentioned earlier, the t-tests, even though in a strict sense they assume normality, are quite robust against departures from normality. In other words, you need to be some way off normality for the  $p$ -values and associated CIs to become invalid, especially with the kinds of moderate to large sample sizes that we tend to see in our clinical trials.

Further, non-parametric methods have several disadvantages:

- With parametric methods, CIs can be calculated, which link directly with the  $p$ -values; recall the discussion in Section 9.8. With non-parametric methods, the  $p$ -values are based directly on the calculated ranks, and it is not easy to obtain a CI in relation to parameters that have a clinical meaning that links with this. This compromises our ability to provide an assessment of clinical benefit.
- Non-parametric methods reduce power. If the data are normally distributed, either on the original scale or following a transformation, the non-parametric test will be less able than the parametric alternatives to detect differences (should they exist).
- Non-parametric procedures tend to be simple two-group comparisons, although there are some extensions to allow stratified analyses (van Elteren tests). Further, there are some general non-parametric approaches akin to analysis of covariance, although they are somewhat complex. The advantages provided by ANCOVA – correcting for baseline imbalances, increasing precision and looking for treatment-by-covariate interactions – are not readily extended to the non-parametric framework.

For these reasons, non-parametric methods are used infrequently within the context of clinical trials, and they are generally considered only if a corresponding parametric approach, either directly or following a data transformation, is not possible.

## 11.7 Outliers

An *outlier* is an unusual data point well away from most of the data. Usually, the outlier in question will not have been anticipated, and the identification of these points and appropriate action should be decided at the *blind review*.

The appropriate method for dealing with an outlier will depend somewhat on the setting, but one or two general points can be made. The first thing that should be done is to check that the value is both possible from a medical perspective and correct. For example, a negative survival time is not possible, and this could be a consequence of recording an incorrect date at randomisation, for example. Hopefully, these problems will have been picked up at the data cleaning stage, but sometimes things slip through. Clearly, if the data point is incorrect, it should be corrected before analysis.

An extreme, large positive value may sometimes be a manifestation of an underlying distribution of data that is heavily skewed. Transforming the data to be more symmetric may then be something to consider.

Analysing the data with and without the outliers may ultimately be the appropriate approach, just to ensure that the conclusions are unaffected by their presence. The ICH E9 provides some guidance on this point.

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'If no procedure for dealing with outliers was foreseen in the trial protocol, one analysis with the actual values and at least one other analysis eliminating or reducing the outlier effect should be performed and differences between their results discussed'.*

## CHAPTER 12

# Equivalence and non-inferiority

### 12.1 Demonstrating similarity

In this chapter, we will move away from superiority trials to look at methods for evaluating equivalence and non-inferiority. The setting in most cases here is the comparison of a new treatment to an active control where we are looking to demonstrate similarity (in some appropriately defined sense) between the two treatments.

It should be clear from our earlier development, especially the discussion in Section 9.10.1, that obtaining a non-significant  $p$ -value in a superiority evaluation does not demonstrate that the two treatments are the same or even similar; a non-significant  $p$ -value may simply be the result of a small trial, with low power even to detect large differences. ICH E9 makes a clear statement in this regard.

#### ***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate'.*

Unfortunately, this issue is not especially well understood within the clinical trial community, and the misinterpretation of non-significant  $p$ -values is all too common. See Jones et al. (1996) for further discussion on these points.

Equivalence trials are routinely used in the evaluation of bioequivalence, and the methodology there is well established; both European and FDA guidelines exist. There are also settings where there is a need to establish therapeutic equivalence, and Ebbutt and Frith (1998) provide a detailed case study in the development of an alternative propellant for the asthma inhaler. The standard metered-dose inhaler has a CFC gas as the propellant, and this causes environmental damage. An alternative propellant was sought to reduce the environmental burden. Trials developing the alternative propellant were all equivalence

trials making comparisons between the two devices: the existing inhaler and the new inhaler with the alternative propellant. It was necessary to show that the new inhaler provided an improvement in lung function that was on average similar to that provided by the existing inhaler. The requirement was to match the effectiveness of the two inhalers as closely as possible, as eventually the new inhaler would be used as a substitute for the existing inhaler. For example, should the new inhaler result in a substantially greater increase in lung function, this could mean the new device was delivering a higher dosage: an unsatisfactory situation that might be associated with safety issues.

In recent years, we have also seen considerable development of biosimilars. In both Europe and the United States, guidelines exist in relation to this area; see, for example, CHMP (2014) ‘Guideline on similar biological medicinal products’ and FDA (2014) ‘Reference Product Exclusivity for Biological Products Filed Under Section 351(a) of the PHS Act’.

For non-inferiority, we are looking to establish that our new treatment is *at least as good as or no worse than* an existing treatment. We of course need to define what we mean by *at least as good as or no worse than* in an operational sense for this to be unambiguous.

As mentioned in Section 1.10, there are several areas where we would want to conduct non-inferiority trials: firstly, where including a placebo is not possible due to either practical or ethical issues and we are therefore looking to demonstrate the efficacy of the new treatment indirectly, by showing similarity to an established active treatment; secondly, where it is necessary to show that there is no important loss of efficacy for a new treatment compared to an existing control treatment in a setting where the new treatment offers advantages outside of efficacy; and finally, where we want to show that a new treatment does not increase the incidence of certain adverse events beyond a certain level. See Section 12.7.4 for a discussion of this final point in the diabetes setting. It must also be noted that within the same trial, there may be a mixture of superiority and non-inferiority comparisons. When we talk about a non-inferiority trial, we are usually referring to the fact that the primary comparison is a non-inferiority comparison. But of course, there may be secondary or other comparisons: for example, in terms of quality of life, tolerability or other efficacy endpoints that are evaluating superiority.

Finally, before we move on to look at statistical methods, it is worth mentioning that there is some confusion regarding the term *non-inferiority*. In a strict sense, any reduction in efficacy means the new treatment is not as good as the existing treatment and so is inferior. We use the term *non-inferiority* in a clinical trial sense, however, to denote a non-zero but clinically irrelevant reduction in efficacy. The expression *one-sided equivalence* is sometimes used as an alternative to *non-inferiority*.

A good overview of various aspects of non-inferiority trials is provided by Kaul and Diamond (2006).

## 12.2 Confidence intervals for equivalence

The first step in establishing equivalence is to define what we mean by *equivalence*. Following Ebbutt and Frith (1998), suppose we are looking to establish the equivalence of a new asthma inhaler device with an existing inhaler device in a trial setting, and further suppose that our clinical definition of *equivalence* in relation to change from baseline in peak expiratory flow (PEF) is 15 l/min. In other words, if the treatment difference in the mean increase in PEF following four weeks of treatment is <15 l/min, we will conclude that the two devices provide a clinically equivalent benefit on average. We may argue whether 15 l/min is the appropriate value, but whatever we do, we must choose a value. The  $\pm 15$  l/min values are the *equivalence margins*, and the interval  $-15$  l/min to  $+15$  l/min is the *equivalence region* (see Figure 12.1).

The next step is to undertake the trial and calculate the 95% confidence interval (CI) for the difference in the means (mean increase in PEF on new inhaler [ $\mu_1$ ] – mean increase in PEF on existing inhaler [ $\mu_2$ ]). As a first example, suppose this CI is  $(-7$  l/min,  $12$  l/min). In other words, we can be 95% confident that the true difference,  $\mu_1 - \mu_2$ , is between  $7$  l/min in favour of the existing inhaler and  $12$  l/min in favour of the new inhaler.

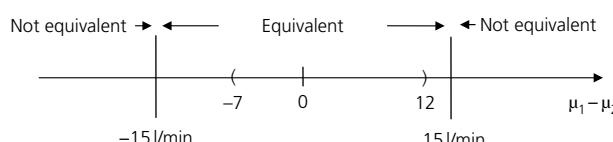
As seen in Figure 12.1, this CI is completely contained between the equivalence margins  $-15$  l/min to  $15$  l/min, and all values for the treatment difference supported by the CI are compatible with the definition of clinical equivalence. In this case, we have established equivalence as defined.

In contrast, suppose that the 95% CI had turned out to be  $(-17$  l/min,  $12$  l/min). This interval is not entirely within the equivalence margins, and the data support potential treatment differences below the lower equivalence margin. In this case, we have not established equivalence.

Note that there are no conventional *p*-values here. Such *p*-values have no role in evaluating equivalence; establishing equivalence for the moment has been based entirely on the use of CIs.

### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Statistical analysis is generally based on the use of confidence intervals. For equivalence trials, two-sided confidence intervals should be used. Equivalence is inferred when the entire confidence interval falls within the equivalence margins'.*



**Figure 12.1** Establishing equivalence

The CIs we have used to date are all two-sided. We will talk later about one-sided CIs.

### 12.3 Confidence intervals for non-inferiority

For non-inferiority, the first step involves defining a non-inferiority margin. Suppose we are developing a new treatment for hypertension, and the reason the new treatment is better may be that it has fewer side effects, although we are not anticipating any improvement in terms of efficacy. Indeed, we may be prepared to pay a small price in terms of efficacy for a reduction in the side effects profile: say, up to 2 mmHg in the mean reduction in diastolic blood pressure.

In Figure 12.2,  $\mu_1$  and  $\mu_2$  are the mean reductions in diastolic blood pressure in the test treatment and active control groups, respectively. If the difference in the means is above zero, the test treatment is superior to the active control; if the difference is zero, they are identical. If the difference falls below zero, the test treatment is not as good as the active control. This, however, is a price we are prepared to pay, but only up to a mean reduction in efficacy of 2 mmHg; beyond that, the price is too great. The non-inferiority margin is therefore set at -2 mmHg.

Step 2 is then to run the trial and compute the 95% CI for the difference,  $\mu_1 - \mu_2$ , in the mean reductions in diastolic blood pressure. In our example, suppose that this 95% CI turns out to be (-1.5 mmHg, 1.8 mmHg). As seen in Figure 12.2, all values within this interval are compatible with our definition of non-inferiority. In this case, the non-inferiority of the test treatment compared to the control treatment has been established. In contrast, had the 95% CI been, say, (-2.3 mmHg, 1.8 mmHg), non-inferiority would not have been established since the lower end of that CI falls below -2 mmHg. Note again that there is no mention of conventional  $p$ -values; they have no part to play in establishing non-inferiority.

To demonstrate non-inferiority, only one end of the CI matters; in our example, it is simply the lower end that needs to be above -2 mmHg. It is therefore not strictly necessary to calculate the upper end of the interval, and sometimes we leave this unspecified. The resulting CI with just the lower end is called a *one-sided 97.5% CI*; the two-sided 95% CI cuts off 2.5% at each of the lower and upper ends, and leaving the upper end undefined results in just 2.5% being cut

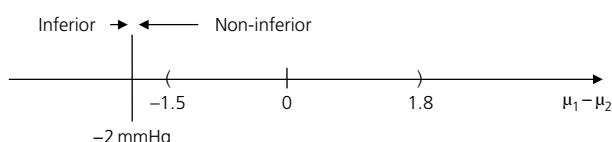


Figure 12.2 Establishing non-inferiority

off at the lower end. The entire CI must be to the right of the non-inferiority margin for non-inferiority to be established. Note that the conclusions we draw when we use a one-sided 97.5% CI in place of a two-sided 95% CI are unchanged as the lower end of the interval is unaffected.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

'For non-inferiority a one-sided confidence interval should be used'.

Example 12.1 provides an application that we will return to later in this chapter.

**Example 12.1** Fluconazole compared to amphotericin B in preventing relapse in cryptococcal meningitis

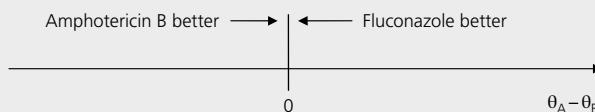
The aim of this study reported by Powderly et al. (1992) was to establish the non-inferiority of the test treatment, fluconazole, compared to an established treatment, amphotericin B, in preventing the relapse of cryptococcal meningitis in HIV-infected patients. It was thought that fluconazole would be less effective than amphotericin B but would offer other advantages in terms of reduced toxicity and ease of administration; fluconazole was an oral treatment, while amphotericin B was given intravenously. The non-inferiority margin was set at -15% in terms of relapse rates:

Let  $\theta_F$  = relapse rate on fluconazole

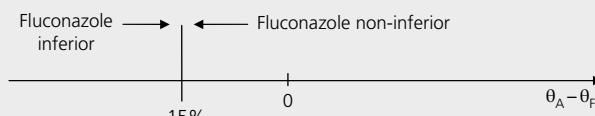
Let  $\theta_A$  = relapse rate on amphotericin B

In Figure 12.3, a positive difference for  $\theta_A - \theta_F$  indicates that fluconazole is a better treatment, while a negative difference indicates that amphotericin B is better.

The non-inferiority margin has been set at -15%. Figure 12.4 displays the non-inferiority region. We need the (two-sided) 95% CI (or the one-sided 97.5% CI) to be entirely within this non-inferiority region for non-inferiority to be established: that is, to the right of -15%.



**Figure 12.3** Difference in relapse rates



**Figure 12.4** Definition of non-inferiority

## 12.4 A *p*-value approach

Although conventional *p*-values have no role in equivalence or non-inferiority trials, there is a *p*-value counterpart to the CI approach. The CI methodology was developed by Westlake (1981) in the context of bioequivalence; Schuirmann (1987) developed a *p*-value approach that was mathematically connected to these CIs, although somewhat more difficult for non-statisticians to understand! It nonetheless provides a useful way of thinking, particularly when we later consider type I and type II errors in this context, and the sample size calculation. We will start by looking at equivalence and use  $\pm\Delta$  to denote the equivalence margins.

Within the framework of hypothesis testing, the null and alternative hypotheses of interest for equivalence when dealing with means are as follows:

$$H_0: \mu_1 - \mu_2 \leq -\Delta \text{ or } \mu_1 - \mu_2 \geq \Delta$$

$$H_1: -\Delta < \mu_1 - \mu_2 < \Delta$$

In this case, the alternative hypothesis states that the two treatments are equivalent; the null hypothesis says that the two treatments are not equivalent. Note that the alternative hypothesis captures the *objective*; we are trying to disprove the null to establish equivalence.

These hypotheses can be expressed as two separate sets of hypotheses corresponding to the lower and upper ends of the equivalence range:

$$H_{01}: \mu_1 - \mu_2 \leq -\Delta \quad H_{11}: \mu_1 - \mu_2 > -\Delta$$

$$H_{02}: \mu_1 - \mu_2 \geq \Delta \quad H_{12}: \mu_1 - \mu_2 < \Delta$$

Undertaking two tests each at the 2.5% level, one for  $H_{01}$  vs.  $H_{11}$  and one for  $H_{02}$  vs.  $H_{12}$ , can be shown to be mathematically connected to the CI approach developed earlier. In particular, if both tests give a statistically significant *p*-value at the 2.5% significance level, then the 95% CI for the difference in the means will be entirely contained within the equivalence margins  $\pm\Delta$ . Conversely, if the 95% CI is contained within the equivalence margins, then each of the two tests will give *p*-values significant at the 2.5% level. These two sets of hypotheses are both one-sided comparisons; the first set looks to see whether the treatment difference,  $\mu_1 - \mu_2$ , is either  $\leq$  or  $> -\Delta$ , while the second set looks to see if  $\mu_1 - \mu_2$  is either  $\geq$  or  $< \Delta$ . The approach using *p*-values is therefore known as the *two one-sided tests approach*. Following on from the earlier quote specifying the role of CIs, ICH E9 states:

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Operationally, this is equivalent to the method of using two simultaneous one-sided tests to test the (composite) null hypothesis that the treatment difference is outside the equivalence margins versus the (composite) alternative hypothesis that the treatment difference is within the margins'.*

With the  $p$ -value methodology, we reject the null hypothesis  $H_0$  in favour of the alternative hypothesis  $H_1$ , provided the two (one-sided)  $p$ -values are  $\leq 2.5\%$ . We have then established equivalence, and we can talk in terms of the treatments being *significantly equivalent*. The terminology sounds almost contradictory but is a correct statement. If either of the two  $p$ -values is above 2.5%, the treatments are *not significantly equivalent*.

For non-inferiority, the one-sided comparison

$$H_0 : \mu_1 - \mu_2 \leq -\Delta \quad H_1 : \mu_1 - \mu_2 > -\Delta$$

yields a  $p$ -value that links with the one-sided 97.5% CI for establishing non-inferiority. If the  $p$ -value from this test is statistically significant at the 2.5% level, the one-sided 97.5% CI will be entirely to the right of the non-inferiority margin  $-\Delta$ , and vice versa. If we see this outcome, we can talk about the new treatment being *significantly non-inferior* to the active control. Alternatively, if we get a non-significant  $p$ -value, the new treatment is *not significantly non-inferior*. Note that it is not possible under this same set of conditions to say that the new treatment is inferior! We do not have enough evidence to conclude in favour of non-inferiority, but nothing stronger than that.

Using a 2.5% significance level for non-inferiority may initially appear out of line with the conventional 5% significance level for superiority. However, a moment's thought should suffice to realise that in a test for superiority, we would never make a claim for a new treatment if that treatment was significantly worse than control. We would only ever make a claim if the new treatment was significantly better than control: so for superiority, we are effectively conducting a one-sided test at the 2.5% level to lead us to a positive conclusion for the new treatment.

In practice, I always prefer using CIs for evaluating equivalence and non-inferiority rather than these associated  $p$ -values. This is because the associated  $p$ -values tend to get mixed up with conventional  $p$ -values for detecting differences. The two are not the same and look at quite different things. The CI approach avoids this confusion and provides a technique that is easy to present and interpret.

## 12.5 Assay sensitivity

One concern with equivalence and non-inferiority trials is that a positive conclusion of equivalence/non-inferiority could result from an insensitive trial by default. If, for example, equivalence is established, this could mean either that the two treatments are equally effective or that they are equally ineffective. If the chosen endpoints are insensitive, dosage of the control drug is too low, patients are recruited who do not have the target condition and the trial is conducted in a sloppy fashion with lots of protocol deviators and dropouts, the treatments will inevitably look similar! Clearly, we must ensure that a

conclusion of equivalence/non-inferiority from a trial is a true reflection of the treatments. Regulatory guidelines (see, for example, ICH E10) talk in terms of *assay sensitivity* as a requirement of a clinical trial that ensures this.

### ***ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'***

*'Assay sensitivity is a property of a clinical trial defined as the ability to distinguish an effective treatment from an ineffective treatment'.*

Of course, assay sensitivity applies in the same way to trials that evaluate superiority, but in many cases, things take care of themselves. A conclusion of superiority, of itself, tells us the trial is a sensitive instrument; otherwise, superiority would not have been detected.

Ensuring assay sensitivity is best achieved through good design and trial conduct to a high standard. The current design should be in line with trials that have historically shown the active control treatment to be effective and more generally the following aspects need to be carefully considered:

- *Entry criteria:* The patients should preferably be at the moderate to severe end of the disease scale.
- *Dose and regimen of the active control:* In line with standard practice.
- *Endpoint definition and assessment:* Use established endpoints.
- *Run-in/washout period to exclude ineligible patients:* Avoid diluting the patient population.

Further, the trial conduct should protect against any compromise of assay sensitivity, and the following, for example, should be avoided:

- Poor compliance
- Use of concomitant medication that could interfere with response to the new and active treatments
- Poor application of diagnostic criteria
- Unblinding
- Unnecessary dropouts and missing data
- Lack of objectivity and consistency in the evaluation of endpoints

It is also possible at the analysis stage to use the trial data to support assay sensitivity. Ebbutt and Frith (1998) investigate the mean change from baseline in the active control group and observe that the magnitude of the change is in line with what one would expect historically in terms of the effect of the active control. In a trial where the primary endpoint is a binary outcome, seeing a response rate in the active control group similar to response rates seen historically supports assay sensitivity. In contrast, if the response rate in the current trial is higher or lower than expected, then assay sensitivity could be drawn into question. A higher rate may indicate a population that is less severe than that used previously, while a lower rate could indicate dosages that are too low, for example.

Finally, one sure way to investigate assay sensitivity is to include a placebo group as a third arm in the trial. This allows direct assessment of assay sensitivity by comparing the active control with placebo where statistically significant differences would need to be seen. However, including a placebo arm is only possible where it is ethically and practically reasonable to do so, and in many equivalence/non-inferiority settings, this will not be the case. A related point is that some therapeutic settings are unsuitable for equivalence/non-inferiority trials unless a placebo arm is included: for example, depression, anxiety, allergic rhinitis and Alzheimer's disease and other dementias, where established effective drugs do not consistently demonstrate effects over placebo. Many therapeutic-specific regulatory guidelines in these areas actively discourage the use of non-inferiority trials in the absence of a placebo group.

***CHMP (2008): Guideline on Medicinal Products for the Treatment of Alzheimer's Disease and Other Dementias***

*'Active control parallel group trials comparing the new treatment to an already approved treatment are needed in order to give the comparative benefit/risk ratio of the new treatment, at least in those treatments intended for symptomatic improvement. However, due to concerns over assay sensitivity, the use of a non-inferiority design versus active control only, will not be accepted as proof of efficacy. Therefore three-arm studies with placebo, test product and active control or a superiority trial are the preferred design options.'*

The CHMP (2012) addendum to the note for guidance on bacterial infections states that for certain infections – for example, acute bacterial sinusitis and superficial skin infections – '*An approval based solely on non-inferiority studies is not currently acceptable*'.

A related concept used in this area is *historical evidence of sensitivity to drug effects*. This idea, introduced initially in ICH E10, refers to the ability of effective treatments to consistently show an advantage over placebo in appropriately designed and conducted clinical trials. As mentioned in the previous paragraph, there are certain therapeutic settings where this is not the case.

## **12.6 Analysis sets**

In superiority trials, the full analysis set is the basis for the primary analysis. As discussed in Section 7.2, the regulators prefer this approach, in part, because it gives a conservative view of the new treatment. In equivalence/non-inferiority trials, however, the full analysis set (FAS) is not conservative and may result in the treatments looking more similar than they are in truth. This is because the full analysis set includes the patients who have not complied with the medication schedules and have not followed the study procedures and including such

patients will, if anything, weaken treatment differences. Therefore, regulators have previously requested analyses based on the per-protocol set (PPS), since these tend to exaggerate the true treatment difference.

***CPMP (2000): 'Points to Consider on Switching Between Superiority and Non-inferiority'***

*'In a non-inferiority trial, the full analysis set and the per-protocol analysis set have equal importance and their use should lead to similar conclusions for a robust interpretation'*

Historically, regulators have considered analyses based on the FAS and PPS as co-primary. There has been a common misconception that the PPS has been primary for equivalence/non-inferiority trials. This was not the case, and both analysis sets have been required to support equivalence/non-inferiority for a robust conclusion.

However, the estimand framework has changed the way we think about the role of analyses based on the FAS and the PPS for equivalence/non-inferiority trials moving forward.

***ICH E9(R1) (2019): 'Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials'***

*'Estimands that are constructed with one or more intercurrent events accounted for using the treatment policy strategy present similar issues for non-inferiority and equivalence trials as those related to analysis of the FAS under the ITT principle. Responses in both treatment groups can appear more similar following discontinuation of randomised treatment or use of another medication for reasons that are unrelated to the similarity of the initially randomised treatments'.*

For equivalence and non-inferiority trials, the treatment policy strategy may make the treatments look more similar. Alternative strategies, such as the hypothetical and composite strategies, can be constructed that may provide the basis for a more appropriate evaluation of 'similarity'. Consider the intercurrent event of taking rescue medication. The hypothetical strategy targets the treatment effect had rescue not been available, while the composite estimand considers taking rescue medication to be failure of the treatment; both approaches lead to an attenuation of the treatment difference that is anti-conservative for concluding similarity.

## **12.7 The choice of $\Delta$**

One of the most difficult aspects of the design of equivalence and non-inferiority trials is the choice of the margin(s). The exception is bioequivalence, where margins are well established, and this area will be dealt with first.

### 12.7.1 Bioequivalence

The equivalence margins for bioequivalence specified by both the FDA (2001) *Statistical Approaches to Establishing Bioequivalence* and the CPMP (2001) *Note for Guidance on the Investigation of Bioavailability and Bioequivalence* require that the ratios of the geometric means,  $\mu_1^{\text{GM}}/\mu_2^{\text{GM}}$ , for the two treatments lie between 0.80 and 1.25. This requirement applies to both AUC and  $C_{\max}$ . A deviation from the rules for therapeutic equivalence is that 90% CIs are used for bioequivalence rather than 95%, leading to a more relaxed requirement.

The reason ratios of geometric means are used in this context follows from what was discussed in Section 11.4: the distributions of AUC and  $C_{\max}$  tend to be positively skewed, and the log transformation is applied to recover normality.

By taking logs, the condition  $0.8 < (\mu_1^{\text{GM}}/\mu_2^{\text{GM}}) < 1.25$  can be translated into the following requirement for  $\mu_1^* - \mu_2^*$ , where  $\mu_1^*$  is the mean AUC (or  $C_{\max}$ ) on the log scale for the test treatment and  $\mu_2^*$  is the mean AUC (or  $C_{\max}$ ) on the log scale for the active control:

$$\ln(0.80) < \mu_1^* - \mu_2^* < \ln(1.25)$$

$$\text{or equivalently } -0.22 < \mu_1^* - \mu_2^* < 0.22$$

Note that on the log scale, the margins are symmetric around zero.

### 12.7.2 Therapeutic equivalence, biosimilars

The rules for therapeutic equivalence are different from those for bioequivalence. The choice of margin is a mixture of statistical and clinical reasoning. Strict equivalence is appropriate when we want to consider essential similarity or where the test treatment is to be used as an exact replacement for the active control. In these cases,  $\Delta$  should be chosen to be a completely irrelevant difference from a clinical point of view. The primary endpoint for the studies covered by Ebbutt and Frith (1998) was PEF rate, a measure of lung function. These authors based their choice of  $\Delta = 15 \text{ l/min}$  on several considerations:

- Previous trials with the beta-agonist salmeterol had given an average difference from placebo in PEF of 37 l/min, while the effect of inhaled steroids was of the order of 25 l/min, and  $\Delta$  was chosen as a proportion of those effects.
- Typically, a mean improvement of 70 l/min would be seen following treatment with a short-acting beta-agonist, and  $\Delta$  was chosen to be about 20% of this.
- Discussion with practitioners suggested that 15 l/min was clinically irrelevant.

In terms of establishing biosimilarity, the margin will be at most the largest difference between the biosimilar and the originator that is viewed as being clinically acceptable. This should also be below the smallest difference seen in trials that established the therapeutic benefit of the originator over placebo. We will present a case study in biosimilarity in Section 12.11

In the next section, we will discuss the non-inferiority setting. Many of the considerations there also apply to the situation of therapeutic equivalence and biosimilarity.

### 12.7.3 Non-inferiority

In the context of using a non-inferiority trial to demonstrate that a test treatment is efficacious, the following provides a statistical approach for the choice of  $\Delta$ . Consider, in the setting of hypertension, the trials that historically compared the active control with placebo. In a meta-analysis (see Chapter 18), suppose that the 95% CI for the active control treatment effect in terms of the fall in diastolic blood pressure is (4.5 mmHg, 10.3 mmHg). The CI tells us that the active control may only be 4.5 mmHg better than placebo, and clearly,  $\Delta$  would need to be chosen to be considerably less than 4.5 mmHg. If we allow a margin equal to 4.5 mmHg in the non-inferiority trial comparing the new treatment to the active control, we are postulating that the new treatment is just as good as the active control if it gets within 4.5 mmHg – and under such circumstances, we could find ourselves simply developing another placebo! Defining  $\Delta$  to be one-half (= 2.25 mmHg) or one-third (= 1.50 mmHg) of the lower bound of this CI would give statistical confidence coming out of our non-inferiority trial with a positive result that the test treatment is at worst either 2.25 mmHg or 1.50 mmHg less efficacious than the active control and that the test treatment therefore still maintains a clear advantage over placebo. This value for  $\Delta$  would need to be additionally justified on clinical grounds: that is, as an irrelevant difference clinically. Example 12.2 provides an application of this way of thinking.

#### **Example 12.2** Prevention of venous thromboembolism after total hip replacement

The RE-NOVATE trial (Eriksson et al., 2007) compared dabigatran etexilate at two doses (150 mg and 220 mg) with enoxaparin 40 mg once daily (active control) in a randomised, double-blind non-inferiority trial. The primary efficacy outcome was a composite of total venous thromboembolic events and all-cause mortality during the four- to five-week treatment period. In the absence of placebo-controlled trials using enoxaparin for four to five weeks, the authors undertook a meta-analysis of placebo-controlled trials that used enoxaparin for one to two weeks. The 95% CI for the difference in event rates was (23.2%, 42.6%), showing that with 95% confidence, enoxaparin shows a reduction in the rate of venous thromboembolic events and all-cause mortality at least as large as 23.2%. From this, a conservative non-inferiority margin of 7.7% was chosen, equal to one-third of the lower end of the 95% CI. A positive result in favour of non-inferiority would therefore support the conclusion that the experimental treatment was no more than 7.7% worse than enoxaparin, still keeping it well away from placebo.

The 95% CIs for the difference between the event rates for the two doses of dabigatran etexilate and enoxaparin were ( $-0.6\%$ ,  $4.4\%$ ) for the 150 mg dose and ( $-2.9\%$ ,  $1.6\%$ ) for the 220 mg dose, in both cases leading to a conclusion of non-inferiority since the upper ends of these CIs are below 7.7%, the pre-specified margin. It is perfectly acceptable and appropriate to claim, with 95% confidence, that the 150 mg and 220 mg doses have event rates at most 4.4% and 1.6% higher than the rate for enoxaparin.

In situations where we are trying to show no important loss of efficacy of a test treatment to an active control, it is not possible to be entirely prescriptive about methods for choosing  $\Delta$ . For example, if a new treatment provides an advantage over the existing treatment in terms of safety, the price we are prepared to pay in terms of efficacy for this advantage will depend on the extent of the safety advantage. In these cases, it is not appropriate to think simply in terms of preserving a proportion of the effect of the active control over placebo. Further, if the active control effect over placebo is large, then preserving only a proportion does not fit with the objectives of the non-inferiority evaluation if we are looking to conclude similarity.

#### **CHMP (2005): 'Guideline on the Choice of Non-Inferiority Margin'**

*'Alternatively the aim may be to provide data to show that there is no important loss of efficacy if the test product is used instead of the reference. This is probably the most common aim of non-inferiority trials. The choice of delta for such an objective cannot be obtained by looking only at past trials of the comparator against placebo. Ideas such as choosing delta to be a percentage of the expected difference between active and placebo have been advocated, but this is not considered an acceptable justification for the choice. To adequately choose delta an informed decision must be taken, supported by evidence of what is considered an unimportant difference in the particular disease area'.*

#### **12.7.4 The 10% rule for cure rates**

Historically, it has been relatively common practice to use a  $\Delta$  of 10% for cure rates and protection when dealing, for example, with anti-infectives and vaccines respectively.

#### **CPMP (1999): 'Note for Guidance on Clinical Evaluation of New Vaccines'**

*'In individual trials,  $\Delta$  can often be set to about 10 per cent, but will need to be smaller for very high protection rates'.*

#### **CPMP (2003): 'Note for Guidance on Evaluation of Medicinal Products Indicated for Treatments of Bacterial Infections'**

*'In most studies with antibacterial agents in common indications this ( $\Delta$ ) should likely be 10 per cent, but may be smaller when very high cure rates are anticipated'.*

The message here appears consistent: 10% is likely to be acceptable except when rates are high, say, >90%; and although these regulatory guidelines are from Europe, the FDA position has been similar. In more recent times, however, the regulators have removed these considerations. For example, in the 2005 update of the 1999 guideline (CHMP [2005]), there is no mention of the 10%. In a rare

disease in which only one or a small number of treatments currently exist, the regulators may be willing to relax the 10% to 12.5% or even 15%. In contrast, for common diseases, the regulators may suggest a tighter  $\Delta$ , arguing that in the interests of public health, the new treatment will only be acceptable if its performance is very close to the active control.

There are one or two other specific therapeutic settings where there has been more guidance on the choice of  $\Delta$ . For example, for patients with ST Segment Evaluation Acute MI:

***CPMP (2003): 'Points to Consider on the Clinical Development of Fibrinolytic Medicinal Products in the Treatment of Patients with ST Segment Evaluation Acute Myocardial Infarction (STEMI)'***

*'In the recent past differences of 14 per cent relative or 1 per cent absolute (whichever proves smallest) have been accepted. These margins were based on "all cause mortality" rates at day 30 close to 6.5–7 per cent'.*

In the area of diabetes, the FDA have expressed concerns regarding cardiovascular safety among treatments effective for glycaemic control. Until recently, there has been a requirement for new treatments to be evaluated for non-inferiority against placebo to eliminate unacceptable increases in major cardiovascular events (MACE). Based on their experience, however, this requirement has been withdrawn and the emphasis shifted: now the focus is on the '*size and nature of the safety databases needed to support drugs for chronic use to improve glycemic control in patients with type 2 diabetes*' (FDA (2020)). However, trials conducted under the pre-2020 framework are illustrative. For example, Cannon et al. (2020) evaluated ertugliflozin in type 2 diabetes in a non-inferiority comparison against placebo where the hazard ratio for the primary endpoint of time to a composite of death from cardiovascular causes, nonfatal myocardial infarction or nonfatal stroke was evaluated against a margin of 1.3, a 30% increase in the hazard for the event.

One final point on the choice of  $\Delta$  that is relevant in all therapeutic settings, and certainly in relation to cure rates and antibiotics, is that the regulators could very well change their minds about what is and is not acceptable for  $\Delta$  if the performance of the active control in the trial deviates from what was expected. For example, suppose that a  $\Delta = 10\%$  was chosen with an expected cure rate of 85% for the active control. If the cure rate for the active control in the trial turns out to be 93%, the regulators may view 10% as too large a value and may suggest a reduction to, say, 5% in this context. This is always a review issue once the data are in hand.

### **12.7.5 The synthesis method**

The method detailed in Section 12.7.3 for determining  $\Delta$  is called the *fixed margin approach*. In this approach, the value for  $\Delta$  is determined from historical data, clinical considerations and the data from the current active controlled trial used to establish (or not) non-inferiority based on this margin.

An alternative approach, the *synthesis approach*, does not require a margin to be pre-specified. This method combines the estimated treatment effect from the active control trial with the estimated effect of the active control compared to placebo from the historical data to provide an indirect estimate of the treatment effect of the experimental treatment compared to placebo.

In the example introduced at the beginning of Section 12.7.3, we were considering a hypothetical example where the primary endpoint was the fall in diastolic blood pressure and historical data on the reference product suggested that it was on average 7.4 mmHg better than placebo. So if  $\bar{x}_R^* - \bar{x}_P^* = 7.4$  mmHg represents the treatment difference between the reference (active control) mean and the placebo mean in the meta-analysis, while  $\bar{x}_T - \bar{x}_R = -0.6$  mmHg equals the observed difference in the test and reference treatments in the non-inferiority trial, then assuming that the conditions of the trials are similar (termed *constancy* – see Section 12.7.6) so we can assume that  $\bar{x}_R \approx \bar{x}_R^*$ , an estimated difference between test treatment and placebo means, is (by adding these two effects together) 6.8 mmHg. The standard error (se) attached to each of these differences, say,  $se_1$  and  $se_2$ , can be combined to give an se for  $\bar{x}_T - \bar{x}_P^*$  using the formula  $se = \sqrt{se_1^2 + se_2^2}$ . The 95% CI for the true test treatment effect (vs. placebo) is then given approximately by  $\bar{x}_T - \bar{x}_P^* \pm 2^* se$ . If the lower end of that CI is greater than zero, we have evidence of efficacy for the test treatment. The magnitude of the point estimate for the difference between the test treatment and placebo and the lower limit of the CI provide evidence in relation to the clinical relevance of the effect of the test treatment. In addition, the 95% CI for the difference between the test treatment and the active control allows an evaluation of the potential loss of efficacy when moving from the active control to the test treatment.

The synthesis method treats the data from the current trial and the data from the historical data as if they came from the same randomised trial to indicate what treatment difference would have been seen had placebo been included as an additional arm in the current trial. Clearly, some very strong assumptions are being made here, which are largely unverifiable. For this reason, this method is less acceptable for regulators who like to see a fixed pre-specified value for  $\Delta$ , justified based on statistical and clinical arguments, independently from the active control trial currently being conducted.

## 12.8 Biocreep and constancy

One valid concern that regulators have is the issue of so-called *biocreep*. Demonstrating that a second-generation active treatment is non-inferior to the active control may well mean the new treatment is slightly less efficacious than the active control. Evaluating a third-generation active treatment to the now established second-generation active treatment may lead to further erosion of

efficacy, and so on, until at some stage a new active treatment, while satisfying the *local* conditions for non-inferiority, is indistinguishable from placebo. The FDA discussed this issue many years ago, specifically concerning anti-infectives (FDA [1992], *Points to Consider on Clinical Development and Labeling of Anti-Infective Drug Products*).

The issue of *constancy* concerns the conditions under which the current active control non-inferiority trial is being conducted compared to the conditions under which the active control was established historically. Things may well have changed. For example, the effectiveness of ancillary care may be such that the active control performs rather differently now than it did when the original placebo-controlled trials were undertaken. This may be true, for example, for antibiotics where populations of patients have developed resistance to certain treatments. If this were the case, then the current non-inferiority trial could lead to a misleading conclusion of effectiveness for the new active when the comparator treatment is ineffective in the context of the current trial.

Both biocreep and constancy elements cause nervousness among regulators – so much so that, for example, the FDA Anti-Infective Drugs Advisory Committee (AIDAC) recommended that the non-inferiority design should no longer be used in trials for acute bacterial sinusitis (CDER Meeting Documents; Anti-Infective Drugs Advisory Committee [29 October 2003], [www.fda.gov](http://www.fda.gov)).

## 12.9 Sample size calculations

We will focus our attention in this section on non-inferiority. Within the testing framework, the type I error in this case is, as before, the false positive (rejecting the null hypothesis when it is true), which now translates into concluding non-inferiority when the new treatment is inferior. The type II error is the false negative (failing to reject the null hypothesis when it is false), which translates into failing to conclude non-inferiority when the new treatment truly is non-inferior. The sample size calculations that follow relate to the evaluation of non-inferiority when using either the CI method or the alternative *p*-value approach detailed in Section 12.4; recall that these approaches are mathematically the same.

The sample size calculation requires pre-specification of the following quantities:

- Type I error; for non-inferiority, this is one-sided and usually set at 2.5%
- Power = 1 – type II error, which is usually at least 80%
- $\Delta$ , the non-inferiority margin

The remaining quantities depend on the primary endpoint and the design. Assume that we are dealing with the parallel group case.

For a continuous endpoint, we would need

- The standard deviation of the endpoint
- The anticipated true difference in the two mean values

For a binary endpoint, we would need

- The response rate in the active control group
- The anticipated true difference in the response rates

Often, we undertake the sample size calculations assuming no difference between the treatments in terms of means (or rates), but this is not always a realistic assumption. It is good practice to at least look at the sensitivity of the calculation to departures from this assumption.

**Example 12.3** Evaluating non-inferiority for cure rates

In an anti-infective non-inferiority study, it is expected that the true cure rates for both the test treatment and the active control will be 75%.  $\Delta$  has been chosen to be equal to 15%. Using the usual approach with a one-sided 97.5% CI for the difference in cure rates, a total of 176 patients per group will give 90% power to demonstrate non-inferiority. Table 12.1 gives values for the sample size per group for 90% power and for various departures from the assumption regarding the cure rates.

When the cure rates are equal, the sample size decreases as the common cure rate increases. When the test treatment cure rate is above the active control cure rate, the test treatment is better than the active control, and it is much easier to demonstrate non-inferiority. When the reverse happens, however, where the true cure rate under the test treatment falls below that of the active control, the sample size requirement goes up; it is much more difficult under these circumstances to demonstrate non-inferiority. It is also worth noting that when the test treatment is truly 15% (the value for  $\Delta$  in this example) below the rate in the active control, planning a trial to demonstrate non-inferiority is simply not possible. Under this condition, as the sample size increases, the CI will converge around 15%, with part of that CI above the 15% value.

**Table 12.1** Sample sizes per group

		Cure rate (test treatment)			
		65%	70%	75%	80%
Cure rate (active control)	65%	213	115	70	46
	70%	460	197	105	63
	75%	1745	418	176	91
	80%	$\infty$	1556	366	150

Note also that, as is the case in superiority trials, the sample size may need to be factored up if there are randomised patients who are being systematically excluded from the full analysis set. This might be the case, for example, in anti-infective trials, where we tend to deal with the population of clinically (or microbiologically) evaluable patients (see Section 7.2.1).

There is a perception that non-inferiority trials are inevitably larger than their superiority counterparts. Under some circumstances, this is true, but it is

by no means always the case. One crucial quantity in the sample size calculation for a non-inferiority trial is  $\Delta$ , which plays a role similar to the clinically relevant difference (crd) in a superiority sample size calculation. The sample size (this is also true for equivalence) is inversely proportional to the square of  $\Delta$ . If  $\Delta$  is small, then the sample size will be large, and the constraints placed upon us by regulators together with the clinical interpretation of *irrelevant differences* tend to make  $\Delta$  small in non-inferiority trials. In a superiority trial, the choice of the crd to detect is an internal, clinical, sometimes commercial decision that is under the control of the trialists, and we are at liberty to power a trial on a large value, larger perhaps than a value that would be viewed as being clinically important. The net effect of these considerations is that non-inferiority trials tend to be larger than trials designed to demonstrate superiority. However, and in contrast to this, if we truly feel that the test treatment is somewhat better than the active control, then assuming such a positive advantage can have the effect of considerably reducing the sample size, as seen in Example 12.3.

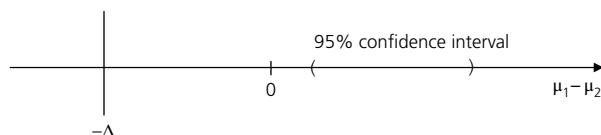
## 12.10 Switching between non-inferiority and superiority

In a clinical trial with the objective of demonstrating non-inferiority, suppose that the data are somewhat stronger than this and the 95% CI is not only entirely to the right of  $-\Delta$  but also completely to the right of zero, as in Figure 12.5; there is evidence that the new treatment is in fact superior.

What conclusions can we draw in this situation? Can we conclude superiority even though this was not the original objective? Well, generally the answer is yes: superiority can be claimed.

**CPMP (2000): 'Points to Consider on Switching Between Superiority and Non-Inferiority'**

*'If the 95 per cent confidence interval for the treatment effect not only lies entirely above  $-\Delta$  but also above zero, then there is evidence of superiority in terms of statistical significance at the 5 per cent level ( $p < 0.05$ ). In this case, it is acceptable to calculate the exact probability associated with a test of superiority and to evaluate*



**Figure 12.5** Concluding superiority in a non-inferiority trial

*whether this is sufficiently small to reject convincingly the hypothesis of no difference. . . Usually this demonstration of a benefit is sufficient for licensing on its own, provided the safety profiles of the new agent and the comparator are similar'.*

No multiplicity arguments impact this switch. Essentially, we can think of the two tests – a test of non-inferiority followed by a test of superiority – as a hierarchy, and we will only consider the second provided the first one gives a statistically significant result. As pointed out by the FDA, there are no multiplicity issues in these considerations.

#### **FDA (2016): *Non-Inferiority Clinical Trials to Establish Effectiveness'***

*'In some cases, a study planned as an NI study may show superiority to the active control. Recommendations in International Conference on Harmonisation guidance E9: Statistical Principles for Clinical Trials (ICH E9) and FDA policy have been that this superiority finding arising in an NI study can be interpreted without adjustment for multiplicity.'*

However, perhaps the safest thing to do in this setting is to pre-specify a hierarchy in the protocol with a test for non-inferiority followed by a test for superiority either immediately or lower down in the hierarchy if other endpoints need to be built into the confirmatory structure; there can then be no dispute about such a switch in the eyes of regulators. Following calculation of the exact  $p$ -value for superiority, the 95% CI allows the clinical relevance of the finding to be evaluated. Presumably, however, any level of benefit would be of value given that at the outset we were looking only to demonstrate non-inferiority; in this regard, the new treatment is likely to have other benefits outside of efficacy.

For superiority, the treatment policy estimand is often the basis for the primary analysis, so the emphasis in the superiority evaluation would need to be based around this.

We will return shortly to Example 12.1 and see a situation where switching from non-inferiority to superiority was possible.

Moving in the opposite direction and concluding non-inferiority in a superiority trial is much more difficult, as this would generally require pre-specification of a non-inferiority margin. Such pre-specification would usually not have been considered in a trial designed to demonstrate superiority. However, if a conclusion of non-inferiority would be a useful outcome, then it could be appropriate to consider such pre-specification.

#### **FDA (2016): '*Non-Inferiority Clinical Trials to Establish Effectiveness'***

*'A study designed primarily to show superiority, however, would yield credible evidence of non-inferiority only if the study had the key features of a NI study. . . An unplanned determination of non-inferiority following failure to show superiority, when the margin was not determined until results of the trial were known, would not be sufficient for demonstrating non-inferiority of the test drug.'*

**Example 12.1** (Revisited) Fluconazole compared to amphotericin B in preventing relapse in cryptococcal meningitis

This example was presented earlier in this chapter. The initial objective was to demonstrate the non-inferiority of fluconazole compared to amphotericin B in the prevention of cryptococcal meningitis in patients with AIDS. The non-inferiority margin was set at -15% for the difference in the relapse rates.

The 95% CI for  $\theta_A - \theta_F$  was in fact (7%, 31%), and this is entirely to the right not only of -15% but also of zero. In this case, the data support a claim for the superiority of fluconazole. The authors concluded that non-inferiority had been established, but additionally, there was evidence that fluconazole was more effective than amphotericin B:

*'These data allow us to conclude that fluconazole was at least as effective as weekly amphotericin B... Indeed, the 19% difference in the probability of relapse at one year... suggests that fluconazole was more effective than amphotericin B in preventing a relapse of cryptococcal disease in this population of patients'.*

Given the various possibilities regarding switching, there may well be a strong argument in many active control comparisons to go for non-inferiority. If the data then turn out to be stronger and support superiority, this additional conclusion can be made. However, there are drawbacks to this way of thinking:

- Non-inferiority trials are more difficult to design; assay sensitivity and the choice of non-inferiority margin are just two of the issues that also need to be considered.
- Non-inferiority trials sometimes require large sample sizes.
- Designing the trial as a non-inferiority evaluation may give a negative perception external to the company and the clinical trial team that can be disadvantageous.

Nonetheless, this strategy may be worth considering under some circumstances.

## 12.11 Biosimilars

There are now extensive guidelines in Europe and the US concerning biosimilars. However, the statistical requirements for the demonstration of biosimilarity across those two jurisdictions are somewhat different, which is brought out in the following case study.

**Example 12.4** Biosimilarity in treatment of age-related macular degeneration

Woo et al. (2021) investigated the safety and efficacy of a proposed biosimilar to ranibizumab for the treatment of neovascular age-related macular degeneration (AMD). The FDA, together with other regulatory authorities, favoured a clinical outcome based on visual acuity

for the evaluation of biosimilarity, and the appropriate primary endpoint was chosen to be the change from baseline to week 4 in best-corrected visual acuity (BCVA). The EMA and several other authorities preferred a pharmacodynamic anatomical outcome, and the primary endpoint was chosen to be the change from baseline in central subfield thickness (CST) at week 4.

The fixed margin approach was used to determine equivalence margins for each of the chosen primary endpoints based on meta-analyses of historical data. For BCVA at week 8, the biosimilarity region was defined as  $-3$  to  $+3$  letters. For CST, the region was  $-36\text{ }\mu\text{m}$  to  $36\text{ }\mu\text{m}$ . Analyses of each of the primary endpoints were based on an analysis of covariance (ANCOVA) model with baseline BCVA (or CST) as a covariate and region and treatment as factors in the model. But a further difference between the regulatory authorities now emerges. The FDA in general accept the use of 90% CIs for demonstrating biosimilarity as in bioequivalence, but the EMA argue in favour of 95% CIs.

Table 12.2 presents the results obtained based on the FAS and the PPS. For both analysis sets, the CIs for BCVA are completely contained within the biosimilarity region,  $-3$  letters to  $3$  letters, and similarly for CST with CIs for both analysis sets within the  $-36\text{ }\mu\text{m}$  to  $36\text{ }\mu\text{m}$  interval.

**Table 12.2** Primary efficacy endpoints for ranibizumab biosimilar trial

Primary endpoint	Analysis set	Number of patients	Least squares (LS) mean change from baseline	Confidence coefficient	Confidence interval
BCVA (number of letters), week 8	FAS	704	-0.8	90%	(-1.8, 0.2)
	PPS	669	-0.8	90%	(-1.8, 0.3)
CST ( $\mu\text{m}$ ), week 4	FAS	698	-8.0	95%	(-19, 3)
	PPS	680	-8.5	95%	(-19, 3)

## CHAPTER 13

# The analysis of survival data

### 13.1 Time-to-event data and censoring

In many cases, an endpoint measures time from the point of randomisation to some well-defined event, such as time to death (survival time) in oncology or time to rash healing in herpes zoster. The data from such an endpoint invariably have a special feature known as *censoring*. For example, suppose the times to death for a group of patients in a 24-month oncology study are as follows:

14 7 24\* 15 3 18 9\* 10 24\* 9...

Here, the first patient died after 14 months from time of randomisation, and the second patient after 7 months. The third patient, however, is still alive at the end of the study, while patients 4, 5 and 6 died after 15, 3 and 18 months, respectively. Patient 7 was lost to follow-up at 9 months, and patient 8 died after 10 months. Patient 9 is also still alive at the end of the trial, while patient 10 died after 9 months. As can be seen, the primary endpoint, survival time, is not available for all patients. It is not that we have no information on patients 3, 7 and 9, but we do not have complete information – we know only that their survival times are at least 24, 9 and 24 months, respectively. These patients provide what we call *censored observations*. Unfortunately, we cannot even do some simple things. For example, it is not possible to calculate the mean survival time. You might say, well, can't we just ignore the fact that these observations are censored and calculate the average of the numerical values? Well, you could, but this would clearly underestimate the true mean survival time since eventually, the actual survival times for patients 3, 7 and 9 would be greater than the numerical values in the list. Can't we just ignore the censored values and calculate the mean of those that remain? Again, this calculation would give an underestimate of the true mean; the censored values tend to come from the patients who survive a long time, and ignoring them would systematically remove the patients that do well and give a biased value for the mean.

This specific feature has led to the development of special methods to deal with data of this kind. If censoring were not present, we would probably just take logs of the patient survival times and undertake the unpaired t-test or one of its

extensions, ANOVA or ANCOVA, to compare treatments. Note that survival time, by definition, is always positive, and frequently the distribution is positively skewed; taking logs is often successful in recovering normality in such cases.

The special methods we are going to discuss were first developed primarily in the 1970s and applied in the context of analysing time to death, and, in general, we refer to the topic as *survival analysis*. However, as time has gone by, we have applied these same techniques to a wide range of time-to-event endpoints. The following list gives some examples:

- Time to rash healing in herpes zoster
- Time to complete cessation of pain in herpes zoster
- Disease-free survival in oncology
- Time to first seizure in epilepsy
- Time to alleviation of symptoms in flu
- Time to first serious adverse event

Throughout this section, we will from time to time adopt the conventions of the field and refer to survival analysis and survival curves, accepting that the methods are applied more widely to events other than death.

Censoring in clinical trials usually occurs because the patient is still alive at the end of the period of follow-up. In the earlier example, if this were the only cause of censoring, all censored observations would be equal to 24 months. But there are other ways in which censoring can occur, such as loss to follow-up or withdrawal. These can sometimes raise difficulties, and we will return to discuss the issues in a later section. Also, in an interim analysis, the period of follow-up for the patients still alive in the trial is variable, and this will produce a whole range of censored event times. Finally in event-driven studies (see Section 13.10) analysis will take place once a certain number of patients with events have been observed and in these cases patients who are event-free will give different censored values depending on the length of time they have been in the trial. Our methodology needs to be able to cope with having censored survival times that are distributed over the period of follow-up.

In the next section, we will discuss Kaplan-Meier (KM) curves, which are used to display the data and to calculate summary statistics. We will then cover the logrank and Gehan-Wilcoxon tests, which are simple two-group comparisons for censored survival data (akin to the unpaired t-test for a continuous endpoint), and extend these ideas to incorporate baseline covariates and factors.

## 13.2 Kaplan-Meier curves

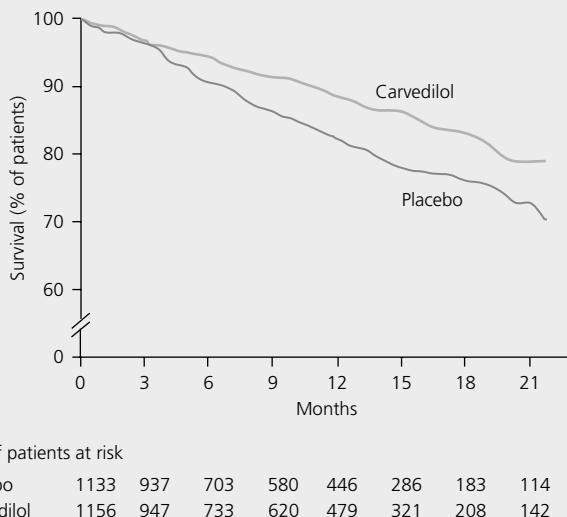
### 13.2.1 Plotting Kaplan-Meier curves

Kaplan and Meier (1958) introduced a methodology for estimating the probability of being event-free as a function of time from censored survival data. If the event is death, then we are estimating the probability of surviving, and the

resultant plots of the estimated probability of surviving as a function of time are called *Kaplan-Meier curves* or *survival curves*.

**Example 13.1** Carvedilol in severe heart failure

A placebo-controlled randomised trial reported by Packer et al. (2001) investigated the effect of carvedilol on survival time in severe heart failure. Figure 13.1 shows the survival curve for each of the two treatment groups following the early termination of the trial at a planned interim analysis.



**Figure 13.1** Kaplan-Meier survival curves in placebo and carvedilol groups. Source: Packer M, Coats AJS, Fowler MB, et al. for the Carvedilol Prospective Randomised Cumulative Survival Study Group (2001). Effect of carvedilol on survival in severe chronic heart failure. *NEJM*, **344**, 1651–1658. Reproduced by permission of Massachusetts Medical Society.

The Kaplan-Meier method looks at the patterns of deaths over time to estimate the probability of surviving. The censored values contribute to this estimation process: if a patient is censored at 12 months, then that patient is involved in estimating the probability of surviving over intervals of time up to and including 12 months, but not beyond those times. More detail is given in Section 13.4.5 regarding this calculation. The form of the estimated survival curves is a so-called *step function*, with steps down occurring at time points where there are deaths. As patients die or are censored, the number of patients remaining alive and continuing in the trial in each treatment group diminishes. Therefore, the probabilities of surviving are estimated from fewer and fewer patients as time goes on, and the estimated curves become less precise. That is why you generally

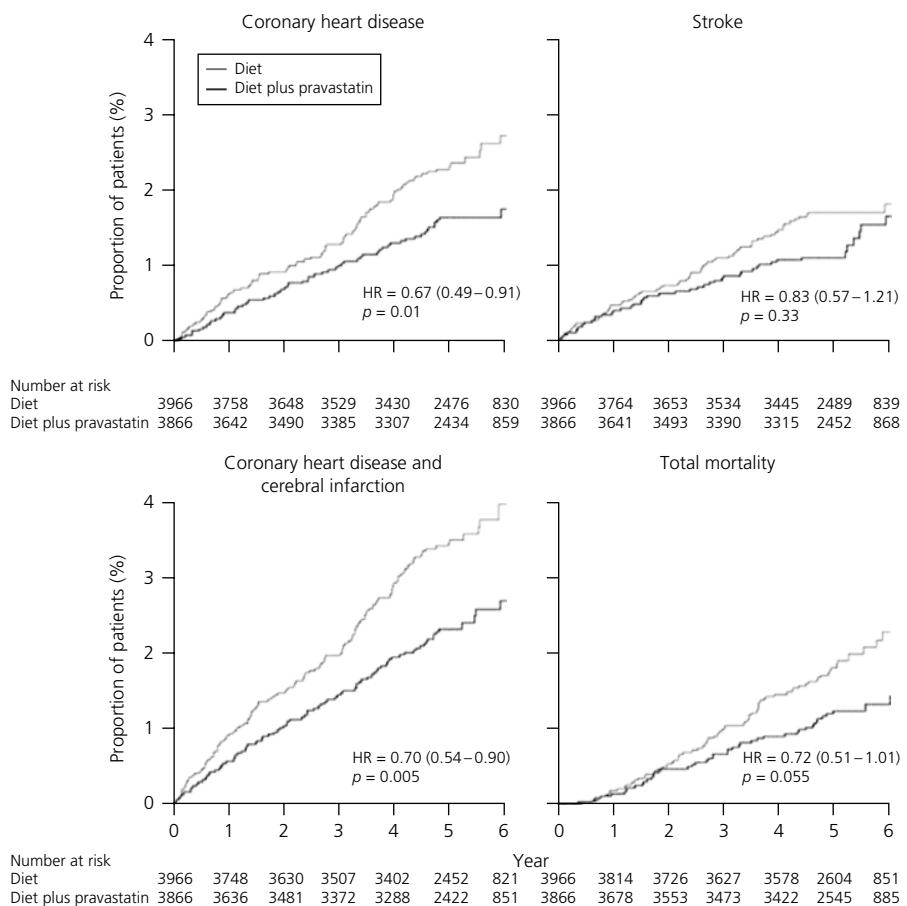
see greater instability in the curves at the longer follow-up times. To give information in relation to this, it is common practice to record the *number of patients at risk* through time; these are the numbers of patients alive and continuing in the trial at certain time points. In the Packer et al. (2001) study, 1133 patients were randomised to placebo and 1156 patients were randomised to carvedilol; at 12 months following randomisation, there were 446 patients at risk in the placebo group compared to 479 in the carvedilol group. This means 687 ( $= 1133 - 446$ ) patients in the placebo group either died before month 12 or gave censored observations that were less than 12 months. Similarly, 677 ( $= 1156 - 479$ ) patients in the carvedilol group either died or were censored prior to month 12. By 21 months, the risk sets (sets of patients at risk in each treatment group) comprised 114 patients in the placebo group and 142 in the carvedilol group.

It is usual to estimate and plot the probability of being event-free, but there are occasions, such as when the event rates are low, where interpretation is clearer when the opposite of this is plotted: cumulative incidence (or cumulative probability of experiencing the event by that time). This is simply obtained as  $1 - \text{probability of being event-free}$ . Pocock, Clayton and Altman (2002) discuss issues associated with the interpretation of these plots. Figure 13.2 provides some examples of these kinds of plots in a trial looking at several cardiovascular events in the primary prevention of cardiovascular disease with pravastatin (Nakamura et al., 2006). When the event rates are low and the Kaplan-Meier curves do not fall too far below 1, it is common in the conventional type of plot to put a break in the *y*-axis, as seen in Figure 13.1. As Pocock et al. point out, this can sometimes visually exaggerate the treatment difference, so take care with interpretation in this case!

### 13.2.2 Event rates and relative risk

It is straightforward to obtain the estimated probability of surviving for key time points from the Kaplan-Meier curves. In the Packer et al. (2001) example, the estimated survival probability at 12 months in the carvedilol group was 0.886 compared to 0.815 in the placebo group, an absolute difference of 7.1% in the survival rates. A formula provided by Greenwood (1926) enables us to obtain standard errors and consequently confidence intervals (CIs) for these individual survival rates and their differences.

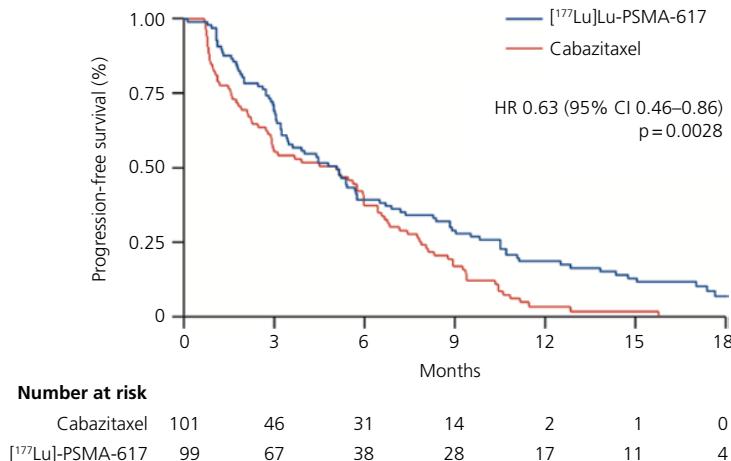
The estimated probability of dying in the first 12 months is then 0.114 ( $= 1 - 0.886$ ) in the carvedilol group compared to 0.185 ( $= 1 - 0.815$ ) in the placebo group. The relative risk at 12 months is the ratio of the risks or probabilities of dying in the first 12 months =  $0.114/0.185 = 0.62$ . The relative risk reduction is 38%, while the absolute risk reduction is  $0.061 = 0.185 - 0.114$  and number needed to treat (NNT) =  $1/0.061 = 17$  when rounded up to the nearest integer. A total of 17 patients need to be treated with carvedilol to prevent one death in the first 12 months. Similar calculations can be undertaken at other time points.



**Figure 13.2** Inverted Kaplan-Meier curves for the primary and secondary endpoints. Source: Nakamura H, Arakawa K, Itakura H, et al., for the MEGA Study Group (2006) 'Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA study): a prospective randomised controlled trial'. *The Lancet*, **368**, 1155–1163. Reproduced by permission of Elsevier.

### 13.2.3 Median event times

We mentioned earlier in this chapter that it is not possible to calculate the overall mean survival time. It is, however, often possible to obtain median survival times from the Kaplan-Meier curves. The median survival time for a particular group corresponds to the time on the *x*-axis when the survival probability on the curve takes the value 0.5. For this statistic to be obtained, the survival curves must fall below the 0.5 value on the *y*-axis. In Example 13.1, this has not happened. In such cases, we use the survival rates at various time points as summary descriptions of the survival experience in the groups. We will see an example later where the curves do fall below the 0.5 point. When the median times are available, it is also possible to obtain associated standard errors and CIs.



**Figure 13.3** Radiographic or PSA progression-free survival HR=hazard ratio. PSA=prostate-specific antigen. PSMA=prostate-specific membrane antigen.  $^{177}\text{Lu}$ =lutetium-177. Source: Hormann MS, Emmett L et al (2021) ' $[^{177}\text{Lu}]\text{Lu-PSMA-617}$  versus cabazitaxel in patients with metastatic castration-resistant prostate cancer (TheraP): a randomised, open-label, phase 2 trial' *The Lancet*. Reproduced by permission of Elsevier

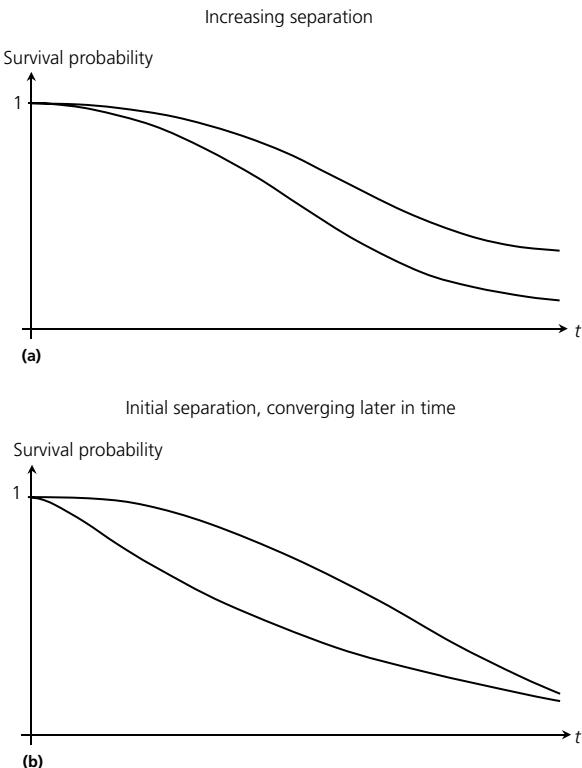
It should be noted that although the medians may be defined based on the Kaplan-Meier curves, they do not always provide good summary measures that capture treatment differences. Figure 13.3 is a good example where this is the case. This is a randomised trial in prostate cancer reported by Hormann and Emmett (2021), and the figure displays Kaplan-Meier curves for progression-free survival. The curves are clearly separated with an overall benefit for the experimental treatment  $[^{177}\text{Lu}]\text{Lu-PSMA-617}$ , but the quoted medians are identical for each of the treatment groups: both equal 5.1 months.

One further cautionary note about medians. If they have been obtained in one or both treatment groups at points in time when the risk sets are small, then, as with the curves themselves at those time points, they are unstable, and it would be a mistake to rely on their difference to provide a precise measure of treatment benefit.

### 13.3 Treatment comparisons

The Kaplan-Meier curves do not of themselves provide a formal  $p$ -value comparison of the treatments. This comparison of the survival curves is usually undertaken using either the logrank test or the Gehan–Wilcoxon test. We will look at these two test procedures in turn.

The *logrank test* was developed by Peto and Peto (1972). The  $p$ -value resulting from this test is not a comparison of the survival curves at a particular time



**Figure 13.4** Pattern of survival curves (a) Increasing separation, (b) Initial separation, converging later in time

point, nor is it a comparison of the medians (although such tests could be constructed), but a test comparing the two complete curves. A statistically significant  $p$ -value indicates that the survival experience in the two groups is different. The *Gehan–Wilcoxon test* (Gehan, 1969) (sometimes referred to as the *generalised Wilcoxon test*) is an alternative procedure for producing a  $p$ -value comparison of two survival curves. Why do we need two different tests? The reasons relate to the basic shape of the curves we are comparing. Both tests provide valid  $p$ -values with control of type I error, but these two tests are designed in different ways to pick up different patterns of treatment difference.

Figure 13.4 provides two sets of hypothetical population survival curves. In the first display, the curves separate gradually over the period of follow-up. In contrast, the curves in the second display separate early on in time but then start to converge later.

The first set of curves represent, over the period of follow-up, a *permanent* treatment effect. There is a long-term advantage in survival for one group compared to the other group. But the interpretation of the second set of curves is different. Here we see a short-term benefit of one treatment group over the other, but as time goes on, this relative benefit diminishes, leaving little

long-term effect. This pattern of differences represents a delay in the occurrence of the event in one group compared to the other group.

The two tests mentioned earlier are designed to look for these different patterns. The logrank test is best able to pick up the longer-term differences, while the Gehan–Wilcoxon test focuses on short-term effects, or a delay. The appropriate test to use depends upon what kind of differences you are expecting to see.

In the Packer et al. (2001) trial in Figure 13.1, we see longer-term differences displayed over the 21-month follow-up period, and the logrank test is entirely appropriate here. The  $p$ -value from the logrank test was 0.00013, a highly statistically significant result.

Generally, we see these kinds of patterns in long-term cardiovascular and oncology trials. In other applications, however, such long-term effects are not anticipated and do not fit with the study's objectives. For example, in flu trials, where the primary endpoint is the time to alleviation of symptoms, we are looking for more rapid resolution of symptoms in the active group compared to placebo. Treatment effects are seen in the first two days or so, with many more patients symptom-free in the active group compared to placebo. By the time we reach seven days, however, most of the patients in each of the two groups have only minor symptoms remaining, so the probability of being event-free (still having symptoms) is the same in the two groups. In this case, the Gehan–Wilcoxon test is best suited to pick up differences. Similar comments often apply to progression-free survival in advanced cancer trials, where the best we can hope for is that the test treatment delays the event (death or progression). Unfortunately, the logrank test dominates this area of statistics, and many applications have failed to detect important short-term differences between survival curves when using a test that is insensitive to detecting them. One of many examples is Okwera et al. (1994), who compared two treatments (thiacetazone and rifampicin) for pulmonary tuberculosis in HIV-infected patients. The survival curves at 300 days following randomisation were separated by more than 10% (77% surviving in one group compared to around 66% in the second group). In other words, 10% more patients, at least, were alive at 300 days in the rifampicin arm compared to the thiacetazone arm, an important benefit of one treatment over the other. At 600 days, however, the survival curves had come together, showing no long-term effect. The quoted  $p$ -value from the logrank test was  $>0.50$ . Now, one could argue whether a short-term benefit is clinically important, but the issue here is that the test used was not sensitive in terms of picking up differences between the curves. The Gehan–Wilcoxon test would have stood a much better chance of yielding statistical significance and perhaps generated some interest in discussing the implication of those short-term effects.

A question often arises: 'Which test should I use when I don't know what kind of effect I am going to see?' My short answer to this question is that in a confirmatory setting, you should know! By the time you reach that stage in the

drug development programme, your knowledge of the disease area and the treatment, in combination with the endpoint, should enable accurate prediction of what is likely to happen. Of course, earlier in the programme, you may not know; and in this exploratory non-confirmatory phase, it is perfectly valid to undertake both tests to explore the nature of the effects.

There are indeed several other options for the statistical comparison of survival curves, including *weighted logrank tests*, and the interested reader is referred to Royston and Parmar (2020) for further details. And in Section 13.5, we will discuss a further approach to the comparison of survival curves.

## 13.4 The hazard ratio

### 13.4.1 The hazard rate

To be able to understand what a hazard ratio is, you first need to know what a hazard rate is. The *hazard rate (hazard function)* is formally defined as the conditional death (or event) rate calculated through time. What we mean by this is as follows. Suppose that in a group of 1000 patients, 7 die in month 1; the hazard rate for month 1 is 7/1000. Now suppose that 12 die in month 2; the hazard rate for month 2 is 12/993. If 15 die in month 3, the hazard rate for month 3 is 15/981, and so on. So, the hazard rate is the death (event) rate over each time interval among those patients still alive at the start of the interval.

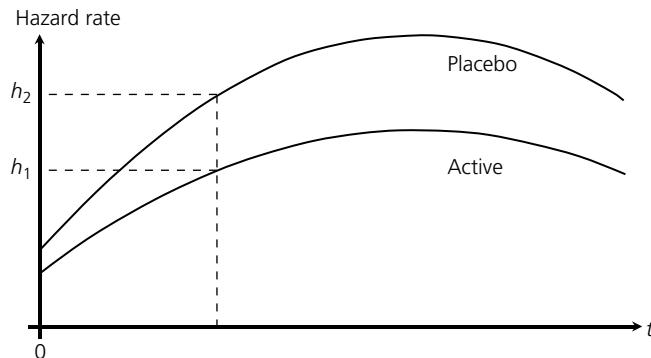
There are several things to note about the hazard rate. Firstly, the hazard rate is unlikely to be constant over time. Secondly, even though we have introduced the concept of the hazard rate as taking values over monthly intervals of time, we can think in terms of small time intervals, weeks or even days, with the hazard rate essentially being a continuous function through time.

The hazard rate can be estimated from data by looking at the patterns of deaths (events) over time. This estimation process takes account of the censored values in ways similar to how such observations were used in the Kaplan-Meier curves.

Figure 13.5 shows a schematic plot of two hazard rates corresponding to two treatment groups in a randomised trial. As can be seen from this plot, the hazard rates in each of the two treatment groups start just after randomisation ( $t = 0$ ) at a modest level, increase to a peak after a certain period of time (say, one year) and then begin to decrease. This tells us that at one year, the death rate in each group is at its maximum; before that, the death rates steadily increase, and following one year, the death rates tail off. It is also clear from the plot that the hazard rate in the placebo group is higher than the hazard rate in the active group at every time point.

### 13.4.2 Constant hazard ratio

Even though the individual hazard rates seen in Figure 13.5 are not constant, it would be reasonable to assume, wherever we look in time, that the ratio of the hazard rates is approximately constant. When this is the case, the ratio of the



**Figure 13.5** Hazard rates for two groups of patients

hazard rates is a single value, which we call the *hazard ratio*. Indeed, this plot was specifically drawn to have this property. We denote this ratio by  $\lambda$  so that  $\lambda = h_1/h_2$ , where  $h_1$  and  $h_2$  are as displayed in the plot.

It is the convention for the hazard rate in the test treatment group to appear as the numerator and the hazard rate in the control group to be the denominator in the definition of the hazard ratio.

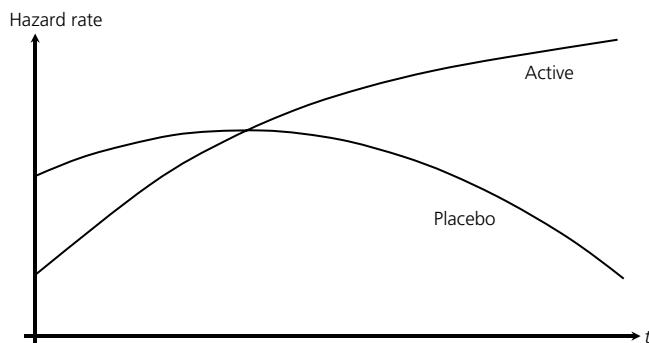
A hazard ratio of one corresponds to exactly equal treatments; the hazard rate in the active group is exactly equal to the hazard rate in the placebo group. If we adopt this convention, and the event is death (or any other undesirable outcome), a hazard ratio of less than one tells us that the test treatment is better. This is the situation we see in Figure 13.5. A hazard ratio greater than one tells us that the test treatment is the poorer treatment.

Even if the hazard ratio is not precisely a constant value as we move through time, it can still provide a valid summary, provided the hazard rate for one of the treatment groups is consistently above the hazard rate for the other group. In this case, the value we get for the hazard ratio from data represents an average of that ratio over time.

CIs for the hazard ratio are straightforward to calculate. Like the odds ratio and relative risk/risk ratio (see Section 4.5.5), this CI is firstly calculated on the log scale and then converted back to the hazard ratio scale by taking antilogs of the ends of that CI.

### 13.4.3 Non-constant hazard ratio

It is not always the case, by any means, that we see a constant or approximately constant hazard ratio. There are situations, as shown in Figure 13.6, where the hazard rate for one group starts lower than the hazard rate for a second group, they then move closer together as we move through time, but at a certain point a switch occurs. The hazard rate for the first group overtakes that of the second group, and they continue to move further apart from that point on. When this



**Figure 13.6** Hazard rates for two groups of patients where the hazard ratio is not constant

happens, the hazard ratio will start below one, increase towards one as we move through time, and then flip over and increase above and away from one. The hazard ratio still exists but is not constant and varies over time. See Kay (2004) for further discussion on these points.

In this case, it clearly makes no sense to assign a single value to the hazard ratio. It is also worth noting that there are tests to assess whether a constant hazard ratio exists. A statistically significant  $p$ -value from this test indicates departures from this so-called *proportional hazards* assumption.

#### 13.4.4 Link to survival curves

In an earlier section, we saw two different patterns for two sets of survival curves. In the upper part of Figure 13.4, the survival curves move further and further apart with increasing time. This pattern is consistent with one of the hazard rates (think in terms of death rates) being consistently above the other hazard rate. This in turn corresponds to a constant hazard ratio, at least approximately the situation we discussed in Section 13.4.2. So, a constant hazard ratio manifests itself as a continuing separation in the two survival curves as in the figure. Note that the higher hazard rate (more deaths) aligns with the lower of the two survival curves.

In the lower part of Figure 13.4, the survival curves move apart very early in time and later begin to converge. This pattern is consistent with one of the hazard rates being above the other initially, which corresponds to the survival curve that decreases most rapidly due to the high death rate early on. For the survival curves to begin to converge, however, a reversal of the death rates in the two groups needs to take place so that the death rate in the group that did well initially begins to increase and overtakes the death rate in the other group. This is the only way the catch-up can take place to ensure that the probability of surviving beyond the end of the observation period is approximately equal in the two groups. This pattern is consistent with a hazard ratio that is not constant. Here, the hazard ratio starts well below one, increases to one and then moves above one through time as the death rates in the two groups reverse. For the pattern

of survival curves in this figure, the hazard ratio does not take a single value because it is not constant. A better way to summarise the relative performance in the two groups in this case would be to use the relative risk at time points of interest. We mentioned in Section 13.3 that the logrank test is specifically designed to provide a *p*-value for a pattern of survival curves, as in the upper part of Figure 13.4. As we have now seen, this pattern corresponds to a constant hazard ratio; indeed, the logrank test can be considered as a test of the hypothesis  $H_0: \lambda = 1$  against the alternative  $H_1: \lambda \neq 1$ , where  $\lambda$  is the hazard ratio.

### 13.4.5 Calculating Kaplan-Meier curves

It is worth considering the calculation of the Kaplan-Meier curve following on from the discussion of the hazard rate, to see, firstly, how censoring is accounted for and, secondly, how the two are linked in terms of the calculation.

As in Section 13.4.1, consider a group of 1000 patients at the point of randomisation. We will now add into the calculation some censored values. Suppose that in month 1, 7 die: the hazard rate for month 1, as before, is 7/1000. Now assume that in addition, 10 are censored at the end of month 1, and among the 983 (= 1000 – 7 [deaths] – 10 [censorings]) remaining in the study at the beginning of month 2, 15 die in that month. The hazard rate for month 2 is 15/983. Next, assume that 8 patients are censored at the end of month 2, so that 960 (= 983 – 15 [deaths] – 8 [censorings]) patients are in the study at the beginning of month 3, and of these, 12 die in that month; the hazard rate for month 3 is 12/960, and so on. The calculation of the hazard rate through time with the introduction of the censored values now considers both the number of patients dying and the number of patients censored. Patients with censored observations contribute to the denominator of the hazard rate calculation in the months before the time at which they are censored, but not after that.

Now to calculate the Kaplan-Meier survival probabilities. Using this example, the estimated probability of surviving beyond month 1 is  $1 - (7/1000) = 0.9993$ . The probability of surviving beyond month 2 is the probability of surviving beyond month 1  $\times$  the probability of surviving beyond month 2 among those alive at the start of month 2; from the data, this is estimated to be  $0.9993 \times (1 - (15/983)) = 0.984$ . Continuing this process, the probability of surviving beyond month 3 is the probability of surviving beyond month 2  $\times$  the probability of surviving beyond month 3 among those alive at the start of month 3; this is estimated to be  $0.984 \times (1 - (12/960)) = 0.972$ . In general, if  $h_i$  is the hazard rate for month  $i$ , then the estimated probability of surviving beyond month  $t$  is equal to

$$(1 - h_1) \times (1 - h_2) \times \dots \times (1 - h_t)$$

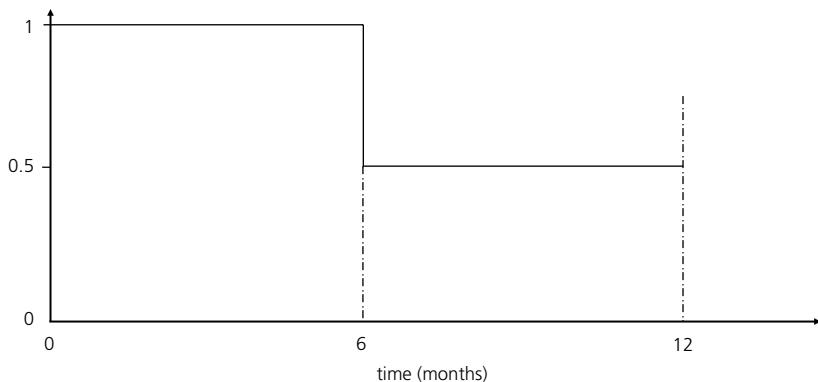
Note also that the denominators 1000, 983 and 960 for the hazard rates are the numbers of patients alive and in the trial in that group at the start of months 1, 2 and 3, respectively: that is, the number at risk.

This calculation in relation to both the hazard rate and the survival probabilities has been undertaken at intervals of one month. In practice, we use intervals that correspond to the unit of measurement for the endpoint itself, usually days, to make use of the total amount of information available.

### 13.5 Restricted mean survival time

At the beginning of this chapter, we mentioned that calculating the mean is not possible in general because of censoring and that survival times/time to event values are not available for all subjects. However, a related measure, the restricted mean survival time (RMST), can be calculated. The RMST is defined as the mean survival time up to a certain point in time. For example, suppose that data were available for a group of 10 subjects, with 5 of those subjects dying after 6 months and the remaining 5 subjects surviving to 12 months (the end of the observation period). The value for RMST would then be 9 months, the average survival time among the 10 subjects over the 12-month observation period. In practice, calculating the RMST is not quite so straightforward, but there is a mathematical process that allows us to do so. Figure 13.7 shows the Kaplan-Meier curve for this example, where five subjects die after 6 months while the remaining five subjects survive to the end of the 12-month observation period.

The RMST can be calculated as the area under the Kaplan-Meier curve. This can be seen for our simple example where the area under the curve (AUC) is  $(1 \times 6) + (0.5 \times 6) = 9$  months. This method of calculation applies in more complex settings. In practice, where, for example, we are comparing two treatment groups, we can calculate the RMST values for each of the groups and the difference between these two RMSTs, which can be viewed as a measure of treatment difference, is then simply the area between the two Kaplan-Meier curves. This



**Figure 13.7** Kaplan-Meier curve for data of section 13.5 and calculation of restricted mean survival time

has some intuitive appeal, at least visually, when we are looking to compare the curves. Standard errors can be calculated for the RMSTs and their difference, leading to a p-value for the treatment difference and an associated CI.

**Example 13.2** Weekly dose-dense chemotherapy in first-line epithelial ovarian, fallopian tube or primary peritoneal carcinoma treatment

Clamp and James (2019) report on a trial in epithelial ovarian, fallopian tube or primary peritoneal carcinoma. Their use of the restricted mean survival time as a basis for a treatment comparison resulted from the violation of the proportional hazards assumption, which prevented them from calculating a hazard ratio:

'Evidence of non-proportional hazards was observed (group 1 vs group 2 p=0·011; group 1 vs group 3 p=0·051; figure 2), which, therefore, means that hazard ratios are not robust over time, and so the restricted mean survival time (RMST) is the most appropriate primary estimate of treatment effect. RMST for progression was 24·5 months (97·5% CI 23·0–26·0) in group 1, 24·9 months (24·0–25·9) in group 2, and 25·4 months in group 3 (23·9–26·9)'.

A good description of the methodology associated with the RMST, including sample size calculations, is provided by Pak et al. (2017). Using the RMST offers several advantages:

- 1 It is an alternative to the median that does not depend on a single point on the Kaplan-Meier curve.
- 2 It can be calculated whether or not the median exists.
- 3 The difference between RMST values can be used as a basis for comparing Kaplan-Meier curves, and this does not involve assumptions regarding the relationship between those curves (for example, proportional hazards).

## 13.6 Adjusted analyses

In Chapter 6, we covered methods for adjusted analyses and analysis of covariance in relation to continuous (ANOVA and ANCOVA) and binary and ordinal endpoints (CMH test and logistic regression). Similar methods exist for survival data. As with these earlier methods, particularly in relation to binary and ordinal data, there are numerous advantages in accounting for such factors in the analysis. If the randomisation has been stratified, then such factors should be incorporated into the analysis to preserve the properties of the resultant *p*-values.

### 13.6.1 Stratified methods

Both the logrank and Gehan–Wilcoxon tests can be extended to incorporate stratification and other factors measured at baseline. These methods provide *p*-value comparisons of treatments, allowing imbalances in baseline factors to be accounted for. Although possible, extensions of these procedures to evaluate

the homogeneity of the treatment effect – that is, the investigation of treatment-by-covariate interactions – is not so straightforward. Consequently, we tend to build baseline factors and covariates into the modelling through methods covered in the following section.

### 13.6.2 Proportional hazards regression

The most popular method for covariate adjustment is the *proportional hazards model*. This model, originally developed by Cox (1972), is now used extensively in the analysis of survival data to incorporate, and adjust for, covariate effects.

The method provides a model for the hazard function. As in Section 6.6, let  $z$  be an indicator variable for treatment taking the value one for patients in the active group and zero for patients in the control group, and let  $x_1, x_2$ , etc. denote the covariates. If we let  $\lambda(t)$  denote the hazard rate as a function of  $t$  (time), the model takes the form

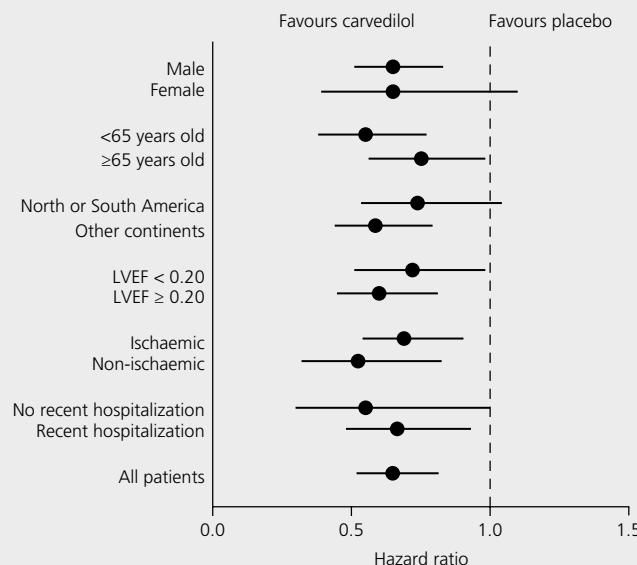
$$\ln(\lambda(t)) = a + cz + b_1x_1 + b_2x_2 + \dots$$

As before, the coefficient  $c$  measures the effect of treatment on the hazard rate. If  $c < 0$ , then the log hazard rate, and therefore the hazard rate itself, in the active group is lower than the hazard rate in the control group. If  $c > 0$ , the reverse is true, and the active treatment is giving a higher hazard rate; and if  $c = 0$ , there is no treatment difference. An analysis using this model largely revolves around testing the hypothesis  $H_0: c = 0$  and subsequently presenting an estimate for the treatment effect. The structure of the model is such that  $c$  is the log of the hazard ratio, and the antilog,  $e^c$ , is the (adjusted) hazard ratio. This, together with a CI for the hazard ratio, gives a measure of the treatment effect, adjusting for baseline factors (covariates).

Treatment-by-covariate interactions can be investigated by including cross-product terms in the model as with binary endpoints and logistic regression, although it is common to also evaluate these potential interactions visually through a Forest plot of treatment effects in subgroups (see Example 13.1 Revisited). The remarks made previously in Section 6.8 regarding regulatory aspects of the inclusion of covariates apply equally well to the survival data setting and the proportional hazards model.

**Example 13.1 (Revisited)** Carvedilol in severe heart failure

The proportional hazards model was fitted to the survival data overall and also, within subgroups, defined according to key baseline prognostic factors to calculate hazard ratios and evaluate the homogeneity of the treatment effect. Figure 13.8 shows these hazard ratios together with 95% CIs. Such plots were discussed earlier in Section 10.8 in relation to subgroup testing. The data in Figure 13.8 indicate a consistency of treatment effect across the various subgroups investigated.



**Figure 13.8** Hazard ratios (and 95% CIs) for death in subgroups defined by baseline characteristics. Source: Packer M, Coats AJS, Fowler MB, et al. for the Carvedilol Prospective Randomised Cumulative Survival Study Group (2001). Effect of carvedilol on survival in severe chronic heart failure. *NEJM*, **344**, 1651–1658. Reproduced by permission of Massachusetts Medical Society.

As the name suggests, the proportional hazards model assumes that the hazard ratio is a constant. As such, it provides a direct extension of the logrank test, which is a simple two-treatment group comparison. Indeed, if the proportional hazards model is fitted to data without the inclusion of baseline factors and just the treatment indicator in the model, the  $p$ -value for the test  $H_0: c = 0$  is very similar to the  $p$ -value arising out of the logrank test. Any differences are due to the subtleties of the slightly different formulas for constructing the test statistic.

### 13.6.3 Accelerated failure time model

The *accelerated failure time model* is an analysis of variance technique that models the survival time itself but on the log scale:

$$\ln T = a + cz + b_1x_1 + b_2x_2 + \dots$$

We mentioned earlier, in Section 13.1, that if we did not have censoring, an analysis would likely proceed by taking the log of survival time and undertaking the unpaired t-test. This model simply develops that idea by now incorporating

covariates through a standard ANCOVA. If we assume that  $\ln T$  is also normally distributed, then the coefficient  $c$  represents the (adjusted) difference in the mean (or median) survival times on the log scale. Note that for the normal distribution, the mean and the median are the same; it is more convenient however to think in terms of medians. To return to the original scale for survival time, we then antilog  $c$  to give  $e^c$ , and this quantity is the ratio (control divided by active) of the median survival times. CIs can be obtained in a straightforward way for this ratio.

It can be shown that this model is a special case of the proportional hazards model and consequently is not considered as an alternative for analysis. It does however provide a modelling framework for time to death that helps us to deal with cross-over in oncology trials in particular where patients who progress in the control group are sometimes switched to the test treatment following progression. This issue will be discussed in Section 13.8.

## 13.7 Independent censoring

One important assumption we have made in the analyses presented so far is that the censoring process is independent of the process associated with the event we are looking at. This is the assumption of *independent* (or *non-informative*) *censoring*. If, for example, patients were being withdrawn from the trial and therefore given censored survival times at the time of withdrawal because their health was deteriorating, this assumption would be violated. Doing this would completely fool us about the true hazard (death) rate. In the extreme, if this happened with every patient in a time-to-death analysis, we would never see any deaths!

Usually, censoring occurs at the end of the observation/follow-up period for each patient, and in these cases, the issue of independent censoring is not a problem; the censoring process is clearly unconnected with the underlying process for the events. Similarly, there are no problems in an interim analysis where censoring occurs only because, for many patients, follow-up is not yet complete and they are alive at that time. In both these settings, we talk about *administrative censoring*. When patients are lost to follow-up or withdrawn, however, there is the potential for bias. Unfortunately, there is no easy way to deal with this. The recommended approach is to analyse the data as they are collected, but then undertake one or more sensitivity analyses. For example, for patients who withdraw for negative reasons (deteriorating health, suffering a stroke), assume that the censoring is in fact a death; for patients who withdraw for positive reasons (disease-free for six months), assign them the maximum length of follow-up and censor them at that maximum time. This is the worst-case scenario for the withdrawals for negative reasons and the best-case scenario for the withdrawals for positive reasons. If the conclusions are essentially unchanged when these assumptions are made, you can be assured that the overall conclusion is robust

to any violation of the assumption of independent censoring. An alternative – and certainly this is possible with a hard endpoint such as death – is to continue follow-up even though the patient has withdrawn from treatment, to obtain the required information on the time to death. The various options here link with our earlier discussion in Chapter 8 on estimands where withdrawal from the treatment or trial would be considered an intercurrent event. Further, the assumption of independent censoring is akin to the assumption of missing at random (see Section 7.3.9).

**Example 13.1 (Revisited)** Carvedilol in severe heart failure

A total of 12 patients (6 in each of the two treatment groups) underwent cardiac transplantation during the study. One issue was how to deal with these patients in the analysis, as receiving a transplant impacts the patient's survival prospects. In the primary analysis, these patients were censored at the time of transplantation. As a sensitivity analysis, the eventual time to death (or censoring at the end of follow-up if they were still alive) was included in the analysis. Parker et al. (2001) reported that the conclusions were essentially unchanged.

## 13.8 Crossover

### 13.8.1 Rank Preserving Structural Failure Time Model

In an oncology study, patients in the control arm are often offered a switch to the experimental treatment following progression or are sometimes switched before progression if, in the view of the investigator, they are performing poorly and would potentially benefit from such a switch. An intention-to-treat (ITT) type analysis of overall survival (OS) would compare the treatment groups ignoring the switch. If the experimental treatment is effective, this analysis will likely underestimate the true treatment effect since one assumes there would be benefit for some patients who switch. In these situations, there is often interest in estimating the true treatment effect had switching not occurred or not been available. This thinking is in line with a hypothetical strategy in the context of estimands (see Chapter 8), while the ITT approach follows the treatment policy strategy.

A simple approach to analysis would censor the patients at the time point where crossover occurs, but a little thought indicates that this constitutes dependent (informative) censoring. These patients are being selectively censored in such an analysis because they either have progressed or are performing poorly, and these patients are more likely to die sooner than similar patients who are progression-free in the control group at the same point in time. This will cause bias in the analysis of OS that will favour the control group since

you are artificially reducing the number of deaths that could be observed in that group.

An alternative analysis based on the Rank Preserving Structural Failure Time Model (RPSFTM) (see, for example, Morden et al. [2001]) provides a method that, under some strict assumptions, estimates the true treatment effect had switching not been available.

This model is fitted to the data as follows. Assume that, for patients randomised to the experimental treatment group and those switching to the experimental treatment, the experimental treatment multiplies OS by a factor  $\psi$  compared to control. For example, if  $\psi = 1.1$ , the experimental treatment increases the time to death by 10%. Apply that factor  $1/\psi$  to 'shrink' the OS values for the patients randomised to the experimental treatment and 'shrink' the OS values for control patients who switch to the experimental treatment from the time of the switch. In Table 13.1, we can see how this is applied. Patient 1 was randomised to the experimental treatment group and had an OS value of 14 months. This value is shrunk by dividing by 1.1 (10% shrinkage factor) to give an adjusted OS value of 12.7 months. Patient 2, also randomised to the experimental group, has a censored value for OS equal to 9 months, and this is also shrunk to give an adjusted censored value of 8.2 months. Patient 3 was randomised to the control (placebo) group and had an OS value of 8 months. The shrinkage factor is not applied to this patient since at no point did they receive the experimental treatment. Finally, patient 4 was also randomised to the control group and had an OS value of 10 months but switched to the experimental treatment at 6 months due to progression. For this patient, the six-month portion of OS is unaffected, but the remaining four months of survival on the experimental treatment are shrunk on division by 1.1 to 3.6 months; this is then added to six to give an adjusted OS value of 9.6 months. The Kaplan-Meier (KM) curves are then calculated based on these adjusted survival times according to treatment group as randomised and compared with the logrank test. The shrinkage factor (1.1 in our example) is then varied until the treatment group KM curves are as close as possible (based on having a test statistic value that is as small as possible). The control group KM curve is then

**Table 13.1** Application of the Rank Preserving Structural Failure Time Model

Patient	Randomised group	OS	Switching?	Adjusted OS (with 10% shrinkage)
1	Experimental	14 months	Not applicable	$14/1.1 = 12.7$ months
2	Experimental	9 months (censored)	Not applicable	$9/1.1 = 8.2$ months (censored)
3	Placebo	8 months		8 months
4	Placebo	10 months	Switched to exp. group at month 6	$6 + 4/1.1 = 9.6$ months

the KM curve we would have seen had none of the patients in that group switched treatment; the process has removed the effect of the experimental treatment on OS from that group. The final part of this evaluation then compares this modified KM curve for the control group with the original KM curve for the experimental group.

The fundamental assumption that underpins this methodology is that the shrinkage factor applies equally to patients who are randomised to the experimental treatment and to those who are randomised to the control group but switch subsequently to the experimental group from the time point when the switch occurs. This assumption is untestable, and as we shall discuss, this is one of the main reasons regulators prefer that this analysis be used only as a sensitivity analysis rather than being central to the regulatory decision.

**Example 13.3** Lenvatinib vs. placebo in radioiodine-refractory thyroid cancer

Schlumberger et al. (2015) report a trial of lenvatinib vs. placebo in radioiodine-refractory thyroid cancer. An independent radiologic review confirmed that disease progression patients who were receiving placebo could choose to cross over to open-label lenvatinib. The primary analysis of OS according to ITT gave a non-significant difference between the treatment groups (hazard ratio [HR] = 0.73, 95% CI 0.50 to 1.07,  $p = 0.10$ ). Following the application of the RPSTM to account for crossover, this difference became statistically significant (HR = 0.62; 95% CI 0.40 to 1.00,  $p = 0.05$ ).

### 13.8.2 Regulatory position

The regulatory position is that the ITT approach (treatment policy strategy in the estimand framework) is the basis for regulatory decision making, while the RPSTM approach (hypothetical strategy in the estimand framework) constitutes a sensitivity analysis. This regulatory position is largely because of the untestable assumptions that underlie the RPSTM and similar approaches.

**EMA (2018): 'Question and answer on adjustment for crossover in estimating effects in oncology trials'**

*'Given that the underlying assumptions of the adjustment methods for crossover described above can in principle not be proven to be true, a positive result from an analysis adjusted for crossover cannot be used to rescue a trial that is negative as per other evidence, or to ascertain that a treatment confers an OS advantage when this is not apparent in an analysis that does not (strongly) depend on unverifiable assumptions, such as an "ITT-analysis" that uses the observed OS outcome for each patient. For these reasons, these analyses may only be useful for regulatory purposes as supportive or sensitivity analyses with (as outlined above) a clearly demonstrated robustness against deviations from the underlying assumptions'.*

**Example 13.3 (Revisited)** Lenvatinib vs. placebo in radioiodine-refractory thyroid cancer

Regulatory approval (EMA and FDA) was granted for lenvatinib (Example 13.3) based on the positive results for progression-free survival (PFS) and response rate and an acceptable safety profile. The ITT analysis of OS is also reported in the label, but the analysis adjusting for crossover (RPSFTM) is not.

Following the hypothetical strategy for economic evaluation can be more important for the evaluation of the experimental treatment, as indicated by Latimer and Abrams (2014), who suggest in particular that ITT has limited value: '*When switching occurs, an "intention to treat" (ITT) analysis – whereby the data are analysed according to the arms to which patients were randomised – of the overall survival (OS) advantage associated with the new treatment will be biased: If control group patients switch treatments and benefit from the new treatment the OS advantage of the new treatment will be underestimated. For interventions that impact upon survival, health technology assessment (HTA) bodies such as the National Institute for Health and Care Excellence (NICE) require that economic evaluations consider a lifetime horizon. This is problematic in the presence of treatment switching, because standard ITT analyses are likely to be inappropriate*'.

## 13.9 Composite time-to-event endpoints

### 13.9.1 Cumulative incidence functions

It is quite common to see time-to-event endpoints that are composites: for example, PFS in oncology and time to a major cardiovascular event (MACE) in cardiovascular trials. These composite endpoints are usually the relevant primary endpoints, but there is also interest in evaluating the different components of the composite; indeed, this is necessary in some cases to comply with regulatory requirements (see Section 10.4.2). With PFS, breaking down progression further, for example, to look at local/regional recurrence and distant metastases, may also help in understanding the mechanism of action of the experimental drug.

Cumulative incidence functions (CIFs) provide a methodology that breaks down the KM curve for the composite into its component parts. In developing CIFs, it is better to think of the KM curve in terms of its inverse ( $1 - \text{KM}$ ): the probability of having the event by time  $t$ , rather than the probability of being event-free at time  $t$ . Suppose there are three components to the composite. Let  $T$  denote the time to the composite event, and let  $j = 1, 2, 3$  denote the three distinct components: for example, for time to the three-point MACE,  $j = 1$  for nonfatal stroke,  $j = 2$  for nonfatal MI and  $j = 3$  for cardiovascular death. The

probability of suffering the composite event before time  $t$  (written mathematically as  $\text{pr}(T \leq t)$ ) can be broken down into three components as follows:

$$\text{pr}(T \leq t) = \text{pr}(T \leq t \text{ and } j = 1) + \text{pr}(T \leq t \text{ and } j = 2) + \text{pr}(T \leq t \text{ and } j = 3)$$

The three components are the cumulative incidence functions, one for each of the components. These functions are usually denoted  $I_1(t)$ ,  $I_2(t)$  and  $I_3(t)$ , with  $I(t)$  denoting  $1 - \text{KM}$  so that

$$\text{pr}(T \leq t) = I(t) = I_1(t) + I_2(t) + I_3(t)$$

The extension of the logrank test known as Gray's test (Gray, 1988) allows a  $p$ -value comparison of treatment groups for each component.

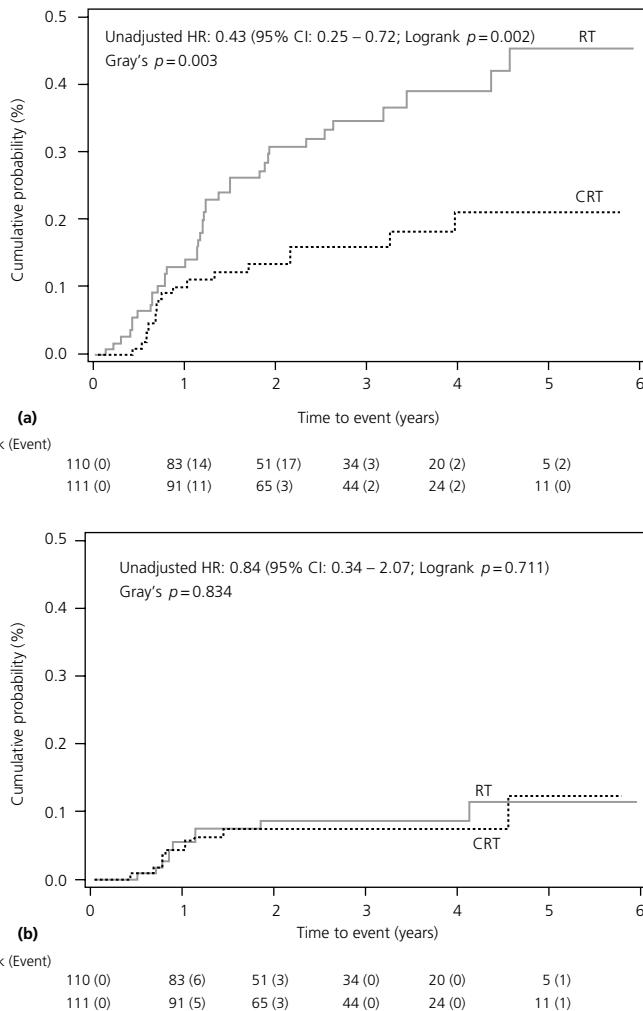
Tai et al. (2011) summarise the theory behind CIFs and include an example from a randomised clinical trial of patients with advanced, non-metastatic, nasopharyngeal cancer (NPC). The trial compared radiotherapy alone (RT) with radiotherapy plus chemotherapy (RTC) in terms of OS. There was, however, secondary interest in time to either distant metastases or loco-regional recurrence, and this secondary outcome is the focus of our discussion. Figure 13.9 presents CIFs that have broken down the time to recurrence in terms of the first occurrence of distant metastases or loco-regional recurrence. The  $p$ -values for Gray's test show statistically significant differences for distant metastases ( $p = 0.003$ ) with a reduction in the incidence of this event for RTC compared to RT but non-significant differences for loco-regional recurrence ( $p = 0.834$ ). It is also possible to obtain hazard ratios for treatment effects in association with CIFs. In the example, the hazard ratio for distant metastases was 0.43 (95% CI [0.25, 0.76]) and for loco-regional recurrence was 0.91 (0.37, 2.24).

There is a complementary methodology to CIFs that focuses on an additive decomposition of the hazard function rather than an additive decomposition of the KM function and revolves around consideration of *cause-specific hazard functions*. The interested reader is referred to Tai et al. (2011) for further details.

### 13.9.2 Regulatory position

As discussed in Section 10.5.2, there is a need from a scientific and regulatory point of view to report on the different components of a composite endpoint to ensure that a positive effect on the composite is not hiding any concerning negative trends in one or more of the components. Cumulative incidence functions allow us to view the components individually, but there is another way of thinking about this breakdown of the composite.

An example given in the FDA 2017 guideline on 'Multiple Endpoints in Clinical Trials' is a randomised trial of losartan in diabetic neuropathy (the RENAAL trial). The primary endpoint in this trial was a composite time-to-event endpoint, with components doubling of serum creatinine, end-stage renal



**Figure 13.9** Cumulative incidence of relapse amongst patients in RT (solid line) and CRT (dashed line) for (a) Event M: Distant metastasis, (b) Event R: Loco-regional recurrence.  $HR = \text{Hazard Ratio}$ ;  $CI = \text{Confidence Interval}$ . Source: Tai et al. / BioMed Central Ltd / Public Domain CC BY

disease (ESRD) or death. The composite endpoint investigates time to the first of these events, and the cumulative incidence functions break this composite into its three components. However, patients can suffer multiple events. For example, a patient can die following a doubling of serum creatinine and/or end-stage renal disease. The FDA recommendation is to present – in addition to the usual kind of analysis of the primary endpoint based on hazard ratios and possibly evaluations based on CIFs – frequencies according to a decomposition of the primary endpoint and frequencies according to the multiple events. Table 13.2 is based on the table provided in the FDA (2017) guidance document, which was

**Table 13.2** Decomposition of endpoint events in RENAAL

Endpoint	Losartan (N = 751)	Placebo (N = 762)	Hazard ratio (95% CI)	p-value
<b>Primary endpoint</b>				
Doubling of serum creatinine, ESRD or death	327	359	0.84 (0.72, 0.97)	0.022
<b>Decomposition of the primary endpoint</b>				
Doubling of serum creatinine	162	198	0.75	
ESRD	64	65	0.93	
Death	101	96	0.98	
<b>Any occurrence of individual components</b>				
Doubling of serum creatinine	162	198	0.75	
ESRD	147	194	0.71	
Death	158	155	1.02	

Source: FDA 2017 / U.S. Department of Health and Human Service / Public Domain

in turn extracted from the FDA Statistical Review ([http://www.accessdata.fda.gov/drugsatfda\\_docs/nda/2002/20-386s028\\_Cozaar.cfm](http://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/20-386s028_Cozaar.cfm)).

For the breakdown of the first event, the ESRD component accounted for 127 (= 64 + 65) patients. However, many more patients (341 = 147 + 194) suffered ESRD at some stage in the study. Further, in the breakdown of the primary endpoint, 197 (= 101 + 96) patients recorded death as the first event on the composite, although during the study, 313 (= 158 + 155) patients died. In terms of interpretation, the decomposition of time to the first event is hiding the positive effect of losartan on a reduction in the occurrence of ESRD, which is seen when the data on any occurrence of the individual components are presented. The corresponding hazard ratios were 0.93 (time to ESRD as a first event) and 0.71 (time to ESRD as either first or subsequent event), and the 95% CI for the latter was (0.57, 0.89), showing a nominally significant difference of the HR from 1.

As the FDA guidance points out, the analysis of so-called competing risks can be complicated:

#### **FDA (2017): 'Multiple Endpoints in Clinical Trials. (Draft) Guidance for Industry'**

'The analysis of any occurrence of an event type, however, can be complicated by the issue known broadly in statistics as competing risks. This is the phenomenon wherein occurrence of certain endpoints can make it impossible to observe other events in the same patient. For example, in the RENAAL trial, patients whose first event was death could never be observed to have doubling of serum creatinine. If one study group had higher early mortality, it could appear to have a favorable profile with respect to other endpoint events simply because fewer patients survived, diminishing the number of patients at risk for the other types of events'.

When reporting results on composite endpoints, it is important to avoid pitfalls in interpretation. It is recommended that results be reported for the composite itself, its components (first occurrence) and all occurrences of all components to give the full picture.

## 13.10 Sample size calculations

The power of a study where the primary endpoint is time to event depends not so much on the total patient numbers but on the number of patients with events. A trial with 1000 patients and 100 deaths has the same power as a trial with only 200 patients but also 100 deaths. The sample size calculation for survival data is therefore done in two stages. Firstly, the required number of patients with events is calculated in the two groups combined; and secondly, this is factored upwards by an assumed proportion of patients who are expected to have events, again in the two groups combined, to give the required number of patients in total for the trial.

**Example 13.4** Sample size calculation for survival data

Suppose that on a standard treatment, it is expected that the three-year survival rate will be 35%. It is desired to detect an increase in this rate to 55% with 90% power in a 5% level test. The required number of deaths to give this level of power is 133. The overall survival rate at three years is expected to be 45% (the average of 35% and 55%), and therefore, we expect to see 45% of patients die. It follows that  $133/0.45 = 296$  patients need to be recruited overall, or 148 patients per group, and followed for three years.

This sample size calculation assumes that all patients are followed for a fixed time: three years in Example 13.4. In practice, patients will enter the study over a prolonged recruitment period and likely be followed up for different periods of time, with an analysis taking place following a minimum period of follow-up: for example, three years. In addition, withdrawals will impact the potential to observe events in all patients. This adds complexity to the calculations, and the reader is referred to Machin et al. (2006) for further details and considerations.

Whatever assumptions are made, there is no guarantee that things will turn out as expected: in our example, even though we require 133 events, 296 patients may not give us those 133 events following three-year follow-up for each patient. An alternative way of designing the trial would be to continue follow-up until the required number of events are observed. Having an *event-driven study* is frequently a safer option to ensure that the required power is achieved, although it does add another element of uncertainty: the duration of the study. Careful management of the whole trial process is then needed.

**Example 13.1 (Revisited)** Carvedilol in severe heart failure

This trial adopted a design that continued follow-up until the required number of events was observed:

'The sample size was estimated based on the following assumptions: the one-year mortality in the placebo group would be 28 per cent, the risk of death would be altered by 20 per cent by treatment with carvedilol and the study would have 90 per cent power (two-sided  $\alpha = 0.05$ ) to detect a significant difference between the treatment groups. Since it was recognised that the estimated rate of events might be too high, the trial was designed to continue until 900 deaths had occurred'. (Packer et al., 2001)

This sample size methodology is usually undertaken assuming proportional hazards and is based around the logrank test as the method of analysis. If this assumption is not appropriate and we do not expect proportional hazards, comparing RMSTs may provide an alternative framework for data analysis. The interested reader is referred to Royston and Parmar (2013) for methods that cover sample size calculations in association with RMSTs.

## CHAPTER 14

# Interim analysis and data monitoring committees

### 14.1 Stopping rules for interim analysis

In Chapter 10, we spoke extensively about the dangers of multiple testing and the associated inflation of the type I error. Methods were developed to control that inflation and account for multiplicity in an appropriate way.

One area that we briefly mentioned was interim analysis, where we look at the data in the trial as they accumulate. The method of Pocock (1977) was discussed to control the type I error rate across the series of interim looks. The Pocock methodology divides the 5% type I error rate equally across the analyses. So, for example, for two interim looks and a final analysis, the significance level at each analysis is 0.022. For the O'Brien and Fleming (1979) method, most of the 5% is left over for the final analysis, while the first analysis is at a very stringent level and the adjusted significance levels are 0.00052, 0.014 and 0.045.

These two methods are the most common approaches seen in pharmaceutical applications. A third method, which we see used from time to time, is by Haybittle (1971) and Peto et al. (1976). Here, a significance level of 0.001 is used for each of the interims, leaving most of the 5% for the final analysis as for the O'Brien-Fleming scheme. For two interims and a final analysis, the adjusted significance level for the final analysis is 0.05 to two decimal places; for three interims, we have 0.049 left over. Clearly, this method has little effect on the final evaluation, but there is also little chance of stopping at an interim stage. Many other schemes are available; any scheme that controls the overall 5% level of significance can be used, provided it has been pre-specified in the protocol in accordance with the usual requirements for the control of multiplicity.

The term *stopping rule* has been used as part of the title of this section to indicate that when the required level of statistical significance is achieved at an interim analysis, the trial can be stopped in terms of further recruitment and statistical significance declared. Table 14.1 presents two hypothetical cases involving two interim analyses and a final analysis using the O'Brien-Fleming scheme. In case 1, a non-significant difference is seen at the first interim analysis after 200 patients have reported on the primary endpoint, and the trial continues. But at the second interim analysis, after 400 patients have reported on the primary endpoint,

**Table 14.1** Examples of interim analyses and stopping rules

	<b>Analysis</b>	<b>Number of patients</b>	<b>Adjusted significance level</b>	<b>p-value</b>	<b>Decision</b>
Case 1	1	200	0.00052	0.073	Non-significant differences; continue trial
	2	400	0.014	0.009	Stop trial and declare statistically significant differences
	3	600	0.045	N/A, analysis not done	
Case 2	1	200	0.00052	0.17	Non-significant differences; continue trial
	2	400	0.014	0.03	Non-significant differences; continue trial
	3	600	0.045	0.048	Non-significant differences

N/A, not applicable.

$p = 0.009$ . This is below the required significance level of 0.014, and the trial can be stopped and statistical significance declared. The final analysis after 600 patients does not take place. In case 2, non-significant differences are seen at each of the two interim analyses according to the required levels of significance. However, the  $p$ -value at the second interim analysis is below the usual level of statistical significance, with  $p = 0.03$ . This is not statistically significant when judged against the required significance level of 0.014, and the trial continues. The final  $p$ -value is 0.048, which again falls below 0.05 (nominal significance) but does not fall below the significance level of 0.045 as required by the interim analysis scheme. This is unfortunate, especially when the  $p$ -value of 0.03 seen at the second interim is below the end-of-trial threshold. Despite all of this, statistical significance cannot be declared in this case.

The methods outlined assume that the analyses take place at equally spaced intervals. So, for example, if the total sample size is 600 and we are planning two interims and a final analysis, the interims take place after 200 and 400 patients have been observed for the primary endpoint for these adjusted significance levels to apply. For practical reasons associated with recruitment, etc., it may be difficult to ensure that this happens precisely. The analyses may end up being conducted, for example, at 208, 390 and 615 patients, deviating from what was pre-planned. Also, for other reasons, we may not want the analyses to be equally spaced; we will see an example later in the chapter. In both these cases, we would need to use a more general methodology,  $\alpha$ -spending functions (Lan and DeMets [1983]) to calculate the adjusted significance levels. Applying the  $\alpha$ -spending methodology requires some sophisticated computer software; several computer packages are available for these calculations.

Finally, for time-to-event endpoints, information is carried through the numbers of patients with events and not directly the number of patients in the

trial (see Section 13.10). For example, if the trial is continuing until we have seen 900 deaths, then equally spaced interim analyses will correspond to when 300 and then 600 deaths have been observed.

## 14.2 Stopping for efficacy and futility

There are several potential reasons we would want to stop a trial at an interim stage:

- *Efficacy*: There is overwhelming evidence of the efficacy of the experimental treatment.
- *Futility*: The data are such that there is little chance of achieving a positive result if the trial were continued to completion.
- *Safety*: There is collective evidence of serious safety problems with the experimental treatment.

We will deal with safety and this final point separately in Section 14.3.

### 14.2.1 Efficacy

Each of the schemes outlined in Section 14.1 for dividing up the 5% type I error can be applied to evaluate efficacy in theory. In practice, however, we would only want to stop for efficacy if the evidence was overwhelming. For this reason, the O'Brien and Fleming scheme appears to be appropriate in that it has a sliding scale of adjusted significance levels starting from very stringent through less stringent to something close to 5% for the final analysis. The Pocock scheme pays a high price for the early looks, and at the final analysis, the adjusted significance level is well below 5%; it would be very unfortunate if, with two interims and a final analysis, the final analysis gave  $p = 0.03$ . Statistical significance could not be claimed under the Pocock scheme because the  $p$ -value has not fallen below the required 0.022. The Haybittle and Peto et al. scheme is also a possibility but gives only a small chance of stopping early.

Note that the schemes we have been discussing are based upon the calculation of standard two-sided  $p$ -values. Clearly, we would only claim overwhelming efficacy if the direction of the observed treatment effect were in favour of the experimental treatment. If the treatment effect were in the opposite direction, we might still stop the trial, but now based on differences in the negative direction, which would constitute harm. Formal stopping rules for harm can be set up, but this is generally not necessary since effects in the direction of harm could be a safety issue and stopping would be considered on those grounds. One issue that can cause problems is when a nominally significant result (that is,  $p \leq 0.05$ ) is seen at an interim analysis for efficacy but the requirement for declaring statistical significance is not met. For example, for the three analyses described in Section 14.1, the second interim required the  $p$ -value to be  $\leq 0.014$  for efficacy to be declared. Suppose at that analysis we see  $p = 0.03$ , as in case 2 in Table 14.1.

Under conventional circumstances, such a  $p$ -value could be viewed as providing clear evidence supporting efficacy. However, in this case, the result cannot be declared statistically significant. Allowing that would destroy the control of the type I error and be unacceptable from a regulatory (and scientific) perspective.

### **14.2.2 Futility and conditional power**

In addition to looking for overwhelming efficacy, there is also the possibility of stopping the trial at an interim stage for futility. For example, the analysis of the interim data may not be especially favourable for the experimental treatment, and were the trial to continue to the end, there would be no realistic possibility of obtaining a positive (statistically significant) result. For commercial reasons, it may be better to abandon the trial to save on additional costs and resources. Also, for ethical reasons, is it appropriate to include patients in a trial that has little potential to achieve its objective?

There are several approaches to evaluating futility. One common method is based on *conditional power*. At the design stage, we may base the sample size calculation on a power of, say, 90% to detect a certain level of effect – the clinically relevant difference,  $d$  – arguing that a difference less than  $d$  is of little or no clinical importance. It is possible at an interim stage to recalculate the power of the trial to detect a difference  $d$ , since at that stage we already have in hand a proportion of the data on which the final analysis will be based. Suppose that this so-called *conditional power* was equal to 20%. In other words, were we to continue the trial (and if the treatment difference observed to date reflected the true treatment difference), there would be only a 20% probability of seeing a significant  $p$ -value at trial completion. Under these circumstances, it may not be worth continuing. In contrast, if the conditional power turned out to be, say, 60%, it might be worth carrying on since we would have a reasonable chance of achieving a positive conclusion if the trial continued. The cut-off should ideally be pre-specified, to avoid ambiguity; it is based on commercial risk/benefit considerations but in general may be around 20–30%.

This calculation of conditional power assumes that the observed treatment difference at the interim stage is the true difference, termed the conditional power *under the current trend*. It is also possible to calculate conditional power under other assumptions: for example, that the true treatment difference in the remaining part of the trial will be equal to  $d$ , the treatment effect assumed initially. If the observed effect at the interim analysis is lower than  $d$ , this represents a much more optimistic assumption and will lead to a larger value for the conditional power. We are saying that even though in the first part of the trial, the treatment difference was less than the difference  $d$  that we were targeting, we still believe that  $d$  is the true treatment difference. These calculations under different assumptions about how the future data will behave can provide a broad basis on which to make judgements about terminating the trial for futility.

From a technical perspective, there is no alpha price to pay for conducting interim analyses for futility. Such analyses are not in any way associated with declaring a positive result, and the false positive elements associated with inflating the type I error rate need not be considered. Regulators however may insist on a small price being paid for such analyses for the following reason. Suppose that at an interim analysis for futility, the observed treatment difference is very positive for the experimental treatment: so much so that it could be viewed as unethical to continue recruiting patients into the study or continuing to treat patients with a clearly inferior control treatment. This places the data monitoring committee (DMC) in a difficult position. On the one hand, there is no signal for futility; but on the other hand, there is a strong signal for overwhelming efficacy for the experimental treatment. Having a Haybittle-Peto scheme with a small  $\alpha$  price of 0.001 attached to an efficacy evaluation at the interim analysis for futility may satisfy regulators and avoid placing the DMC in that difficult position while not spending too much alpha for efficacy at the interim look.

### 14.2.3 Some practical issues

In general, a trial is not required to have an interim analysis for either efficacy or futility. In most long-term trials, however, where there is the opportunity for an interim evaluation, it may be worth putting one or more in place. The interim can involve only efficacy, only futility, or both and may involve other things, such as a re-evaluation of sample size (see Section 8.5.3). But one aspect must be borne in mind: stopping a trial early for overwhelming evidence of efficacy based on the primary endpoint may compromise the ability to gain good information on important secondary endpoints, detailed efficacy information in subgroups and, finally, enough data on the safety of the experimental treatment. If this is a possibility, then introducing interim analyses with associated stopping rules may not be the sensible thing to do. See Example 14.1 for a situation that relates to these points.

#### Example 14.1 Lapatinib plus capecitabine for HER2-positive advanced breast cancer

This was a phase III trial in HER2-positive breast cancer comparing lapatinib plus capecitabine vs. capecitabine monotherapy (Geyer et al., 2006). The planned sample size was 528, with a final analysis of time to progression scheduled to take place after 266 progression events. A single interim analysis was prospectively included for this endpoint after approximately 133 progression events, using a one-sided significance level according to an O'Brien–Fleming scheme. The interim analysis took place following 146 events; at that time, 324 patients had been recruited. The O'Brien–Fleming adjusted significance level for this analysis to preserve an overall one-sided significance level of 0.025, calculated using an alpha-spending function, was 0.0014. The observed  $p$ -value for this interim analysis for time to progression was 0.00016, well below the required cut-off for statistical significance, and the independent DMC recommended stopping the trial based on this evidence for overwhelming efficacy. Recruitment was stopped, and patients in the control arm were offered treatment with lapatinib.

While the data for time to progression were convincing, regulators expressed concern that this early stopping effectively prevented the collection of long-term survival data. The results for overall survival at the time of the interim gave a non-significant two-sided  $p$ -value of 0.72; there were 36 deaths in the lapatinib group compared to 35 in the control arm. Following extensive discussions with regulators, these data nonetheless led to the successful filing for lapatinib in this indication.

A related point worth raising is the issue of *overrun*. An interim analysis is based on a data cut at a particular point in time. By the time those data are cleaned, analysed and presented to the DMC and a decision is taken on whether to stop or continue the trial, things will have moved on. More patients will have been recruited and more data obtained on those patients already recruited. If the trial is stopped, the final database may be somewhat different from the database on which the decision to stop was based. The concern is that statistical significance according to the required  $\alpha$ -level at the interim analysis may have been lost with an analysis based on the final database. Is this a concern? Not really. Regulators are aware that this can happen and, while requiring all data to be presented and analysed, consider the  $p$ -value from the interim analysis on which the decision to stop was taken as correct for formal inferential purposes.

For practical reasons, the number of interims should be kept to a minimum. Undertaking interims adds cost to the trial, and they also need to be very carefully managed. In particular, the results of each interim must be made available in a timely way for go/no-go decisions to be made in good time. Remember that the trial does not stop to allow the interims to take place; recruitment and follow-up continue. So do not overburden the clinical trial with lots of interim analyses; two at most is often my recommendation!

In some situations, trial recruitment may be complete once the results of the interim analysis are known. Is the interim analysis therefore of any value? Well, commercially and clinically, it could be, in that it could enable a regulatory submission to be made earlier than would have been the case had the trial run its full course and effective treatments could be made available to patients earlier than they would otherwise have been.

The results of interim analyses are not formally binding for the sponsor, but it would be a very brave decision to continue a trial when the decision coming out of the interim was to stop. If the trial data at the interim give a statistically significant result, there would clearly be ethical problems in continuing randomising patients or treating patients with a treatment that has been demonstrated to be inferior. The investigators will likely not be willing to continue with the study in its current form. For futility, this can be less of an issue; there may be no major ethical problems with continuing the trial with two treatments that appear somewhat similar – although, as mentioned earlier, is it ethical to continue a trial that is going nowhere?

It is almost self-evident that all analyses of the kind we are discussing here must be pre-planned in a detailed way. The regulators are very unhappy with unplanned interims or interims that are ill-defined. Such situations give rise to major problems at the time of submission.

From time to time, sponsors are interested in conducting so-called *administrative* interim analyses. These analyses are not intended for inferential purposes, and no stopping rules are involved. If, for example, they are conducted in a blinded way to monitor recruitment or overall event rates, there are no concerns. But if they involve unblinding and comparative analyses, there will likely be concerns from regulatory authorities, primarily because of potential problems with operational bias caused by the dissemination of those comparative interim results. These issues are discussed further in the broader context of adaptive designs in Section 16.2.3. Sankoh (1995), a former FDA reviewer, makes these comments: '*Planned or unplanned administrative interim analyses in confirmatory clinical trials that involve formal statistical methods that compare the relative treatment group differences and the dissemination of analysis results (either dictated by forces outside the clinical trial operations or not) should be treated as interim analyses*'.

#### **14.2.4 Point estimates and confidence intervals**

When interim analyses for efficacy are incorporated, a trial will stop for efficacy at an interim stage or continue through to the planned maximum sample size. In either case, calculating a confidence interval (CI) for the treatment difference is not as straightforward as when no interim analyses are involved. In this situation, it is incorrect to calculate a CI in the conventional way. If we were to do that, then the 95% CI, for example, would not have its usual 95% coverage interpretation; it would not contain the true difference 95% of the time. In addition, the usual point estimate of the treatment difference, such as the difference in the two means or the hazard ratio, is also subject to bias, although these biases are more theoretical than of practical consequence. Methods are available that take account of the interim analyses and avoid this bias, but these are beyond the scope of this book; the interested reader is referred to Jennison and Turnbull (2000) for a technical development.

### **14.3 Monitoring safety**

In addition to considerations of efficacy and futility, it is appropriate in most long-term trials to monitor safety in an ongoing way. This is usually facilitated through a DMC, and we will consider various aspects of the committee's structure and conduct later. For the moment, we will focus on associated statistical methodologies. Ongoing safety monitoring is done on a regular basis by looking at various aspects of safety: adverse events (serious and non-serious), vital signs,

key laboratory parameters, physical examination, ECGs, etc., both in individual cases and overall, for each of the treatment groups. I hesitate to say that this is done in an informal way, because it is taken very seriously; but there is usually little formal statistical structure wrapped around the process. Yes, in some cases (although it is the exception rather than the rule) *p*-values are put on the adverse events, suitably grouped, but these are simply used as flags for potential problems. A discussion ensues, and members of the DMC take decisions. Producing *p*-values for comparisons of adverse events between treatment groups can be problematic due to the multiplicity involved in this process, and the general recommendation is not to calculate them. If they are produced, we should be very careful not to jump every time we see a statistically significant *p*-value. Conversely, there may be safety issues of concern even though formal statistical significance has not been achieved.

There is further discussion on these points in Section 19.3.

## 14.4 Data monitoring committees

### 14.4.1 Introduction and responsibilities

In this section, we cover several aspects associated with DMCs, particularly in relation to statistical issues. This is not meant to be a comprehensive discussion of the area, and the reader is referred to the book by Ellenberg, Fleming and DeMets (2019) for an excellent and exhaustive coverage and a plethora of case studies. There are two guidelines – one from the FDA (2006), ‘Establishment and Operation of Clinical Trial Data Monitoring Committees’; and one from CHMP (2005), ‘Guideline on data monitoring committees’ – that outline the roles and responsibilities of DMCs in the regulatory environment. DMCs are also referred to as data monitoring boards (DMBs), data and safety monitoring committees/boards (DSMCs/DSMBs) and safety review committees (SRCs).

It is important that the trial sponsor remains blind to the accumulating data within the separate treatment groups, and an important reason for having a DMC is to enable trial data to be looked at without compromising that blinding. The responsibilities of a DMC vary depending on the circumstances. The main responsibility, however, is always to protect the trial participants’ safety and ensure that the trial is conducted ethically. There may also be additional responsibilities associated with interim analyses for overwhelming efficacy and/or futility. The DMC may or may not have access to evolving efficacy data, as the committee may only be looking at safety. This can cause problems: it is difficult for DMCs to make recommendations based on safety in isolation, and the absence of efficacy data makes it impossible to make a risk/benefit judgement. There should therefore be provision for the committee to have access to efficacy data, either routinely or on request.

**CHMP (2005): 'Guideline on data monitoring committees'**

*'In most cases, safety monitoring will be the major task for a DMC. Even if the safety parameters monitored are not directly related to efficacy, a DMC might need access to unblinded efficacy information to perform a risk/benefit assessment in order to weigh possible safety disadvantages against a possible gain in efficacy'.*

The DMC also has responsibilities in relation to protecting the scientific integrity of the trial. In particular, the DMC must ensure that all interim analyses and safety monitoring activities are conducted appropriately to protect blinding and control type I error. There is also a responsibility to oversee the trial's overall conduct to ensure that it will fulfil its objectives.

**FDA (2006): 'Establishment and Operation of Clinical Trial Data Monitoring Committees'**

*'A DMC will generally review data related to the conduct of the study (that is, the quality of the study and its ultimate ability to address the scientific questions of interest), in addition to data on effectiveness and safety outcomes. These data may include, among other items:*

- *Rates of recruitment, ineligibility, non-compliance, protocol violations and dropouts, overall and by study site;*
- *Completeness and timeliness of data;*
- *Degree of concordance between site evaluation of events and centralized review;*
- *Balance between study arms on important prognostic variables;*
- *Accrual within important subsets'.*

It is important that interaction between the sponsor and the DMC is kept to an absolute minimum to avoid any inadvertent communication of unblinded information. Sponsor exposure to unblinded interim results, except in terms of the communication of go/no-go decisions associated with interim analyses, can cause operational bias and seriously compromise the scientific validity of the study. The sponsor is in the position of being able to influence the future conduct of the trial, and exposure to interim results could influence aspects of that, leading to bias. There can also be pressure from the sponsor to provide interim results for planning purposes: taking decisions about future production facilities, agreeing on budgets for further trial activity and so on. Such pressure should be resisted, as it can seriously undermine the integrity of the trial. Where this need is compelling, communication should be managed in a very tight way. Further discussion of this point was provided earlier (Section 14.2.3). See also the FDA (2006) guideline (Section 6.5).

A DMC is usually needed in long-term trials in life-threatening diseases and sometimes in non-life-threatening diseases where there are potential safety concerns. It may also be necessary to have DMCs in studies in specific and

vulnerable or fragile populations such as children, pregnant women, or the very elderly, but DMCs are not usually necessary in phase I and early phase II trials or in short-term studies where the goal is relief of symptoms.

#### **14.4.2 Structure and process**

The independence of the committee from the sponsor is important, and there should also be no conflicts of interest among the participants: for example, holding equity in the sponsor's company or a direct competitor's company. The members of the committee should also not be otherwise involved in the study: for example, as investigators. The DMC consists of at least three participants, one of whom is a statistician; the remaining participants are clinicians with expertise in relevant clinical disciplines associated with the disease under study or potential side effects. If the trial is an international study, it is advisable to have committee members from the spread of geographical regions in which the trial is to be conducted. This final point often determines the ultimate size of the DMC.

At least one other statistician is also involved closely with the activities of the DMC: the statistician who supplies data tables, listings and figures (TLFs) to the committee for their deliberations. This 'independent' statistician should also not be otherwise involved in the trial as they will access unblinded information. In the way these things tend to be organised, this individual may be part of a clinical research organisation (CRO) that is providing this service (and potentially other services) to the sponsor. See Pocock (2004) and the FDA (2006) guideline for further discussion on this and related points. The DMC also receives details of individual patient serious adverse events (SAEs), usually in the form of narratives, and these are often supplied directly from the sponsor. An alternative is to provide summary tables for SAEs periodically (for example, monthly), with an option to see narratives if needed. These patients can be unblinded by the statistician who controls the supply of unblinded TLFs to the DMC if this has not already been done by the sponsor's pharmacovigilance group.

Data tables produced for the DMC should contain separate summaries by treatment group, with the treatment groups labelled A and B (partially blinded). A separate password-protected electronic file should be provided to the members with decodes for A and B to enable the DMC members to be completely unblinded. This may seem an elaborate process, but it protects against inadvertent unblinding.

#### **FDA (2006): 'Establishment and Operation of Clinical Trial Data Monitoring Committees'**

*'A common approach is presentation of results in printed copy tables using codes (for example, Group A and Group B) to protect against inadvertent unblinding should a report be misplaced, with separate access to the actual study arm assignments provided to DMC members by the statistical group responsible for preparing DMC reports'.*

From time to time, the DMC may request additional data, and there should be a process for communicating these needs to those supplying the unblinded TLFs. Preferably such requests should not involve the sponsor; otherwise, the requests themselves may reveal specific concerns among the DMC members, and this form of unblinding has the potential for operational bias.

The activities of the DMC should be covered by a charter, prepared in advance of running the trial. The charter should detail the participants, their responsibilities, the format and conduct of meetings, communication pathways with the sponsor, decision making, confidentiality, indemnity and conflict-of-interest issues. Details regarding the format of the data supplied to the DMC by the independent statistician and the supply of other data (such as details of SAEs) should also be included in the charter. It is advisable to involve the DMC members as early as possible in the review and construction of this document to gain clear buy-in to their role in the conduct of the trial.

#### **14.4.3 Meetings and recommendations**

The DMC should meet at a set of predefined time points during the trial, typically following the completion of a proportion of the patients. For example, four meetings could be organised following completion of 25%, 50% and 75% of the patients and finally at trial completion. If the trial is to involve interim analyses for efficacy and futility, then some of the meetings will revolve around those. TLFs should be supplied in conjunction with all meetings.

Meetings of the committee are usually organised in open and closed sessions. The open sessions, which also involve, for example, members of the sponsor company, cover general issues such as recruitment rates, timelines and the presentation of summary demographic and other baseline tables. During these open sessions, details can also be given on progress and results from other trials within the drug development programme and other trials and data from outside of the programme that potentially impact the current trial: for example, where they provide new critical information on the performance of the control treatment. The closed sessions involve only the members of the DMC plus the independent statistician supplying data to the DMC. It is important that the independent statistician attends to answer questions that inevitably arise with the data supplied, its format, conflicts across the different tables and so on. Minutes of both closed and open sessions of these meetings should be kept and the closed minutes stored securely and confidentially until trial completion. A good model for the closed minutes is for them to be written by the independent statistician and filed securely by them once reviewed and agreed upon by the DMC members. These minutes need not be overly extensive; they should simply capture the main points of discussion and the final decision. Electronic copies of the data sets on which interim analyses are based should also be retained in conjunction with the closed minutes.

Outside of the regular planned meetings, details of all SAEs can be supplied to the DMC in real time or at regular intervals, and the DMC members should

arrange to discuss these by email or teleconference as and when they feel necessary.

Recommendations to the sponsor coming out of the regular DMC meetings will be one of the following:

- Trial to continue unchanged
- Modification of the protocol to protect the safety of the trial participants
- Temporary pause to recruitment while other data are considered
- Termination of the trial

Clearly, if the recommendation is anything other than *continue unchanged*, additional information needs to be supplied to the sponsor to support the recommendation, taking care to protect blinding as much as possible. The recommendations are not binding on the sponsor, although, as mentioned earlier in conjunction with interim analyses for efficacy and futility, it would be very unusual to see such recommendations ignored or overturned. Alternatively, the sponsor may come back to the DMC following such recommendations with an alternative solution that in some cases could allay DMC concerns.

# CHAPTER 15

## Bayesian statistics

### 15.1 Introduction

The methods for statistical inference that we have discussed so far in this book fall under the heading of *classical* or *frequentist* methods. These methods are usually applied in the analysis of clinical trial and other data within the pharmaceutical industry. They are, however, not the only methods available; Bayesian methodology provides an alternative way of thinking about statistical inference. In this section, we will outline the Bayesian approach and contrast it with the frequentist way of thinking. We will then set down how Bayesian methods can be applied in the analysis of data and discuss implications for trial design. Finally, we will discuss the application of these methods within a regulatory context.

Frequentist methods centre on making treatment comparisons using *p*-values and estimating treatment effects through point estimates and confidence intervals (CIs) in relation to the value of unknown parameters. In comparing two treatment means  $\mu_1$  and  $\mu_2$ , for example, frequentist statisticians view these means as fixed unknown numerical values. They are the mean values you would see if you were able to treat the complete population of patients with drug A ( $\mu_1$ ) or drug B ( $\mu_2$ ). But Bayesian statisticians view these means differently: as variables about which we can make probability statements. As we will see in the next section, it is possible within this framework to develop methods akin to *p*-values and CIs with different and, many people would say, more straightforward interpretations.

The Bayesian approach also allows the incorporation of *beliefs* (called *prior beliefs*) about parameters based on knowledge (and opinions) and information from outside the clinical trial under consideration. This is one aspect of Bayesian statistics that causes concern among regulators, and we will address this point in subsequent sections.

## 15.2 Prior and posterior distributions

### 15.2.1 Prior beliefs

To develop the methodology underpinning the Bayesian approach, consider the following simple example. Suppose we are planning to run a single-arm phase II study to gain information about the cure rate of a particular new antibiotic; and suppose, based on our knowledge of the disease setting, the population under study, the preclinical performance of this agent and its therapeutic class, we feel that the cure rate will be around 40%. Assume that the probability distribution in Figure 15.1 captures our *prior beliefs* regarding the likely values for the cure rate. This distribution (distribution A) is just one possibility; we will discuss other possibilities later. The mean of this distribution = 0.4; we are suggesting that the cure rate will be around that value, but we are not ruling out that the cure rate could be as low as 0.1 or as high as 0.8, although these values are less likely based upon our (prior) knowledge before we run the study.

Figure 15.2 presents alternative prior distributions. Distribution B, which has a mean of 0.4, is less spread out than distribution A. Distribution C, which is the so-called uniform distribution (known as a *vague* prior distribution), gives equal prior probability to any value between 0 and 1.

### 15.2.2 Prior to posterior

Now let's suppose that our study recruits 20 patients, and at the 21-day test-of-cure visit, 10 of those patients are cured. The observed cure rate is 50%, which is a little higher than we expected based on our prior beliefs. The Bayesian analysis combines the observed data with the prior beliefs to produce *posterior beliefs*

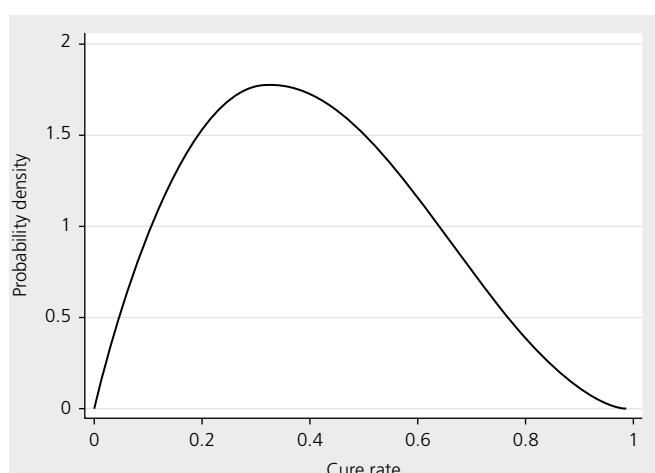
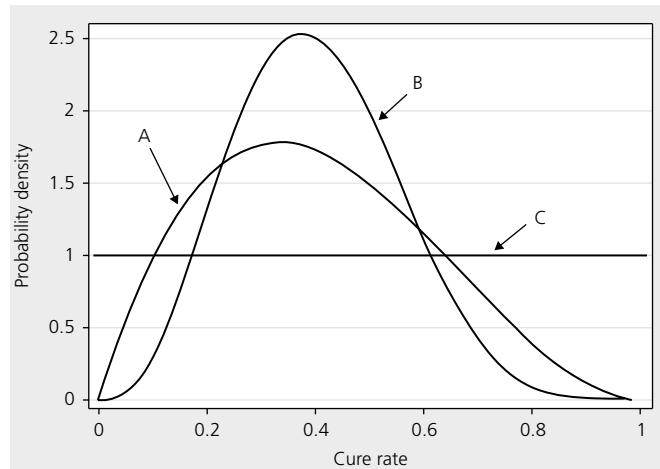
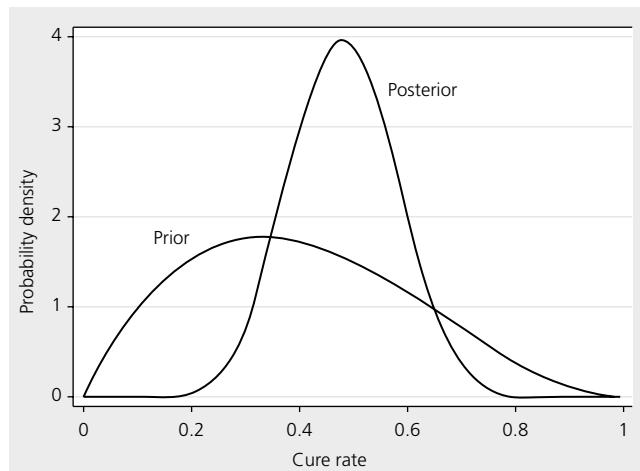


Figure 15.1 Prior distribution A for cure rate



**Figure 15.2** Prior distributions A, B and C

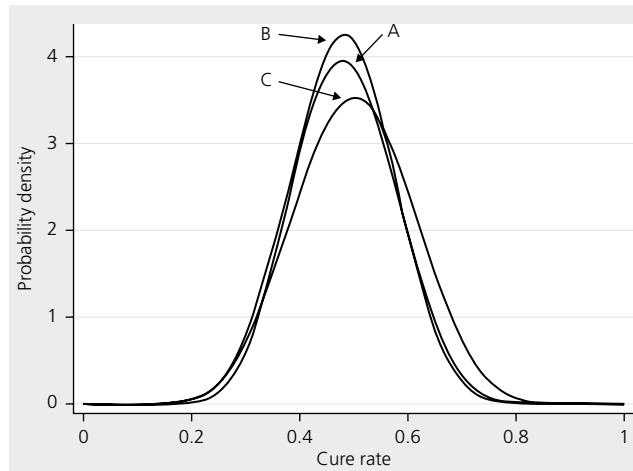


**Figure 15.3** Prior distribution A and corresponding posterior distribution

in the form of a second probability distribution, the *posterior distribution*. We will discuss later in this section how this combination is done.

Figure 15.3 shows the posterior distribution for these data together with our original *prior* distribution A. Note that our posterior beliefs are now centred at a value between 0.4 and 0.5; the observed data have influenced our prior beliefs and moved them in the direction of 0.5.

Figure 15.4 shows the posterior distributions A, B and C for the alternative prior distributions A, B and C, respectively. The data have strongly influenced our beliefs (as they should) about the cure rate. Indeed, the posterior beliefs have very much come together, and this is generally what happens; the data,



**Figure 15.4** Posterior distributions A, B and C

especially with larger sample sizes, ultimately dominate our views. The influence of our prior beliefs would have been greater had the sample size been smaller.

### 15.2.3 Bayes theorem

The term *Bayesian statistics* derives from the mathematical theorem known as *Bayes' theorem*, developed by Thomas Bayes (more later), which uses data to convert prior beliefs into posterior beliefs.

The simple form of Bayes' theorem (also called *Bayes' rule*) is as follows. If  $E_1$  and  $E_2$  denote any two events, then

$$pr(E_1|E_2) = \frac{pr(E_2|E_1) \times pr(E_1)}{pr(E_2)}$$

Here, for example,  $pr(E_1)$  denotes the probability of event  $E_1$ , while  $pr(E_1|E_2)$  denotes the probability of event  $E_1$  given that event  $E_2$  has already occurred, called the *conditional probability of  $E_1$  given  $E_2$* .

To take a simple example, consider throwing two perfectly balanced six-sided dice, and suppose that event  $E_1$  is getting a 6 on the first throw and  $E_2$  is getting a total score of at least 9. The outcomes that give a total score of at least 9 (event  $E_2$ ) are 3 6, 4 5, 4 6, 5 4, 5 5, 5 6, 6 3, 6 4, 6 5 and 6 6, and each of these is equally likely to occur when throwing two fair dice. Of these 10 outcomes, 4 involve getting a 6 on the first throw, so if you know you have a total of at least 9, the probability that you got a 6 on the first throw is  $pr(E_1|E_2) = \frac{4}{10}$ . Let's now evaluate the right-hand side of Bayes rule and see if this gives the same answer. The probability of getting a total of at least 9, given that the first throw gives a 6, is

$pr(E_2|E_1) = \frac{4}{6}$  since 4 of the 6 outcomes (6 1, 6 2, 6 3, 6 4, 6 5 and 6 6) that involve a 6 on the first throw result in a total of 9 or above. Further, the probability of a 6 on the first throw is  $pr(E_1) = \frac{1}{6}$  and the probability of a total of at least 9 on the two throws is  $pr(E_2) = \frac{10}{36}$ . The right-hand side of the Bayes rule is then

$$\frac{pr(E_2|E_1) \times pr(E_1)}{pr(E_2)} = \frac{4/6 \times 1/6}{10/36} = \frac{4}{10}$$

So, the equation works!

Further mathematical detail is beyond the scope of this book, but simply speaking, if  $pr(\theta)$  denotes the equation for the prior distribution and  $pr(\theta|y)$  denotes the equation for the posterior distribution where  $y$  are the data (10/20 patients cured in the example in Section 15.2.2), then Bayes rule tells us that

$$pr(\theta|y = 10) = \frac{pr(y = 10|\theta) \times pr(\theta)}{pr(y = 10)}$$

This equation allows us then to calculate the posterior distribution  $pr(\theta|y = 10)$  based on the prior distribution  $pr(\theta)$  and probabilities associated with the observed data,  $y = 10$ .

## 15.3 Bayesian inference

### 15.3.1 Frequentist methods

The statistical methods we have used to make inferences within the frequentist framework are  $p$ -values and CIs. Let's recap the way these quantities are defined in the context of comparing two treatments in a superiority study.

The  $p$ -value is the probability associated with the observed difference (or a larger difference) when the null hypothesis is true. The precise nature of this  $p$ -value was introduced in Chapter 3 and discussed in detail in Section 3.3.1. For many who work in the field of clinical research, this definition is not easy to grasp, and the  $p$ -value is often misinterpreted as the probability that the null hypothesis is true. In the next section, we will see how the Bayesian equivalent to the  $p$ -value can be interpreted this way.

A 95% CI for the difference in the treatment means/proportions contains the true difference on 95% of occasions when the trial is repeated. So, on any single occasion, we can be 95% confident that the true difference is within the range given by the CI. See Section 3.1.1 for further discussion. The corresponding

quantity coming out of the Bayesian framework is a 95% *credible interval*, which has a much simpler and, some would say, more intuitive and useful interpretation.

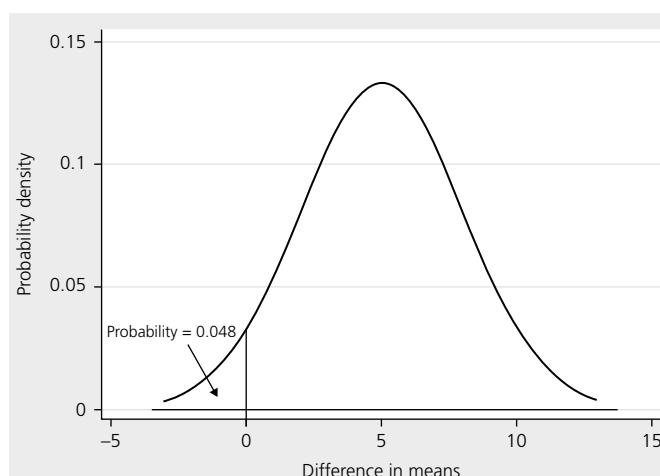
### 15.3.2 Posterior probabilities

Bayesian methods allow us to calculate probabilities that are of interest to us based on the posterior distribution. For example, when evaluating an experimental treatment for blood pressure lowering in a placebo-controlled trial, there may be interest in calculating the probability that the mean reduction in blood pressure in the active group is less than or equal to the mean reduction in the placebo group: in other words, that the active drug is ineffective. If  $\mu_1$  and  $\mu_2$  denote the mean reductions in the active and placebo groups, respectively, then we are looking to obtain

$$pr(\mu_1 \leq \mu_2) = pr(\mu_1 - \mu_2 \leq 0)$$

Figure 15.5 presents a typical posterior distribution for  $(\mu_1 - \mu_2)$  in this setting; the marked area is this probability.

This probability (0.048 in the example) is akin to a (one-sided) *p*-value within the frequentist framework. If this is small, we are thinking in terms of rejecting the null hypothesis that the active treatment is ineffective. We could, of course, calculate this the opposite way: the probability that the new treatment



**Figure 15.5** Posterior distribution for the difference in means and probability that the drug is ineffective

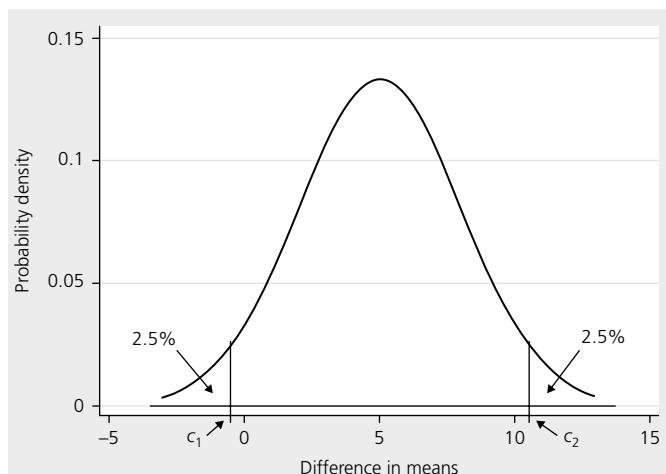
works,  $pr(\mu_1 - \mu_2 > 0) = 1 - 0.048 = 0.952$ . We could also obtain other probabilities of interest from this posterior distribution. Let's suppose that 4 mmHg is viewed as being a difference in the means that is of clinical importance. We can calculate  $pr(\mu_1 - \mu_2 \geq 4)$  to address this issue, the probability that the new treatment has efficacy of clinical relevance. For this example, this probability is 0.63 (or 63%).

### 15.3.3 Credible intervals

Credible intervals are the Bayesian equivalent of CIs and can be produced directly from the posterior distribution. For example, if we want a 95% *credible interval*, we calculate the lower end of that interval as the value  $c_1$  on the  $x$ -axis that cuts off the lowest 2.5% probability. Figure 15.6 shows this graphically. Similarly, the upper end of this interval is the value  $c_2$  on the  $x$ -axis that cuts off the upper 2.5% probability. The 95% credible interval is then  $(c_1, c_2) = (-0.9, 10.9)$  in our example. The interpretation of this 95% credible interval is very straightforward; we can say there is a 95% probability that the difference in the means  $\mu_1 - \mu_2$  lies between -0.9 and +10.9: that is, up to 0.9 mmHG in favour of placebo through to 10.9 mmHG in favour of active.

It is straightforward to extend this to other coverage probabilities: for example, 99%, where we would cut off the lower and upper 0.5% probabilities.

Wijeyesundara et al. (2008) provide an extended discussion of the interplay between frequentist and Bayesian methodologies.



**Figure 15.6** Posterior distribution for difference in means and 95% credible interval ( $c_1, c_2$ )

## 15.4 Case study

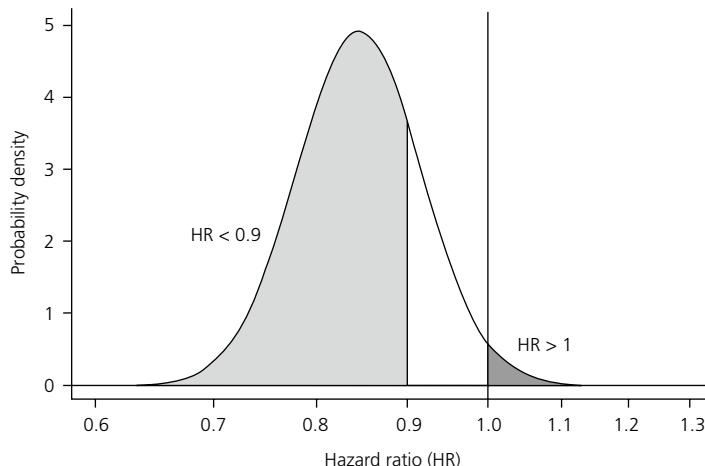
In this section, we will consider a specific example and contrast the frequentist and Bayesian approaches to the analysis of data and the associated interpretations. Wijeyasundera et al. (2008) report a clinical study evaluating the effectiveness of paclitaxel in addition to platinum-based chemotherapy as second-line treatment in women with ovarian cancer originally reported by the ICON and AGO Collaborators (Palmer et al., 2003). The trial was designed to detect an improvement in overall survival in the paclitaxel group with a hazard ratio of 0.71.

In the classical frequentist analysis reported in the trial publication, the hazard ratio was 0.82 with a 95% CI of (0.69, 0.97) and a statistically significant one-sided  $p$ -value of 0.012. The CI tells us that the true hazard ratio is, with 95% confidence, somewhere in the range 0.69 to 0.97. The one-sided  $p$ -value tells us that the probability of seeing a hazard ratio (in favour of paclitaxel) at least as low as 0.82 by chance is only 1.2% and it is on this basis that a statistically significant difference is declared.

Wijeyasundera et al. (2008) reanalysed these data from a Bayesian point of view. They assumed a normal distribution prior for the hazard ratio on the log scale with a mean of zero and a standard deviation equal to 0.21. This prior assumes that the most likely value for the hazard ratio is, *a priori*, equal to 1 (no treatment benefit) with only a 5% probability of a treatment benefit given by a hazard ratio of 0.71 or better. This therefore represents a very conservative prior distribution; we would call it a *sceptical prior*. The posterior distribution, again on the log hazard ratio scale, given the data, also turns out to be normal but with a mean of  $-0.17$  and a standard deviation of 0.081 (Figure 15.7).

We can now use this posterior distribution to calculate probabilities of interest. In particular, the probability that the hazard ratio is  $\geq 1$  (no benefit from paclitaxel) is equal to 0.019. This is the Bayesian equivalent of the classical  $p$ -value. However, its interpretation is much more straightforward; we can say that the probability of no benefit is 1.9%. Conversely, we can conclude that there is a 98.1% probability that paclitaxel treatment is beneficial in terms of improving overall survival. The 95% credible interval for the hazard ratio based on the posterior distribution is calculated to be (0.72, 0.99), so we can say with 95% probability that the true hazard ratio is within this range. Further probabilities can also be calculated. It may be that clinicians would only view this treatment as being beneficial clinically, possibly because of associated risks, if the hazard ratio were 0.90 or lower, for example. The Bayesian analysis allows us to calculate this probability from the posterior distribution, and its value is 0.78; there is a 78% probability that paclitaxel has a ‘clinically relevant’ beneficial effect in terms of overall survival.

Note that in one sense, the Bayesian and classical analyses are not too dissimilar. The classical  $p$ -value is 0.021, and the Bayesian posterior probability of no effect is 0.019. The 95% CI for the hazard ratio is (0.69, 0.97), the 95% Bayesian credible interval is (0.72, 0.99). In part, the assumption of a sceptical prior has given marginally less impressive results for the Bayesian analysis compared to the classical



**Figure 15.7** Estimation of the probabilities for specified treatment effects using a Bayesian posterior distribution. The posterior distribution is scaled such that the area under the curve is 1. Probabilities for specified treatment effects are determined using the area under the curve. For example, the probability of hazard ratio  $<0.90$  is the area under the curve to the left of 0.90. Similarly, the probability of no benefit, namely a hazard ratio  $>1$ , is the area under the curve to the right of 1. Source: Wijeyesundara DN, Austin PC, et al. (2009) ‘Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials’ *Journal of Clinical Epidemiology*, **62**, 13–21. Reproduced by permission of Elsevier

analysis. But that is not really the point. The main distinction between the two approaches is in the interpretation. The Bayesian approach has given quantities that are much easier to interpret; we can present and discuss probabilities of interest. This kind of interpretation is not available to us under the classical approach.

## 15.5 History and regulatory acceptance

Thomas Bayes (1702–1761) first set down the basis for what we now know as Bayes’ theorem, but Pierre-Simon Laplace (1749–1827) developed the ideas and applied them to problems in medical statistics. So, these methods have been around for an awfully long time! Their widespread application, however, has until very recently been constrained by the practicalities of doing the calculations. Until around 15 years ago, strict mathematical assumptions regarding the prior distribution were needed before problems could be solved. But in recent years, the development of the so-called Markov Chain Monte Carlo (MCMC) methods and computer algorithms that apply these methods has enabled tremendous flexibility in the range of problems that can be addressed within the Bayesian framework; mathematical constraints are no longer an issue. WinBUGS (<https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs>) is a well-established computer package that applies the Bayesian methodology to the design and analysis of clinical trials.

How do regulatory authorities view Bayesian methods?

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Because the predominant approaches to the design and analysis of clinical trials have been based on frequentist statistical methods, the guidance largely refers to the use of frequentist methods when discussing hypothesis testing and/or confidence intervals. This should not be taken to imply that other approaches are not appropriate: the use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust'.*

The choice of the prior distribution is, by definition, based on prior beliefs, which is the basis of one concern over the use of the Bayesian methods; those beliefs potentially have an element of subjectivity. The rationale for the choice for the prior distribution should be set down and, in line with general regulatory thinking, must be pre-specified before the experiment. Further, sensitivity analyses based on other choices for the prior distribution should essentially lead to the same results and conclusions.

In some cases, prior distributions are based entirely on data, perhaps from a previous trial or some other alternative source external to the trial. When the prior distribution is obtained in this way and does not involve any elements of subjectivity, we talk in terms of *empirical Bayes methods*. This overcomes to a certain extent the issues associated with subjectivity, although the choice of the data on which to base the calculation of the prior is still subjective. Nonetheless, following the empirical Bayes pathway can alleviate some concerns.

Another area of potential concern to regulators is when Bayesian methods are applied to the sequential (interim) analysis of data. The Bayesian approach in this setting takes a prior distribution at the start of the study and updates this to produce a posterior distribution based on the prior and the data up to that first interim. This posterior distribution then becomes the prior distribution following the first interim analysis, which is itself updated at the second interim analysis using the new data collected between the two interims to produce a second posterior distribution. This again then becomes the prior at that point in time, and so on, until the end of the study. In general, this methodology pays no price for many looks at the data; there are no type I errors in Bayesian statistics. However, there are hybrid approaches in this setting, while using Bayesian methods to convert priors into posteriors, also pays a price for the multiple looks associated with the interim analyses. The interested reader is referred to Spiegelhalter, Abrams and Myles (2004), Section 6.6.5, for further discussion.

There is one further area where Bayesian methods have gained greater acceptance: in medical devices. The FDA has produced a guideline (FDA, 2010) entitled *Guideline on the Use of Bayesian Statistics in Medical Device Clinical Trials*. The document points out that Bayesian statistics are generally acceptable in these kinds of trials.

**FDA (2010): 'Guideline on the use of Bayesian Statistics in Medical Device Clinical Trials'**

*'When good prior information on clinical use of a device exists, the Bayesian approach may enable this information to be incorporated into the statistical analysis of a trial. In some circumstances, the prior information for a device may be a justification for a smaller-sized or shorter-duration pivotal trial. Good prior information is often available for medical devices because of their mechanism of action and evolutionary development. The mechanism of action of medical devices is typically physical. As a result, device effects are typically local, not systemic. Local effects can sometimes be predictable from prior information on the previous generations of a device when modifications to the device are minor'.*

## 15.6 Discussion

There has been some resistance in the past to using Bayesian methods. This has been linked firstly to concerns about incorporating prior beliefs into the inferential process, especially when the methodology was constrained by the mathematical limitations of undertaking calculations of posterior probabilities and credible intervals and where, therefore, prior beliefs have been chosen based on mathematical convenience. Secondly, regulators and others have concerns about the Bayesian treatment of multiple testing where the pure Bayesian view does not incorporate the idea of a false positive. The false positive is essentially a frequentist concept built around the frequency of drawing an incorrect conclusion based on repeated looks at the data. In Bayesian statistics, the focus is on expressing information in terms of posterior distributions rather than on the control of false-positive rates.

The development of specialist statistical software has essentially eliminated the first of these concerns in that realistic priors can be specified, and the robustness of conclusions can be easily evaluated by choosing a range of different priors. The second concern can be addressed by using hybrid solutions that, on the one hand, do calculate Bayesian posterior probabilities and credible intervals but, on the other hand, incorporate multiplicity considerations.

In some circumstances, as we have mentioned, it is possible to base prior beliefs entirely on data from, for example, other trials that have already been conducted. Using such *empirical* priors can alleviate some of the regulatory concern regarding the basis for prior beliefs.

A major benefit of Bayesian methods relates to how results can be expressed and interpreted. It is true that many practitioners misinterpret *p*-values, no matter how hard we as statisticians try to explain them. Similar comments apply to CIs.

# CHAPTER 16

## Adaptive designs

### 16.1 What are adaptive designs?

#### 16.1.1 Advantages and drawbacks

There has been a tremendous amount of interest in the last 10 years or so, both from statisticians and non-statisticians, in so-called *adaptive* or *flexible* designs where certain aspects of the clinical trial design can be changed based on accumulating data from within the trial. These designs have the potential, at least in theory, to improve the efficiency of decision-making by reducing sample size, reducing costs, avoiding waste and shortening timelines. In practice, however, things are not quite so simple, and very careful thought should be given to the use of such designs. In particular, the validity and integrity of the trial must be preserved in terms of maintaining control of type I error and avoiding bias in the eventual treatment comparison. These aspects cannot be compromised; if they are, this can destroy the trial in the eyes of regulators and the scientific community in general. This is rather a negative statement to start with, but such cautionary words are justified. There has been a tendency to view adaptive designs as the Holy Grail with an ability to recover failing trials or to avoid careful planning at the design stage. The CHMP is clear on this point in their guidance on adaptive designs:

**CHMP (2007): ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’**

‘... adaptive designs should not be seen as a means to alleviate the burden of rigorous planning of clinical trials’.

Regulators make a clear distinction between exploratory and confirmatory trials, and adaptations in a confirmatory trial may destroy the ability of that trial to make confirmatory statements and claims.

**CHMP (2007): 'Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design'**

'Adaptations to confirmatory trials introduced without proper planning will render the trials to be considered exploratory'.

The FDA talk of the '*learn vs. confirm*' paradigm (Wang, Hung and O'Neill, 2011); the exploratory phase II trial is used for proof of concept, gaining information about the statistical aspects of endpoints of interest, identifying appropriate inclusion/exclusion criteria and choosing a single dose (or a limited number of doses) for phase III. Having gained this information, we then finalise our plans for the subsequent phase III studies; and in that confirmatory phase, we establish the efficacy of the drug and evaluate its risks. Allowing uncontrolled design adaptations within a phase III clinical trial clearly conflicts with the confirmatory nature of the trial.

Nonetheless, there are certain adaptations that, if carefully planned, do not undermine the ability of the trial to provide confirmatory conclusions, and it is these adaptations that we will explore in this chapter.

### 16.1.2 Restricted adaptations

We have already considered adaptations to the design of a phase III trial of two kinds. Firstly, in Section 9.5.3, we discussed the re-assessment of sample size based on a blinded evaluation of the variance of the primary endpoint or the overall event rate, for example. In general, such adaptations have at most only minor effects on the control of type I error, provided that blinding is maintained (Friede and Kieser, 2006), and regulators are relaxed about their use. This is of course in marked contrast to their position in relation to unblinded, comparative evaluations of data. The FDA in their guidance make these issues clear:

**FDA (2019): 'Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics'**

'In general, adequately prespecified adaptations based on non-comparative data have no effect or a limited effect on the Type I error probability'.

'... in contrast to adaptations based on non-comparative data, adaptations based on comparative data often directly increase the Type I error probability and induce bias in treatment effect estimates'.

Secondly, in Section 14.2, we considered group sequential designs and stopping for both futility and overwhelming efficacy. Various schemes were considered, using strict rules to preserve the type I error rate. Again, if these rules are followed, there should be no regulatory concerns.

We will refer to adaptations of these kinds as *restricted adaptations*. They are restricted in the sense that they only consider blinded data or do not provide any

flexibility to adapt the trial based on unblinded data except in relation to a go/no-go decision.

A further potential for adapting the design occurs when certain aspects of a trial are changed based on external data. For example, it may be that the trial was designed based on only rudimentary information on the variance of the primary endpoint, and additional information from external sources regarding the variance may become available as the trial is ongoing. If the information is that the variance used in the sample size calculation is too small, then the trial is underpowered, and a revised sample size calculation can be undertaken based on this new information for the variance and the sample size adjusted upwards. This does not compromise the study's validity, provided the blind is completely maintained. From a regulatory point of view, it is always better to have such considerations pre-planned. Possibly, we may know at the design stage that new information on the variance will become available during the conduct of the trial, and this reconsideration of sample size could then be pre-planned. Unfortunately, if this is not the case, there is always the suspicion from regulators that the sample size is being reconsidered based on some unblinding, and it is often difficult to dispel those concerns completely.

### **16.1.3 Flexible adaptations**

We will call the adaptations that build in additional amounts of flexibility *flexible adaptations*, and these will be the main focus of our consideration in this chapter. There are many different types of possible adaptation, but for our purposes in this book, we will focus primarily on two settings that are the most common.

Firstly, we will consider the re-assessment of sample size at an interim time point based on unblinded data, where we compare the treatments, and resize the trial based on that interim result.

Secondly, there is the seamless phase II/III trial, where we start by considering several dose levels and placebo in the dose-finding phase II part and where the plan is to take a reduced number of dose levels (often one) through to the phase III part. At the end of the study, data from the phase II and phase III parts are used to evaluate statistical significance as a basis for a confirmatory claim.

## **16.2 Minimising bias**

### **16.2.1 Control of type I error**

As we have discussed at various points in this book, it is a requirement in general in a clinical trial to control the potential for the two-sided type I error at 0.05. It will be better in this chapter to consider one-sided tests in which we control the type I error rate at 0.025. The rules are unchanged when we look at non-inferiority studies (Section 12.4), where again, the type I error – falsely concluding non-inferiority, when in truth the new treatment is inferior – is controlled at

0.025. The requirement to control the type I error for an adaptive design is one essential element that underpins the validity of the adaptation.

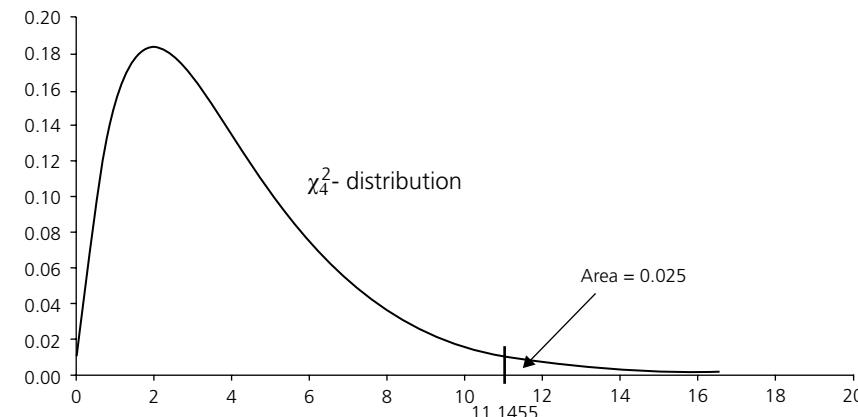
**CHMP (2007): '*Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*'**

*'A minimal prerequisite for statistical methods to be accepted in the regulatory setting is the control of the pre-specified type I error . . .'*

Consider a simple example. Suppose, in a blood pressure lowering placebo-controlled trial, we are looking to detect a difference of 4 mmHg in the mean change from baseline between the two groups. Assuming a standard deviation of 8 mmHg, a sample size of 86 per group will give 90% power to detect such a difference. In the planned interim analysis, after 43 patients per group have provided data on the primary endpoint, a difference of only 3 mmHg is observed. Is it valid to increase the sample size to 151, which gives a power of 90% to detect a difference of 3 mmHg, continue to the end of the study and then use a conventional analysis and *p*-value calculation to compare the groups? The short answer is no – it is not. It has been shown (see, e.g. Proschan, Lan and Wittes [2006], Section 11.4) that such a procedure potentially more than doubles the type I error rate – you might say we are *chasing* a significant result!

However, there is a statistical solution to this problem that does preserve the type I error. This is associated with calculating one-sided *p*-values ( $p_1$  and  $p_2$ , respectively) from the two parts of the trial, before and after the adaptation, and combining them in a particular way. There are several different ways of combining; we will primarily use a method derived from Fisher's combination test (Bauer and Köhne, 1994), which involves simply multiplying the *p*-values together. If the product  $p_1 \times p_2 \leq 0.0038$ , we have statistical significance overall at the one-sided 0.025 level. The result on which this test is based is that under the null hypothesis, either in a superiority trial or in a non-inferiority trial,  $-2\ln(p_1 p_2)$  follows a chi-square distribution on four degrees of freedom. Figure 16.1 illustrates the calculation with  $-2\ln(0.0038) = 11.1455$ .

Two examples of how Fisher's combination test works are given in Table 16.1. In the first example (row 1), the part 1 one-sided *p*-value is 0.041, while the part 2 one-sided *p*-value is 0.03. Fisher's combination test works by multiplying the two *p*-values to give 0.00123. Since this value falls below 0.0038, we have statistical significance at the one-sided 2½% level for the full trial data. The level of statistical significance is obtained by calculating  $-2\ln(p_1 p_2) = 13.4015$ . According to the chi-square distribution on four degrees of freedom, 13.4015 cuts off a proportion of 0.009 of the right-hand tail of this distribution. This is the reported level of statistical significance, the 'real' *p*-value. In the second example, the part 1 one-sided *p*-value is also 0.041, but now the part 2 one-sided *p*-value is 0.15, which is not nearly as strong in support of rejecting the null hypothesis. The product of the *p*-values is



**Figure 16.1** Chi-square distribution on 4 degrees of freedom. (Note: The 0.025 area in the right-hand tail is cut off by 11.1455.)

**Table 16.1** Examples of Fisher's combination test

One-sided $p$ -value from part 1, $p_1$	One-sided $p$ -value from part 2, $p_2$	$p_1 \times p_2$	Product of $p$ -values $\leq 0.0038?$	Overall one-sided $p$ -value from two parts combined
0.041	0.03	0.00123	Yes	0.009
0.041	0.15	0.00615	No	0.037

0.00615, which is not less than 0.0038, and overall statistical significance has not been achieved. The  $p$ -value for reporting the level of statistical significance for the complete trial is 0.037 since  $-2\ln p_1 p_2 = 10.1826$  cuts off an area of 0.037 in the right-hand tail of the chi-square distribution on four degrees of freedom.

A practical application of the procedure is given in Example 16.1 in the context of a non-inferiority study.

**Example 16.1** Acute treatment of moderate to severe depression with hypericum extract WS 5570

Szegedi et al. (2005) report a randomised, double-blind, non-inferiority trial of hypericum extract WS 5570 vs. paroxetine as the active control treatment. The primary endpoint was the change from baseline in the 17-item Hamilton Depression Scale (17-HAMD). The pre-specified non-inferiority margin was set at 2.5 points on the 17-HAMD for the difference in the means. A total of 100 patients were recruited into part 1, which under the assumption of a true treatment difference in the means of zero and a standard deviation for the primary endpoint of 6 points gave 90% power to achieve a one-sided significance level of 0.025. The data analysed at that point gave a  $p$ -value for non-inferiority of 0.084. As set down earlier, to achieve statistical significance at the end of the trial using Fisher's combination test, the product of the part 1 and

part 2 *p*-values for non-inferiority,  $p_1, p_2$ , needed to be  $\leq 0.0038$ , and for this to be achieved, the requirement was that  $p_2$  should be  $\leq 0.045$ . A total of 150 patients recruited into stage 2 gave 80% power, assuming a true difference between the treatments of zero points, to demonstrate non-inferiority with this adjusted significance level of 0.045. The data at the end of the study gave statistical significance for non-inferiority with  $p_1, p_2 \leq 0.0038$  and a positive confirmatory conclusion that hypericum extract WS 5570 is non-inferior to paroxetine.

An alternative method for combining the *p*-values from the two parts of the study is best expressed in terms of combining the corresponding *z*-scores. Recall from Section 3.3.3 that the *z*-score, which is equal to the signal/standard error ratio, leads directly to the *p*-value. The *inverse normal method* (Lehmacher and Wassmer, 1999) calculates an overall *z*-score as an *average* of the *z*-scores ( $z_1$  and  $z_2$ , respectively) from each of the two parts as follows:

$$z = \frac{z_1 + z_2}{\sqrt{2}}$$

This combined *z*-score is then used to obtain the one-sided *p*-value for the complete trial with the conventional cut-off of 0.025 for statistical significance.

It is important to note that these *p*-value calculations are valid, irrespective of the adaptation made at the interim point providing the method used to obtain the *p*-values (and *z*-scores), Fishers combination test or the inverse normal method, is pre-specified.

### 16.2.2 Estimation

In addition to using a correct method to calculate the *p*-value in an adaptive design, there is also a need to estimate the magnitude of the treatment difference and provide an associated confidence interval (CI). In Section 14.2.4, we briefly discussed that calculating point estimates and CIs in trials with interim analyses for efficacy is not straightforward. Similar difficulties arise in trials with an adaptive design, and special methods are needed to provide these quantities, especially if these involve interim stopping rules for efficacy and possibly futility.

**CHMP (2007): 'Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design'**

'Corresponding methods to estimate the size of the treatment effect and to provide confidence intervals with pre-specified coverage probability are required in addition to the presentation of the *P*-value'.

There is a considerable amount of ongoing theoretical research to develop good methods for controlling the properties of the estimation procedures: that is, bias

and correct coverage. These methods are beyond the scope of this book, but the interested reader is referred to the appendix of the paper by Bhatt and Mehta (2016) for some further discussion.

### 16.2.3 Operational bias

Controlling the type I error rate and providing correct estimation methods are required to produce valid confirmatory conclusions. However, these are not the only aspects of an adaptive design that must be carefully handled. A further issue that is equally important is the control of the dissemination of information regarding the data at the interim point on which an adaptation has been based and, indeed, the nature of that adaptation. The CHMP, in their guideline, talk about the importance of confidentiality of interim results. If the results of the interim analysis are revealed to personnel involved in the study, to investigators or even to patients, this could cause bias in various aspects of the study from that point on. Potential sources of bias include recruitment of different kinds of patients following a positive *trend* at the interim stage, specific changes in the administration of the intervention, endpoint assessment and so on. Of course, these concerns regarding dissemination and operational bias are not just confined to adaptive designs but also more widely to any clinical trial containing an interim unblinded look at the data. Regulators have the potential to ask for (or to undertake, in the case of the FDA) an analysis of data before and after an interim analysis to check for consistency. Any inconsistencies could seriously undermine their confidence in drawing clear conclusions.

**CHMP (2007): ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’**

*‘Studies with interim analyses where there are marked differences in estimated treatment effects between different study parts or stages will be difficult to interpret. It may be unclear whether the estimated treatment effects differ just by chance, as a consequence of the intentional or unintentional communication of interim results, or for other reasons. This problem can be even greater if the study design has been changed as a result of an interim analysis’.*

**Example 16.2** The 2NN study

This was an open-label, parallel-group, randomised trial in patients with chronic HIV-1 infection reported by van Leth et al. (2004). The primary endpoint was treatment failure, a composite endpoint based on virology, disease progression or therapy change. Initially, patients were randomised 1:1:1 to one of the following three groups:

- 1 N<sub>1</sub> – nevirapine (once daily)
- 2 E – efavirenz
- 3 N<sub>1</sub> + E – combination of nevirapine (once daily) and efavirenz

Five months into the trial (388 patients randomised), another study concluded that the effectiveness of nevirapine was related to the minimum concentration, so the 3NN trial was adapted to include a fourth arm: nevirapine twice daily ( $N_2$ ). Further, the randomisation ratio was changed so that randomisation to  $N_1$ , E,  $N_1 + E$  and  $N_2$  was in the ratio 1:2:1:2. The final sample size was 1216, and in the final analysis, data from before and after the adaptation were pooled. Is this appropriate and supported by the data? The adaptation was driven by external data, although note that the trial was not blinded. In the treatment groups used throughout the study, the failure rates and 95% CIs shown in Table 16.2 were seen.

The issue here is the lack of homogeneity in the results before and after the change. There was a 12.2% absolute difference between nevirapine (once daily) and efavirenz before the design change, and there was virtually no difference between these two arms after the change. No explanation of why this occurred was given in the publication. This lack of consistency would cause regulators major concerns.

**Table 16.2** Failure rates in the 2NN study and 95% CIs

	Before addition of fourth arm (%)	After addition of fourth arm (%)
$N_1$	46.6 (37.8, 55.5)	39.3 (29.1, 50.3)
E	34.4 (26.3, 43.2)	39.4 (33.5, 45.5)
$N_1 + E$	51.6 (42.5, 60.6)	55.4 (44.1, 66.3)

As a final point, it should be noted that statistical methods of themselves cannot correct for this operational bias; if it is present, its impact is unmeasurable.

## 16.3 Unblinded sample size re-estimation

### 16.3.1 Product of p-values

Re-estimating sample size based on unblinded data at an interim point has already been introduced through Example 16.1 in Section 16.2.1. In this setting, the trial data from the first part of the trial, the observed difference between the treatment groups and possibly the within-group standard deviations, together with the required significance level according to Fisher's combination test, can be used in the calculation of the final sample size for the study. In Example 16.1, the assumptions regarding the treatment difference and standard deviation were not changed for the sample size re-assessment. However, this could be a consideration in cases where the observed treatment difference at the interim point is smaller than that anticipated in the sample size calculation or if the observed standard deviation is much larger than initially assumed.

### 16.3.2 Weighting the two parts of the trial

The combination tests give equal weight to the two parts of the trial; the individual  $p$ -values,  $p_1$  and  $p_2$ , carry the same weight when multiplied in Fisher's combination test. Similarly, in the inverse normal method, the separate  $z$ -scores from the two parts of the trial are given equal weight in the test. This particular property of these tests for adaptive designs is one of the reasons why some commentators have concerns about the efficiency of such designs. In the usual (non-adaptive design) analysis of data, each patient is given equal weight. For example, if 50 patients had been randomised to group A before the adaptation and 100 patients randomised to group A after the adaptation, then those 150 patients would, ignoring the adaptation, contribute equally to the calculation of the mean for that group with  $\bar{x} = \frac{\Sigma x}{150}$  and contribute equally to the calculation of the  $p$ -value for the treatment comparison. In the analysis for the adaptive design, however, this does not happen: in this example, the 100 patients in group A recruited after the adaptation would collectively only be given the same weight as the 50 patients recruited before the adaptation. In short, in this example, a patient recruited after the adaptation is given only half the weight of a patient recruited before the adaptation.

This flies in the face of certain fundamental statistical principles for the analysis of data, and although it may seem a particularly technical point, it does have implications for the power of these adaptive designs to detect treatment differences (or to establish non-inferiority, if that is the objective). A further issue concerns the estimated treatment benefit and the associated CI. Should we simply use the conventional mean values and the difference between them, or should we incorporate the same weighting used in the  $p$ -value calculation? In Section 16.2.2, we argued that the former approach is biased. The FDA is clearly concerned about the extent of this bias:

**FDA (2019): 'Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics'**

'... a conventional end-of-trial treatment effect estimate such as a sample mean that does not take the adaptations into account would tend to overestimate the true population treatment effect. This is true not only for the primary endpoint which formed the basis of the adaptations, but also for secondary endpoints correlated with the primary endpoint. Furthermore, confidence intervals for the primary and secondary endpoints may not have correct coverage probabilities for the true treatment effects'.

There is a way around the issue of unequal weighting in the calculation of  $p$ -values: the interested reader is referred to the work of Mehta and Pocock (2011). However, this modified method, in which certain restrictions are applied to the choice of the final sample size, is not without controversy, as set down in a commentary on that paper by Emerson, Levin and Emerson (2011).

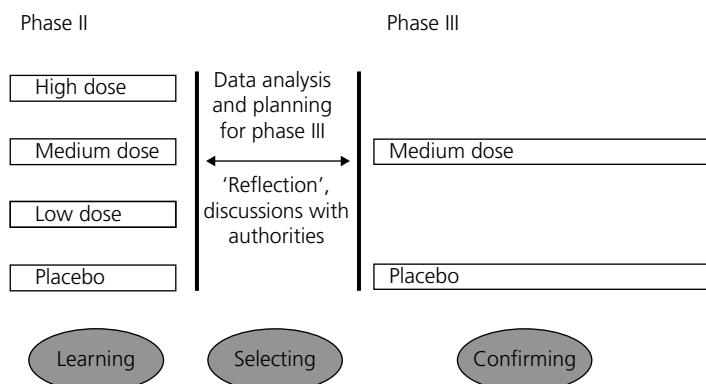
### 16.3.3 Rationale

Under what circumstances might one be interested and willing to increase the sample size at an interim stage in this way? There are two aspects. Firstly, one could imagine a situation where budget constraints result in a trial with only a modest sample size, perhaps with a level for the power less than ideal for a confirmatory study. Results at an interim look may be promising enough to support additional investment in the project and an increase in the sample size. Secondly, if indeed the observed treatment difference was lower than anticipated and one were increasing the sample size to give sufficient power to detect a smaller level of effect, it would be necessary to argue that the reduced level of treatment benefit was going to be worthwhile from a clinical relevance point of view.

## 16.4 Seamless phase II/III studies

### 16.4.1 Standard framework

The standard framework for a dose-finding phase II trial and a confirmatory phase III trial taking forward a single or maybe two doses from phase II is displayed in Figure 16.2. The white space between the two phases provides time for the analysis of data (both efficacy and safety), reflection, planning, discussions with authorities, etc., and decisions regarding how to move forward with phase III are built on those considerations as mentioned earlier according to the FDA '*learn-verses-confirm*' paradigm (Wang, Hung and O'Neill, 2011). The statistical analysis of the phase III study is undertaken separately from the analysis of the phase II study, and there is no sense in which the data from the two phases are formally combined.



**Figure 16.2** Traditional framework for phase II/III

A so-called *seamless* phase II/III study has two distinct aspects:

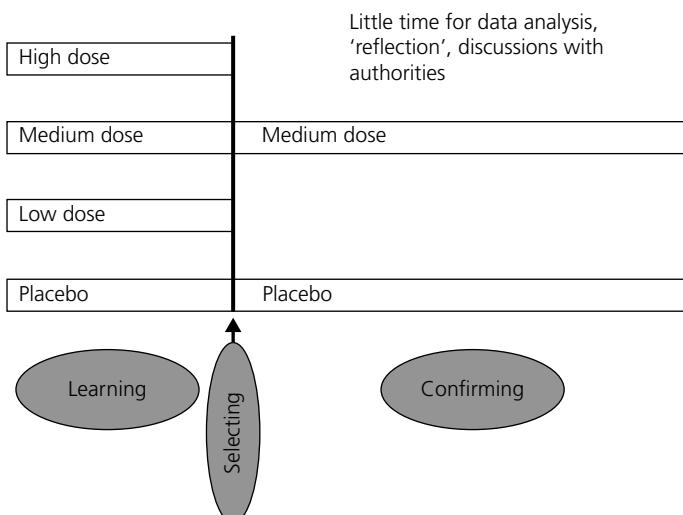
**1 Inferentially seamless:** This refers to the fact that the data from the phase II and phase III parts are combined as the basis for the final statistical analysis and inferences regarding the treatment being evaluated.

**2 Operationally seamless:** Here the gap between the phase II and phase III parts is eliminated, and recruitment into the trial continues under the same protocol.

We will primarily discuss the inferentially seamless adaptive design elements, although the operational aspects impact the statistical issues. Having the trial organised in an operationally seamless way in theory saves time but requires clear decision rules to be formulated in advance. Are we looking to reduce the three doses to two doses or just a single dose, or is that to be left open? Is that decision to be made purely on efficacy grounds, perhaps because we have no concerns about safety, or is it going to be a combination of efficacy and safety? And is there a possibility to increase the sample size depending on the phase II treatment differences vs. placebo? Usually, these kinds of considerations are precisely what fill the '*white space*' in the conventional approach. In an adaptive framework, the white space is lost (Figure 16.3) and algorithms covering all possible actions for the end of the phase II part need to be well defined for a seamless transition to work in practice. It is a tall order – some would say too tall! Nonetheless, there remains substantial interest in these kinds of designs.

### 16.4.2 Multiplicity

Before we discuss how the data are combined for the two parts of the seamless design, we will return to one aspect of our earlier multiplicity discussion in Section 10.4. In that section, we discussed various approaches for adjusting the



**Figure 16.3** Adaptive design for seamless phase II/III

significance level to account for multiplicity, namely the Bonferroni, Holm, and Hochberg procedures. In the case of three dose levels and placebo for the dose-finding phase II part, we could, for example, use a Bonferroni correction and compare each dose level to placebo based on a one-sided adjusted significance level equal to 0.0083 (= 0.025/3). In a combination test for the utilisation of data from both the phase II and phase III parts, we need to rescale the *p*-values from the phase II part so that they can be judged against a significance level of 0.025. For example, suppose the three one-sided *p*-values for the three dose levels low (L), medium (M) and high (H) vs. placebo are as follows:

$$p_{1L} = 0.09, p_{1M} = 0.01, p_{1H} = 0.04$$

Multiplicity implies that each of these needs to be judged against a threshold for statistical significance of 0.0083. Equivalently, they can be judged against a threshold of 0.025 by multiplying each of them by 3. These *multiplicity-adjusted p*-values are then

$$p^*_{1L} = 0.27, p^*_{1M} = 0.03, p^*_{1H} = 0.12$$

These *p*-values now have the potential to be used in the analysis of the combined phase II/phase III data.

### 16.4.3 Incorporating the phase II data

Based on the phase II outcomes for efficacy, together with the safety and tolerability profiles, suppose we decide to move forward into the phase III part of the trial with the medium dose only and placebo, perhaps with an increase in sample size in view of more information on the magnitude of treatment benefit at the medium dose level and measures of variability. Following on from the example in the previous section, suppose we see a *p*-value for the medium dose vs. placebo in the phase III part of  $p_{2M} = 0.04$ . We now use Fisher's combination test based on the product of the medium dose vs. placebo *p*-values from phase II and phase III:

$$p^*_{1M} \times p_{2M} = 0.03 \times 0.04 = 0.0012$$

Other combination tests could also be used, such as the inverse normal method. Note that the *p*-value for the medium dose vs. placebo comparison for the phase II data is the multiplicity-adjusted *p*-value.

This is a statistically significant result since the value of the product of the *p*-values is  $\leq 0.0038$ , the required threshold according to Fisher's combination test. The level of statistical significance for this outcome is 0.0093, obtained using the chi-square distribution on four degrees of freedom as seen in Section 16.2.1.

As previously mentioned, the Bonferroni correction is not the most efficient, and the Holm and Hochberg procedures are more sensitive to detecting treatment

differences. Other procedures also have advantages over Bonferroni. Any of these methods can be used to adjust the phase II *p*-values for multiplicity. Bonferroni was chosen for our example because of its simplicity. The alternative methods add complexity to the considerations and calculations, and the reader is referred to Posch et al. (2005) for a technical development in the seamless phase II/III setting.

#### **16.4.4 Logistical challenges**

Several logistical challenges need to be overcome if a seamless trial is to be successful, as discussed by Maca et al. (2006). Some of these problems are listed here, and the reader should refer to Maca et al., who suggest ways of dealing with them:

- The endpoints on which decisions relating to the adaptation are to be based need to be endpoints that become available in a short period of time. Otherwise, the trial will run on in terms of recruitment, and efficacy data in dosage groups that are not taken forward into part 2 will be wasted.
- Drug supply packaging could be difficult because the choice regarding the dose for part 2 will need to be made at a particular point in time with supplies then available immediately. Some drugs for dose levels not taken forward will likely be wasted.
- It is somewhat inevitable that the sponsor will need to be involved in the decision-making at the interim point; it is very difficult to give watertight algorithms to an independent data monitoring committee (IDMC) that cover all possible scenarios. This conflicts with the need to control the dissemination of interim results and potentially makes a timely decision more difficult.

### **16.5 Other types of adaptation**

#### **16.5.1 Changing the primary endpoint**

It is difficult to envisage a setting where changing the primary endpoint in an ongoing trial based on evidence from within the trial would be acceptable. The CHMP is particularly uncomfortable with this aspect.

***CHMP (2007): 'Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design'***

*'A change in the primary endpoint is difficult to justify: primary endpoints are chosen to describe a clinically relevant treatment effect. It is acknowledged that, when a certain primary endpoint is chosen at the planning stage, practicality and feasibility sometimes also play important roles. Once the study is ongoing it is difficult to imagine any situation where the perception of what constitutes a relevant clinical benefit should change based on interim results, especially as primary endpoints are usually not selected to differentiate between treatment and control group'.*

Just to reiterate, there is one aspect that is a common theme running through the statistical principles we are following in this chapter. Having chosen a more appropriate primary endpoint for part 2, it may seem acceptable to go back and calculate a standard *p*-value for this new primary endpoint for the complete trial, ignoring the adaptation or, indeed, accounting to some extent for the adaptation by using a combination test based on the separate one-sided *p*-values from part 1 and part 2 for the new primary endpoint. What is the problem with these approaches? It is all to do with cherry-picking. Presumably, the endpoint chosen to be the new primary at the adaptation point is selected because it shows promise in part 1 of the trial. Maybe the *p*-value associated with it is quite small, with a trend towards statistical significance – and there lies the problem. We have introduced bias into the analysis of the trial overall with a part 1 *p*-value chosen because it is small.

Other manifestations of this same problem with the primary endpoint include the following:

- Changing the order in a pre-specified hierarchy of endpoints
- Promoting a secondary endpoint to become primary
- Removing a component from, or adding a component to, a primary composite endpoint
- Defining a *responder* and using it as the basis for the primary endpoint rather than the pre-specified continuous endpoint

These considerations also apply to secondary endpoints if these are to be structured to provide confirmatory conclusions: for example, in a hierarchy. If there were to be a change in the primary endpoint at the interim stage, one valid option from a statistical point of view would be to combine the *p*-value for the original part 1 primary endpoint with the *p*-value for the part 2 revised primary endpoint in a combination test (Fisher's combination test or the inverse normal) to gain an overall test of significance. However, the issue here is in the interpretation. What hypothesis is being tested if the two parts of the study have different primary endpoints? CHMP provide a suitable comment.

**CHMP (2007): '*Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*'**

*'The mere rejection of a global hypothesis combining results from different endpoints will not be sufficient as proof of efficacy'.*

### **16.5.2 Enrichment**

In the age of targeted medicine, there is a lot of interest in developing treatments for subgroups of patients, defined by pathophysiological, genetic or biomarker characteristics, who are more likely to respond to a particular treatment. In part 1 of the study, a broad population of patients are enrolled through to an interim analysis with the potential to focus on a subgroup of the population following the interim defined according to a specific set of patient

characteristics. These characteristics need to be defined at the design stage of the trial: for example, in terms of a positive/negative value of a particular biomarker. The objective of the interim analysis is to identify whether the biomarker positive subgroup benefits differentially compared to the full population. If that is the case, only the biomarker-positive subgroup are enrolled for part 2. If not recruitment continues for the full population. As with our seamless phase II/III discussion, two aspects require consideration. Firstly, there is multiplicity in part 1, where we consider the full population and the subgroup. This same multiplicity may or may not be present in part 2 depending on whether, based on the interim analysis, we remain interested in the full population or just the subgroup. Secondly, the two parts of the trial need to be kept separated for the calculation of  $p$ -values, which will then be combined for final inference using Fisher's combination test, for example. Jenkins, Stone and Jennison (2010) provide an example in oncology that considers these aspects.

Detailed guidance on strategies for enrichment designs is given by the FDA (2012).

### **16.5.3 Dropping the placebo arm in a non-inferiority trial**

We discussed in Section 12.5 and elsewhere in Chapter 12 that including a placebo arm in a non-inferiority study can be of value in terms of providing a basis to evaluate assay sensitivity. Expected differences between each active treatment and placebo may be such that relatively few patients are needed for those comparisons, with more patients needed to establish the non-inferiority of the experimental treatment compared to the active control. In an adaptive design, there is an opportunity to terminate recruitment to the placebo arm at the interim stage and, from that point on, recruit only into the two active groups.

In principle, this should not cause problems provided the trial is planned carefully, although there are risks attached. The comparisons of each active arm to placebo need to be focused on the part 1 data only where those comparisons are based on simultaneous randomisation into the two groups; it would not be correct to use all patients recruited into the active groups from both parts of the trial for comparison with placebo. The non-inferiority evaluation, though, uses all data from the active groups from the complete trial. The risk here relates to the nature of the patients recruited.

#### ***CHMP (2007): 'Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design'***

*'It is well known that clinical trials do not recruit random samples of potential patients. It might be that different types of patients would be recruited into a two-arm trial comparing an experimental treatment with placebo, a three-arm trial including experimental, reference and placebo arms, and a two-arm, non-inferiority, trial comparing the experimental with the reference treatment (e.g. placebo-controlled trials*

*may include a patient population with less severe disease compared to a trial including active treatments only). The treatment effect may then differ to an extent that may make the combination of results from different stages impossible where the placebo arm has been stopped’.*

To prevent such issues from an operational point of view could complicate the running of the trial with concealment of the decision taken at the interim. As the CHMP go on to say, an alternative and maybe better approach from the outset is to run a standard study with unbalanced randomisation: for example, 2:2:1, with fewer patients randomised to placebo.

## **16.6 Further regulatory considerations**

In some therapeutic settings, it is acceptable to make a regulatory application based on a single, phase III confirmatory study. In this situation, data from a phase II study can often provide some element of independent confirmation of efficacy; having a combined phase II/phase III study would compromise that independence. Regulators recommend that adaptive designs not be considered in such cases.

***CHMP (2007): ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’***

*‘ . . . a major prerequisite for an application with one pivotal trial in phase III has always been that a sufficient body of evidence from phase II is already available so that phase III can be limited to simply replicating these findings in an independent setting’.*

An exception, however, is in the study of orphan diseases, where the availability of patients is limited. Logistically, in such situations, it may be difficult to run an extensive phase II programme followed by an even more extensive phase III. Regulators accept these difficulties and the use of adaptive designs in these cases.

***CHMP (2007): ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’***

*‘Investigation of drugs for the treatment of orphan diseases is a difficult task and specific requirements apply. A single phase II/phase III combination trial may be justified if such an approach is more efficient to display the totality of available information that can be derived from a limited number of patients’.*

The CHMP makes a more general point somewhat linked to this and other difficult experimental situations where the use of adaptive designs meets with less regulatory resistance:

**CHMP (2007): ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’**

*‘Instead, adaptive designs would be best utilized as a tool for planning clinical trials in areas where it is necessary to cope with difficult experimental situations’.*

We have covered general aspects of adaptive designs in this chapter. When using such designs, it is important to proceed with caution. An ill-judged adaptive design can in fact slow down the development process and lead to increases in costs when the trial fails to provide confirmatory evidence as a consequence of its adaptive structure. Some would argue that standard group sequential designs coupled with tight control of multiplicity, without the accompanying risks of ending up with a trial that does not have a robust interpretation, are the better option.

# CHAPTER 17

## Observational studies

### 17.1 Introduction

#### 17.1.1 Non-randomised comparisons

It is widely accepted that the randomised controlled trial has the potential to provide the best evidence for the effectiveness of a new treatment. However, in some special circumstances, conducting a randomised trial is problematic, and an external control group provides the only possibility to evaluate efficacy and safety. But such non-randomised studies are susceptible to a range of different biases, and in this chapter, we will explore the nature of those biases and how to address them from a statistical point of view. It should be pointed out at a very early stage, though, that if a randomised controlled study is possible, ethically and logically, this is the design that should be used. Only if such a design is not possible should the alternative of a non-randomised study be considered.

The focus here is primarily, although not exclusively, on confirmatory phase III and late phase II studies. In early phase II, it is perfectly possible to run studies without a control group: – for example, by simply looking at change from baseline on active treatment only – to give an early impression of the potential for treatment efficacy. However, the findings coming out of such studies cannot be confirmatory for efficacy but may provide information based on which considered choices can be made to guide the development programme in relation to endpoints, inclusion criteria and other aspects of trial design.

Observational studies are also used in a post-authorisation setting where there is a need to compare treatments that have not been compared directly in randomised trials based on data collected through, for example, registries and insurance claims and other databases. In recent years, for instance, there has been a lot of activity in comparing various non-vitamin K antagonist oral anti-coagulants (NOACs) in the treatment of atrial fibrillation based on observational data (see, e.g. Noseworthy et al. [2016]).

### 17.1.2 Study types

In evaluating an experimental intervention in the non-randomised setting, the control group may be chosen either historically, based on a group of subjects who have been treated retrospectively with an alternative intervention, or concurrently, where some subjects receive the experimental intervention, while the remaining subjects receive the control intervention. An example of a historically controlled study is the treatment of severe hepatic veno-occlusive disease described by Richardson et al. (2012). These authors pointed out that although using historical controls was not ideal, the absence of any other effective medication in this life-threatening condition limited the options for a randomised study. Using a historical control group was viewed as the best approach, both ethically and practically. A total of 102 patients were assigned to receive the experimental treatment, defibrotide, while a medical review committee was tasked with selecting a control group with baseline characteristics in line with the inclusion/exclusion criteria set down for recruitment into the defibrotide treatment programme. Defibrotide was designated an orphan medicine by the EMA in 2004; this study was part of the evidence for efficacy and safety submitted to the EMA, which recommended authorisation for defibrotide in 2013.

As an example of a concurrent control group, Bellingan et al. (2013) report a phase II study evaluating the efficacy and safety of intravenous interferon-beta-1a in relation to respiratory distress syndrome mortality. The treated group consisted of 37 patients, while the non-randomised concurrent control group contained 59 patients who entered the study but did not receive treatment either because of delays in obtaining informed consent or because of recruitment during 14-day safety windows where recruitment was stopped for the group receiving intravenous interferon-beta-1a for ongoing safety evaluation.

Deeks et al. (2003) provide a taxonomy of non-randomised studies. Two of these, labelled *concurrent cohort study* and *historical cohort study*, are the two most common non-randomised designs used in the pharmaceutical industry. Other designs can also be useful under certain circumstances. A *case-control study*, discussed later in this chapter, can provide information on, for example, the safety of a medicine following authorisation. Rosiglitazone was approved in the EU in 2000 and in the United States in 1999 for the treatment of type II diabetes mellitus, but marketing authorisation was withdrawn in the EU in 2010 based on the increasing risk of cardiovascular outcomes. The evidence for this increased risk came from a range of observational studies, including several case-control studies.

Another type of non-randomised study is the *cross-sectional study*. In such a study, a large cohort of subjects is identified, and at a specific point in time, subjects are asked to provide information on numerous risk factors and health outcomes. Cross-sectional studies provide a basis for assessing association but cannot provide information on causation. For example, suppose that data from such a study showed an association between drinking milk and peptic ulcer. Can we

conclude that drinking milk causes peptic ulcer? No, we cannot. This association could be the result of other relationships. For example, it could be that peptic ulcer sufferers drink milk to relieve their symptoms; or it could be that drinking milk is more likely among farm workers who have greater exposure to pesticides, and it is the pesticides that cause peptic ulcer. There are a host of possible explanations for this association, and the cross-sectional study cannot distinguish between them. This type of study is little used in the pharmaceutical industry as it cannot provide information on how treatment affects outcome, although in principle it can be the basis for generating hypotheses.

The ICH E2E (2005) *Note for Guidance on Planning Pharmacovigilance Activities* lists three study types – cohort (both retrospective and prospective), case–control and cross-sectional – as potential designs for use in pharmacovigilance.

### 17.1.3 Sources of bias

The major problem with non-randomised, externally controlled trials is bias.

#### ***ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'***

*'Inability to control bias is the major and well-recognised limitation of externally controlled trials and is sufficient in many cases to make the design unsuitable. . . . Blinding and randomization are not available to minimize bias when external controls are used'.*

The main sources of bias in such studies (Deeks et al., 2003) are as follows:

- Selection bias
- Attrition bias
- Detection bias
- Performance bias

*Selection bias* is caused by systematic differences in the baseline characteristics of the two groups being compared. In a randomised trial, the randomisation itself takes care of this bias: any baseline imbalances under those circumstances are simply due to chance. Not so in the non-randomised case, where such differences can (and do) frequently undermine the ability of the study to reach a valid conclusion. Note that some authors use the term *selection bias* to refer to the biased selection of the sample of subjects to be included in the study from the population defined by the inclusion/exclusion criteria. Here, we are using the term only to refer to the biased allocation of subjects to the treatment groups.

*Attrition bias* occurs as a result of dropout. This type of bias can also occur in randomised trials but is likely to be less of an issue than in a non-randomised study, where the follow-up information on subjects in the control group could well be collected with less rigour than would be the case in a randomised trial (especially when the control group is constructed retrospectively).

*Detection bias* is a consequence of the lack of standardisation in assessing outcomes between the treatment groups. In a blinded randomised trial, outcome assessment is, by definition, standardised. In an open-label randomised study, however, there is the potential for a greater amount of detection bias, which should be controlled by protocol. In a non-randomised study, such bias is likely to be even greater.

Finally, *performance bias* occurs if there is a lack of consistency in the application and recording of the interventions. In a randomised study, this should not happen under a tight protocol in either blinded or open-label settings.

To have a successful non-randomised comparison, it is essential to address each of these sources of bias and design the study so that their potential impact is minimised. It is also important, especially with selection bias, to employ statistical methods that help account for the sources of bias. We will discuss this issue further in subsequent sections.

### **17.1.4 An empirical investigation**

Deeks et al. (2003) undertook an exercise to investigate the impact of selection bias in using both concurrent and historical controls. These investigations were based on two large randomised multi-centre trials. The first trial was the International Stroke (IS) trial, which looked at the use of aspirin and heparin in the treatment of ischemic stroke. Patients were randomised to one of four treatment groups: placebo, heparin alone, aspirin alone and heparin plus aspirin. The efficacy outcome considered by these authors was dead or dependent at six months; for the purposes of their investigations, they combined across the heparin groups and compared those patients taking aspirin (with or without heparin), considered as the experimental group, with those not taking aspirin (with or without heparin), considered as the control group. The second trial was the European Carotid Surgery (ECS) trial, which evaluated the benefits of carotid endarterectomy in relation to stroke prevention; the primary outcome for the ECS trial was death or major stroke.

For the IS trial, these authors undertook a resampling (simulation) exercise to mimic the conditions that would be present firstly in a concurrent cohort study and secondly in a historical cohort study. In addition, they undertook corresponding resampling exercises to mimic conventional randomised trials. This trial recruited almost 20,000 patients in 467 centres in 36 countries. Fourteen geographical regions were constructed as a basis for the resampling, and the data were also split according to the date of recruitment to give *early recruits* (recruited up to and including 15 January 1995) and *late recruits* (recruited after 15 January 1995).

A concurrent cohort study was then constructed by sampling data from the experimental group in one region and comparing it with data sampled from the control group in another region, with regions chosen at random. A corresponding randomised controlled trial was constructed for comparison purposes by sampling data from the experimental group in one region and comparing with data sampled from the control group in that same region.

Similarly, a historically controlled trial was constructed by comparing data sampled from late recruiters in the experimental group in a particular region with data sampled from early recruiters in the control group in the same region. A corresponding randomised controlled trial was constructed for comparison by sampling data from late recruiters in the treated group in one region and comparing with data sampled from late recruiters in the control group in that same region.

For each treatment group, samples of size 100 were used (giving a study sample size of 200), and the sampling for the concurrent control and corresponding randomised control and for the historical control and corresponding randomised control settings was repeated 1000 times for each of the 14 regions. In each of the four settings and for each region, this gave 14,000 ( $= 14 \times 1000$ ) studies. The odds ratio (OR) was calculated in each case, aspirin vs. no aspirin, and the distributions of these ORs were analysed.

The procedure for the ECS trial was similar, but as this was a smaller trial, only 8 regions were chosen (rather than 14) with group samples of size 40 (rather than 100). For further details of the setup for these and other evaluations, the reader is referred to the paper by Deeks et al. (2003). The randomised comparisons in these simulations can be considered to give unbiased results. In what follows, we will compare the results from those comparisons with the results from the comparisons from the corresponding concurrently controlled studies and historically controlled studies.

For these analyses, note that an  $OR < 1$  indicates a benefit for the experimental treatment compared to control, while an  $OR > 1$  indicates a harmful effect. Also note that the results reported by Deeks et al. (2003) do not relate to the original analyses reported in the publications for these trials. What we are reporting here, and what these authors undertook, is an empirical investigation of only parts of the data for the sole purpose of investigating bias in non-randomised studies.

### 17.1.5 Selection bias in concurrently controlled studies

Table 17.1 provides the results of the concurrently controlled studies compared to those that were randomly controlled.

**Table 17.1** Concurrently controlled vs. randomly controlled studies

Study		Average OR	SD, $\ln(\text{OR})$	% of studies with $p < 0.05$	
				Benefit	Harm
IS	Randomised control	0.91	0.34	7	2
	Concurrent control	0.91	0.85	29	21
ESC	Randomised control	1.01	0.68	3	5
	Concurrent control	1.02	0.69	3	6

OR, odds ratio; SD, standard deviation.

The mean OR for the 14,000 concurrently controlled studies simulated from the IS trial data set was 0.91. The mean OR for the corresponding 14,000 randomised controlled trials was also 0.91, indicating no bias on average. This is what was expected, given that, in the concurrently controlled simulation, one of the simulated studies compared treated patients from Region 4 with control patients from Region 11, for example, while another simulated study compared treated patients in Region 11 with control patients in Region 4, and, on average, these would give the same *results* as the randomised comparisons comparing treated patients in Region 4 with control patients in Region 4 and similarly for Region 11. However, the variation was much greater for the concurrently controlled studies. The standard deviation for the log of OR ( $\ln\text{OR}$ ) was 0.85 across the concurrently controlled studies compared to only 0.34 for the randomised comparisons, and this has implications for the numbers of spurious results. For the randomised comparisons, 7% gave statistical significance ( $p < 0.05$ ) in favour of aspirin and 2% gave statistical significance in favour of no aspirin, compared to 29% and 21%, respectively, for the concurrently controlled comparisons. If we take the randomised comparisons as being reflective of the true situation in this investigation, the concurrently controlled comparisons give many more statistically significant results than we would expect, in both directions.

The results for the ECS data set were somewhat different. The mean ORs were again in close agreement; but in this example, the standard deviations were also approximately equal – 0.69 for the concurrently controlled studies and 0.68 for the randomised trials – as were the percentages of studies giving statistically significant results in both directions.

### 17.1.6 Selection bias in historically controlled studies

Table 17.2 provides the results of the historically controlled studies compared to those that were randomly controlled.

For the evaluation of bias in historically controlled studies, Deeks et al. reported a mean OR of 0.88 for the 14,000 simulated historically controlled studies based on the IS trial data compared to a mean of 0.89 for the randomised

**Table 17.2** Historically controlled vs. randomly controlled studies

Study		Average OR	SD, $\ln(\text{OR})$	% of studies with $p < 0.05$	
				Benefit	Harm
IS	Randomised control	0.89	0.35	9	2
	Historical control	0.88	0.44	16	4
ESC	Randomised control	1.23	0.83	4	12
	Historical control	1.06	0.85	6	10

OR, odds ratio; SD, standard deviation.

controlled trials. Again, we see no bias on average in this case. There was, however, a difference in the standard deviations (on the *In*OR scale), with values of 0.44 for the historically controlled studies and 0.35 for the randomised trials. Although this difference was not as marked as that for the concurrently controlled trials, it still resulted in a larger number of significant findings in the historically controlled studies compared to the randomised setting. For the randomised trials, there were 9% statistically significant results in favour of aspirin and 2% in favour of no aspirin compared to 16% and 4%, respectively, for the historical studies. Again, we see more spurious findings for the historically controlled trials.

The results for the ECS trial data again gave a rather different picture. The mean OR for the randomised controlled trials was 1.23, while the mean OR for the historically controlled studies was only 1.06, indicating a strong bias. The randomised trials suggest, on average, a 23% increase in the odds of death or major stroke with treatment compared to only 6% for the corresponding odds for the historically controlled trials. Again, note that these findings do not reflect the true situation with this treatment in this setting; we are simply reporting a simulation exercise to evaluate the properties of historical controls compared to randomised controls. The standard deviations for *In*OR were similar for the randomised controlled and historically controlled studies: 0.83 and 0.85, respectively. But the difference in the mean OR values resulted in slightly more positive findings ( $OR < 1$ ) for carotid endarterectomy in the historical controlled studies compared to the randomised controlled trials (6% vs. 4% statistically significant results) and marginally fewer negative findings ( $OR > 1$ ) for carotid endarterectomy (10% vs. 12% statistically significant results).

### 17.1.7 Some conclusions

This resampling exercise indicates the unpredictability of results when using both concurrently controlled studies and historically controlled studies compared to the randomised setting.

For the concurrently controlled studies, we saw no bias on average, and this was a consequence of how the resampling was undertaken; but the increased variability associated with the OR in the case of the IS data and the corresponding increase in the number of statistically significant results indicate that at the study level, we would tend to produce many more statistically significant findings than we would with a randomised study in the same setting. This is one manifestation of selection bias. The comparability of the patients in the two treatment groups in a concurrently controlled study is not guaranteed, and these differences have caused an increase in the numbers of statistically significant findings. The authors did not see this bias in the ECS study, probably because the patients recruited across regions were more comparable.

For the historically controlled studies, again, there was no bias on average with a small increase in the variability for the OR in the case of the IS data, and this increase resulted in more statistically significant findings compared to the

randomised trials. There was, however, bias on average for the ECS data set, with a reduced OR across the historically controlled studies. Selection bias in historically controlled studies is a consequence of underlying time trends in the characteristics and behaviour of patients in the trial and in the changing trial conditions and patient management. In the case of the ECS data, these trends worked against treatment differences.

These findings have clear implications for how non-randomised studies of this kind are conducted. It is essential that the subjects assigned to experimental and control groups are comparable and are treated within a common environment as much as possible. This is more likely to be achieved in a concurrently controlled study than in a historically controlled study.

## 17.2 Guidance on design, conduct and analysis

### 17.2.1 Regulatory guidance

When should non-randomised comparative studies be considered? There appear to be several situations:

- When the scientific and clinical community strongly believes that the test therapy is superior to all other available therapies. Under these circumstances, it would be extremely difficult to conduct a randomised study that would attract sufficient investigators who would be willing to recruit patients into the study.
- When the course of disease is highly predictable based on clinical evidence.

#### ***ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'***

*'An externally controlled trial should generally be considered only when prior belief in the superiority of the test therapy to all available alternatives is so strong that alternative designs appear unacceptable and the disease or condition to be treated has a well-documented, highly predictable course'.*

- In orphan indications where the population size is limited and using, for example, a historical control group may be the only way to gain enough statistical power to demonstrate a treatment difference.

#### ***CHMP (2006): 'Guideline on clinical trials in small populations'***

*'Although internal controls are the preferred option for comparative trials, under exceptional circumstances external controls may be acceptable. Historical controls (using patients treated with "current" therapies, or not treated at all) might, in some circumstances (even if not routinely), be acceptable to demonstrate efficacy, safety, ease of administration and so on, of a new treatment'.*

The ICH E10 guideline includes comments on the control of bias and associated requirements for successful non-randomised studies:

- The study endpoint(s) should be objective.
- The persuasiveness of findings from non-randomised studies is greater when there is a high degree of statistical significance and much larger estimated differences than we would normally expect from a randomised trial.
- Covariates influencing outcome should be well characterised. We will talk in later sections about adjusting for baseline covariates to help account for selection bias; having well-established predictors of outcome is necessary for this to be helpful.
- The control group and the experimental treatment group should be highly comparable in terms of all relevant baseline factors and concomitant treatments received.
- When no clear control group is available, it is of value to study a range of possibilities and show superiority of the experimental treatment to the most favourable control group.
- It can be useful to involve an independent set of reviewers to reassess endpoints in a blinded way. It can also be useful to involve an independent committee in selecting the control group, again working in a blinded way, but this time blind to outcome. The defibrotide example discussed in Section 17.1.2 had such a committee.

### **17.2.2 Strengthening the Reporting of Observational Studies in Epidemiology**

The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) initiative began in 2004 and was reported in 2007 (von Elm et al., 2007). This initiative developed recommendations for the accurate and complete reporting of observational studies covering the three main study designs: cohort, case-control and cross-sectional. These recommendations are provided in terms of a 22-item checklist, some of which are general and others are specific to the type of design. The authors clearly point out that STROBE does not provide guidance on design and conduct. Nonetheless, following the checklist for reporting the design, conduct and results of a study will help researchers think through some important elements of those aspects. The authors also point out that STROBE is not a tool for assessing the quality of observational studies. Tools for assessing quality are available, and the reader is referred to Deeks et al. (2003) and Viswanathan et al. (2013) for more details.

### **17.3 Assessing the presence of baseline balance**

Selection bias will result in imbalances in known and unknown baseline factors. In a randomised trial, the randomisation itself protects against imbalance, at least in large studies. Methods discussed in Chapters 5 and 6, such as ANCOVA

and logistic regression, can be used to adjust for any imbalances that do occur in relation to known and measured baseline factors. In a non-randomised study, baseline imbalances will be greater, and one key aspect of assessing the validity of such a study is an evaluation of the balance of baseline factors across the experimental and control groups. Of course, it is only possible to undertake this evaluation for known factors that have been measured in patients in both treatment groups. This is one of the drawbacks of a non-randomised study compared to a randomised study. In a randomised study, there is a guarantee that factors that have not been measured will on average be balanced across the groups. In non-randomised studies, that is not the case.

How should we best do this? It is tempting to use *p*-values at baseline to identify factors that are statistically significantly different between the treatment groups and perhaps adjust for those. We have already indicated that this is of no value in randomised studies; 5% of such tests will be significant purely by chance. See Section 6.9 for further discussion on this point. Such significance testing is also of little help in non-randomised studies. There may be important factors, unbalanced at baseline, that are highly predictive of outcome yet give non-significant *p*-values. This could be because the imbalance is only modest yet very important clinically, while the sample size is not large enough to give sufficient power to detect statistical significance. Conversely, we could see statistically significant differences for a factor that has no impact on patient outcome, in which case the imbalances are irrelevant and will not influence our ability to obtain a valid treatment comparison. As discussed in Section 9.8, *p*-values tell us nothing about magnitude; and the magnitude of the difference, and whether this is for a factor that predicts outcome, determines whether that difference is relevant.

The best way to evaluate baseline imbalances is to consider them from a clinical perspective. The focus should be on those factors likely to influence outcome, and clinical judgment is then required to assess whether any observed imbalances could potentially undermine the validity of the treatment comparisons. If in doubt, adjust, at least as a sensitivity analysis; adjusting in general is never a bad thing to do. The only limitations are in relation to the number of factors that can be included in an adjustment model with small sample sizes, and in those cases, more consideration is needed.

## **17.4 Adjusting for selection bias: stratification and regression**

In Chapter 5, we presented a detailed development of methods for adjusting or stratifying the analysis to account for baseline imbalances in randomised trials. These same techniques can also be used in non-randomised studies. The limitation of these methods, in that only a small number of baseline factors can be

included, was raised in Section 6.1. In non-randomised studies, we are likely dealing with imbalances that are, not only more severe, but also affect several important predictors of outcome. The ANCOVA methods detailed in Chapter 6 are better equipped to deal with many baseline factors, and these methods potentially have even greater value in non-randomised comparisons. Section 6.5 dealt with continuous (and score) endpoints, and this is where ANCOVA was introduced. In Section 6.6.1, we covered logistic regression, a methodology that enables corrections to be made for baseline imbalances when the endpoint is binary. In Section 6.6.2, we discussed the negative binomial model, which provides a corresponding methodology for a count endpoint. Finally, Section 13.5 presented methods for adjustment when the endpoint is time to event. The proportional hazards model enabled us to adjust for multiple factors for endpoints of this kind. To summarise, in any non-randomised study, we usually want to adjust for a range of baseline factors, and ANCOVA (continuous, score endpoints), logistic regression (binary and ordered categorical endpoints), negative binomial regression (count endpoints) and the proportional hazards model (time-to-event endpoints) provide appropriate approaches.

## 17.5 Adjusting for selection bias: propensity scoring

### 17.5.1 Defining propensity scores

Propensity score methods were first introduced by Rosenbaum and Rubin (1983) and provide an alternative way to adjust for baseline factors, especially where these are large in number. D'Agostino (1998) presents an extensive discussion of the technique when used for comparing a test treatment group to a non-randomised control group. The methodology is best presented through an example.

Austin and Mamdani (2006) undertook a comparison of several different propensity score methods in a study evaluating the use of statins in reducing all-cause mortality in patients discharged from hospital who had been admitted previously with a diagnosis of acute myocardial infarction (MI). The cohort of patients consisted initially of 11,524 individuals admitted to hospitals in Ontario, Canada, with a diagnosis of acute MI between 1 April 1999 and 31 March 2001. Of these, 1137 died during hospitalisation and 1283 were excluded from further consideration due to incomplete data on baseline factors collected during hospitalisation. A total of 9104 were therefore included in the analysis. Of these, 3049 (33.5%) were prescribed statins on discharge, while 6055 were not. The objective of the analysis was to evaluate the effectiveness of statins in reducing three-year mortality. This was not a randomised study, and it may well be that the statins were prescribed only for certain kinds of patients as decided by the physician treating the patient at the time of discharge. Without considering any kind of adjustment for baseline imbalances, the three-year mortality among patients

receiving statins was 14.2% compared to 25.3% for those patients who did not receive statins, with an OR of 0.49 and  $p < 0.0001$ . This result is highly statistically significant, with almost a 50% reduction in the odds of dying within three years. However, this result is potentially influenced by selection bias caused by differences in known (and unknown) factors measured at baseline. Data were available on 24 baseline factors including age, gender, presenting characteristics (shock, acute chronic heart failure/pulmonary oedema), acute MI risk factors (family history, diabetes, etc.), co-morbidities (angina, cancer, etc.), vital signs on admission (diastolic and systolic blood pressure, heart rate, respiratory rate) and laboratory parameters (white blood count, haemoglobin, etc.).

The basic idea behind the propensity score is to estimate the probability that a patient would receive a statin (as opposed to not receiving a statin) based on these baseline factors. We will denote this quantity by

$$\text{Prob}(\text{patient with baseline characteristics } x_1, x_2, \dots, x_{24} \text{ receives a statin})$$

Here  $x_1$  denotes age,  $x_2$  denotes whether the patient is male or female, and so on, through all the 24 baseline factors. To give a flavour of how this is done, let's suppose we were adjusting for just two factors, gender and whether a patient was diabetic, and consider the (artificial) data set down in Table 17.3.

In this hypothetical example, overall, 1190 patients received statins and 2500 did not. We have split the data according to the two baseline factors: gender and diabetic/non-diabetic. Of the 590 male diabetics, only 210 received a statin – that is, 36% – so the probability that a male diabetic receives a statin is 0.36. Similarly, the probabilities associated with the other factor combinations are given in the final column. This probability is what we term the *propensity score* – it is the probability of receiving the intervention. So, across the whole dataset, each patient has a propensity score that differs from patient to patient depending on their baseline characteristics. With many more baseline factors to account for, it is impossible to divide the data according to all combinations of baseline factors and do what we have done in Table 17.3, but the principle is the same. With lots of baseline factors, we do this by fitting a logistic regression model for the probability of receiving the intervention. See Section 6.4 for more

**Table 17.3** Constructing propensity scores based on two characteristics (hypothetical)

	<b>Statin</b>	<b>Non-statin</b>	<b>Total</b>	<b>Probability (patient receives statin)</b>
Male, diabetic	210	380	590	$210/590 = 0.36$
Male, non-diabetic	660	1200	1860	$660/1860 = 0.35$
Female, diabetic	90	300	390	$90/390 = 0.23$
Female, non-diabetic	230	620	850	$230/850 = 0.27$
	1190	2500		

discussion on logistic regression. In this logistic model, the *outcome* variable is the binary variable that takes the value 1 if the patient receives the statin and 0 if the patient does not receive the statin. Fitting this model then gives a propensity score for every patient in the sample as the probability that the specific patient, with baseline factors  $x_1, x_2, \dots, x_{24}$ , would have received a statin. This probability assignment is based on which patients are more or less likely to receive a statin according to the patterns in the data. In a randomised study, each of these probabilities is equal to 0.50; every patient has a 50/50 chance of receiving the experimental treatment! In a non-randomised study, however, investigator or patient preferences or both will take these probabilities away from 0.50.

How do we choose which baseline factors to include in the propensity score model? Two kinds of factors are important: firstly, those factors that have the potential to influence outcome; and secondly, those on which the treating physician has based their decision to give one treatment rather than the other. It will often be the case that knowledge of the prognostic and influencing factors is limited, and to be conservative, all reasonable possibilities should be included. Propensity score methodology can only ever balance on the factors included in the model and never on omitted factors.

### 17.5.2 Propensity score stratification, regression and matching

So, how do we use these propensity scores? Well, there are several possibilities. Firstly, we could compare the statin and non-statin patients in terms of their three-year mortality rates by stratifying by the propensity score. For example, we divide the patients according to their values of the propensity score – < 0.20,  $\geq 0.20$  but < 0.40,  $\geq 0.40$  but < 0.60,  $\geq 0.60$  but < 0.80 and  $\geq 0.80$  – to create a structure as in Table 17.4. The *n*'s in the table refer to the numbers of patients within each category according to treatment received.

Stratifying by the propensity score produces subgroups of statin and non-statin patients with similar propensity scores. A stratified analysis calculates a treatment difference for each of these subgroups. A hypothetical example is provided in Table 17.4. Stratification gives five treatment differences, and the stratified analysis then averages these (as in Section 5.2) and compares the average to 0 if we are looking at absolute differences (in means, for example) or to 1 if the

**Table 17.4** Stratified analysis by propensity score (hypothetical)

Propensity score	Statin	Non-statin
< 0.20	<i>n</i> = 30	<i>n</i> = 180
$\geq 0.20$ but < 0.40	<i>n</i> = 38	<i>n</i> = 150
$\geq 0.40$ but < 0.60	<i>n</i> = 62	<i>n</i> = 121
$\geq 0.60$ but < 0.80	<i>n</i> = 144	<i>n</i> = 63
$\geq 0.80$	<i>n</i> = 205	<i>n</i> = 26
Total	479	540

treatment difference is expressed as a ratio (e.g. an odds ratio, as in the statin example). The basic idea behind these considerations is that patients with the same (or similar) propensity score have the same (or similar) probability of receiving a statin, and whether they do or not is simply *chance*. What we have effectively done is to mimic a randomised setting where each patient within each stratum has a similar probability of getting the experimental treatment one is *randomised* to statin and one to non-statin. Of course, there is no randomisation, but we can think of things conceptually in those terms. It is standard practice to use five groupings. These groupings can be based on a fixed split of the propensity score, as in Table 17.4, or chosen according to the quintiles across the treatment groups combined. So, the first stratum would include patients with the lowest 20% of propensity scores, the second stratum the next 20% of patients according to their propensity scores, and so on.

This stratified analysis has adjusted for propensity score, but this is not the only way to adjust. A second approach uses the propensity score as a covariate in an adjusted (model-based) analysis. Finally, a third approach takes each patient in the statin group and matches to a patient in the control group with a propensity score that is the closest (within a so-called *caliper* of, for example, 0.01); if a match cannot be found, that subject in the statin group is ignored. The two resulting groups (by definition, of the same size) are then compared using a simple two-group test (such as the chi-square test, in this case). Table 17.5 shows the ORs from each of these three methods for the statin example together with the OR from the simple comparison without adjustment and an analysis that uses all 24 of the baseline factors in an adjusted (model-based) analysis that does not involve propensity scoring (Austin and Mamdani, 2006).

It is noticeable that all the adjusted analyses give ORs closer to 1 compared to the unadjusted analysis. This indicates that the unadjusted analysis overestimates the benefit of statins in reducing three-year mortality. Each of the adjusted analyses gives an OR between 0.75 and 0.85, indicating between a 15% and 25% reduction in the odds of death within three years. All results are statistically significant at the 5% level, but the matching method has a much less significant *p*-value. This is in part due to the reduced sample size. The matching method gave an overall sample size of only around 67% of the sample size used for the other adjustment methods.

**Table 17.5** Analysis of the statin observational study by propensity score and other methods

	OR	<i>p</i> -value
Unadjusted	0.49	< 0.0001
Model-based adjustment using 24 factors	0.75	< 0.0001
Stratified by quintiles of the propensity score	0.77	0.0003
Model-based adjustment with propensity score as a covariate	0.84	0.0033
Matching on propensity score	0.85	0.037

One advantage of matching is its transparency. Baseline tables can be produced for the matched groups, and the extent to which the matching has achieved balance for the baseline factors can be observed. Researchers often find this especially comforting! In the statin example (Austin and Mamani, 2006), an excellent balance was achieved. For example, the mean ages were 63.33 (no statin group) and 63.30 (statin group) in the matched sample, compared to 68.11 (no statin group) and 63.36 (statin group) in the complete cohorts. Similarly, the percentages of females in the matched samples were 29.8% and 30.9%, respectively, vs. 37.0% and 29.1%, respectively, in the full cohorts. *P*-value comparisons were made to assess the baseline comparability of the matched samples, and of the 24 variables that were the basis of the matching, none were statistically significantly different (at the 5% level) between the statin and non-statin matched groups.

It is also worth noting that the stratification and regression approaches answer a different question than that addressed by propensity score matching. The matched analysis looks at the treatment effect in patients treated with statins. The stratification and regression approaches look at the treatment effect among the population of subjects who are eligible for treatment with statins, because those analyses are based on the complete population of patients in the two cohorts.

## 17.6 Comparing methods that correct for selection bias

Deeks et al. (2003) undertook an empirical investigation of the ability of various methods to adjust for selection bias in both the concurrently controlled studies and the historically controlled studies to bring the results back in line with the results of the randomised studies. They used those studies that were sampled previously in their evaluation of selection bias as described in Section 17.1. As these authors point out, methods of baseline adjustment are looking to achieve, through analysis, what could not be achieved by design.

A total of 10 relevant baseline factors, in terms of factors that were viewed as potentially predicting outcome, were considered for the IS trial, while 7 factors were considered for the ECS trial. For the IS trial, these included sex, age, atrial fibrillation (yes/no), infarct visible on CT scan, etc., while for the ECS trial, factors included sex, age, previous MI, degree of stenosis, etc. For our purposes, we will consider the unadjusted analysis and then only a subset of the adjustment methods considered by Deeks et al. (2003):

- Logistic regression adjusting for all baseline factors
- Propensity score stratification based on five strata
- Propensity score adjustment based on using the propensity score as a covariate
- Propensity score adjustment-based matching

**Table 17.6** Comparison of methods of adjustment for baseline imbalance: concurrently controlled studies resampled from the IS trial

<b>Method</b>	<b>% of studies with <math>p &lt; 0.05</math></b>	
	<b>Benefit</b>	<b>Harm</b>
Randomised control	7	2
Unadjusted	29	21
Logistic regression, all factors	23	18
Propensity score; stratified	19	15
Propensity score; logistic regression	19	15
Propensity score; matched	15	12

Source: Adapted from Deeks et al., 2003.

**Table 17.7** Comparison of methods of adjustment for baseline imbalance: concurrently controlled studies resampled from the ECS trial

<b>Method</b>	<b>% of studies with <math>p &lt; 0.05</math></b>	
	<b>Benefit</b>	<b>Harm</b>
Randomised control	3	5
Unadjusted	3	6
Logistic regression, all factors	4	6
Propensity score; stratified	5	3
Propensity score; logistic regression	5	3
Propensity score; matched	2	6

Source: Adapted from Deeks et al., 2003.

The results of these investigations for the concurrently controlled studies are shown in Table 17.6 for the IS study and Table 17.7 for the ECS study. Similar tables for the historically controlled studies are Table 17.8 for the IS data and Table 17.9 for the ECS data. Note that in each case, the results for the randomised control and the unadjusted analysis have already been discussed in Sections 17.1.5 and 17.1.6. The main issues centre around whether the methods of adjustment have improved on the unadjusted methods in terms of bringing the results in line with the unbiased results from the randomised studies. In each evaluation, the number of significant findings tells us how much more likely we are to be deceived compared to the *true* situation as reflected by the randomised study results.

In Table 17.6, for the IS data, we see that there are statistically significant findings in 9% of the randomised studies, 7% in favour of aspirin (benefit) and 2% against (harm), while the unadjusted analysis for the concurrently controlled studies gave a total of 50%: 29% in favour of aspirin and 21% against. All the adjustment methods improve on the unadjusted analyses. The biggest improvement comes from the propensity score matching, which gives a total of 27% statistically significant studies split 15% in favour of aspirin and 12%

**Table 17.8** Comparison of methods of adjustment for baseline imbalance: historically controlled studies resampled from the IS trial

Method	% of studies with $p < 0.05$	
	Benefit	Harm
Randomised control	9	2
Unadjusted	16	4
Logistic regression, all factors	13	3
Propensity score; stratified	9	2
Propensity score; logistic regression	9	2
Propensity score; matched	7	2

Source: Adapted from Deeks et al., 2003.

**Table 17.9** Comparison of methods of adjustment for baseline imbalance: Historically controlled studies resampled from the ECS trial

Method	% of studies with $p < 0.05$	
	Benefit	Harm
Randomised control	4	12
Unadjusted	6	10
Logistic regression, all factors	12	14
Propensity score; stratified	12	10
Propensity score; logistic regression	12	10
Propensity score; matched	5	10

Source: Adapted from Deeks et al., 2003.

against, although some bias remains, as can be seen with many more comparisons giving statistical significance compared to the randomised setting.

Table 17.7 presents results for the concurrently controlled studies based on the ECS trial data. A total of 8% of the randomised studies give statistically significant results: 3% showing a benefit for carotid surgery and 5% showing harm. In this case, as seen in Section 17.1.5, the unadjusted analyses performed similarly, showing little bias. Here, all the adjusted methods seem to leave things pretty much unchanged, although the propensity score-matched analyses are closest to the randomised analyses.

Table 17.8 presents the results of the adjusted analyses for the historically controlled studies for the IS data. The unadjusted analyses give 20% statistically significant studies: 16% in favour of aspirin and 4% against. This is compared to only 11% for the randomised studies; 9% in favour and 2% against. All of the propensity score adjustment methods worked well here to correct the results to be in line with the randomised studies with little to choose between them.

Table 17.9 presents the analyses for the ECS trial data. Here, 16% of studies give statistically significant results for the unadjusted analyses: 6% in favour of

carotid surgery and 10% against. This contrasts with the randomised studies, where 4% show a benefit and 12% show harm. Each adjustment method made things worse compared to the unadjusted analyses except the propensity score adjustment-based matching method, where the results were closest to those for the randomised studies.

This empirical investigation gives somewhat mixed results in terms of firstly the value of adjustment in correcting for selection bias and secondly which of the adjustment methods considered gives the best results. For the IS trial, all the methods of adjustment considered improved on the unadjusted analyses, but the propensity score method using matching made the biggest improvement and, for the historically controlled studies, brought the results back in line with those from the randomised studies.

For the ECS trial, the results for the unadjusted analyses were already similar to those for the randomised studies, and little was gained through adjustment. In some cases, in particular for the historically controlled studies, some adjustment methods made things worse, however the propensity score method based on matching was the approach that gave results most similar to the unadjusted analyses and closest to the results from the randomised studies. It was argued earlier that the homogeneity in patient characteristics and potentially the standardisation of the environment in which patients were treated across the regions in the ECS trial resulted in the unadjusted analyses being close to the randomised analyses for the concurrently controlled studies.

In terms of general recommendations, two points are worth making:

- When comparability between the two groups is achieved by design, adjustment does not appear to be especially useful, although it does no harm. As mentioned earlier in this section, comparability is likely easier to achieve with a concurrent control group than with a historical control group.
- When comparability between the groups cannot be or has not been achieved, adjustment using propensity scores based on matching is the preferred approach.

Methods based on propensity scoring can be effective only if the key baseline characteristics that determine the choice of treatment group and are prognostic for the outcome being evaluated have been recorded and included in the model. If key factors are omitted from the propensity score model, the methodology will fall short of adjusting completely for the baseline imbalances. In the next section, we will discuss how a generalisation of propensity score matching in many settings can increase sensitivity.

## 17.7 Inverse propensity score weighting

One clear limitation of matching based on propensity scoring is sample size, which directly affects the power of the comparison. Firstly, the sample size for the matched groups is driven by the smaller of the two cohorts. In the earlier

example, the sample size for the non-statin cohort was 6055 and for the statin cohort was 3049. There is the possibility in principle to choose a 2:1 ratio for matching and so, for example, have twice as many control patients as treated patients, although this would not be feasible for the statin example.

Secondly, the caliper value can be relaxed to avoid discounting too many patients for whom a match cannot be found. In the statin example, the matched propensity score analysis involved 2348 statin patients, so there was a loss of 23% of the statin patients from the original cohort of 3049 patients.

However, there is an alternative way of using the propensity scores, which involves weighting each patient's outcome value across the complete cohort. Consider a patient in our example with a propensity score of 0.25 who received a statin. According to the propensity score, this kind of patient was relatively unlikely to receive a statin and will be underrepresented in that group. So, give that patient a weight of  $1/0.25 = 4$  in the analysis. Consider now a patient with a propensity score of 0.80 who also received a statin. This patient has a relatively high probability of receiving a statin, so give that patient a weight of only  $1/0.80 = 1.25$  in the analysis. Next, consider two patients who did not receive a statin, and suppose the first one had a propensity score of 0.15. This patient had a probability of not receiving a statin of  $0.85 (= 1 - 0.15)$  and is therefore well represented in the non-statin group. Give that patient a weight of  $1/0.85 = 1.1765$  in the analysis. Finally, suppose the second patient in the non-statin group had a propensity score of 0.75 and thus a probability of 0.25 to not receive a statin. This patient is relatively underrepresented in that group, so give a weight of  $1/0.25 = 4$  to the patient. In general, if  $p$  is the propensity score, the method gives a weight of  $1/p$  to each patient in the statin group and a weight of  $1/(1-p)$  to each patient in the non-statin group. This technique is known as *inverse propensity score weighting* (IPSW).

What do we mean when we say, 'give a subject a weight of say 4 in the analysis'? Suppose we code the outcomes in the statin example with  $y = 0$  if the patient dies within three years and  $y = 1$  if the patient survives beyond three years. A simple comparison of the proportions of survivors across the statin and non-statin cohorts would be calculated by averaging the  $y$  values in the statin

(T) group  $\left( \frac{\sum y}{n_T} \right)$  and averaging the  $y$  values in the non-statin (C) group  $\left( \frac{\sum y}{n_C} \right)$

where  $n_T$  and  $n_C$  are the sizes of the statin and non-statin cohorts. These quantities are the proportions of survivors in each of the two treatment groups. The weighing is brought in by calculating the weighted average in the statin group

$(p_T = \left( \sum \frac{y}{p} \right) / \left( \sum \frac{1}{p} \right))$  in place of the simple proportion in that group and, similarly,

the weighted average in the non-statin group  $(p_C = \left( \sum \frac{y}{1-p} \right) / \left( \sum \frac{1}{1-p} \right))$  in

place of the simple proportion in that group. These weighted proportions can then be compared.

In the matched case, it was straightforward to assure ourselves that balance had been achieved for baseline factors by providing tables of summary statistics in the matched sample of patients. With IPSW, it is not quite so straightforward but can be achieved through the weighting. We use the weights to calculate weighted means

for continuous baseline factors ( $\bar{x}_T = \left( \sum \frac{x}{p} \right) / \left( \sum \frac{1}{p} \right)$ ,  $\bar{x}_C = \left( \sum \frac{x}{1-p} \right) / \left( \sum \frac{1}{1-p} \right)$ )

and express the difference between the means in standardised form as follows;

$$\text{standardised difference (means)} = 100 \times \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$$

where  $s_T$  and  $s_C$  are the standard deviations associated with the weighted variable values in statin and non-statin groups respectively.

For the binary factors, we calculate weighted proportions based on a 0/1 coding of the factor:  $x = 0$  if the factor is absent,  $x = 1$  if the factor is present. The weighted proportions are given by  $q_T = \left( \sum \frac{x}{p} \right) / \left( \sum \frac{1}{p} \right)$  and  $q_C = \left( \sum \frac{x}{1-p} \right) / \left( \sum \frac{1}{1-p} \right)$ , and the difference between the proportions in standardised form is given by

$$\text{standardised difference (proportions)} = 100 \times \frac{q_T - q_C}{\sqrt{\frac{q_T(1-q_T) + q_C(1-q_C)}{2}}}$$

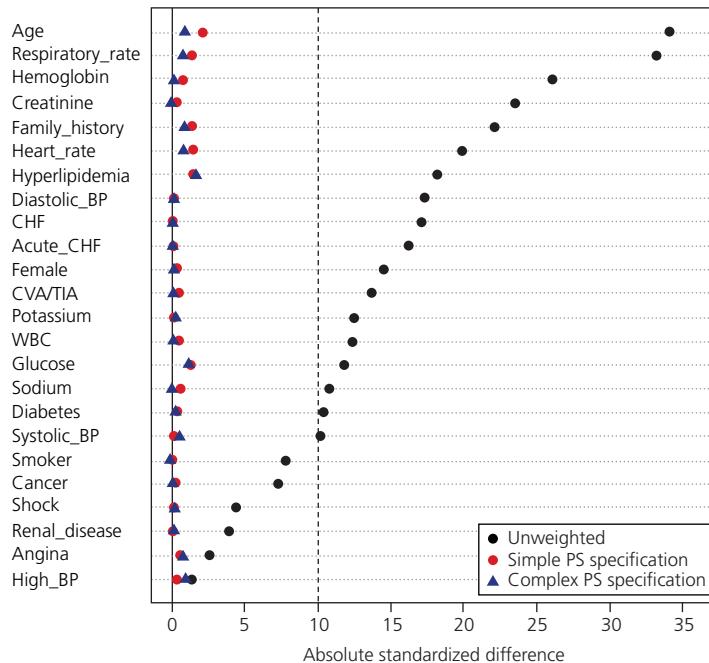
The standardised differences are known as *balance statistics*. Figure 17.1 (Austin and Stuart, 2015) shows the balance statistics before any weighting and the balance statistics following the weighting according to the inverse of the propensity score for patients in the statin (T) group and according to the inverse of 1 minus the propensity score in the non-statin (C) group. The weighting has successfully provided balance across the statin and non-statin groups for these baseline factors.

## 17.8 Case-control studies

### 17.8.1 Background

The purpose of this section is to comment on some statistical aspects associated with case-control studies. Case-control studies have a long history, and such a study (Doll and Bradford-Hill, 1950) was the basis for establishing the causal effects of smoking on lung cancer. In this example, the cases were subjects with the disease being investigated, lung cancer, while controls were subjects without the disease.

In general, each case is matched to one or more controls based on sex, age and possibly geographical location (e.g. each case may be matched to a control



**Figure 17.1** Absolute standardized differences in unweighted and weighted samples

**Table 17.10** Case-control study investigating smoking and lung cancer (hypothetical data); one control to each case

	Lung cancer	No lung cancer	Total
Smoker	170	50	220
Nonsmoker	30	150	180
Total	200	200	400

from the same GP practice). Controls can be thought of as subjects who do not have the disease under study; had they had the disease, they would have been included as cases. The disease under study is considered in this case as the *outcome*. We now ascertain for each case and each control whether they have the *attribute* we are investigating (smoking, in our example). Research suggests that between one and four controls per case can be used depending on practicality. More controls will to a certain extent increase the power, but this should not be at the expense of choosing controls who are not good matches, as that will have the opposite effect and weaken the size of the effect we are trying to estimate. Table 17.10 provides some hypothetical data for the smoking example for illustrative purposes.

In this example, 200 cases were chosen, and there was a 1-to-1 matching to give 200 controls. Of the 200 cases, 170 were smokers, while 30 were nonsmokers.

For the 200 controls, 50 were smokers, while 150 were nonsmokers. The OR for lung cancer is calculated as

$$\text{OR} = \frac{170/50}{30/150} = 17$$

Note that the odds for the smokers, the attribute we are investigating, are the numerator, while the odds for the nonsmokers provide the denominator. This is a large OR and would give a highly statistically significant result. In the original investigation (Doll and Bradford-Hill, 1950), there were 649 cases (lung cancer) and 649 controls (no lung cancer) for the males. Among the cases, 647 were smokers and 2 were nonsmokers, and among the controls, 622 were smokers and 27 were nonsmokers. The OR was 14.04 with  $p = 0.00000064!$  For the females (60 cases and 60 controls), the OR was less dramatic (OR = 2.47) but nonetheless statistically significant,  $0.01 < p < 0.02$ . These authors clearly established through this case-control study that smoking had a causal effect on lung cancer. The Doll and Bradford-Hill study was conducted at 20 London hospitals, and these hospitals were asked to include all cases of lung cancer. Each case was matched with a control patient with a disease other than cancer who was of the same sex, closely comparable in terms of age (within a five-year age band) and from the same hospital or, if this was not possible, from a neighbouring hospital. This study remains the classic example of a well-designed and well-conducted case-control study.

One critical aspect of a case-control study is the population from which the controls are taken. There has been considerable discussion over many years about the risk of soft tissue sarcoma following exposure to phenoxy herbicides and chlorophenols, and a number of case-control studies have been conducted. See Smith and Christophers (1992) for a general discussion. In some studies, controls have been taken from the general population, while in others, the controls have been patients with other cancers. It may be that those studies with controls from the general population overestimate the risk because of poor recall regarding exposure, while studies using controls with other cancers underestimate the risk if phenoxy herbicides and chlorophenols are linked to other cancers. The true risk may be somewhere between the two.

Nowadays, within the pharmaceutical industry, case-control studies are used primarily within the context of safety evaluation and pharmacovigilance. In these settings, the serious side effect of interest is the outcome, while the drug being evaluated is the attribute. As mentioned earlier, a series of such studies were conducted to investigate the risk of cardiovascular adverse effects associated with rosiglitazone in patients with type II diabetes.

### 17.8.2 Odds ratio and relative risk

In Section 4.5, we discussed both ORs and relative risks and distinguished the two. We also discussed the difficulties in interpreting ORs as opposed to relative risks. In this section, we will explore this distinction again in the context of a

**Table 17.11** Case-control study investigating smoking and lung cancer (hypothetical data); two controls for each case

	Lung cancer	No lung cancer	Total
Smoker	170	100	270
Nonsmoker	30	300	330
Total	200	400	600

case-control study. Looking at the hypothetical data in Table 17.10, we can, at least in theory, think in terms of calculating a relative risk: that is, the risk of lung cancer among smokers divided by the risk of lung cancer among nonsmokers.

This relative risk is  $RR = \frac{170/220}{30/180} = 4.64$ . But suppose we had decided to have, for example, two matching controls rather than one for each case. The data would then be as in Table 17.11.

In this case, the relative risk is  $RR = \frac{170/270}{30/330} = 6.93$ . The relative risk has

changed merely by choosing a different design, with two controls to each case instead of one control per case. This is clearly unsatisfactory and is why relative risks should not be used in case-control studies. We must remember that this is not a randomised comparison with subjects randomised to be smokers or non-smokers, and we do not have control over the proportions of these types of subjects in the study. However, the OR for Table 17.11 is  $OR = \frac{170/100}{30/300} = 17$ , unchanged from the OR for Table 17.10. The OR is unaffected by changing the ratio of controls to cases. This is one reason behind the importance and usefulness of the OR as a suitable statistic for the evaluation of binary data in that in some settings, the relative risk cannot be estimated. A further point is that with large sample sizes and rare events, the relative risk and OR have similar numerical values (see Section 4.5.6); in many case-control studies, we are in this situation. It is possible in these settings to calculate the OR but then, at least conceptually and for ease of interpretation, talk in terms of relative risk, risk reduction/risk increase as these quantities would be similar numerically.

## CHAPTER 18

# Meta-analysis and network meta-analysis

### 18.1 Definition

A *systematic review* addresses a well-defined research question by collecting, evaluating and summarising all evidence that fits with the pre-specified eligibility criteria. A *meta-analysis* is the statistical core of a systematic review that brings together in a quantitative way the results from separate studies to provide an overview regarding a particular treatment or intervention. Meta-analysis involves the combination of head-to-head studies of two treatments, and the chapter will begin by covering those types of analyses. In recent years, the methodology has been extended to incorporate indirect comparisons of treatments and the analysis of networks of treatments in particular therapeutic settings through *network meta-analysis* to allow a broader basis for comparisons. These methods will be considered in the second part of the chapter.

Meta-analysis should not be confused with simple pooling. *Pooling* is a related procedure that puts all the data together and treats the data as if they came from a single study. Meta-analysis does not do this; it recognises study-to-study variation and, indeed, generally involves an assessment of whether the data from the different studies give a consistent answer. Meta-analysis is preferred to pooling, as Example 18.1 illustrates.

#### Example 18.1 Adverse event rates (hypothetical)

In two randomised trials, each with 300 patients, comparing two active treatments in terms of the incidence of a particular adverse event, the data were as shown in Table 18.1.

**Table 18.1** Adverse event rates in two trials

	<b>Treatment A</b>	<b>Treatment B</b>	<b>Difference (A – B)</b>
Trial 1	10/100, (10.0%)	16/200 (8.0%)	2.0%
Trial 2	7/200 (3.5%)	2/100 (2.0%)	1.5%

Pooling the data gives an overall adverse event rate on treatment A of 5.7% (17/300) compared to 6.0% (18/300) on treatment B, a pooled difference ( $A - B$ ) of  $-0.3\%$  with a higher rate on treatment B. This is clearly misleading since in each trial there is a higher adverse event rate on treatment A. A more appropriate measure of the treatment difference is given by the average absolute difference of  $1.75\%$ . This is what is done with meta-analysis: each trial provides an estimated treatment difference, and the meta-analysis then averages these to give an overall difference.

The fundamental problem is that the simple pooled analysis is a non-randomised comparison. In Example 18.1, two-thirds of the patients receiving treatment A are from trial 2, while two-thirds of the patients receiving treatment B are from trial 1 and the underlying SAE rate is much higher in trial 1 than in trial 2. Meta-analysis respects the randomisation by calculating a treatment difference within each trial, and these differences, based on randomised comparisons, are then averaged. This discrepancy between pooling and meta-analysis, known as *Simpson's paradox*, illustrates the dangers of simple pooling. This quotation is taken from Gibbons and Amatya (2016):

*'Another bad approach to synthesizing evidence across studies is to simply pool them and perform an analysis on the combined data without taking the clustering of observations within studies into account. This is an all too common practice. Not only can this provide incorrect standard errors and tests of hypotheses, it can also produce biased results'.*

The term *pooling* in the literature is often used loosely and covers both pooling as described here and meta-analysis. In this book, we will use these terms to mean different things, as described in this section. Take care when interpreting results!

It is recognised that the studies combined in a meta-analysis usually are not identical and do not share a common protocol. The dosages may be different, the precise definition of the endpoints may be different, treatment duration may differ, the precise nature of both the treatment and the comparator may be different and so on. Clearly, the more specific the research question being addressed by the meta-analysis, the more closely the studies must match, but the breadth of the studies that are combined will depend on the breadth of the question being asked.

## 18.2 Objectives

Meta-analysis is used in numerous contexts and both within and outside of the regulatory setting. The technique can quantify the current state of knowledge regarding a particular treatment in terms of safety and efficacy. The *Cochrane*

*Collaboration* uses the methodology extensively within systematic overviews of treatments in particular therapeutic areas and to answer general health questions (Higgins et al., 2019). Meta-analysis can also be a useful way to combine the totality of data from studies in relation to a particular treatment, perhaps as the basis for a marketing campaign.

In addition, combining studies can very effectively increase power for primary and secondary efficacy endpoints and subgroups. Individual studies are unlikely to be powered for secondary endpoints and subgroups, and meta-analysis can be a way of increasing power in relation to these. For primary endpoints, increasing the power improves precision (reduces the standard error) and gives narrower confidence intervals (CIs), enabling clinical benefit to be more clearly understood. In evaluating safety, there may simply not be enough information in any single trial, and a meta-analysis is the only way to make comparative inferences and identify safety problems. This is a useful approach within an integrated safety evaluation at the regulatory submission stage but is also of potential value retrospectively when safety concerns arise.

A further area of application for meta-analysis is in the choice of the non-inferiority margin,  $\Delta$ . As mentioned in Section 12.7,  $\Delta$  is often chosen as some proportion of the active control treatment effect (over placebo), and meta-analysis can be used to obtain an estimate of that treatment effect with an associated CI across placebo-controlled trials.

Combining studies can be useful in resolving apparently conflicting results. For example, Mulrow (1994) reported a meta-analysis of trials of intravenous streptokinase for treatment in acute myocardial infarction. The complete analysis involved a total of 33 trials reported between 1959 and 1988, and Mulrow presented a *cumulative meta-analysis* that combined the trials chronologically over time. Of the eight trials reported between 1959 and the end of 1973, five gave odds ratios (ORs) that favoured intravenous streptokinase (two were statistically significant), while three trials gave ORs favouring control (none were statistically significant). In one sense, a confusing picture was emerging, with six negative/inconclusive trials out of the first eight conducted. It was only at the end of 1988, when sufficient trials pointed in the same direction, that a firm conclusion regarding the benefit of intravenous streptokinase was reached. However, the meta-analysis combination at the end of 1973 gave a clear result, with an OR of around 0.75 in favour of streptokinase and a statistically significant  $p$ -value of 0.0071. As Mulrow (1994) pointed out, '*This cumulative type of review indicated that intravenous streptokinase could have been shown to be life saving almost 20 years ago, long before its submission to and approval by the United States Food and Drug Administration and its general adoption in practice*'.

Meta-analysis can also address whether studies provide a consistent result. Exploring heterogeneity is a key element of any meta-analysis: if heterogeneity

is present, that of itself can be insightful regarding the efficacy and safety of treatments in different settings.

## 18.3 Statistical methodology

### 18.3.1 Methods for combination

Each trial included in the meta-analysis provides a measure of treatment effect (difference). For a continuous endpoint, this could be the mean response on the active treatment minus the mean response in the placebo arm. Alternatively, for a binary endpoint, the treatment effect could be captured by the difference in the cure rates, for example, or by the OR. For a count endpoint, the rate ratio captures the treatment difference, while for a time-to-event endpoint, the hazard ratio is usually the measure of treatment effect.

Assume that we have decided on the best measure for the treatment effect. If this is expressed as a difference (for example, in the means), there will be an associated standard error measuring the precision of that difference. If the treatment effect is captured by a ratio (an OR, a rate ratio, or a hazard ratio), there will be an associated standard error on the log scale: the standard error of the log OR, the log rate ratio, or the standard error of the log hazard ratio.

Again, whichever measure of treatment effect is chosen, the meta-analysis combination proceeds in a standard way. We average the treatment effect over the  $m$  studies being combined. It is not the straight average but a weighted average, weighted according to the precision of each individual study, and this precision is captured by the standard error. This is very similar to what we did in Section 5.2 when the analysis for comparing treatments was adjusted for baseline factors. In our development here, *stratum* is replaced by *study*. For study  $i$ , let  $d_i$  be the treatment effect with associated standard error  $se_i$ . The overall estimate of the treatment effect is then

$$d = (w_1 d_1 + w_2 d_2 + \dots + w_m d_m) / (w_1 + w_2 + \dots + w_m)$$

where  $w_i = 1/se_i^2$ . Essentially, weighting by the inverse of the standard error this way is weighting primarily by the sample size, so larger, more precise studies are given more weight.

If the treatment effect in each individual trial is the difference in the mean responses, then  $d$  represents the overall adjusted mean difference. If the treatment effect in the individual trials is the log OR,  $d$  is the overall adjusted log OR, and so on. In the case of overall estimates on the log scale, we generally antilog the estimate to give a measure back on the original scale (for example, the OR scale). This is similar to the approach we saw in Section 4.5.5 when we looked at calculating a CI for an OR. Note that for ratios, we work on the log scale because that is the scale on which we have a formula for the standard error and where the estimation process is more stable.

### 18.3.2 CIs

The methods of the previous subsection give us a combined estimate,  $d$ , for the treatment effect. We now need to construct a CI around this estimate. Doing so initially involves obtaining a standard error for  $d$ , the overall effect:

$$se = \frac{1}{\sqrt{w_1 + w_2 + \dots + w_m}}$$

From this, it is easy to obtain a 95% CI for the overall treatment effect as  $(d - 1.96se, d + 1.96se)$ .

If this CI is on the log scale – for example, with the OR – then both the lower and upper confidence limits should be converted by using the antilog to give a CI on the original OR scale.

### 18.3.3 Fixed and random effects

The *fixed effects model* assumes that the studies are essentially replicates of each other and share the same underlying true treatment effect. An alternative approach considers the collection of studies included in the meta-analysis as a series of studies, each with its own true underlying treatment effect. This results in a slightly changed methodology: the *random effects model*. The random effects model allows the true treatment effect to differ for the trials. For example, let  $\theta_1$  be the true difference in the means in trial 1,  $\theta_2$  be the true difference in trial 2, and so on, to trial  $m$  with a true difference in the means of  $\theta_m$ . The model then assumes that the  $m$  different values  $\theta_1, \theta_2$ , etc. are drawn from a normal distribution with mean  $\theta$  and variance  $\tau^2$ . The random effects model implicitly allows for study-to-study variation, and the resulting analysis provides estimates of the overall mean effect  $\theta$  and the variance  $\tau^2$ . The mathematics for fixed and random effects models is a little different. For the random effects model, we use different weights:  $w_i$  in the fixed effects formula for  $d$  is replaced by  $1/(se_i^2 + \tau^2)$ , where  $\tau^2$  is estimated as the between-study variance of the study treatment effects. See Borenstein et al. (2010), for example, for an overview of the methodology and the mathematical techniques.

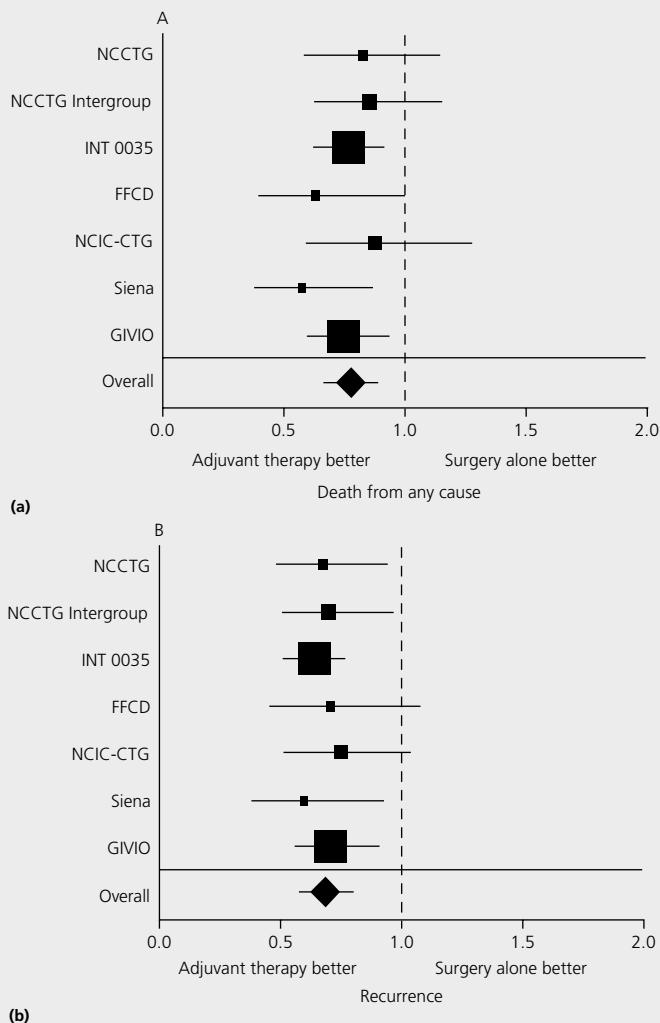
*Random effects* meta-analysis gives wider CIs with a treatment difference that can increase or decrease compared to a *fixed effects* analysis. Random effects analysis is a more conservative approach with more realistic assumptions. It is difficult to envisage a setting where it can be assumed that the trials being combined are simply replicates of each other, which is the assumption that underpins a fixed effects analysis.

### 18.3.4 Graphical methods

An extremely useful addition to the formal method for combining the studies is to represent the data from the individual studies, together with the combination, in a *forest plot*. Example 18.2 displays this kind of plot. Note that the CIs in Figure 18.1 are not symmetric around the estimated hazard ratio. This is because

**Example 18.2** Meta-analysis of adjuvant chemotherapy for resected colon cancer in elderly patients

Sargent et al. (2001) undertook a fixed-effects meta-analysis of seven phase III randomised trials involving a total of 3351 patients that compared the effects of fluorouracil plus leucovorin (five trials) or fluorouracil plus levamisole (two trials) with surgery alone in patients with stage II or stage III colon cancer.



**Figure 18.1** Hazard ratios and 95% CIs for (a) death from any cause and (b) recurrence by treatment group. Source: Sargent DJ, Goldberg RM, Jacobson SD, MacDonald JS, et al. (2001). A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients. *NEJM*, **345**, 1091–1097. Reproduced by permission of Massachusetts Medical Society.

CIs for hazard ratios and ORs – and indeed, ratios in general – are symmetric only on the log scale: once the upper and lower confidence limits are transformed back onto the ratio scale, symmetry is lost. Sometimes we see plots where the  $x$ -axis is plotted on the log scale, although the scale is calibrated in terms of the ratio itself, and in this case, the CIs will appear symmetric.

The studies with the highest precision are those with the narrowest CIs. Usually, these aspects of the different trials are emphasised by plotting squares at the estimated values whose size is related to the precision within that trial. Each square has an area proportional to the weight,  $w_i$ , given to study  $i$  in the meta-analysis. Recall that this is closely related to the sample size of the study. This helps visually to identify which studies provide the most precise information: they have the most prominent squares. The diamond gives the overall combined result.

### 18.3.5 Detecting heterogeneity

A key element of any meta-analysis is to look for heterogeneity across the studies. This is akin to looking for treatment-by-factor interactions in an adjusted analysis. Here, we are looking for treatment-by-study interactions. This can be done by calculating Cochran's Q statistic (Cochran, 1954)

$$Q = w_1(d_1 - d)^2 + w_2(d_2 - d)^2 + \dots + w_m(d_m - d)^2$$

and comparing the resulting value with the  $\chi^2_{m-1}$  distribution, the distribution of this statistic when the different study outcomes are simply a result of random variation, to obtain a  $p$ -value. If the individual studies give very different results so that the  $d_i$  values are not comparable, then, on average, the differences  $d_i - d$  will be large, and  $Q$  will be statistically significant on the  $\chi^2_{m-1}$  scale. A significant  $p$ -value tells us that there is evidence for trial-to-trial heterogeneity. Unfortunately, this test lacks power: even a large amount of heterogeneity can easily go undetected with this test, especially if there are only a small number of studies. Higgins et al. (2003) have developed a related statistic or index,  $I^2$ , that takes values between 0 and 100%, with 0% denoting no heterogeneity and increasing values indicating increasing heterogeneity. If there is no heterogeneity, it can be shown that, on average, the Q statistic equals  $m - 1$  (recall that  $m$  is the number of studies being combined). The  $I^2$  statistic is expressed as a percentage and is related to  $Q$  as follows:

$$I^2 = 100 \times \frac{Q - (m - 1)}{Q}$$

$I^2$  is the proportion of the study-to-study variation that cannot be explained by randomness.

If  $Q$  is close to  $m - 1$ ,  $I^2$  will be close to zero. The convention is to replace negative values of  $I^2$  by zero so the statistic is always in the range 0–100%. If there is a large amount of heterogeneity,  $Q$  is large, and  $I^2$  increases towards 100%. As a guide, Higgins et al. (2003) suggest that values of 25%, 50% and

75% indicate, respectively, low, moderate, and high amounts of heterogeneity. In analyses, it is common to quote  $I^2$  together with the  $p$ -value from the Q statistic but to rely more on the numerical value of  $P$  rather than the  $p$ -value in terms of picking up heterogeneity.

If heterogeneity is detected, it is important to investigate what is causing it. In some studies, such as those recruiting only older patients, there may be a different treatment effect compared to studies recruiting across the full age range. Such considerations may already be built into the overall investigation plan. Typically, studies may be classified according to the characteristics of the patients recruited into the studies, dosage level of the treatment being evaluated, duration of treatment, specific drug if the meta-analysis looks at a class of drugs and so on. In these cases, it may be necessary to appropriately classify the studies and conduct separate meta-analyses, comparing the overall results from the separate meta-analyses and evaluating homogeneity again within each of these.

One possible strategy that has been proposed is to use Cochran's Q-test or the  $P$  statistic (or both) to decide whether to stay with a fixed effects model or use a random effects model where study-to-study variation is built in. Some practitioners use random effects models all the time to allow at least some level of heterogeneity from the outset. In one sense, a fixed effects model is a special case of a random effects model where  $\tau^2$ , the between study variance over and above random variation, is equal to zero. It is still possible to see heterogeneity, even within a random effects meta-analysis, but at a different level. Recall that a random effects model assumes that the study treatment effects are normally distributed around some overall average effect, and the variation of the observed study effects may not align with that expected from a normal distribution. For example, there may be an additional level of heterogeneity, with the studies recruiting only older patients having a different average effect from those recruiting across the full age range. In such cases, it is important to separate these two groups of studies to further investigate heterogeneity, as mentioned earlier.

### 18.3.6 Robustness

Within any meta-analysis, some trials will be larger than others; and because of how the trials are combined, the larger trials – those with higher precision – will tend to dominate. It is helpful, therefore, to assess the robustness of the overall conclusion by omitting the largest study (or studies) to see if the result remains qualitatively the same. If it does, the result is robust. If it does not, the robustness of the result is drawn into question as it is driven by the largest trial (or trials).

### 18.3.7 Rare events

In meta-analyses dealing with safety issues – for example, based on the occurrence of a specific serious adverse event (SAE) – the number of events in the individual trials being combined is often low. At the extreme, there may be no events in one or both groups in a particular trial. If the number of events in the

control group is zero in a trial, then the relative risk and the OR are not defined. If the number of events in the control group is not zero, while the number of events in the experimental group is zero, then the relative risk and the OR are both zero. In either case, it is not possible to include these studies routinely in the analysis when using the relative risk and ORs as summary measures. Using the risk difference is a little better, provided both groups do not have zero counts. Omitting studies with zero counts causes bias. A large study where the event counts are zero in both groups conveys the information that the experimental treatment in that trial is not increasing the incidence of the event, and to exclude that study from the meta-analysis will potentially lead to an overestimation of the extent of the safety problem.

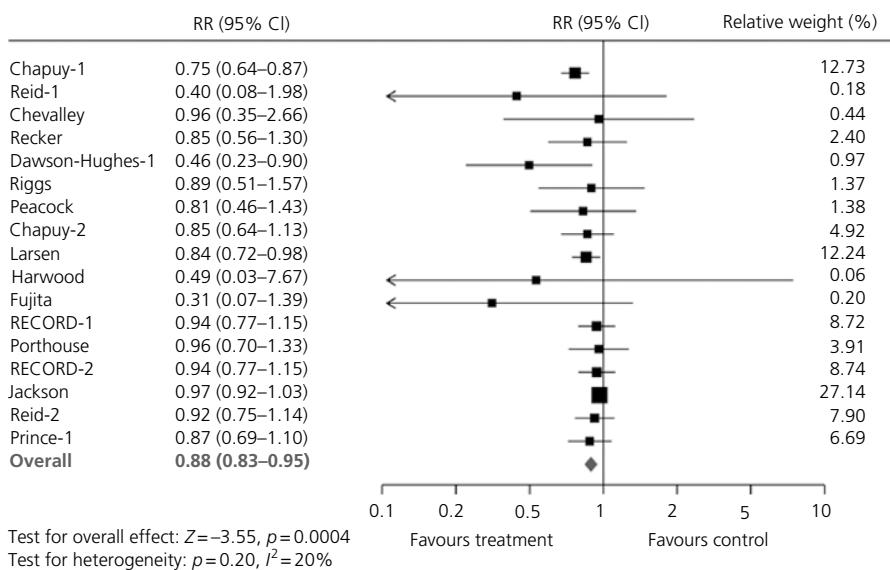
A simplistic way to account for these zero values is to include a continuity correction to the data in that study. This typically involves adding 0.5 to each of the four counts (group 1, patients with the SAE, patients without the SAE; group 2, patients with the SAE, patients without the SAE) in the  $2 \times 2$  contingency table for that trial. This effectively adds a patient to each treatment group and splits them 50:50 across the two outcome categories, event/no event. This is a common, albeit ad hoc, *fudge factor* that enables the maximum amount of information to be included in the final meta-analysis. See Example 19.2 in the next chapter for a use of this method. There are alternative ways of dealing with this problem, but they are beyond the scope of this discussion. The interested reader is referred to Bradburn et al. (2007).

### 18.3.8 Individual patient data

When conducting a retrospective meta-analysis, trials may be combined using summary statistics from publications and summary trial reports without access to the underlying individual patient data. Within a regulatory submission, of course, this is not an issue; but outside of that, it often is. Access to the patient-level data is much better for two main reasons. Firstly, it can help to achieve consistency for the analysis in terms of defining the primary endpoint in a precise way, constructing estimand strategies in a consistent way, using methods for dealing with missing data that are the same for every trial, having a common approach to defining and excluding protocol violators from the analysis and so on. Secondly, it facilitates more sophisticated methods of analysis using regression models that enable a wider range of approaches to be considered for the analysis. Having access to individual patient data is always better; there are no exceptions to this rule.

## 18.4 Case study

It will be useful to review the methods described in this section through a case study. This case study will also be of value in cementing some of the ideas that are to follow. Tang et al. (2007) report a random effects meta-analysis looking at



**Figure 18.2** Forest plot displaying relative risks and 95% CIs for the occurrence of fractures in studies contained in the meta-analysis. Source: Tang BMP, Eslick GD, Nowson C, et al. (2007). Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis. *The Lancet*, **370**, 657–666. Reproduced with permission from Elsevier.

using calcium or calcium plus vitamin D supplementation to prevent fractures and bone loss in people over 50 years of age. The 16 randomised controlled trials that were eventually included in the meta-analysis for the binary endpoint of fracture (yes/no) are displayed in Figure 18.2 and combined in terms of the relative risk for fractures. Note that one trial, the RECORD study, is included twice. There were four treatment groups in this study, including calcium alone, calcium plus vitamin D and placebo. The two relative risks included in the analysis were for the calcium vs. placebo comparison and the calcium plus vitamin D vs. placebo comparison. In fact, including the placebo group twice is incorrect. The relative risks being combined are not independent since the placebo group appears in both, which violates one of the assumptions on which the methods for combination are based. Simply combining the two active groups, giving a single calcium or calcium plus vitamin D group, and calculating a single relative risk would have been a better approach. Despite this shortcoming, we will use this case study to highlight some of the methodological points already made. It is worth noting that the authors were able to obtain individual patient data and did not have to rely on summary statistics from the trials on which to base their meta-analysis.

The  $x$ -axis is on the log scale; consequently, the CIs displayed are symmetric on that non-linear scale. The overall relative risk is 0.88 with a 95% CI of

0.83–0.95 and a highly statistically significant  $p$ -value of 0.0004. There is strong evidence from this analysis that calcium and calcium in combination with vitamin D prevents fractures in this over-50s population. Heterogeneity was assessed using Cochran's Q statistic, which gave a non-significant  $p$ -value of 0.20, and the value of  $I^2$  was 20%, suggesting a low level of heterogeneity. These authors also defined a set of 12 variables for subgroup analyses and divided the data according to, for example, sex, previous fractures (yes/no), clinical setting (community/institutionalised), age (50–69, 70–79,  $\geq 80$ ) and so on. They then undertook separate meta-analyses within each subgroup, producing separate relative risks and a  $p$ -value for treatment-by-factor interactions based on comparing those relative risks across the different age groups. Several of these were statistically significant. For example, the relative risks for the three age groups were 0.97 (age 50–69), 0.89 (age 70–79) and 0.76 (age  $\geq 80$ ), with a  $p$ -value for interaction of 0.003. This is a quantitative interaction (see Section 5.4.2): in each of the age groups, calcium/calcium plus vitamin D had a beneficial effect according to the relative risks, although the effect was greater for the older patients. There were several other statistically significant treatment-by-factor interactions, but all were quantitative, and calcium and calcium in combination with vitamin D were seen to have a consistent beneficial effect, at least directionally, across the population.

## 18.5 Ensuring scientific validity

### 18.5.1 Planning

To ensure that a meta-analysis is scientifically valid, it is necessary to plan and conduct the analysis appropriately and rigorously. It is not sufficient to retrospectively go to a bunch of studies that you like the look of and stick them together!

Ideally, the meta-analysis within a regulatory submission should be pre-planned as part of the development plan for the complete programme, and the rules regarding which trials are to be combined and in what way should be set down in advance of running the trials.

The CPMP (2001) *Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study* indicates that it is good practice to write a protocol for the meta-analysis:

*'When a meta-analysis is included in an application it should be performed in accordance with a protocol . . .'.*

This document then goes on to list the issues that should be covered by that protocol:

- Objective of the meta-analysis
- Criteria for inclusion/exclusion of studies

- Hypotheses and endpoints
- Statistical methods
- Approaches to ensure consistent quality of the studies and how to handle poor quality studies
- Evaluating homogeneity and robustness

A meta-analysis addressing an emerging safety problem or a general therapeutic question also requires careful planning. The first step in any such analysis is clearly defining the research question. For the case study outlined in Section 18.4, the research question was to assess ‘the effect of calcium, or calcium in combination with vitamin D supplementation, on osteoporotic fractures and bone-mineral density, in adults aged 50 years and older’. Having set down a clear research question, the second step is to specify search criteria to find the trials that are potentially relevant to that question and define study eligibility criteria, methods of data extraction and, finally, the methods for statistical analysis.

The search criteria will involve searching electronic databases such as Medline, Embase and the Cochrane Database of Systematic Reviews while also looking for unpublished and ongoing trials. The study eligibility criteria often restrict trials to be included to those with a control group, those giving treatment at a certain target dose over an appropriate treatment period as well as those specifying in a precise way the patient population being studied. Data extraction is often undertaken by at least two individuals independently, with disagreements resolved through discussion.

The research question usually sets down the endpoints of interest, at least approximately. Specifying the statistical methods involves clearly defining those endpoints and the summary measures to be used for measures of treatment effect. Further points that need consideration include whether the analysis will use fixed effects or random effects (or both), how to assess heterogeneity and robustness, the extent of the subgroup analyses and how to deal with any specific problems such as rare events.

### 18.5.2 Assessing the risk of bias

Many publications based on meta-analysis talk about the methodological quality of the trials to be included in the analysis, possibly excluding those that fail to meet certain quality standards. In the past, it has been common to use checklists to assess quality. More recently, the emphasis has been on assessing the risk of bias. The Cochrane Collaboration developed a tool for assessing the risk of bias (Higgins et al., 2011); it is summarised in Table 18.2.

Various forms of bias are listed in this table, with a description of their source. In evaluating the potential for bias, each of these items is assessed as being low risk, high risk or risk unclear. Trials with a high risk of various biases may well be excluded from the meta-analysis. Further, and as part of the evaluation of robustness, it might also be appropriate to exclude some trials based on a potential for bias in sensitivity-type analyses.

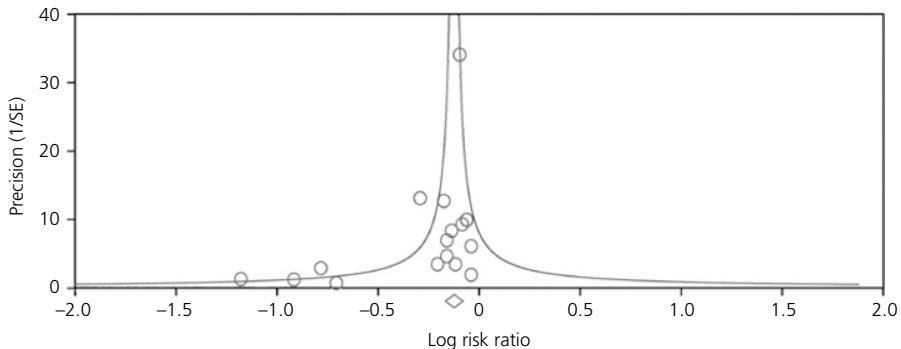
**Table 18.2** Cochrane Collaboration's tool for assessing risk of bias

Bias domain	Source of bias	Assess as low, unclear or high risk of bias
Selection bias	Random sequence generation	Inadequate generation of a randomised sequence
Selection bias	Allocation concealment	Inadequate concealment of allocations before assignment
Performance bias	Blinding of participants and personnel	Knowledge of allocated interventions by participants and personnel during study
Detection bias	Blinding of outcome assessment	Knowledge of the allocated interventions by personnel assessing outcome
Attrition bias	Incomplete outcome data	Amount, nature or handling of incomplete outcome data
Reporting bias	Selective reporting	Selective outcome reporting
Other bias	Anything else	Bias due to problems not covered elsewhere

### 18.5.3 Publication bias and funnel plots

One particular issue concerns meta-analyses based on data obtained through a literature search. It is certainly true that a study that has given a statistically significant result is more likely to be reported and accepted for publication. So, if we only focused on published studies, we would get a biased view of the totality of the available studies. Eggar and Smith (1998) discuss various aspects of this *publication bias* and its causes. There have been many calls over the years for registries of studies to be set up. In early 2005, the European Federation of Pharmaceutical Industries and Associations (EFPIA), the International Federation of Pharmaceutical Manufacturers & Associations (IFPMA), the Japan Pharmaceutical Manufacturers Association (JPMA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) issued a joint statement committing to increasing the transparency of research by setting up a Clinical Trial Registry. Although it is a voluntary initiative, most companies are following this guidance. The registry is maintained by the National Library of Medicine in the USA and can be found at [www.clinicaltrials.gov](http://www.clinicaltrials.gov). In theory, the registry makes it possible to identify all studies sponsored by the industry and potentially avoid any publication bias.

A graphical technique introduced by Eggar et al. (1997), called a *funnel plot*, helps to detect the presence of publication bias by plotting the treatment effect (e.g. the difference in the means or the OR) in each study on the x-axis against the precision as measured by  $1/se$  on the y-axis. Smaller studies tend to give more variable results in terms of the observed treatment difference, while larger studies should give more consistent results. The resultant plot with all studies included should then appear like a funnel, with the wide part of the funnel at the bottom and the narrow part of the funnel at the top of the plot. However, if there is publication bias, the non-significant studies – very often those with smaller sample sizes (less precision) – will be under-represented, and either the



**Figure 18.3** Funnel plot for meta-analysis of trials. Source: Tang BMP, Eslick GD, Nowson C, et al. (2007). Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis. *The Lancet*, **370**, 657–666. Reproduced with permission from Elsevier.

lower left-hand part of the plot or the lower right-hand part of the plot will be missing, depending on how the active compared to placebo difference is measured. Once this has been detected, it can be compensated for in the statistical analysis. Visually inspecting the funnel plot this way is somewhat informal, although the technique can provide some reassurance regarding the absence of publication bias. It is not a substitute however for relentlessly tracking down all studies and all data.

Figure 18.3 is a funnel plot for the case study described in Section 18.4 (Tang et al., 2007). Note that the funnel is sketched on for guidance; this is not the result of a curve-fitting exercise. Inspection of the plot suggests an absence of small studies with a log risk ratio above zero. These would be studies where the risk ratio is greater than 1, in which calcium and calcium plus vitamin D supplementation increases the fracture risk. The key issue now is, does it make a difference? If such studies – presumably, conducted but not published – were included, would that negate the overall meta-analysis result? One way to address this point is to calculate the *fail-safe number* (Rosenthal, 1979): the number of existing studies with a risk ratio of 1 (or a risk ratio giving a statistically significant result against calcium and calcium plus vitamin D) that would be needed to negate the statistically significant result in favour of calcium and calcium plus vitamin D in the meta-analysis. In the case study, approximately 100 studies with a risk ratio of 1 or 22 studies with a statistically significant negative result for calcium and calcium plus vitamin D would be needed to negate the overall effect. Tang et al. (2007) conclude that it is very unlikely that such a large number of unpublished studies exist and that therefore, their results were not materially affected by publication bias.

#### **18.5.4 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)**

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021) provides recommendations for reporting a meta-analysis in the form of a checklist. Although this statement does not directly talk about the planning and conduct of a meta-analysis, it does say a lot about these issues. For example, the statement talks in detail about searches, providing a clear rationale for conducting the meta-analysis, including specifying the research question, describing the study selection or eligibility criteria, data extraction and assessing the risk of bias. Anyone planning to conduct a meta-analysis should read this statement in detail.

### **18.6 Regulatory aspects of meta-analysis**

In a regulatory setting, a pre-planned meta-analysis will always be more convincing. Often, however, a meta-analysis is envisaged either part way through a development programme or at the end of the trial activity when the submission is being put together. It is interesting to note that within the regulatory context, meta-analysis has frequently caused problems for regulators:

*'Meta-analysis has long been a source of regulatory discomfort, mostly because of the poor quality of some meta-analyses submitted in applications for licences'. (Lewis, 2002)*

However, the regulators recognise that pre-planning is not always possible:

**CPMP (2001): 'Points to consider on application with 1. meta-analysis; 2. one pivotal study'**

*'A retrospective specification when the results from all or some of the studies are known should be avoided . . . There are, however situations where the need for a meta-analysis becomes apparent after the results from some or sometimes all studies are known. This is the case when there is a need to put seemingly conflicting results into perspective, or in the exceptional situation where a meta-analysis seems to be the only way to provide reliable proof of efficacy'.*

Even in the retrospective setting, it is still important to write a protocol so that the meta-analysis can be performed as objectively as possible. The CPMP 'Points to Consider' paper lists the prerequisites for such a retrospective analysis.

**CPMP (2001): 'Points to consider on application with 1. meta-analysis; 2. one pivotal study'**

*'Prerequisites for a retrospective meta-analysis to provide sufficient evidence for a claim include:*

- Some studies clearly positive
  - Inconclusive studies showing positive trends in the primary variable
  - No statistically significant heterogeneity
  - Pooled 95 per cent confidence interval well away from zero (or unity for odds ratios, or the pre-defined margin for non-inferiority trials)
  - A justification that a biased selection of studies and/or endpoints is unlikely
  - A sensitivity analysis demonstrating robustness of the findings
- For meta-analyses where these requirements are not fulfilled it will prove difficult to get a regulatory acceptance'.*

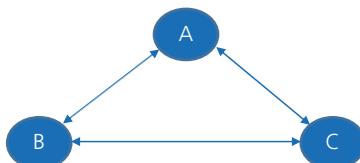
Meta-analysis was a component of the regulatory submission for Valdoxan (agomelatine) in the treatment of major depressive disorder. It is worthy of note that of the six pivotal placebo-controlled trials included in the meta-analysis, not all were positive in the sense of showing statistically significant differences: three were positive, and three failed to reach statistical significance:

*'In a meta analysis including six of the pivotal studies (i.e. all the placebo-controlled short-term studies) whether positive (CL2-014, CL3-042, CL3-043) or not (CL3-022, CL3-023, CL3-024), with 1210 patients receiving agomelatine 25 and 50 mg but also sub-therapeutic doses, 1 and 5 mg, and 805 patients receiving placebo, an overall treatment effect of about 1.5 on the HAM-D in favour of agomelatine over placebo was observed. Using a model with study included as a fixed effect, the estimated difference was 1.51 with a 95% confidence interval of [0.80, 2.22]. When the study was modelled as a random effect, the estimated difference was 1.55 [0.61, 2.48]'.*  
 (CHMP 'Assessment Report for Valdoxan': Procedure No.EMEA/H/C/000915  
[https://www.ema.europa.eu/en/documents/assessment-report/valdoxan-epar-public-assessment-report\\_en.pdf](https://www.ema.europa.eu/en/documents/assessment-report/valdoxan-epar-public-assessment-report_en.pdf))

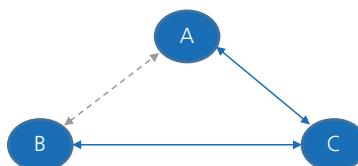
## 18.7 Introduction to network meta-analysis

A standard meta-analysis combines several head-to-head studies as the basis for a treatment comparison. However, there may be only a small number of trials that can provide the basis for such a comparison, and there could be a need to bring in trials that do not compare two treatments directly but compare each of those treatments to a common third treatment that can then be used as a pivot to gain some understanding of the relative merits of the two treatments of interest. In some circumstances, there may indeed be no direct comparisons between the two treatments of interest, yet there is still a need to get some sense of their relative efficacy, tolerability and safety. *Network meta-analysis* (NMA) extends the meta-analysis concept to provide treatment comparisons for a network of treatments based on both direct and indirect evidence.

Figure 18.4 shows these two settings. In Case 1, we are supplementing the direct treatment A vs. treatment B comparisons with treatment A vs. treatment C and treatment B vs. treatment C comparisons, where C acts as the common factor that enables this to happen. In Case 2, there are no direct treatment A vs.

**Case 1: Direct A vs. B comparison supplemented by A vs. C and B vs. C comparisons**

**Use data from A vs. C and B vs. C as indirect evidence to supplement A vs. B comparison**

**Case 2: No direct A vs. B comparison**

**Use data from A vs. C and B vs. C to indirectly compare A vs. B**

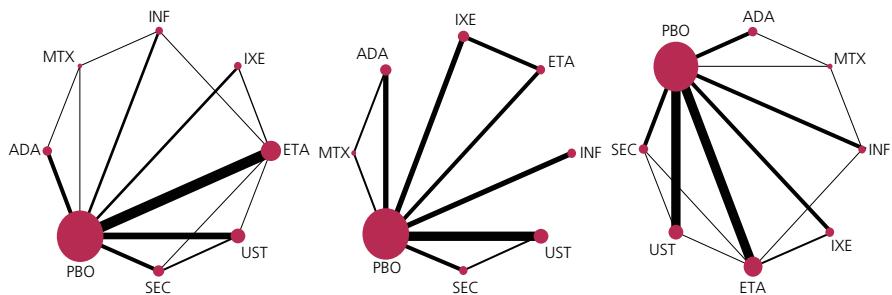
**Figure 18.4** Direct and indirect comparisons

treatment B comparisons, and the comparisons A vs. C and B vs. C provide the only sources of data on which to judge the relative effects of treatments A and B.

In general, an NMA involves combining studies based on published summary statistics from those studies; rarely will it involve the analysis of individual patient data, due to the confidentiality of those data. This limits the potential for a thorough exploration of the data and our ability to drive consistency in the populations of patients under study, the definition of endpoints, methods of statistical analysis and the treatment of missing data. Other concerns relate to the validity of indirect comparisons and assumptions regarding treatment effect modifiers in the population of patients under study. We will discuss this final point later in the chapter.

## 18.8 Case Study

Jabbar-Lopez et al. (2017) report a systematic review and NMA of biologic therapy options for psoriasis. Seven licensed medications were identified and evaluated for efficacy, quality of life and tolerability/safety. The efficacy outcome measure was the binary endpoint clear/nearly clear (Psoriasis Area and Severity Index [PASI] > 90 or 0 or 1 on Physicians Global Assessment [PGA]). Quality of life (QoL) was measured based on the change from baseline in the Dermatology Life Quality Index (DLQI), and tolerability/safety was evaluated based on withdrawal due to adverse events. The duration of treatment in all trials under consideration was 12 to 16 weeks, and the efficacy and QoL measures related to the end of the planned treatment duration. Figure 18.5 is a *network map* displaying the treatments being evaluated and their connections to other treatments in the network. Each treatment is displayed as a *node* with *edges* connecting those treatments that were compared directly. The therapies considered were all compared to placebo in at least one randomised trial and in some cases compared to other



**Figure 18.5** Network maps for outcome measures considered in an NMA of biologic therapy options for psoriasis. Source: Jabbar-Lopez et al., 2017. Reproduced by permission of Elsevier.

biologic therapies. There were randomised comparisons: for example, for etanercept with infliximab, ixekizumab, ustekinumab and secukinumab in addition to placebo for efficacy and tolerability/safety, but only with ixekizumab and placebo for QoL. The nodes and edges are weighted according to the number of trials in which each specific treatment was involved.

The network map provides an essential display of the structure of a network.

## 18.9 Indirect treatment comparisons

### 18.9.1 Cross-trial calculations

Suppose we want to compare the response rates of two treatments and think in terms of the structure in Case 2 in Figure 18.4. For the trials comparing treatment A with treatment C, we have a difference in response rates  $r_A - r_C$ . Similarly, for the trials comparing B with C, we have a difference in response rates  $r_B - r_C$ . The difference  $r_A - r_B$  is then obtained using

$$r_A - r_B = (r_A - r_C) - (r_B - r_C)$$

The response rate on treatment C has essentially been cancelled out in the calculation. Clearly, some strong assumptions underpin the validity of this calculation; we will come to those shortly. Had we been expressing treatment differences through a ratio such as an OR, the calculation would have been slightly different:

$$\text{OR}_{A/B} = \text{OR}_{A/C} / \text{OR}_{B/C}$$

These calculations are the basis of how we extract information from the indirect comparisons. It is straightforward to obtain standard errors for these comparisons based on the following:

$$SE(r_A - r_B) = \sqrt{SE^2(r_A - r_C) + SE^2(r_B - r_C)}$$

where  $SE(r_A - r_C)$ , for example, is calculated from those trials comparing treatment A with treatment C.

For the OR and other ratio measures (relative risk, hazard ratio, rate ratio), the se and CI are initially calculated on the log scale, with the confidence limits converted back onto the ratio scale for presentation. When there is a mixture of direct and indirect comparisons, they are combined through a model to produce an overall comparison.

### 18.9.2 Effect modifiers

There are two situations where an indirect comparison produces an unbiased estimate of the true treatment difference between treatments A and B:

- 1 *The studies that are being combined are similar, on average, in all important factors that are predictive of the outcomes under consideration – known as transitivity.*

*or*

- 2 *There are no effect modifiers: the magnitude of treatment differences is similar across the population under study for the treatments that are being compared.*

To understand the impact of any violation of these conditions, consider those data displayed in Table 18.3. Case 1 presents hypothetical data on three treatments in terms of the response rates according to whether a patient is high risk or low risk. For these data, there are no effect modifiers. Treatment A is better than treatment B by 10%, treatment A is better than treatment C by 20% and treatment B is better than treatment C by 10%; these differences are unaffected by baseline risk. In Case 2, the data have been modified slightly, and treatment differences for A vs. B and for A vs. C are impacted by baseline risk.

Suppose we have two trials. Trial 1, which consists only of high-risk patients, compares A with C; and trial 2, which has recruited 50% high-risk and 50% low-risk patients, compares B with C. Suppose there is no trial comparing A with B, and we are looking to obtain an indirect comparison between A and B. Considering Case 1, the difference between treatments A and C in trial 1 is 20%, and the difference between treatments B and C in trial 2 is 10%. The indirect comparison between treatments A and B is then

$$r_A - r_B = (r_A - r_C) - (r_B - r_C) = 20\% - 10\% = 10\%$$

By inspection of the first two rows of Table 18.3, this is clearly the correct answer.

Consider now Case 2. The indirect treatment difference given by the previous expression remains at 10% since trial 1 (A vs. C) contains only high-risk patients, while for trial 2, the difference (B vs. C) is unaffected by baseline risk and is equal to 10%. Is this correct? Well, no, it is not. The indirect 10% difference for A vs. B only applies to the high-risk patients; for the low-risk patients, the difference is 20%. Thinking about the mix of patients who make up the indirect A vs. B comparison, it could be argued that by combining the two trials,

**Table 18.3** Indirect comparisons for three treatment groups (hypothetical data)

Case 1: No effect modifiers		
Response rates	High-risk patients	Low-risk patients
Drug A	60%	70%
Drug B	50%	60%
Drug C	40%	50%
A vs. B	10%	10%
A vs. C	20%	20%
B vs. C	10%	10%

Case 2: Effect modifiers		
Response Rates	High-risk patients	Low-risk patients
Drug A	60%	80%
Drug B	50%	60%
Drug C	40%	50%
A vs. B	10%	20%
A vs. C	20%	30%
B vs. C	10%	10%

there is a mix of 75% high-risk and 25% low-risk, for which the A vs. B difference should be 12.5% ( $0.75 \times 10\% + 0.25 \times 20\%$ ). The indirect comparison, which has given a calculated difference of 10%, is misleading.

It can be seen from this example that violation of condition 2 results in biased estimates of treatment differences. It is also worth pointing out that had condition 1 been satisfied – for example, with a 50/50 mix of high-risk and low-risk patients in the two trials – things would have worked out fine for Case 2, even though there are treatment effect modifiers. We would have seen a 25% difference in trial 1 (A vs. C) and a 10% difference in trial 2 (B vs. C), giving a calculated A vs. B difference of 15%, which is correct for the mix of patients.

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) (Janssen et al., 2014) give guidance on how to critique an NMA and comments as follows on the issue of effect modifiers:

*'Although the indirect comparison or network meta-analysis is based on RCTs, randomisation does not hold across the set of trials used for the analysis because patients are not randomised to different trials. As a result, systematic differences in the distribution of patient characteristics across trials can ensue. In general, if there is an imbalance in study and patient characteristic-related effect modifiers across the different types of direct comparisons in a network meta-analysis, the corresponding indirect comparisons are biased'.*

### 18.9.3 Critique

As mentioned at the end of the previous section within the ISPOR quotation, comparisons based on an NMA are not randomised, and balance in baseline characteristics is not guaranteed. This contrasts with a meta-analysis, which does

respect randomisation by calculating treatment differences at the study level and combining those differences. See Section 18.2 for further discussion on this point in relation to meta-analysis and pooling.

If both direct and indirect comparisons between two treatments are available, there is the possibility to compare the two sources of evidence. If these agree, we talk in terms of *coherence*. But if they disagree, we have *incoherence*, which undermines the validity of combining these two sources of evidence as the basis for an overall treatment comparison. When incoherence is present, the direct comparisons provide more reliable data on which to base conclusions. Chou et al. (2006) report on a particular example where the network lacked coherence:

*'In the direct meta-analysis, NNRTI-based regimens were better than PI-based regimens for virological suppression (OR 1·60, 95% CI 1·31–1·96). . . . By contrast, in indirect analyses NNRTI-based HAART was worse than PI-based HAART for virological suppression (OR 0·26, 95% CI 0·07–0·91)'.*

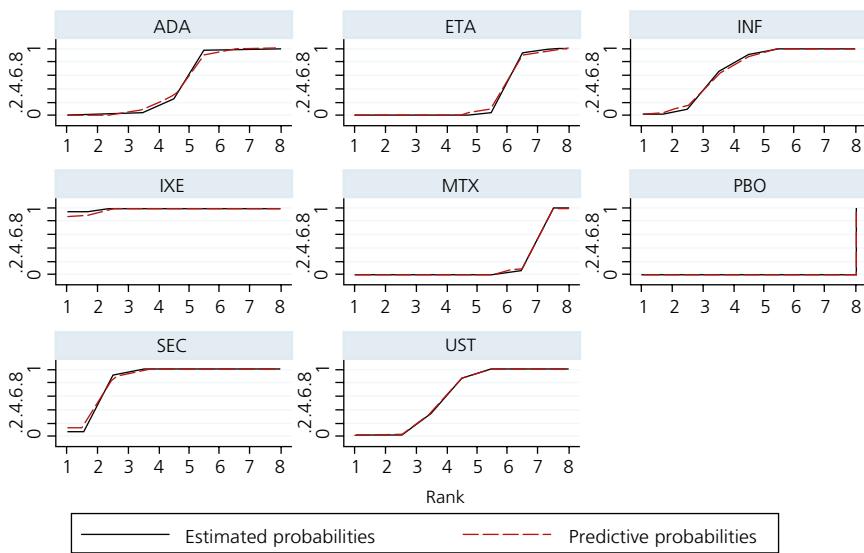
These authors go on to conclude:

*'Our study accords with observations by other investigators that head-to-head trials are the preferred method for investigating the comparative effectiveness of interventions and suggests that indirect comparisons could be particularly unreliable for complex and rapidly evolving interventions such as HAART. Because of the potential for clinically significant discrepancies between indirect and direct treatment comparisons, providers should be cautious about making treatment decisions based on indirect analyses, which should always be verified as sufficient direct evidence becomes available'.*

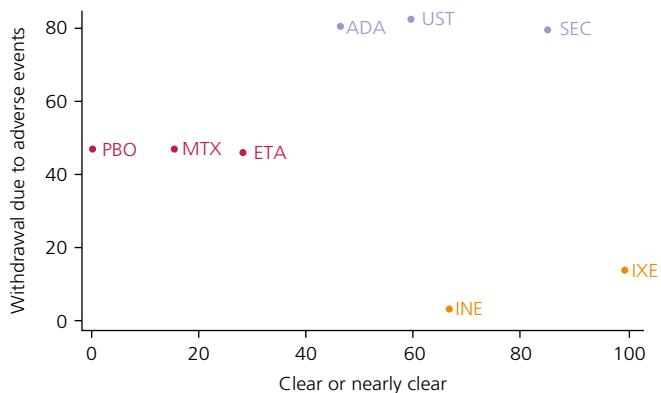
## 18.10 Bayesian rank analysis

Often, the main objective of an NMA is to rank the  $m$  treatments under consideration from best to worst in relation to each of the outcome measures being considered. This ranking is often done in a Bayesian framework. The prior assumption is that each treatment has an equal probability of being ranked 1 to  $m$  for each outcome measure or, equivalently, that the difference between each treatment's mean outcome (or proportions, if the outcome is binary) is zero. The data are then used to compute posterior distributions that each treatment ranks first, second, third and so on.

These are the *rank probabilities*: they can be plotted in a cumulative ranking probability plot as shown in Figure 18.6 for the case study in psoriasis (Jabbar-Lopez et al., 2017) for the outcome clear/nearly clear at 12/16 weeks. The probabilities based on the Bayesian approach are labelled 'predictive probabilities' in the plots. The plots suggest that ixekizumab followed by secukinumab are the two best treatments in the network for efficacy. A good summary measure associated with these curves is the area under the curve, called the *surface under the cumulative ranking curve* (SUCRA).



**Figure 18.6** Cumulative ranking probability plot for efficacy outcome at 12/16 weeks.  
Source: Jabbar-Lopez et al., 2017. Reproduced by permission of Elsevier.



**Figure 18.7** Bivariate plot of SUCRA for efficacy and safety/tolerability. Source: Jabbar-Lopez et al., 2017. Reproduced by permission of Elsevier.

Although ixekizumab looks to be the best treatment for efficacy, it does not rank so highly for safety/tolerability and QoL. Bivariate plots of the SUCRA values – for example, for efficacy and safety/tolerability, as shown in Figure 18.7 (Jabbar-Lopez et al., 2017) – help to shed light on the trade-off between these two measures.

There appear to be three clusters of treatments, with ixekizumab and infliximab doing well on efficacy but performing poorly for safety/tolerability. Adalimumab, ustekinumab and secukinumab perform well for efficacy and have few safety/tolerability concerns. Finally, methotrexate and etanercept perform poorly for efficacy but well for safety/tolerability. Broad conclusions can then be drawn from these considerations.

## CHAPTER 19

# Methods for safety analysis, safety monitoring and assessment of benefit-risk

### 19.1 Introduction

#### 19.1.1 Methods for safety data

The majority of statistical methods we have developed and presented in this book have been focused on the evaluation of efficacy data. In this chapter, we will consider methods related to analysing safety and tolerability data. We will start by looking at the routine analysis and presentation of safety data within a single clinical trial and more broadly within the regulatory package. We will then discuss methods for safety monitoring and pharmacovigilance post-approval. There will also be a discussion of methods for evaluating the benefit-risk balance.

In some circumstances, there may be specific safety and tolerability parameters within the confirmatory structure of the trial. For example, in an oncology study, we may be interested in assessing the potential for a drug to reduce the rate of Common Terminology Criteria Adverse Events (CTCAE) grade 3/4 neutropenia ([https://ctep.cancer.gov/protocoldevelopment/electronic\\_applications/docs/ctcaeversion3.pdf](https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcaeversion3.pdf)) associated with a standard chemotherapeutic regimen. This could be the trial's main focus or at least an important secondary consideration. In this case, there is a specific hypothesis to test, and this binary endpoint (grade 3/4 neutropenia, yes/no) would be incorporated into the confirmatory testing strategy and treated the same as an efficacy endpoint with formal control of multiplicity within that strategy. But this is not what we are discussing primarily in this chapter. Here, our attention will be on gaining an overview of the totality of the safety information to identify specific safety signals and concerns to allow evaluation of the benefit-risk ratio both at the time of approval and beyond. *P*-values in this setting convey little information and, indeed, can be misleading and are not recommended. Non-significant *p*-values may simply be because the overall event rate for a particular adverse event (AE) is low; hence, the power to detect differences is small. But even a rare, very serious AE can have major implications if it occurs more frequently in the experimental group. Statistically significant *p*-values, on the other hand, may be a consequence of multiplicity. We could well be looking at

several hundred distinct AEs together with other safety parameters, and significant *p*-values will inevitably occur purely by chance. We do sometimes calculate 95% confidence intervals (CIs), but these are not to be used to make formal inferences in the statistical significance sense. These intervals are there to provide information on the potential magnitude of effects that are supported by the data.

***ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'***

*'In most trials the safety and tolerability implications are best addressed by applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals wherever this aids interpretation. It is also valuable to make use of graphical presentations in which patterns of adverse events are displayed both within treatment groups and within subjects'.*

Note the final point here on the use of graphical presentations; this will be a particular focus in this chapter.

We will use the following terms in our discussions:

- *Safety* – The medical risk of the product (drug, device, treatment) to the subject
- *Tolerability* – The degree to which the subject can tolerate the adverse effects of a product (drug, device, treatment)
- *Benefit* – The established therapeutic efficacy of a product (drug, device, treatment)
- *Risk* – The probability of being harmed by the product (drug, device, treatment)
- *Harm* – The extent of damage caused by the product (drug, device, treatment)

### **19.1.2 The rule of three**

ICH E1 makes several statements about what would normally be expected in terms of characterising AEs within the clinical trial programme through to phase III.

***ICH E1 (1995): 'Population Exposure: The Extent of Population Exposure to Assess Clinical Safety'***

*'It is expected that short-term event rates (cumulative 3-month incidence of about 1%) will be well characterized. . . . The safety evaluation during clinical drug development is not expected to characterize rare adverse events, for example, those occurring in less than 1 in 1000 patients'.*

Suppose that the incidence of a particular AE occurring with a certain drug is 1%. If we were to study  $n = 300$  patients receiving that drug, then the probability that we would see no events is  $0.99^{300} = 0.049$ . It follows that the probability that we would see at least one patient suffering the event is  $1 - 0.049 = 0.951$ , around 95% in percentage terms. This type of calculation leads to the probabilities in Table 19.1.

**Table 19.1** Number of patients needed according to incidence rate

Probability	Incidence				
	1%	0.5%	0.1%	0.05%	0.01%
95%	300	600	2995	5990	29,956
80%	161	313	1608	3128	16,094

The probabilities in this table give measures of our ability to detect an AE according to its incidence. So, for example, if we are dealing with a rare AE with an incidence of only 0.1% (1 in 1000), we will need a trial (or trials) recruiting 2995 (almost 3000) patients to have a 95% probability of detecting its presence. Turning this around, it follows that if in a trial of 3000 patients, we do not see a certain AE, we can be 95% confident that its incidence is less than 1 in 1000. This leads to the *rule of 3* (Jovanovic and Levy, 1997); if we fail to see a particular AE in a trial with  $n$  patients, we can be 95% confident that the incidence of that AE is less than  $3/n$ .

The recommendations coming out of the ICH E1 guidance for the size of the safety database are as follows:

- It is usually sufficient to have safety data on 300 to 600 patients treated for 6 months. This should be adequate to characterise patterns of AEs over this time period,
- To gauge longer-term AEs, safety data are needed on 100 patients for 12 months. According to the rule of 3, if a particular AE is not seen, we have reasonable assurance (95%) that the incidence is less than 3/100 (3%).

Further details on these and other recommendations are contained in that guidance.

## 19.2 Routine evaluation in clinical studies

Within the context of a single clinical trial, the analyses of safety and tolerability are usually based on the safety set. In Section 7.2, we defined the full analysis set (FAS) and the per-protocol set, and these were the analysis sets on which analyses of efficacy were undertaken. See also the discussions in Chapter 8 on the choice of population and in Chapter 9 on additional considerations associated with estimands. The *safety set* is defined as the set of subjects who have received at least one dose of the study medication.

### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'For the overall safety and tolerability assessment, the set of subjects to be summarized is usually defined as those subjects who received at least one dose of the investigational drug'.*

This analysis set will on many occasions coincide with the FAS, but not always. Further, the make-up of the treatment groups in the FAS may differ from the make-up in the safety set. For example, a subject who is randomised to receive treatment A but receives treatment B by mistake could well be included in group A for the analysis of efficacy to comply with the randomisation but would be included in group B for the analysis of safety and tolerability.

### 19.2.1 Types of data

The kinds of data we collect to evaluate the safety and tolerability of a product include AEs and serious adverse events (SAEs), data on laboratory parameters, ECGs, and vital signs.

AEs and SAEs are classified based on the Medical Dictionary for Regulatory Activities (MedDRA). This dictionary is organised by System Organ Class (SOC) and then divided into High-Level Group Terms, High-Level Terms, Preferred Terms (PTs) and Lower-Level Terms. Data presentations in tabular form usually only use the SOC and PT as a basis for classification. In some cases, *AEs of special interest* may be identified as potentially associated with the drug/disease under study, possibly because of early phase data for the drug or simply because of experience of drugs within that same class. These need to receive special attention, and this is one area where formal inference may be associated with a pre-specified hypothesis and the calculation of a *p*-value could be justified. There is a substantial clinical overlap across individual PTs in the MedDRA system, and some grouping of the terms may be required to get a full picture of a particular area of concern. However, we will not address this grouping issue further in our discussions.

In all clinical trials and development plans, there is the routine collection of a range of laboratory parameters, both clinical chemistry and haematology in serum, with associated normal ranges, and possibly some parameters obtained via a urinalysis. Normal ranges can be age and gender specific and can also differ according to the laboratory performing the testing. The list of parameters measured depends on the disease setting. A specific issue of concern with all drugs is the potential for liver damage as indicated by values for the liver transaminases, alanine transaminase (ALT) and aspartate transaminase (AST). Raised levels of these parameters can be a precursor of acute liver damage and failure. *Hy's law* is now widely used to signify altered liver function with the potential to cause a fatal drug-induced liver injury (DILI) and is defined in the FDA guideline (FDA, 2007) on Drug-Induced Liver Injury as cases where there is

- A threefold or greater elevation above the upper limit of the normal range (ULN) for ALT or AST
- Elevation of the serum total bilirubin of twofold or greater than the ULN without any findings of cholestasis (serum alkaline phosphatase  $>2 \times$  ULN)

- No other reason found for these increases such as the presence of hepatitis or pre-existing liver disease

Electrocardiography (ECG or EKG) provides information on the electrical activity of the heart, and the QT interval, when prolonged, is associated with increased risk of ventricular arrhythmia and potentially sudden death. The QT interval is usually corrected for heart rate and denoted  $QT_c$ . There are more sophisticated methods of correction: Bazett's formula ( $QT_B$ ) and Fridericia's formula ( $QT_F$ ). The QT interval, corrected for heart rate, has separate normal ranges for males and females, but in data analysis, the change from baseline is generally the focus. Increases of 30 ms and 60 ms are considered levels for concern.

Finally, vital signs include body temperature, heart rate, systolic and diastolic blood pressure (sBP and dBP) and possibly respiratory rate. Acceptable ranges for each of these parameters are available and vary with age, although the ranges tend not be used formally for analysis in the same way as normal ranges for laboratory parameters are.

### **19.2.2 Adverse events**

AEs and SAEs are usually summarised by SOC and PT in tables reporting incidence (both absolute and percentage). Incidence is expressed as the number of subjects suffering the event divided by the number of subjects in that treatment group. Note that incidence is not usually expressed in terms of the number of events divided by the number of subjects; if a subject suffers two headaches during the trial, this subject will only count once in the calculation of incidence. These incidences are calculated for the individual PTs but are also totalled for each SOC. At that level, though, there is the potential for some double counting. For example, the SOC gastrointestinal disorders includes the events nausea and vomiting as separate PTs. If a patient suffers three bouts of nausea and vomiting on two occasions, they will be counted once for the PT nausea, once for the PT vomiting and once for the SOC gastrointestinal disorders.

It is also common to produce similar tables for *AEs related to study medication* according to the information on relationship to treatment provided by the investigator and *severe AEs* according to predefined levels of severity.

If the duration of treatment differs between the treatment groups, the simple calculation of incidence may not be a sound basis on which to compare treatments in relation to AEs. Duration may differ by design: for example, one treatment given for a fixed period while the other treatment continues through to the end of the study. This differential exposure can happen in oncology trials, where the standard chemotherapeutic regimen is given for, say, six three-week cycles, while in the experimental arm, the add-on treatment continues until disease progression. Duration may also differ at the subject level, where subjects withdraw

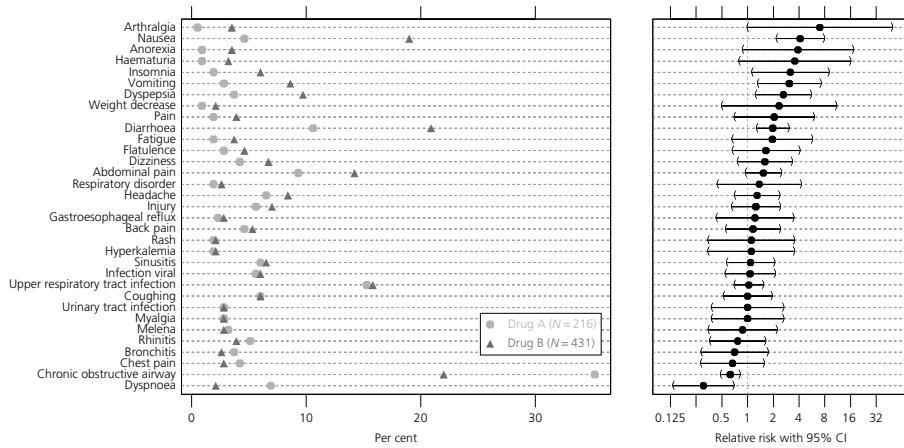
from treatment for safety reasons. In both cases, it can be of value to calculate incidence rates in addition to the incidence itself. This is defined as follows:

$$\text{Incidence rate} = \frac{\text{number of subjects suffering the event}}{\text{total exposure}}$$

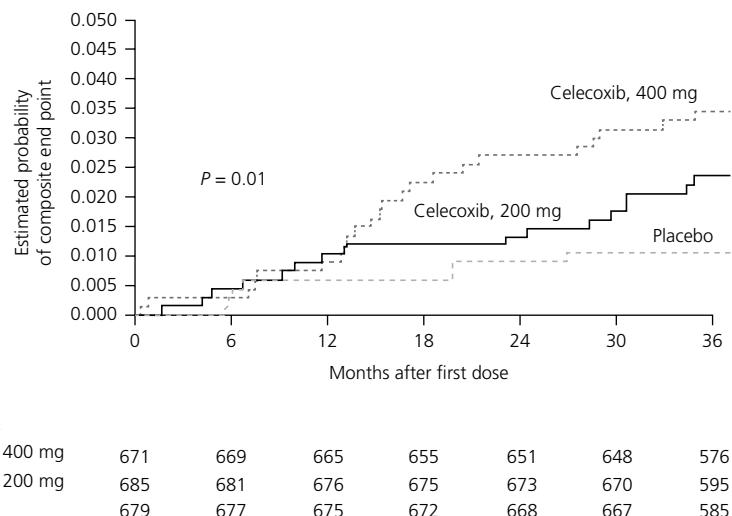
Total exposure, expressed in years, is calculated by summing up the exposures for all subjects in that treatment group. When dealing with rare events, it may also be of value to multiply this rate by 1000 to express it as the rate per 1000 subject-years. It is also common when considering rates in this way to replace the *number of subjects suffering the event* in the numerator by the *number of events suffered over all subjects*. This accounts for multiple occurrences of an event. Whether this is appropriate or not depends on the context.

While tables are extremely important and useful, it can be of value to provide a graphical presentation of the differences in incidence rates for commonly occurring events. Figure 19.1 is an example of such a display and is taken from Amit et al. (2008). The events listed down the left-hand side are those occurring most frequently in the experimental group. The circles and triangles on the left-hand side of the display are the incidences by treatment group. The right-hand side of the display shows the relative risks with 95% CIs. The events are ordered according to the value of the relative risk. AEs for which there is an increased risk are immediately apparent. In some cases, for example, when there are zero events of a particular type in the control group, it may be preferable to plot the risk difference rather than the relative risk; with a zero observed risk in the control group, the relative risk is not defined. Ordering according to relative risk or risk difference is natural in these cases, but ordering by absolute risk in the experimental arm or absolute risk in both arms combined might also be appropriate.

We mentioned earlier that there may be events of special interest, and in addition to possibly presenting separate tables that provide information for only these events, it can be useful to look at the time to an event (or the onset of an event) of special interest. This will give information not only on the frequency of these events in the treatment groups but also on whether the event occurs earlier in one of those groups. Figure 19.2 is taken from Solomon et al. (2005), who report on a retrospective evaluation of an increased cardiovascular risk associated with celecoxib for the prevention of colorectal adenomas. Patients were randomised to two doses of celecoxib (200 mg or 400 mg twice daily) or placebo, and the endpoint evaluated was a composite of death from cardiovascular causes, myocardial infarction, stroke or heart failure. This investigation was evaluating a particular hypothesis due to there being heightened awareness of the possibility that COX-2 inhibitors could be associated with increased cardiovascular risk, and for this reason, a formal *p*-value comparison was undertaken to test this hypothesis. The Kaplan–Meier plots of time to the composite event and the statistically significant *p*-value of 0.01 comparing the three groups collectively support that there is an increased risk associated with celecoxib and that this is dose related.



**Figure 19.1** Most frequent on-therapy events ordered by relative risk. Source: Amit O, Heidberger RM and Lane PW (2008) ‘Graphical Approaches to the Analysis of Safety Data from Clinical Trials’ *Pharmaceutical Statistics*, 7, 20–35. Reproduced with permission from John Wiley & Sons.

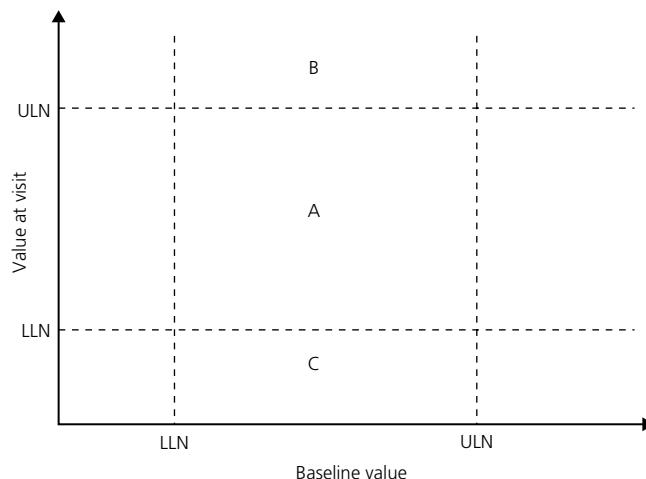


**Figure 19.2** Kaplan–Meier plots for time to composite safety endpoint. Source: Solomon S, McMurray J JV, Pfeffer MA, et al. (2005). Cardiovascular Risk Associated with Celecoxib in a Clinical Trial for Colorectal Adenoma Prevention. *NEJM* **352**, 1071–1080. Reproduced by permission of Massachusetts Medical Society.

### 19.2.3 Laboratory data

The most useful tabular presentations for laboratory data are *shift tables*. These tables display separately the numbers of subjects in the treatment groups who move from having values within/outside the normal range at baseline to having values outside/within the normal range at each visit. A graphical version of this table is a series of scatter plots, one for each visit, with the baseline value on the *x*-axis and the corresponding value at the visit on the *y*-axis, with vertical and horizontal lines drawn at the upper and lower limits of the normal range as shown in schematic form in Figure 19.3.

Points within the box formed by the horizontal and vertical lines (Region A) correspond to patients with values that start within the normal range and remain within the normal range at the visit. Values in Region B correspond to patients whose values start within the normal range but then go above the upper limit of normal at the visit. Patients in Region C also have a value within the normal range at baseline, but their value at the visit falls below the lower limit of normal. However, the graphical version works in a straightforward way only if the normal range is the same across the study. If there were separate normal ranges for males and females, the graphs would need to be produced by sex. If there were different normal ranges for different centres in the study, these graphs would need to be modified by standardising the measurements across centres (and for males and females, if required). If LLN and ULN represent



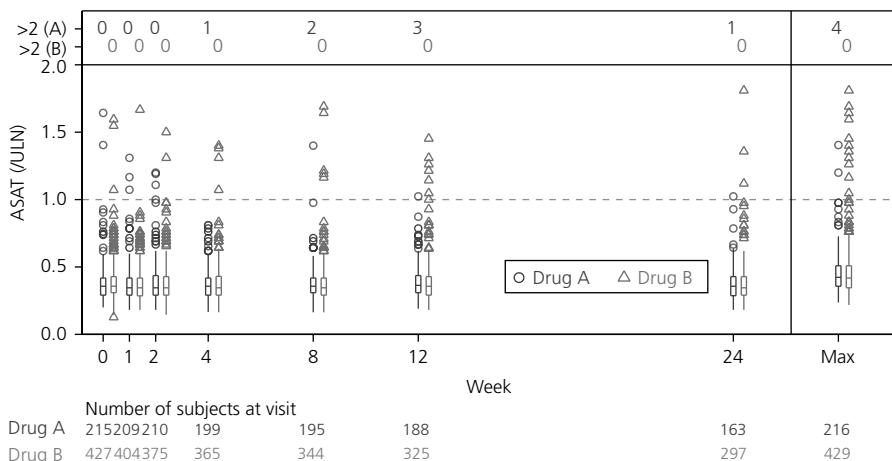
**Figure 19.3** Schematic for baseline and visit values in a shift plot for a typical laboratory parameter. LLN, lower limit of normal; ULN, upper limit of normal required.

the lower and upper limits of the normal range, the standardised measurement,  $y$ , is then

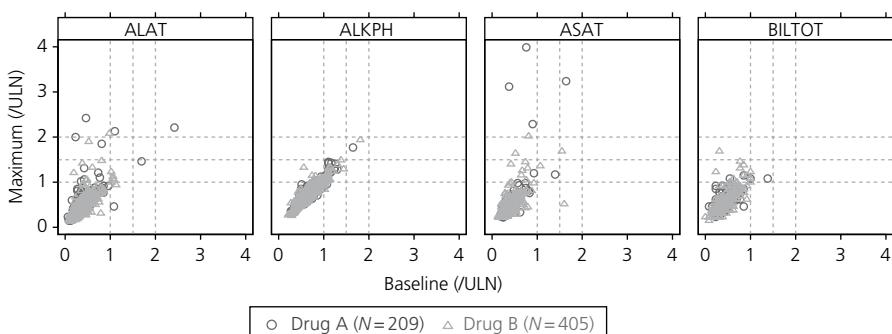
$$y = \frac{x - LLN}{ULN - LLN}$$

where  $x$  is the observed laboratory value. This gives a transformed laboratory value between 0 and 1 when the actual laboratory value is within the normal range, greater than 1 when the untransformed value is above the ULN and less than 0 when the untransformed value is below the LLN. In many cases it is only exceeding the ULN that is clinically relevant and plotting a simple scaled value  $y = x/ULN$  is more useful. The horizontal and vertical lines in Figure 19.3 would then be drawn at 0 and 1.

An alternative graphical method to display laboratory data is the box plot (see Section 2.2.5). Figure 19.4, taken from Amit et al. (2008), displays data on AST (labelled ASAT in the graph) using box plots at baseline and by on-treatment visit. The especially large values, which are often those of greatest interest and concern, are easily seen at each visit in this plot. In the right-hand panel, there is also a plot of the maximum on-treatment value. Horizontal lines have been drawn as reference lines at  $2 \times ULN$ . In plots for other laboratory parameters, it may be more appropriate to draw lines at LLN (or zero) and ULN. Note that the units on the  $y$ -axis are in terms of proportions of the upper limit of normal. Figure 19.4 also contains the numbers of subjects at each visit and the numbers of subjects in each group with values above twice the upper limit of normal. Certain laboratory parameters display highly skewed distributions, and it can sometimes be difficult to plot values on a linear scale while maintaining



**Figure 19.4** Distribution of AST by time and treatment. Source: Amit O, Heidberger RM and Lane PW (2008). Reproduced with permission from John Wiley & Sons.



**Figure 19.5** Trellis plot of a liver function test (LFT) shifts from baseline to maximum by time. Note: ALAT, ALKPH and ASAT, the clinical concern level is 2 ULN; BILTOT, the CCL is 1.5 ULN; where ULN is the upper level of normal range. Source: Amit O, Heidberger RM and Lane PW (2008). Reproduced with permission from John Wiley & Sons.

separation of the individual points at the lower numerical end of the scale. In these circumstances, it can be advantageous to plot parameter values on the log scale. Judgement is needed to decide when to do this.

As we have mentioned, the liver transaminase parameters invariably receive special attention, and the maximum values through the study period are of particular concern. We have already seen in Figure 19.4 a representation of one of these parameters. It is also useful to look at these parameters jointly. Figure 19.5, taken from Amit et al. (2008), shows alkaline phosphatase (ALKPH) and total bilirubin (BILTOT) in addition to AST and ALT (labelled ASAT and ALAT, respectively, in this figure), again using values for these parameters expressed as proportions of the upper limit of normal. The figure presents bivariate plots of

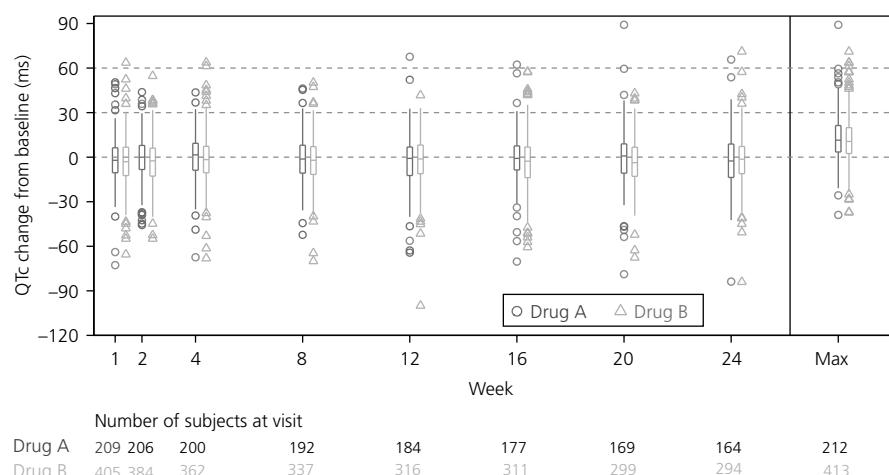
maximum value vs. baseline value with lines on the plots at ULN,  $1.5 \times$  ULN and  $2 \times$  ULN. Each of these component plots is as in Figure 19.3 but now focuses on the maximum value post-baseline. The concerning data points are those that lie in the upper left-hand quadrants and correspond to subjects whose values at baseline are within the normal range but whose values post-baseline go above ULN,  $1.5 \times$  ULN and  $2 \times$  ULN.

### 19.2.4 ECG data

The analysis of the  $QT_c$  interval usually focuses on two aspects:

- Overall increase in the average value through the study
- Exceeding thresholds that are of particular concern

Tables showing median values through time, numbers of subjects exceeding the threshold values of 450 ms, 480 ms and 500 ms, and numbers of subjects with change from baseline  $>30$  ms and  $>60$  ms, which are the concerning values as set down in the CHMP guidance (CHMP, 2005), give the information that is of interest. Box plots as shown in Figure 19.6, also taken from Amit et al. (2008), can address both aspects visually: trends over time and numbers exceeding threshold values. This plot looks at change from baseline, but a similar plot could be produced for the absolute values. The horizontal line within each box is the median value, and this is provided for each treatment group and for each visit. These values can additionally be joined across time within each group to help pick up trends. Note that trends are much easier to identify in a plot than in a table, especially when a comparison between treatment groups is needed. It is important that evaluation of trends is based on such a comparison as regression towards the mean can give a false impression for any one group in isolation.



**Figure 19.6** Box plot of change from baseline in  $QT_c$  by time and treatment. Increase  $<30$  ms 'Normal', 30–60 ms 'Concern',  $>60$  ms 'High'. Source: Amit O, Heidberger RM and Lane PW (2008). Reproduced with permission from John Wiley & Sons.

Regression towards the mean in this context would occur if only subjects with normal (low) values for the QT<sub>c</sub> interval were recruited into the study, as there would be a tendency for these to increase over time due to random fluctuations.

Looking at absolute values rather than change from baseline in the plots would enable us to see numbers above absolute thresholds.

### 19.2.5 Vital signs

Tables and plots for vital signs are similar to those for the QT<sub>c</sub> interval. Both change from baseline and absolute values are of interest, and we focus on average trends compared between treatment groups and, in addition, the numbers of subjects with values outside of what could be considered *normal*. For body temperature, anything above 38.5 °C would indicate fever. For blood pressure, the norms are 80 mmHg for dBP and 120 mmHg for sBP. Heart rate is very much age dependent and even within age brackets can vary enormously from subject to subject so that looking at extremes in terms of change from baseline rather than absolute values may be of greater value. Similar comments apply for respiratory rate.

### 19.2.6 Safety summary across trials

The methods for presenting AE and SAE data covered earlier in this section in relation to single clinical trials can also be used across a collection of trials within the context of a regulatory package by simply totalling over the trials. However, some consideration must be given to dose level and duration of treatment, with tables separated accordingly. Note that with dose level, this also applies to a single trial. Data for the collection of trials within the package will be more extensive, and this will also allow individual data tabulations and plots within subgroups of patients: males/females, <65/≥65 years and so on. We discussed in Section 18.1 the dangers of simply pooling data over separate studies and that doing so can be misleading under some circumstances. The same applies here: we need to be cautious, especially if we have unbalanced randomisation in some studies, very different populations being studied in different trials and varying rates of AEs between those trials. As discussed in that earlier section, the preferred approach to pooling is a meta-analysis. We also discussed in Section 18.3.7 the problem of rare events, a problem that can arise in many analyses of SAEs, both prospective and retrospective.

Given that the trials within the package are likely to differ somewhat in their design, it is generally not useful to provide tables and graphs that look at individual visits in relation to laboratory parameters, ECGs and vital signs. Here, it is better to focus on maximum changes from baseline in each of the studies and collectively summarise those across all trials or across groups of trials that had similar design.

The summary of safety also focuses on specific issues that arise during the programme. For example, if liver toxicity has been an issue, then in addition to

the tables and graphs for the liver transaminases, there will be a presentation of the numbers of Hy's law cases, possibly also looking at time to occurrence in Kaplan-Meier curves.

### 19.2.7 Specific safety studies

At the approval stage and as a condition for approval, the sponsor may be requested to conduct a safety study to better characterise the incidence rates for specific AEs. This could be a randomised study but more likely will be an open-label single group study potentially involving an external control group. We discussed observational studies in Chapter 17, but for the moment, let's consider what statistical arguments might be brought to bear to choose an appropriate sample size. As an example, suppose that the incidence rate for a particular event in the target population is generally around 2%, and there is concern that this rate may increase to an unacceptably high level (say, 10%) on the experimental treatment. If the safety study is to be in a single group receiving the experimental treatment, we need to calculate the sample size required to rule out an incidence rate on the experimental treatment of 10% with, for example, 90% power. Assuming a true rate on the experimental treatment of 4%, a sample size of 200 will satisfy these conditions with a 2.5% type I error. The analysis approach here is based around non-inferiority, where the margin is set at 10%, and we are looking to show that the incidence rate on the experimental treatment is below that level. With the data in hand, we can calculate a one-sided 97.5% CI for the rate,  $\theta$ . If the upper confidence limit is below 10%, then we have ruled out an increase to that level.

## 19.3 Data monitoring committees

Data monitoring committees (DMCs) (DSMCs, DSMBs, SRCs) are tasked with reviewing accumulating safety data to protect the safety of the subjects in the study. We have already spoken about the responsibilities of such committees in Section 14.4 and the things they might evaluate from a safety point of view in Section 14.3. Here, we will further consider the kinds of safety data they should look at and the form in which those data should be presented.

The tables and graphs reviewed by DMCs will be in line with those discussed earlier in this chapter. In fact, it is often the case that the sponsor provides the table templates that are the basis for the safety tables for the final clinical study report (CSR) to the independent statistician, who produces those same tables for the DMC in an ongoing way. Graphs are especially useful and enable the committee to quickly gain an overview of evolving trends and safety concerns.

It must be accepted that although some data cleaning will have occurred, it will be limited, and the tables and graphs will not be 100% clean. Committee members should not be too alarmed at seeing discrepancies across the tables and graphs, although if they are substantial, the committee might want to ask

questions about the cleaning process. There is always a conflict between wanting data that are timely and wanting data that are clean; the adage ‘do you want it now, or do you want it clean?’ applies! The data supplied to the DMC will come from the clinical database, which is based on data collected on the case report forms (CRFs). There will be a delay in downloading those data from the CRFs to the clinical database (although with electronic data capture, the time for this should be short), cleaning the data in the database, sending those data to the independent statistician, producing the tables and graphs and delivering these to the DMC members. The tables and graphs then need to be reviewed by the committee, which meets to discuss their content and issues arising. All of this takes time, and the data may well be several weeks out of date by the time the committee reviews them. In light of this, it is of value to at least have more current SAE data. These data come from the safety database and are supplied by the sponsor’s pharmacovigilance group (or their designee, such as a clinical research organization (CRO) employed to take responsibility for pharmacovigilance) and should be completely up to date. The independent statistician may undertake some analysis in terms of summaries and descriptive statistics, but essentially these data should be available to the DMC in real time. Members of the committee may also receive individual SAE reports in an ongoing way.

The DMC usually only reviews tables and graphs that contain descriptive statistics. However, if there are events of special interest, some more formal analysis may be required. Ninety-five percent CIs for treatment differences can be produced periodically and form part of the evaluation of those events at the meetings of the DMC.

## 19.4 Assessing benefit-risk

### 19.4.1 Current approaches

Balancing benefits and risks is a key element of the regulatory evaluation of all new products, and both the EMA and FDA use structured approaches for that evaluation. The EMA structure for this evaluation follows under a series of headings:

- Benefits
- Beneficial effects
- Uncertainty in the knowledge about the beneficial effects
- Risks
- Unfavourable effects
- Uncertainty in the knowledge about the unfavourable effects
- Balance (*between beneficial effects and unfavourable effects*)
- Importance of favourable and unfavourable effects
- Benefit–risk balance
- Discussion on the benefit–risk assessment

**Table 19.2** FDA benefit–risk framework

Decision factor	Evidence and uncertainties	Conclusion and reasons
Analysis of condition		
Current treatment options		
Benefit		
Risk and risk management		
	Conclusions regarding benefit–risk	

Note: The first two factors are therapeutic area specific, while the final two factors relate to the drug.

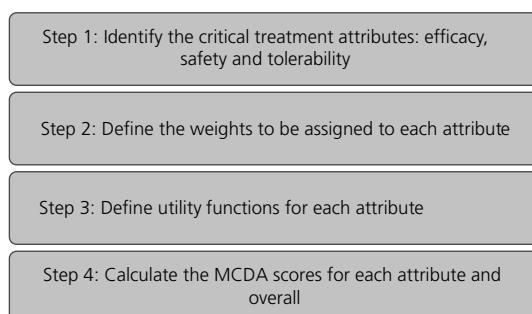
The FDA approach for conducting and communicating benefit–risk assessments is captured in the Benefit Risk Framework for new drug review. This is summarised in Table 19.2.

Neither the EMA nor the FDA historically have used any formal way of numerically combining the evidence for therapeutic benefit with the evidence for risk or harm. Each of those aspects has been kept separated and their balance assessed intuitively through extensive discussion and consideration.

CHMP in 2008 produced a Reflection Paper on benefit–risk assessment methods (CHMP, 2008) and set up a Benefit–Risk Methodology Project to review the existing approach with a view to developing new methodologies. In a series of Work Packages (EMA, 2010, 2011a, 2011b, 2012), they evaluated a number of those methodologies, and this led to the recommendation of *multi-criteria decision analysis (MCDA)* as a quantitative methodology for the assessment of benefit–risk. Only recently, however, following a lot of discussion, has this methodology been formally adopted by the EMA and FDA (Chisholm et al., 2022). MCDA will be described in the next section through an example.

### 19.4.2 Multi-criteria decision analysis

Work Package 4 of the EMA Benefit–Risk Methodology Project sets down the MCDA methodology from the point of view of the regulatory assessment team. There are four steps, and these are displayed in Figure 19.7.

**Figure 19.7** Steps in the calculation of MCDA scores

Step 1 of this approach sets down the efficacy benefits, termed *favourable effects*, and risks and harms, termed *unfavourable effects*, as a list of attributes with associated estimated effects. The example that follows is based on Caprelsa (300 mg) in the treatment of inoperable thyroid cancer and is adapted from Work Package 4, which used this as one of a series of hypothetical examples. The attributes and estimated effects were taken from a single phase III placebo-controlled trial but more generally would be based on the totality of confirmatory evidence from the regulatory package as a whole:

- Favourable effects:
  - Primary endpoint, progression-free survival (PFS): median = 30.5 months vs. 19.3 months on placebo
  - Secondary endpoint, objective response rate (ORR): proportion of complete or partial responders = 45% vs. 13% on placebo
- Unfavourable effects:
  - Diarrhoea CTC grades 3–4: 10.8% vs. 2.0% on placebo
  - QT<sub>c</sub>-related CTC grades 3–4: 13.4% vs. 1.0% on placebo
  - Infection CTC grades 3–4: 49.8% vs. 36.4% on placebo

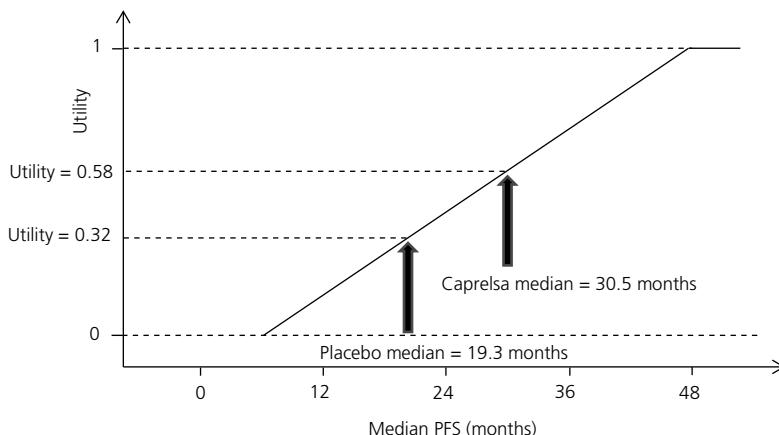
The recommendation is to choose at most about 8 to 10 attributes, and these should encompass the key efficacy endpoints and those unfavourable effects that are most relevant for the benefit–risk balance.

Step 2 involves giving weightings to the various attributes in stages that reflect their importance to the benefit–risk balance. These weights are subjective and are chosen based on discussions within the assessment team. Sensitivity analyses (to be discussed later) look at how the final benefit–risk balance is affected by changing these weights. Given the regulatory emphasis on safety, weights of 40 and 60% were assumed for efficacy and safety, respectively. Within the efficacy attributes, weights of 75% for PFS and 25% for ORR were assigned to reflect their relative importance; and within safety, the assigned weights were 20% for diarrhoea, 40% for QT<sub>c</sub> and 40% for infections. Combining these weights gives relative importance weights as follows:

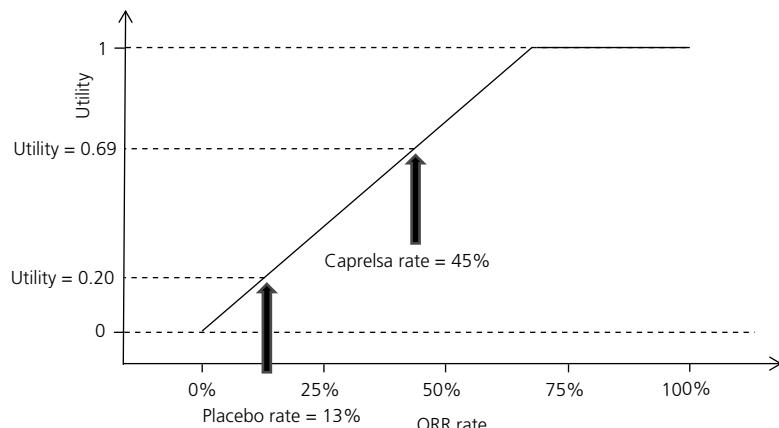
- PFS: 40% × 75% = 30%
- ORR: 40% × 25% = 10%
- Diarrhoea: 60% × 20% = 12%
- QT<sub>c</sub>: 60% × 40% = 24%
- Infections: 60% × 40% = 24%

Note that these weights add up to 100%.

Step 3 requires the assignment of *utility* values to each of the attributes. These reflect the value to be placed on summary outcomes of various orders of magnitude. For PFS, a median of below 6 months is viewed as being of little clinical value, and a median greater than 48 months is considered theoretically very unlikely. Figure 19.8 provides a reflection of possible utility values. Again, this element of the model is subjective and would be formulated through a discussion with clinical experts in the field. The observed median PFS on Caprelsa is



**Figure 19.8** Utility values for PFS



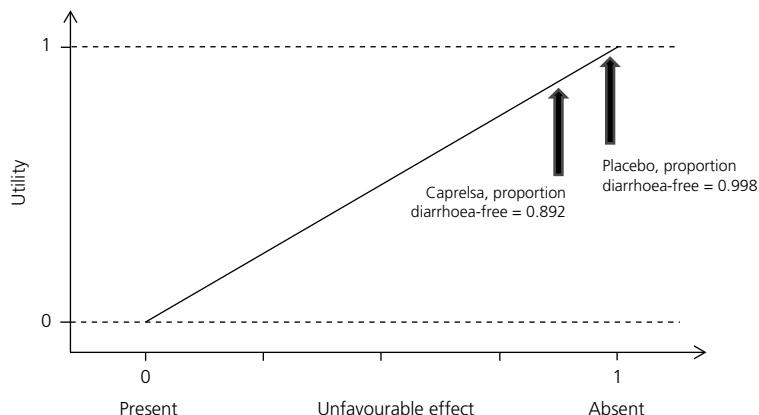
**Figure 19.9** Utility values for ORR

30.5 months, giving a utility of 0.58, and the median PFS on placebo is 19.3 months, giving a utility of 0.32. Note that the higher the utility, the better.

Figure 19.9 provides a corresponding utility function for ORR. The assumption here is that an ORR of up to 0.65 is theoretically possible and the utility is a linear function of the ORR up to that value. According to the utility function, Caprelsa has a utility value of 0.69 and placebo a utility value of 0.20.

In step 4, we calculate the MCDA utility scores. The overall MCDA score (SFE, Summary of Favourable Effects),  $SFE_{Caprelsa}$ , for the favourable effects of Caprelsa is then

$$SFE_{Caprelsa} = 30\% \times 0.58 + 10\% \times 0.69 = 24.3$$



**Figure 19.10** Utility values for unfavourable effects (diarrhoea)

For placebo, the MCDA score is

$$SFE_{\text{placebo}} = 30\% \times 0.32 + 10\% \times 0.20 = 11.6$$

The utility function for each of the unfavourable effects is assumed to be linear, taking the value 1 if the effect is absent and 0 if the effect is present (Figure 19.10). Again, higher values are better. We then use the incidence rates to calculate the utility values.

In order for larger values to correspond to desirable effects, as for PFS and ORR, we calculate 1 minus the incidence rate, which is the incidence rate for the absence of the effect. So for the effects of interest, the utility values are

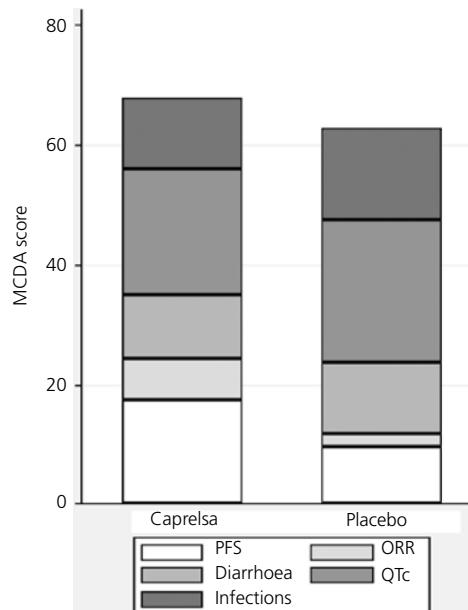
- Caprelsa
  - Diarrhoea =  $1 - 0.108 = 0.892$
  - QT<sub>C</sub> =  $1 - 0.134 = 0.866$
  - Infections =  $1 - 0.498 = 0.502$
- Placebo
  - Diarrhoea =  $1 - 0.02 = 0.98$
  - QT<sub>C</sub> =  $1 - 0.01 = 0.99$
  - Infections =  $1 - 0.364 = 0.636$

The MCDA scores (SUFU, Summary of Unfavourable Effects) for the unfavourable effects are then

$$SUFU_{\text{Caprelsa}} = 12\% \times 0.892 + 24\% \times 0.866 + 24\% \times 0.502 = 43.5$$

$$SUFU_{\text{placebo}} = 12\% \times 0.98 + 24\% \times 0.99 + 24\% \times 0.636 = 50.8$$

The total MCDA scores are then the sums corresponding to the favourable and unfavourable effects, giving final scores of  $24.3 + 43.5 = 67.8$  for Caprelsa and  $11.6 + 50.8 = 62.4$  for placebo. The fact that the Caprelsa score is higher than the



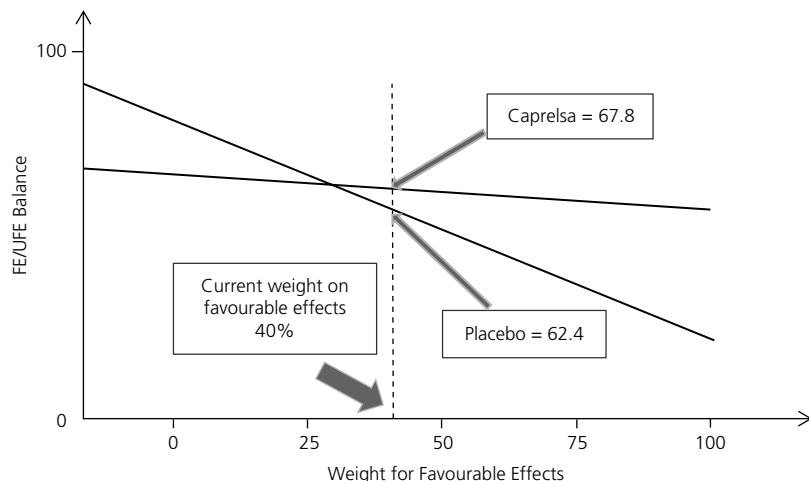
**Figure 19.11** Stacked bar graph for MCDA scores

placebo score indicates a positive benefit–risk balance for Caprelsa. These scores are usually placed into a stacked bar graph, as in Figure 19.11, to allow comparisons of the overall scores and also the individual components that make up those scores.

Clearly, several components of the MCDA model have an element of subjectivity. In particular, the differential weights attached to the favourable and unfavourable effects (40 and 60%, respectively, in the example) may differ between regulators and payers on the one hand and physicians and patients on the other. Physicians and patients may give more weight to the favourable effects, arguing that the unfavourable effects can be managed. Changing these weights will change the MCDA scores. Greater weight given to the favourable effects, and less to the unfavourable effects, usually favour the experimental treatment if it is efficacious. In the earlier example, assigning 50% to both favourable and unfavourable effects results in MCDA scores of 66.7 for Caprelsa and 56.8 for placebo, an even greater benefit–risk balance. Figure 19.12 displays the sensitivity of the results to the chosen weights.

The positive benefit–risk balance is lost only if one is prepared to give a weight of 28.7% or less to the favourable effects. The sensitivity of the overall scoring to the weights assigned to the separate attributes within the favourable and unfavourable effects can be explored in a similar way by changing those weights.

For the moment, no account has been taken of uncertainty in the estimated treatment benefits (that is, the median PFS values and the ORRs) and the



**Figure 19.12** Sensitivity analyses for favourable–unfavourable weightings

observed frequencies of undesirable effects (that is, the frequencies of CTC grade 3–4 diarrhoea, QT<sub>c</sub> prolongation and infections). A full sensitivity analysis involves changing the estimated treatment differences in line with the uncertainty associated with those estimates and seeing what impact that has on the benefit–risk balance. It is straightforward to show that the difference in the MCDA scores (Caprelsa – placebo) is given by

$$\text{MCDA}_{\text{Caprelsa}} - \text{MCDA}_{\text{Placebo}} = 40\% \left\{ 75\% \times \left( \frac{m_c - m_{pl}}{42} \right) + 25\% \times \left( \frac{\text{ORR}_c - \text{ORR}_{pl}}{0.65} \right) \right\} \\ + 60\% \left\{ 20\% \times (D_{pl} - D_c) + 40\% \times (QT_{c_{pl}} - QT_{c_c}) + 40\% (\inf_{pl} - \inf_c) \right\}$$

where  $m_c$  and  $m_{pl}$  are the median PFS values for Caprelsa and placebo; ORR denotes the objective response rates; and D, QT<sub>c</sub> and inf are the incidence rates of CTC grade 3–4 diarrhoea, QT<sub>c</sub> prolongation and infections, respectively. Note that 1 minus the rates for the unfavourable effects is used in the MCDA score calculation: this is why differences between those incidence rates in this equation are calculated by taking the Caprelsa incidence rate away from the placebo incidence rate, whereas for the favourable effects, differences are taken by subtracting the placebo value from the Caprelsa value. It might then be of interest, for example, to see what the impact would be if the actual difference in the PFS medians had been smaller than the observed 11.2 months. Undoubtedly, taking the difference in the medians much lower would remove the positive benefit–risk balance, but these aspects need to be explored based on the sampling variability around the observed treatment differences as expressed by the CIs.

As mentioned, the MCDA methodology contains certain subjective elements: the choice of which favourable and unfavourable effects to focus on, the weights

assigned to those effects and the utility functions. We can of course evaluate the robustness of the conclusions by varying each of these aspects, but nonetheless, these elements of subjectivity have resulted in criticism of the technique. It is also the case that regulators, payers, clinicians and patients may have quite different views on which favourable and unfavourable effects are important, what weights should be placed on these and the utility functions to be assigned. In a sense, this criticism is understandable; but in my view, it is a little misplaced. The evaluation of the benefit–risk balance is subjective anyhow, and again, different people view this balance differently. At least with the MCDA method the various subjective elements are out in the open and can form a sound basis for discussion.

The MCDA methodology has been discussed from the point of view of assessing benefit–risk (favourable–unfavourable effects) within a regulatory submission. This same technique can also be used for internal decision-making within a company, perhaps for deciding which dose to take forward from phase II to phase III, where a corresponding trade-off between benefits and undesirable effects is needed. It can also be used to compare a product with a competitor product to allow better positioning in the marketplace, and in health technology assessment (Angelis and Kanavos, 2017).

### **19.4.3 Quality-adjusted time without symptoms or toxicity**

Another method, *quality-adjusted time without symptoms or toxicity* (*Q-TWiST*), is used in oncology and balances benefits in terms of overall survival (OS) against the toxic effects of treatment and disease progression. The method was introduced by Glasziou et al. (1990) and gives positive weight to increasing survival time that is free from both the toxic effects of treatment and progression.

The basic idea is to consider the time from randomisation in a clinical trial as consisting of three components labelled TOX, TWiST and REL. TOX measures the duration of toxicity as the time that the patient is suffering CTC grade 3 or 4 events prior to progression, TWiST is the duration of survival that is free of toxicity and prior to disease progression and REL is the duration of survival time following progression. The mean value of each quantity is calculated from the patient data in each group. The mean value for TOX is the duration of toxicity up to progression (or death, if that occurs before progression) for each patient. This duration may be zero if the patient does not suffer any CTC grade 3 or 4 events or may consist of two (or more) distinct periods if the patient suffers a CTC event (or several events) over separate periods of time.

The mean values for PFS and OS are calculated from the Kaplan–Meier curves as restricted means up to a certain follow-up period (see Section 13.5 for a discussion of restricted mean survival time). These means are calculated as the areas under the Kaplan–Meier curve for PFS and OS, respectively. Note that all the same arguments can be used when looking at relapse in place of progression and also with TOX including time with severe symptoms in addition to the CTC grade toxicities. There are three mean values for each treatment group: mean

time with toxicity (mean TOX), (restricted) mean PFS and (restricted) mean OS. Based on these, we can calculate the duration of TWiST and REL as follows:

$$\text{Duration of TWiST} = \text{mean PFS} - \text{mean TOX}$$

$$\text{Duration of REL} = \text{mean OS} - \text{mean PFS}$$

We then attach a utility to each of these states:  $u_{\text{TOX}}$ ,  $u_{\text{TWiST}}$  and  $u_{\text{REL}}$ . Generally, we put  $u_{\text{TWiST}} = 1$  as that is the best state (time without symptoms and toxicity) and give  $u_{\text{TOX}}$  and  $u_{\text{REL}}$  values below 1. The so-called *base case* puts  $u_{\text{TOX}} = u_{\text{REL}} = 0.5$ ; this says that one day in TWiST has the same value as two days in either the TOX or REL state. Sensitivity, or *threshold*, analyses varies the values of these utilities to evaluate the robustness of the conclusions to the choice of utility values.

A Q-TWiST analysis calculates the value of Q-TWiST for each of the treatment groups and looks at the difference where

$$\text{Q-TWiST} = u_{\text{TOX}} \times \text{duration TOX} + \text{duration TWiST} + u_{\text{REL}} \times \text{duration REL}$$

A positive difference for an experimental treatment indicates a positive quality-adjusted survival for that treatment. The bootstrap method (Section 3.2.4) can be used to obtain CIs for this difference and give a *p*-value based on the ratio of the Q-TWiST difference divided by the bootstrap standard error. Example 19.1 provides an application of this methodology.

**Example 19.1** Panitumumab plus best supportive care vs. best supportive care in patients with wild-type KRAS metastatic colorectal cancer

Wang et al. (2011) report a Q-TWiST analysis in a comparison of panitumumab plus best supportive care (BSC) vs. BSC in patients with wild-type KRAS metastatic colorectal cancer (mCRC). The mean durations (in weeks) in the various health states were 3.47, 13.26 and 9.35 for TOX, TWiST and REL in the panitumumab group and, correspondingly, 1.09, 8.01 and 16.15 in the BSC-alone group. Note that the mean TOX was greater in the experimental arm, as might be expected. The mean time in the preferred state without symptoms or toxicity (TWiST) was also greater for the experimental treatment, while the mean time in the relapse state was less. In the base case, the balance was such that there was a Q-TWiST advantage of 3.03 weeks with a 95% bootstrap CI of (0.86, 5.20). This difference was statistically significant; note that the CI excludes 0. These authors also collected data on a quality of life (QoL) scale (EQ-5D) and used that to derive utility weights of 0.60, 0.77 and 0.63 for the TOX, TWiST and REL states in the experimental arm and corresponding weights equal to 0.44, 0.66 and 0.64 in the BSC-only arm. Note here that the utility weights were allowed to differ across treatment groups. Using these weights, the advantage of panitumumab + BSC over BSC alone in terms of Q-TWiST was 6.5 weeks (12.3 weeks vs. 5.8 weeks). A threshold analysis showed a numerical advantage for panitumumab except for some cases where the REL state was given a utility weight of 1. A utility weight of 1 for the REL state, as well as for the TWiST state, does not distinguish the REL state in terms of QoL from the TWiST state, and to a certain extent, this could be viewed as being unrealistic.

The Q-TWiST approach gives a broad-based measure for the balance between a specific favourable effect, prolonging life, and the unfavourable effects of toxicity and progression/relapse. Clinicians find that this measure has intuitive appeal in that it down-weights days living with toxicity and days post-progression/relapse.

## 19.5 Pharmacovigilance

### 19.5.1 Post-approval safety monitoring

Following regulatory approval for a drug, adverse drug reactions (ADRs) are reported to the regulatory authorities in a variety of different ways. In Europe, for example, the EMA has set up the EudraVigilance Data Analysis System that stores data generated through spontaneous case reports of individual ADRs and enables stakeholders to analyse those data using various statistical techniques. We are looking here for *signals*: that is, early hints at *unintended drug effects*. In this section, we will be focusing on a particular approach to signal detection, the *proportional reporting ratio (PRR)*. This is the most frequently used measure and is the recommended approach in the EudraVigilance Expert Working Group guidance on the use of statistical signal detection methods within their system (EMA, 2008). Before going on to discuss the PRR, it is worth mentioning two other statistical techniques that can be of value in post-approval safety monitoring.

It may be that information comes to light that generates a hypothesis about safety concerns for a particular drug. In these settings, it can be of value to conduct a systematic review and meta-analysis to test out this hypothesis; see Example 19.2.

**Example 19.2** Risk of cardiovascular SAEs associated with varenicline use for tobacco cessation

Prochaska and Hilton (2012) report on an evaluation of the risk of cardiovascular SAEs associated with varenicline use for tobacco cessation. Concern had been expressed for some time regarding this issue, and one other systematic review had hinted at the potential for a problem. Prochaska and Hilton included a meta-analysis of 22 double-blind, placebo-controlled trials and reported a trend for increased risk in those subjects taking varenicline, although this increase was not statistically significant ( $p = 0.11$ ). The event of interest in this analysis was rare, and across all trials involving a total of over 9000 subjects, there were only 52 subjects with cardiovascular SAEs. A total of eight of the trials reported no cardiovascular SAEs across both groups, while a further eight trials reported no events in just one treatment group. The low event rates make the statistical analysis more challenging. See Section 18.3.7 for discussion on this issue. The eight trials with zero events were excluded from the meta-analysis. For the eight trials that had zero counts in one of the treatment groups, a continuity correction was applied, and 0.5 was added to each of the four counts in the  $2 \times 2$  contingency table for that trial; so in a hypothetical example where there are 85 patients per group and the four contingency table entries are 0 and 85 in group 1 and 2 and 83 in group 2, 0 becomes 0.5, 2 becomes 2.5, 85 becomes 85.5, and 83 becomes 83.5.

The MCDA methodology was presented in Section 19.4.2 in the context of balancing benefit and risk at the time of regulatory submission. As more data become available on both efficacy and safety post-approval, it is also possible and useful to revisit such an analysis to judge whether the benefit–risk ratio has changed sufficiently to be concerning.

### 19.5.2 Proportional reporting ratios

The Proportional Reporting Ratio (PRR) has become a standard metric for the evaluation of safety signals. It is calculated for a drug–event pair and uses ideas associated with the disproportionality of reporting. The basic argument is that if a particular event ( $E$ ) is indeed more prevalent with a particular drug ( $D$ ), that event should be reported more frequently in conjunction with drug  $D$  than in conjunction with other drugs. Table 19.3 is a  $2 \times 2$  contingency table based on the number of events  $E$  associated with the drug of interest and all other drugs in a pharmacovigilance database. In this table:

- $a$  is the number of subjects taking drug  $D$  who suffer the adverse event  $E$ .
- $b$  is the number of subjects taking drug  $D$  who suffer other adverse events (not  $E$ ).
- $c$  is the number of subjects taking other drugs (not  $D$ ) who suffer the adverse event  $E$ .
- $d$  is the number of subjects taking other drugs (not  $D$ ) who suffer other adverse events (not  $E$ ).

The PRR is defined as follows:

$$\text{PRR} = \frac{a / (a + b)}{c / (c + d)}$$

If event  $E$  is more prevalent in patients taking drug  $D$  compared to other drugs, then  $a/(a + b)$  will be larger than  $c/(c + d)$  and the PRR will be much greater than 1. This is the idea behind PRR with values above 1 signifying an association between the drug and the event. The formula for the standard error associated with the PRR is expressed on the log scale:

$$se(\ln \text{PRR}) = \sqrt{1/a + 1/c - 1/(a+b) - 1/(c+d)}$$

**Table 19.3** Contingency table for calculating a PRR

	Event ( $E$ )	All other events	Total
Drug ( $D$ )	$a$	$b$	$a + b$
All other drugs	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

A 95% CI on the log scale is then

$$(lnPRR - 1.96 \times se, lnPRR + 1.96 \times se)$$

Once this is calculated, the reverse transformation (antilog) can be applied to the ends of the interval to give a 95% CI back on the PRR scale.

The examples contained in Example 19.3 are taken from EMA (2008).

### Example 19.3 Calculating PRR

Here we are looking for an association between a drug  $D$  and the event nausea.

#### Case 1

Suppose that there are 100 reports of patients with AEs for drug  $D$ , and of these, 5 are of nausea. Then

$$a = 5, b = 95 \text{ and } a + b = 100$$

Suppose there are 100,000 reports of patients with AEs for all other drugs in the database (not including drug  $D$ ), and of these, 5000 are of nausea. Then

$$c = 5000, d = 95,000 \text{ and } c + d = 100,000$$

In this case,  $\frac{a}{a+b} = \frac{c}{c+d} = 0.05$  and  $PRR = 1$ , and we do not have a signal that links drug  $D$

with nausea. The 95% CI for the PRR is (0.43, 2.35).

#### Case 2

Suppose that there are 100 reports of patients with AEs for drug  $D$ , and of these, 15 are of nausea. Then

$$a = 15, b = 85 \text{ and } a + b = 100$$

Suppose there are 100,000 reports of patients with AEs for all other drugs in the database (not including drug  $D$ ), and of these, 5000 are of nausea. Then

$$c = 5000, d = 95,000 \text{ and } c + d = 100,000$$

In this case,  $\frac{a}{a+b} = \frac{15}{100}, \frac{c}{c+d} = \frac{5000}{100000}$  and  $PRR = \frac{0.15}{0.05} = 3$ , and we do have a signal that

links drug  $D$  with nausea. The 95% CI for the PRR in this case is (1.88, 4.79).

The ideas behind the PRR were developed by Evans et al. (2001), who suggested that a PRR greater than 2 indicates a drug–event combination that is worthy of further investigation. In addition to the PRR, they suggested two other criteria that should be applied as a guide to the existence of a signal: namely, that there should be at least three or more cases of the event  $E$  associated with drug  $D$ , and that the chi-square (test) statistic in the  $2 \times 2$  table (Table 19.3) should be at least 4. The chi-square statistic is given by

$$\frac{N \times (a \times d - b \times c)^2}{(a+b)(a+c)(c+d)(b+d)}$$

Note that this statistic was expressed differently (as  $\sum \frac{(O-E)^2}{E}$ ) in Section 4.4.1

when we developed the chi-square test based on observed and expected frequencies, but these two quantities are numerically equal. Also, having a chi-square statistic of at least 4 would only occur a proportion of 0.0455 (less than 1 in 20) of the time by chance, so effectively this requirement is that the *p*-value in the contingency table is statistically significant at the 5% level. There has been an update on this three-item detection system for the decision-making criteria. If the 95% CI is calculated for the PRR, then rather than simply taking the value 2 as the threshold for a signal, we look at the lower end of that CI and base a signal on that value being above 1.

Using the method based on the three criteria (with the condition that the PRR be > 2 rather than the lower confidence limit above 1), Evans et al. retrospectively examined 15 newly marketed drugs in the UK Yellow Card database, which had the highest levels of ADR reporting at that time. The method identified 481 signals. Further evaluation showed that 339 (70%) were known adverse reactions, 62 (13%) were signals linked to the underlying disease and 80 (17%) were signals requiring further investigation. Of the 80 new signals identified, many went on to receive a detailed review, while 11 signals were, in fact, for dextroamphetamine, which was withdrawn from the market during the period covered by the study.

It is accepted that there are several potential biases associated with these methods. The spontaneous reporting system on which the validity of the methodology depends has clear limitations in terms of collecting complete information in an unbiased way.

***EMA (2008): 'Guideline on the Use of Statistical Signal Detection Methods in the EudraVigilance Data Analysis System'***

*'The results of quantitative methods should be interpreted with caution and bearing in mind the limitations of the spontaneous reporting system databases. . . Consequently there is a scientific consensus that SDRs (Signals of disproportionate reporting) identified with quantitative methods should always be medically assessed'.*

Also, there will inevitably be false signals, in part because of multiplicity, although methods that take account of that are available. Despite this, and given the development of methods that account for some but not all of the problems, the PRR methodology has been widely adopted and over the years has proved to be a reliable way to detect safety signals in pharmacovigilance.

### 19.5.3 Bayesian neural networks

Suppose drug  $D$  and adverse event  $E$  are of particular interest in terms of their association. Let  $pr(D)$  and  $pr(E)$  be the relative frequencies (proportions, probabilities) associated with  $D$  and  $E$  separately, and let  $pr(D,E)$  be the relative frequency (proportion, probability) associated with  $D$  and  $E$  occurring together. Bayesian neural networks consider the *information component* (IC) defined as

$$IC = \log_2 \left( \frac{pr(D,E)}{pr(D) \times pr(E)} \right)$$

If there is no association between  $D$  and  $E$ , the proportion of times both  $D$  and  $E$  are reported together will equal the proportion of times  $D$  is reported  $\times$  the proportion of times  $E$  is reported and

$$\frac{pr(D,E)}{pr(D) \times pr(E)} = 1 \text{ so that } IC = 0$$

Some mathematics shows that if the PPR is equal to 1, then  $\frac{pr(D,E)}{pr(D) \times pr(E)}$  is approximately equal to 1 and the IC is approximately equal to zero, and vice versa. Example 19.4 illustrates this connection. Note also that with the  $\log_2$  function, if there is a doubling of  $\frac{pr(D,E)}{pr(D) \times pr(E)}$ , then the IC increases by 1. Bayesian

methods assume vague prior distributions for all of the  $pr(D)$ ,  $pr(E)$  and  $pr(D,E)$  values and periodically (quarterly) updates the IC (posterior distribution) as new data are collected. The posterior distribution after year 1, Q1, for example, becomes the prior distribution for year 1, Q2, and the process continues. Ninety-five percent credible intervals can be constructed for the IC and inferences based on these.

**Example 19.4** Roxicodone for the management of moderate/severe pain

Roxicodene is an opioid used in the management of acute pain. Following approval, concerns were raised about the occurrence of cardiac arrest as a serious adverse drug reaction. In the first quarter of 2010, the FDA received 213,488 reports of serious drug-ADR combinations (Gibbons and Amatya, 2016). Of those, 6360 reported cardiac arrest. Of the 100 ADRs reported for Roxicodone, 31 were of cardiac arrest. The resulting  $2 \times 2$  contingency table is given in Table 19.4.

**Table 19.4** Roxicodone and the incidence of cardiac arrest

	<b>Cardiac arrest</b>	<b>Other ADRs</b>	<b>Total</b>
Roxicodone	31	69	100
Other drugs	2988	210,400	213,388
Total	3019	210,469	213,488

ADR, adverse drug reaction.

$$PRR = \frac{31/100}{2988/213,388} = 22.13$$

$$\frac{pr(D,E)}{pr(D) \times pr(E)} = \frac{\frac{31}{213,488}}{\frac{100}{213,488} \times \frac{3019}{213,488}} = 21.92$$

As can be seen, the PRR is very high at 22.13 and is close numerically to the ratio of probabilities that forms the basis of the calculation of the information component.

Example 19.5 provides an example of how Bayesian methods were used in the investigation of rosiglitazone, licensed for treatment of type 2 diabetes.

#### **Example 19.5** Rosiglitazone, type 2 diabetes and cardiac arrest

Rosiglitazone (Avandia) was approved by the EMA and FDA in May 1999 for the treatment of type 2 diabetes. Concerns were expressed, however, regarding the occurrence of cardiac arrest. The FDA tracked the information component using data from their AE Reporting System (AERS) from late 1999 onwards (Gibbons and Amatya, 2016). In Q1, 2000, IC = 2.55, already well above zero but with a wide credible interval. In Q3, 2007, the information component had increased to 2.67, but now with a narrow credible interval; and in November 2007, FDA issued a black box warning. In Q3, 2010, the information component had increased to 2.89, further justifying that FDA had taken the correct decision.

## CHAPTER 20

# Diagnosis

### 20.1 Introduction

In this chapter, we will discuss the use of statistical methods to evaluate diagnostic tests. To begin with, we will assume that we have a definitive (or something very close to definitive) diagnosis of a disease through some gold standard (*standard of truth*), such as a biopsy; regulators talk in terms of a *surrogate standard of truth*.

#### **CHMP (2009): ‘Guideline on clinical evaluation of diagnostic agents’**

*‘Surrogate standard of truth: a diagnostic test or a combination of tests or follow-up which has been shown to provide a very good approximation to the true disease state or value of measurement’.*

We will then discuss aspects of the design of trials looking to develop new diagnostic tests. Throughout the development, we will use a specific example as a case study. de la Taille et al. (2011) evaluated the clinical utility of the PCA3 (prostate cancer gene 3, PROGENSA) assay in supporting biopsy decisions in prostate cancer. Higher values of the PCA3 score are indicative of the presence of cancer. The study recruited 516 men who recorded serum total prostate-specific antigen (tPSA) between 2.5 ng/ml and 10 ng/ml and were therefore scheduled for an initial biopsy.

There are also many situations where a standard of truth or surrogate standard of truth does not exist; methods then focus on assessing the agreement between different diagnostic tests. We will look at methods based on the kappa statistic, which can measure the extent of the agreement. The final section in this chapter will look at the emerging field of the development of companion diagnostics.

## 20.2 Measures of diagnostic performance

### 20.2.1 Sensitivity and specificity

Sensitivity and specificity are the primary measures we use to judge a diagnostic test's performance. Table 20.1 is based on the de la Taille et al. data. Initially, these authors used 50 as the cut-off score on the PCA3 scale potentially indicating the presence of cancer.

The rows relate to the result of the PCA3 assay, while the columns denote the biopsy result that definitively tells us about the presence (biopsy positive) or absence (biopsy negative) of prostate cancer. The symbols in the table denote the following:

- TN = true negative. Patients had  $\text{PCA3} < 50$  and a negative biopsy.
- FN = false negative. Patients had  $\text{PCA3} < 50$  and a positive biopsy.
- FP = false positive. Patients had  $\text{PCA3} \geq 50$  and a negative biopsy.
- TP = true positive. Patients had  $\text{PCA3} \geq 50$  and a positive biopsy.

*Sensitivity* is defined as the proportion of (true) positives (those who have cancer) who have been correctly diagnosed using the diagnostic test. That is,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} = \frac{104}{207} = 0.50 \text{ or } 50\%$$

Of the patients who have cancer, 50% had a positive result on the diagnostic test.

*Specificity* is defined as the proportion of (true) negatives (those who do not have cancer) who have been correctly diagnosed using the diagnostic test. That is,

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{256}{309} = 0.83 \text{ or } 83\%$$

Of the patients who do not have cancer, 83% had a negative result on the diagnostic test.

For a diagnostic test to be effective, both measures need to be high. Sensitivity is about the ability of the test to detect true cases of cancer (true positive rate); we do not want to miss real cases. Specificity is all about correctly eliminating those patients who do not have cancer (true negative rate).

**Table 20.1** PCA3 and biopsy results

	Biopsy negative	Biopsy positive	Total
$\text{PCA3} < 50$	$\text{TN} = 256$	$\text{FN} = 103$	$\text{TN} + \text{FN} = 359$
$\text{PCA3} \geq 50$	$\text{FP} = 53$	$\text{TP} = 104$	$\text{FP} + \text{TP} = 157$
Total	$\text{TN} + \text{FP} = 309$	$\text{FN} + \text{TP} = 207$	$n = 516$

TN, true negative; FN, false negative; FP, false positive; TP, true positive;  $n$  = total sample size.

Source: Data from de la Taille et al. (2001).

### 20.2.2 Positive and negative predictive value

The *positive predictive value* (PPV) is the proportion of true positives among those patients testing positive:

$$\text{PPV} = \frac{\text{TP}}{\text{FP} + \text{TP}} = \frac{104}{157} = 0.66 \text{ or } 66\%$$

This quantity addresses the key question ‘If I have a positive diagnostic test result, what is the likelihood that I have the disease?’

The *negative predictive value* (NPV) is the proportion of true negatives among those patients testing negative:

$$\text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}} = \frac{256}{359} = 0.71 \text{ or } 71\%$$

This quantity addresses the key question ‘If I have a negative diagnostic test result, what is the likelihood that I do not have the disease?’ Conversely,  $1 - \text{NPV}$  addresses the question ‘If I have a negative diagnostic test result, what is the likelihood that I could still have the disease?’

### 20.2.3 False positive and false negative rates

The *false positive rate* (FPR) is the proportion testing positive among those patients who do not have the disease:

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{53}{309} = 0.17 \text{ or } 17\%$$

The *false negative rate* (FNR) is the proportion testing negative among those patients who do have the disease:

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} = \frac{103}{207} = 0.50 \text{ or } 50\%$$

The FPR records the number of false alarms, while the FNR records the number of missed cases. Note that these two quantities are directly related to specificity and sensitivity, with  $\text{FPR} = 1 - (\text{specificity})$  and  $\text{FNR} = 1 - (\text{sensitivity})$ .

### 20.2.4 Prevalence

In the example taken from de la Taille et al., the proportion of patients with prostate cancer (*prevalence*) is  $207/516 = 0.40$ , or 40%. Suppose we had been assessing the performance of the PCA3 assay (with a cut-off at 50) in a slightly different population, one where the prevalence was, for example, 60%. Table 20.2 provides data for this setting. Note that these data are artificial and were generated simply by changing the prevalence from 40 to 60%.

The values for sensitivity and specificity for the diagnostic test remain unchanged at 50% and 83%, respectively, as do the FPR and FNR. However, the

**Table 20.2** PCA3 and biopsy results, prevalence = 60%

	<b>Biopsy negative</b>	<b>Biopsy positive</b>	<b>Total</b>
PCA3 < 50	TN = 171	FN = 154	TN + FN = 325
PCA3 ≥ 50	FP = 35	TP = 156	FP + TP = 191
Total	TN + FP = 206	FN + TP = 310	n = 516

TN, true negative; FN, false negative; FP, false positive; TP, true positive; n = total sample size.

values for PPV and NPV change: with a prevalence of 60%,  $PPV = 156/191 = 0.82$ , or 82%, and  $NPV = 171/325 = 0.53$ , or 53%. These changes reflect the fact that when prevalence increases, we are much more likely to say that the disease is present *a priori* than we are to say that the disease is absent, so the PPV goes up while the NPV goes down.

Sensitivity and specificity are quantities that objectively measure how good the test is. Altman and Bland (1994) further discuss the connections between PPV, NPV, sensitivity, specificity and prevalence.

### 20.2.5 Likelihood ratio

The so-called *likelihood ratio* (LR) in connection with diagnostic tests is defined as

$$LR = \frac{\text{sensitivity}}{1 - \text{specificity}}$$

This is the ratio of the estimated probability of having a positive diagnostic test result among patients with the disease compared to the estimated probability of also having a positive diagnostic test result among patients who do not have the disease. We want this ratio to be large for a good diagnostic test, and for our case study with a cutpoint of 50, the likelihood ratio is equal to 2.9.

### 20.2.6 Predictive accuracy

If we simply look at *overall predictive accuracy* (PA), it is  $PA = \frac{TP + TN}{n} = \frac{256 + 104}{516} = 0.70$ , or 70%. Note that this quantity also depends on prevalence. For the artificial data in Table 20.2, where the prevalence is increased from 40 to 60%, the PA is now only 63%. The reason is that the PCA3-based test is better at identifying true negatives than it is at identifying true positives. As the prevalence of disease increases, there are fewer negatives, and the diagnostic test performs less well overall.

Most of the measures defined in this section are based on proportions. It is straightforward to obtain standard errors and hence confidence intervals (CIs) for those proportions, as discussed in Section 2.5.2.

### 20.2.7 Choosing the correct cutpoint

The performance of the diagnostic test relies on the cut-off, which is chosen on the underlying continuous scale to classify an observed value on that scale as either positive or negative. In the case study, a score of 50 on the PCA3 scale was the chosen cutpoint. One issue then is, is this the best cutpoint? If we were to change the cutpoint, how would the performance of the test change? Table 20.3 contains values for sensitivity and specificity using several different cutpoints.

**Table 20.3** Sensitivity and specificity of the PCA3 assay

Cutpoint	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)
20	84 (78, 88)	55 (50, 61)
35	64 (57, 71)	76 (71, 81)
50	50 (43, 57)	83 (79, 87)

Source: Data from de la Taille et al. (2011).

The best cutpoint is chosen to balance the trade-off between sensitivity and specificity. Are we more concerned about the true positive rate (identifying true cases) with the emphasis on sensitivity or the true negative rate (eliminating negative cases) with the emphasis on specificity?

#### **CHMP (2009): 'Guideline on clinical evaluation of diagnostic agents'**

*'In case the test decision is based on a cut-off value the trade-off between sensitivity and specificity requires careful analysis with respect to intended applications of an experimental test and their implications on patient care'.*

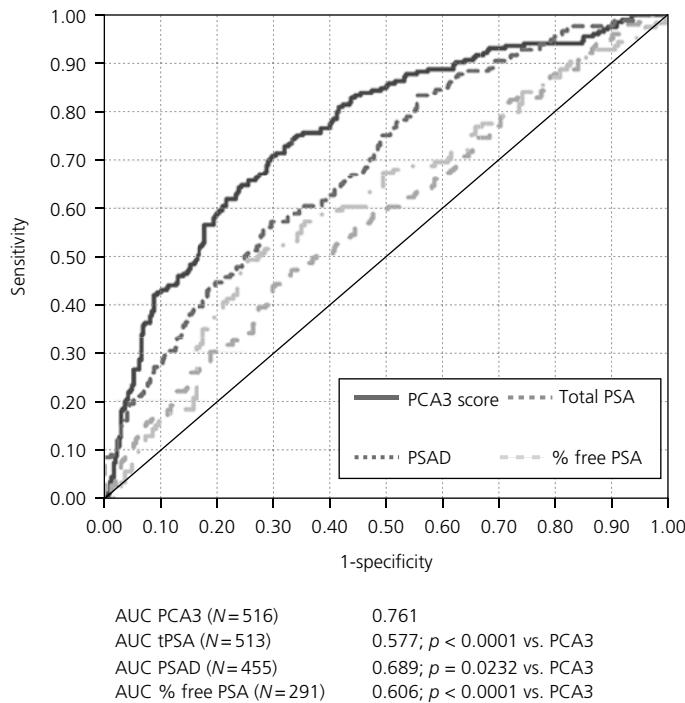
A cutpoint of 35 gives sensitivity of 64% and specificity of 76%, and de la Taille et al. argue that this value provides the optimal balance between the two measures based on clinical considerations.

## 20.3 Receiver operating characteristic curves

### 20.3.1 Receiver operating characteristic

Receiver operating characteristic (ROC) curves plot sensitivity on the  $y$ -axis against  $1 - \text{(specificity)}$  on the  $x$ -axis as the cutpoint (threshold for defining a positive result) is varied. These curves overall provide a summary of the potential for the underlying measure (PCA3 in our example) to form the basis for a good diagnostic test. Figure 20.1, taken from de la Taille et al. (2011), shows four ROC curves together with an ROC curve at a line of  $45^\circ$ , which serves as a reference curve. The PCA3 measure is the basis of one of those curves.

The ROC curve that is the  $45^\circ$  line depicts a diagnostic test with no discriminating ability. Whatever cutpoint is chosen, such a measure gives a test with



**Figure 20.1** ROC curves for PCA3 and other diagnostic tests. Source: de la Taille A, Irani J, Graefen M, et al. (2011) Clinical Evaluation of the PCA3 Assay in Guiding Initial Biopsy Decisions. *Journal of Urology*, **185**, 2119–2125. Reproduced with permission from American Urological Association Education and Research, Inc.

50% sensitivity and 50% specificity; in other words, the test provides a diagnosis based on the toss of a coin. Diagnostic tests with value have associated ROC curves that are to the left of the  $45^\circ$  line – and the more to the left, the better.

A numerical measure of the potential of the diagnostic test is given by the *area under the curve* (AUC). For the test to be useful, the AUC needs to be above 0.5, because a test that assigns based on chance alone achieves that. In theory, it is possible to see the AUC below 0.5; such a test systematically provides a wrong diagnosis. In practice, we rarely come across this situation. The AUC for the PCA3 ROC curve in the case study is 0.761 (de la Taille et al., 2011). It is possible to construct a test of the hypothesis (one-sided) that the diagnostic test has AUC of at least 0.5 ( $H_0$ : AUC  $\leq 0.5$  vs.  $H_1$ : AUC  $> 0.5$ ). Further details of this procedure are provided in Zhou et al. (2002, Chapter 4).

### 20.3.2 Comparing ROC curves

It is of interest to compare ROC curves based on distinct diagnostic procedures. In addition to the PCA3 assay, de la Taille et al. (2011) investigate three other methods: total prostate-specific antigen (tPSA), % free total specific antigen (%free

PSA) and prostate-specific antigen density (PSAD). The AUCs for these alternative measures were 0.577 (tPSA), 0.606 (%free PSA) and 0.689 (PSAD), indicating poorer performance compared to PCA3; the ROC curves are shown in Figure 20.1.

DeLong et al. (1988) have constructed a non-parametric statistical test that formally compares the different measures through their AUC values. For the case study data, the *p*-values comparing PCA3 with each of the three other methodologies were < 0.001 when comparing with tPSA (*n* = 513) and %free PSA (*n* = 291) and 0.023 when comparing with PSAD (*n* = 455), indicating significantly improved performance from the PCA3 assay. The *n*s in brackets here tell us the numbers of patients who gave observations on the measures under consideration.

## 20.4 Diagnostic performance using regression models

There may be variables such as age, prostate volume, etc., that provide additional information (additional to PCA3) on which to base a diagnosis. Regression modelling allows us to take these additional factors into account and assess the extent to which they improve the performance of the diagnostic test. Conversely, these variables of themselves may provide the basis for a diagnosis. The key question then might be, does PCA3 provide any additional discriminating ability? The variables considered may also include the results of other diagnostic tests, and we want to see if the new tests provide additional diagnostic discrimination.

The four variables considered in the modelling exercise for the case study were

- $x_1$  Age (in years)
- $x_2$  Digital rectal examination (DRE), classified as suspicious or unsuspicious
- $x_3$  PSA
- $x_4$  Prostate volume

A logistic regression model (Section 6.4) was fitted to the data by de la Taille et al. This model used prostate cancer/no prostate cancer, as determined by the biopsy, as the binary outcome, with each of these variables plus PCA3 considered as potential *x*-variables in a model for the odds in favour of having prostate cancer as follows:

$$\ln(\text{odds}) = a + b_1 x_1 + b_2 x_2 + \dots$$

The model is initially fitted to the observed data to obtain estimated values of the coefficients  $a$ ,  $b_1$ ,  $b_2$ , and so on. Using this fitted model, it is then possible to predict the odds in favour of having prostate cancer based solely on the *x*-variable values for every patient. If these odds are > 1, then the model predicts that the patient has prostate cancer. If the odds are < 1, then the opposite prediction – that they do not have prostate cancer – is being made for that patient. This then produces a series of individual patient predictions. Based on these, it is possible to form a  $2 \times 2$  table, as we have seen in Tables 20.1 and 20.2, and calculate

**Table 20.4** Logistic regression analysis of potential diagnostic variables

	Base model		Base model plus PCA3		Base model plus PCA3 (cutpoint 35)	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Age	1.07	< 0.001	1.05	0.001	1.04	0.01
DRE	2.30	0.001	2.24	0.003	2.16	0.006
PSA	1.15	0.006	1.15	0.013	1.13	0.013
Prostate volume	0.96	< 0.001	0.96	< 0.001	0.97	< 0.001
PCA3			1.01	< 0.001	4.38	< 0.001
Predictive accuracy	0.737		0.780		0.792	

DRE, digital rectal examination, classified as suspicious or unsuspicious; PCA3, prostate cancer antigen 3; PSA, prostate-specific antigen.

Source: Data from de la Taille et al. (2011).

various summary statistics to measure the model's performance in predicting the outcome. In this discussion, we will focus on predictive accuracy as a measure of performance, although equally, one could look at sensitivity and specificity or any other suitable measure. One thing to note is that the prevalence here affects only the value of  $a$  in the model, not the values of  $b_1$ ,  $b_2$ , etc., and therefore does not impact the observed contribution of each of the variables being considered.

Table 20.4 provides the coefficients of the  $x$ -variables together with  $p$ -values assessing the statistical significance of each variable in the model. Note again that the coefficients ( $b_1$ ,  $b_2$ , etc.) and the associated  $p$ -values are unaffected by the underlying prevalence in the population under study. The *base model* is the model containing the four potential predictors listed earlier in this section. As can be seen, each of those variable coefficients is statistically significantly different from 0, indicating that each variable is providing information as to whether or not a patient has prostate cancer. The base model's overall predictive accuracy for the case study data is 73.7%. PCA3, measured on a continuous scale, is then introduced into the model. The coefficient of PCA3 is significantly different from 0, suggesting that this variable significantly impacts predicting the outcome in addition to the other variables. The statistical significance of all the other variables in the model is also maintained, and the predictive accuracy with PCA3 in the model increases to 78.0%. Finally, a model is fitted with PCA3 included as a binary variable using 35 as the cutpoint. The coding of this variable is 0 if  $\text{PCA3} < 35$  and 1 if  $\text{PCA3} \geq 35$ . Again, all variables in this model are statistically significant, but more importantly, the PA is increased to 79.2% using this model.

The predictive accuracy using PCA3 with 35 as the cutpoint without using the logistic modelling and additional variables, calculated as in Table 20.1, was 71.1%. The modelling, where additional information has been included based on the four key variables (age, DRE, PSA and prostate volume) has increased the

predictive accuracy to 79.2%. It is not uncommon to see increases resulting from a modelling exercise that includes key additional variables and associated diagnostic information. Modelling of this kind can lead to the construction of a *diagnostic index*, which in this case would be the right-hand side of the fitted logistic model earlier. If this index's value was positive for a particular patient, then this would lead to odds of prostate cancer  $> 1$ , while a negative value for the index would lead to odds of prostate cancer  $< 1$ .

## 20.5 Aspects of trial design for diagnostic agents

Trials evaluating new diagnostic tests and comparisons with an existing diagnostic procedure are usually within-patient designs: for example, a crossover, with each patient receiving the investigational diagnostic agent, a comparator agent (or agents) and a definitive diagnosis that establishes the standard of truth. Including the comparator agent allows assessment of superiority to that agent or possibly non-inferiority if the investigational agent offers some advantages such as cost, convenience or safety.

### **CHMP (2009): 'Guideline on clinical evaluation of diagnostic agents'**

*'A trial may be designed to show that an investigational agent is not inferior to a comparator (and thus could be an alternative to this comparator), usually by means of a within-patient comparison. For example, if the study endpoint is the presence or absence of disease, the sensitivities and specificities of the investigational agent and the comparator will be compared (both values are obtained by reference to the standard of truth). The statistical hypothesis may be non-inferiority, superiority or both. However, if superiority fails to be shown, the switch to non-inferiority is not possible'.*

The final sentences in this quotation relate to the switching between superiority and non-inferiority. The reader is referred to Section 12.10 for further discussion on this point.

It is important to avoid *information carry-over* in a crossover design so that the second test is evaluated without knowledge of the results of the preceding diagnostic test. See Section 1.7 for a discussion on carry-over effects in crossover trials. If the number of diagnostic tests that can be performed for a subject is limited by their invasive nature or there is the potential for carry-over, the alternative is a parallel-group design.

In the parallel-group design setting, superiority of the investigational agent over the comparator can be established using the logistic regression techniques detailed in the previous section. Here, PCA3 played the role of the investigational agent, while PSA played the role of the comparator agent. Including PSA in the

model assessing the additional impact of PCA3 means we are considering the investigational agent as an add-on to an existing diagnostic workup. If we were looking to replace an existing diagnostic agent by the new agent, we would compare two models, one including PSA (but not PCA3) with one including PCA3 (but not PSA). This latter scenario could alternatively be evaluated by comparing the ROC AUCs using the DeLong et al. (1988) method, although this would not be in the presence of the additional three variables age, DRE and prostate volume.

The methodology using logistic modelling does not directly evaluate sensitivity and specificity, and it may be appropriate to work more closely with these quantities. Indeed, the regulatory guidelines suggest that these measures be considered as co-primary endpoints. Methods that use these measures directly are available but beyond the scope of this book, and the reader is referred to Chen et al. (2003) for further details. The methods of that paper primarily focus on demonstrating the non-inferiority of the investigational agent vs. the comparator agent, but these methods are easily adapted to deal with superiority by considering the non-inferiority margin to be zero.

When no standard of truth is available, developing new diagnostic agents is more challenging, and it may be necessary to conduct trials focusing on clinical outcomes; does the investigational agent lead to better clinical outcomes in the target population when evaluated against the comparator agent?

#### **CHMP (2009): '*Guideline on clinical evaluation of diagnostic agents*'**

*'Studies assessing patient outcomes may be required if there is no standard of truth to compare to'.*

In these circumstances, an evaluation of the agreement between the investigational and comparator agents in addition may be of value. This is the topic of the next section.

## **20.6 Assessing agreement**

### **20.6.1 The kappa statistic**

The kappa statistic developed by Cohen (1960) is a measure of agreement (*concordance*) between two diagnostic tests. The tests, for example, may be a radiologist's analysis of an X-ray and a computer analysis of the same X-ray. Neither test is perfect, in the sense that neither gives a definitive answer in terms of a tumour being malignant or benign; therefore, neither can be considered the standard of truth. There will inevitably be some level of agreement purely by chance, and the kappa statistic accounts for this chance element in the calculation. Table 20.5 presents some hypothetical data on 94 X-rays, with classification in terms of benign/malignant being made by both the radiologist and the computer. There is quite a lot of agreement between the radiologist and the computer. For a total

**Table 20.5** Classification by radiologist/computer in relation to benign/malignant disease

		Computer		
		Benign	Malignant	Total
Radiologist	Benign	61	2	63
	Malignant	6	25	31
	Total	67	27	<i>n</i> = 94

*n*, total sample size.

of 61 X-rays, both methods result in a diagnosis of benign, while for 25 X-rays, both give a diagnosis of malignant. There are also eight disagreements.

Some agreements will occur purely by chance even if both the radiologist and computer simply guess at the diagnosis. Overall, the radiologist declares benign in 63 out of 94 X-rays (67.0%), while the computer declares benign in 67 out of 94 X-rays (71.3%); there is a slight tendency for the computer to be more likely to result in a diagnosis of benign. If both were diagnosing based on chance, the proportion of cases that would result in a diagnosis of benign would be  $0.670 \times 0.713 = 0.478$ . So, for 47.8% of the X-rays, there would be an agreed diagnosis of benign by chance. The radiologist declares malignant in 31 out of the 94 X-rays (33.0%), while the computer declares malignant in 27 out of the 94 X-rays (28.7%). By chance, we would expect them to agree with a diagnosis of malignant on  $0.330 \times 0.287 = 0.095$ , that is, 9.5% of occasions. Overall agreement will therefore happen purely by chance on  $0.478 + 0.095 = 0.572$ , that is, 57.2% of occasions. The observed level of agreement is  $(61 + 25)/94 = 0.915$ , and the *kappa statistic* is a relative measure of how much better we are compared to chance alone:

$$\kappa = \frac{0.915 - 0.572}{1 - 0.572} = 0.801$$

Higher values for  $\kappa$  signify better agreement. Landis and Koch (1977) set down various thresholds for interpreting the magnitude of this statistic and the quality of the agreement:  $\leq 0.20$  = poor,  $> 0.20$  but  $\leq 0.40$  = fair,  $> 0.40$  but  $\leq 0.60$  = moderate,  $> 0.60$  but  $\leq 0.80$  = good and  $> 0.80$  = very good agreement.

It is possible in theory to get a value for  $\kappa$  that is below zero; this will happen if the level of agreement is worse than chance. Fortunately, such a value is rarely seen in practice.

It is also possible to obtain a CI for  $\kappa$  in the usual way based on assuming large sample normality for the distribution of the kappa statistic. The standard error ( $se$ ) formula is

$$se(\kappa) = \sqrt{\frac{p(1-p)}{n(1-p_e)^2}}$$

where  $p$  is the observed proportion of agreements and  $p_e$  is the expected proportion of agreements that would be obtained by chance. Note that using this notation, the kappa statistic is defined as  $\kappa = (p - p_e)/(1 - p_e)$ . In the example earlier,  $p = 0.915$  and  $p_e = 0.572$ , giving an  $se$  of 0.067 and a 95% CI for kappa of

$$0.801 \pm (1.96 \times 0.067) = (0.67, 0.93)$$

In practice, if the calculated upper limit of this CI is above 1, it is replaced by 1 in the reported CI.

One incorrect approach to evaluating concordance is to calculate a correlation coefficient and a  $p$ -value associated with a test of the hypothesis that the correlation coefficient is zero. A statistically significant  $p$ -value in this case, as mentioned previously in Section 6.10, simply tells us that the two diagnostic methods are not working in totally independent ways; it is telling us nothing about the strength or magnitude of the agreement.

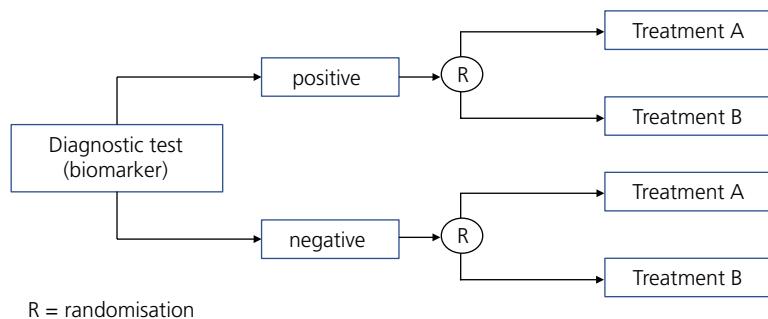
### 20.6.2 Other applications for kappa

The kappa statistic is used outside of diagnosis to measure inter-rater and intra-rater agreement. *Inter-rater agreement* is the agreement between two independent raters or observers. *Intra-rater agreement* concerns agreement between separate assessments by the same observer.

In this context, it is not uncommon for an observer to provide an outcome rating on an ordinal rather than a binary scale. For example, two observers may be rating the general health of an individual as poor, fair, good, and excellent based on some general guidelines. To evaluate the performance of the guideline and the training given to the observers, it may be necessary to ascertain the level of agreement between the two observers. However, not all disagreements are alike. If one observer records a poor rating for an individual while the other observer rates the same individual as excellent, then we have a major disagreement. Had the second observer given a rating of fair, that might not be quite as bad. *Weighted kappa* considers the level of disagreement by assigning weights to the various disagreement possibilities. One simple way of doing this might be to give a weight of 1 to disagreements that are one category apart, a weight of 2 to disagreements that are two categories apart and a weight of 3 to the most major disagreements (those that are three categories apart). Alternatively, we may view the most major disagreements as very problematic and assign a weight of 4 (or even 5) rather than 3. The interested reader is referred to Altman (1991, Sections 14.3) for further details and formulas for calculating weighted kappa.

## 20.7 Companion diagnostics

There has been a substantial amount of recent activity in the development of companion diagnostics to accompany the licensing of novel therapeutic products targeting specific populations of patients. An early companion diagnostic was a



**Figure 20.2** All comers design for testing a therapeutic product in conjunction with a companion diagnostic

test to detect amplification of the HER-2/NEU gene in breast cancer tissue for use with trastuzumab (Campbell, 2021). The purpose of the diagnostic test in this context is to identify a population of patients who can be more effectively treated by the therapeutic product. The diagnostic test here acts as a prognostic or predictive biomarker.

A common design for a clinical trial that is developing a targeted therapeutic product in conjunction with a companion diagnostic is displayed in Figure 20.2 (FDA, 2016).

This design enables a treatment A vs. treatment B comparison to be made in both the diagnostic test (biomarker) positive and diagnostic test (biomarker) negative subpopulations. If the diagnostic test positive identifies a subpopulation of patients with a differential (usually higher) probability for the specific event being targeted, the biomarker is *prognostic*. If the treatment difference with respect to the occurrence of the event is larger in the biomarker-positive group than in the biomarker-negative group, the biomarker is *predictive*. An alternative way to think about this is that if there is a treatment  $\times$  biomarker positive/negative interaction (see Section 10.8), then the biomarker is predictive. If there is no interaction between the treatment and the biomarker, but patients in the biomarker-positive group are nonetheless at differential risk of the event, the biomarker is prognostic.

The event being targeted could be death, for example, and the biomarker-positive subgroup could be at a higher risk of dying. Or the event could be response, with the biomarker-positive subgroup having a higher probability of responding. Finally, the biomarker-positive subgroup could be at high risk of suffering a serious adverse event (SAE) from the control treatment being evaluated. In each of these settings, the biomarker is prognostic for the outcome, and the experimental treatment may either reduce the incidence of the event (death, SAE) or increase the incidence of the event (response) compared to control in a clinically meaningful way. If the biomarker is predictive, the experimental

treatment shows a differential effect compared to control in the incidence of these events, which could be of clinical utility.

In the discussion so far, we have considered the diagnostic test as definitive for identifying biomarker positive/negative subgroups. This is unlikely to be the case, and there will be misclassifications. Further, there may well be several competing diagnostic tests for identifying biomarker positive/negative subgroups, with no 'standard of truth'. In such cases, it is not possible to calculate sensitivity, specificity, positive predictive value or negative predictive value on which to base comparisons, and evaluation of the competing diagnostic tests requires careful consideration. As pointed out by Campbell (2016), there are at least 10 companion diagnostics for the oncogene HER2 to guide the use of trastuzumab in the treatment of breast cancer. The interested reader is referred to the Campbell article for further discussion on these points and reference to regulatory guidelines related to companion diagnostics.

## CHAPTER 21

# The role of statistics and statisticians

### 21.1 The importance of statistical thinking at the design stage

A clinical trial is an experiment. Not only do we have to ensure that the clinical elements fit with the objectives of the trial, but we also have to design the trial in a tight scientific way to make sure it is capable of providing valid answers to the key questions in an unbiased, precise and structured way. This is where statistics comes in, and statistical thinking is a vital element of the design for every clinical trial.

The following list of areas where statistical thinking is required is not exhaustive but is meant to give a flavour of the sorts of things that need to be considered:

- What are the key prognostic factors, and how, if at all, should these be used to stratify the randomisation?
- Should the randomisation be stratified by centre or by some higher-level factor such as region or country?
- What are the implications for block size to ensure balance and prevent inadvertent unblinding?
- How should we choose primary and secondary estimands in line with the clinical objectives for the trial?
- What are the expectations regarding withdrawals, and what can be done to minimise/prevent those withdrawals?
- How should we deal with the missing data to align with the primary and secondary estimands?
- What methods can be used to control variability to increase precision both practically and from a statistical modelling perspective?
- Which objectives form part of the confirmatory strategy for the trial, and which elements are purely exploratory?
- What statistical testing strategy will provide valid answers to the range of questions being asked, particularly in terms of controlling multiplicity in the confirmatory setting?

- For each of the comparisons being considered, is the focus superiority, equivalence or non-inferiority? In the latter two cases, how should we choose the margins?
- How will the homogeneity of treatment effect be assessed, and what sub-groups will be evaluated in this regard?
- Is it appropriate to build in an interim analysis given the nature of the trial? Is this practical, and if so, how should the interim analysis be structured?
- How many patients are needed to provide answers to the questions being asked, and what are the assumptions upon which this calculation is based?
- Do we need to revisit the sample size calculation at some interim stage in the trial?
- How are safety data being collected and evaluated and will the trial be sufficiently large to allow the appropriate assessment of safety?
- How should results from statistical analyses be presented to enable efficient and correct interpretation?

These points relate to each individual trial, but there will be similar considerations at the level of the development plan. For the overall ordered programme of clinical trials to be scientifically sound, there needs to be a substantial amount of commonality across the trials in terms of endpoints, estimands, definitions of analysis sets, recording of covariates and so on. This will facilitate the use of integrated summaries and meta-analysis for the evaluation and presentation of the complete programme or distinct parts of that programme and, outside of that, will allow a consistency of approach to evaluating the different trials.

At both the trial and development plan levels, statisticians should take time to review the case report forms (CRFs) to make sure, especially, that the data being collected will be appropriate for the precise, unambiguous and unbiased evaluation of primary and secondary estimands. Other aspects of the data being collected should also be reviewed in light of how they will be used in the analysis. For example, baseline data will provide information on covariates to be used in any adjusted analyses, and intermediate visit data may be needed for the use of methods to deal with missing data: for example, mixed models for repeated measures (MMRM).

## 21.2 Regulatory guidelines

Statistical thinking and practice in the pharmaceutical industry is very much determined by the regulatory guidelines that are in place. ICH E9, *Statistical Principles for Clinical Trials*, published in 1998, sets down the broad framework within which we operate. In 2001, we saw the publication of ICH E10, *Choice of Control Group*, containing advice on the appropriate choice of a concurrent control group and introducing the concept of assay sensitivity (see Section 12.5) in active control, non-inferiority trials. An addendum to ICH E9, dealing with estimands, came into effect in 2020. This guideline grew out of concerns about the use of simple methods for missing data (such as last observation carried forward [LOCF]), the proliferation of new methods and what clinical questions such

methods address; also the increasing drift away from intention-to-treat and the implications for bias that result from that drift without explicitly defining the population under consideration; and a somewhat unstructured approach to the specification of sensitivity analyses. This guideline is beginning to have a substantial impact on the way we structure our questions, how we deal with intercurrent events, which often lead to missing data or data that are not relevant for the clinical question in hand, and the associated analyses that specifically address the robustness of the primary methods of analysis.

Each of the main agencies – EMA in Europe, FDA in the United States and PMDA in Japan – have produced guidelines dealing with specific statistical topics, and many of these have been mentioned in earlier chapters of this book. Those that have been used in our discussions have been placed in the references list. The regulatory landscape is ever-changing, and further guidelines undoubtedly will be issued, while the existing ones will be revised as time passes.

In previous editions of this book, this section has contained a list of the key guidelines from ICH, EMA and FDA that provide guidance on statistical methods. Given the ever-changing nature of such guidance, such a list would, at best, provide a snapshot. For more relevant up-to-date information, the reader is referred to the respective web pages ([www.ich.org](http://www.ich.org), [www.ema.europa.eu](http://www.ema.europa.eu), [www.fda.gov](http://www.fda.gov)) and, in relation to Japan, [www.pmda.go.jp](http://www.pmda.go.jp).

Finally, most therapeutic-specific guidelines contain recommendations that directly impact statistical considerations: for example, in terms of defining endpoints, the requirement for more than one primary endpoint in some settings, defining analysis sets, choosing estimands, and establishing the non-inferiority margin. In a particular therapeutic setting, it is self-evident that the requisite guidelines should be studied carefully to extract relevant information for statistical aspects of design and analysis.

## 21.3 The statistics process

We have already discussed the role that statistics and statisticians play in the design of clinical trials and programmes of clinical trials. In this section, we will look at the results of that planning in terms of the statistical methods section of the protocol and, following on from that, what happens from a statistical standpoint once the trial is ongoing through to the final reporting of that trial and construction of the regulatory package.

### 21.3.1 The statistical methods section of the protocol

The statistical methods section of the protocol sets down the main aspects of both design and analysis. This section should contain the following:

- Justification of the sample size (including the possible re-evaluation of the sample size once the trial is ongoing).

- Method of randomisation, including stratification (although block size will not be specified).
- Clear definition and delineation of the primary and secondary estimands to include population, variable and intercurrent events.
- How the primary and secondary estimands link with the objectives of the trial.
- Which aspects of the analysis are to be viewed as confirmatory, and which are to be viewed as exploratory.
- How multiplicity will be dealt with within the confirmatory parts of the analysis.
- Detail regarding the methods of analysis for the primary endpoint(s), including specification of covariates to be the basis of any adjusted analyses.
- How missing data will be handled, and consideration of sensitivity analyses.
- Overview of statistical methods for the analysis of secondary endpoints.
- Methods for handling safety and tolerability data. Safety data being coded using the *Medical Dictionary for Regulatory Activities* (MedDRA; [www.meddramsso.com](http://www.meddramsso.com)) coding system.
- Interim analyses and how the type I error will be divided across these.
- Software to be used for statistical analysis.

Only methods set down in the protocol can be viewed as confirmatory, so it is very important to get this section right; mistakes and ambiguities can be costly.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Only results from analyses envisaged in the protocol (including amendments) can be considered as confirmatory'.*

#### **21.3.2 The statistical analysis plan**

The *statistical analysis plan* (SAP) is a more detailed elaboration of the statistical methods of analysis contained in the protocol. The SAP is written as the trial is ongoing but before database lock and the breaking of the treatment codes to unblind those involved in analysing the data. The SAP for open-label studies should be finalised before the statistics team involved in analysing and reporting the data has access to any part of those data, to avoid bias potentially introduced by exposure to treatment-specific data.

The SAP sometimes also contains table templates that set down the precise way in which the statistical analysis will be presented.

#### **21.3.3 The data validation plan**

Once the CRF is finalised, the data management team puts together a *Data Validation Plan*, which sets down the data checks that will be made in conjunction with the data entry process; for example, are the visit dates in chronological order, are the ages of the patients within the range specified in the inclusion criteria, and so on. It is useful for a statistician to review this plan, especially regarding

issues related to the definition of endpoints, for two reasons. Firstly, it is important that the statistics team is aware of what data checks are being undertaken so that at the data analysis stage, they can rule out potential data problems and be assured of a certain level of data quality. Secondly, the statistician may be able to suggest other specific checks that will help increase the quality of those data and the subsequent analysis.

#### **21.3.4 The blind review**

There is one final opportunity to revisit the proposed statistical analysis methods prior to breaking the blind or, in an open-label trial, before the statistics group tasked with analyzing the data, have seen study data. This so-called *blind review* usually takes place around the time of database lock, and the following are some of the aspects of analysis that are generally considered:

- Precise definition of analysis sets: specifically, which patients are to be included and which are to be excluded
- Finalisation of the algorithms for handling missing data
- Finalisation of algorithms for combining centres, should this be required
- Outlier identification and specific decisions on how these are to be handled

Under normal circumstances, the blind review should take place over a 24- or 48-hour period to limit delays in beginning the statistical analysis. The blind review should be documented, detailing precisely what was done.

Sometimes, the blind review reveals data issues that require further evaluation by the data management group, with data queries raised that may result in changes to the database. This sequence of events can cause major headaches and delays in statistical analysis and reporting. It is important to get the data validation plan correct in the planning phase so that issues can be identified and dealt with in an ongoing way.

#### **21.3.5 Statistical analysis**

The SAP details the precise methods of analysis and presentation and should ideally be finalised well before database lock. This enables work to begin in good time on the programming of the analyses. These programs are tested on *dirty* data from the trial so that they can be pretty much finalised before the trial ends, enabling (at least in theory) a rapid turnaround of the key analyses.

This is not always as simple as it sounds. Working with dirty data can bring its own problems, including illogical data values that the programs cannot handle. Also, when the final data arrive, specific issues and data problems may arise that the earlier *dry runs* did not pick up. Nonetheless, these aspects of planning and program development and validation are essential in order for us to complete the statistical analyses and analysis presentations quickly.

The analyses and tables are a joint effort involving statisticians and statistical programmers. Quality control (QC) is an essential component of this part

of the process, and double programming is frequently undertaken: that is, every analysis and all table entries are reproduced independently by a second programmer and cross-checked against the original. Data listings are also produced and checked, although the level of checking may not be as rigorous as with the tables. Figures and graphs require a different kind of QC, but the points on these figures and graphs should be verified independently by a second programmer.

### 21.3.6 Reporting the analysis

ICH E3 (1995), *Note for Guidance of the Structure and Content of Clinical Study Reports*, sets down the structure required within the regulatory setting for reporting each study, down to the numbering of the sections and precisely what goes in each section. Medical writers work with statisticians and the clinical team to put these reports together.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'... statistical judgement should be brought to bear on the analysis, interpretation and presentation of the results of a clinical trial. To this end the trial statistician should be a member of the team responsible for the clinical study report, and should approve the clinical report'.*

The statistician contributes to several areas in the clinical report. In particular, Section 11, 'Efficacy Evaluation', and Section 12, 'Safety Evaluation', require statistical oversight. Section 16 of the report contains the appendices, and Subsection 16.1.9, 'Documentation of Statistical Methods', is usually written by the trial statistician.

Within Section 11, Subsection 11.4.2 ('Statistical/Analytical Issues') contains a series of items covering many of the areas of complexity within statistical analysis:

- Adjustment for covariates
- Handling of dropouts or missing data
- Interim analyses and data monitoring
- Multi-centre studies
- Multiple comparisons/multiplicity
- Use of an efficacy subset of patients
- Active control studies intended to show equivalence
- Examination of subgroups

Each of these clearly requires input from the statistician.

### 21.3.7 Pre-planning

A common theme running across almost everything we do within our statistical analyses is the need for pre-planning.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principle features of its proposed analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial'.*

This pre-planning, in terms of both conduct and analysis, is predominantly set down in the trial protocol. Pre-planning is one key aspect of how we design and run our trials that helps reduce bias. It would be entirely inappropriate to take decisions about methods of analysis based on unblinded looks at the data. Pre-planning also enables us think through in advance how we will handle the data. This is good discipline and can help us to anticipate problems. A final benefit of pre-planning is very practical: once the trial is complete and the database is locked, there is inevitably a *mad dash* to analyse the data and look at the results. Only with pre-planning and effective pre-programming and testing of those programs can the statistical analyses be undertaken quickly and without major hitches.

As the trial is ongoing, there is also an opportunity to change some of the planned methods of analysis; for example, information that a particular covariate could be important or that a different kind of effect could be seen in a certain subgroup may have become available based on external data from a similar trial that has been completed and reported. Of course, such decisions should be made blind to the trial data to avoid concerns regarding bias. Such changes can be incorporated by modifying the SAP; if they represent major changes to the analysis – for example, if they were associated with the analysis of the primary endpoint – then a protocol amendment will need to be issued. The reason, as mentioned earlier, is that only methods specified in the protocol, including amendments, can be viewed as confirmatory, and issuing an amendment ticks that box.

In a limited way, there may also be changes in the design as the trial is ongoing, such as resizing the trial. Such changes represent major design modifications, and protocol amendments are needed to align with regulatory guidance.

**ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report . . . The potential need for re-estimation of the sample size should be envisaged in the protocol whenever possible'.*

The final sentence again emphasises the need for pre-planning.

A change in the statistical methods at the data analysis stage, such as including additional covariates in the statistical model or using a transformation of the primary endpoint when one was not planned, is problematic. The danger here is that a method may be chosen that affects the resulting magnitude of the treatment effect. The choice of statistical method should be pre-specified in the SAP and possibly modified at the blind review. However, changes to these methods will be more acceptable in conjunction with a clearly defined algorithm. For example, the logrank test may be the planned method of analysis for survival data, but if the assumption of proportional hazards is not valid according to some pre-defined assessment of that assumption, then treatment comparisons could be based on the Gehan–Wilcoxon test. Alternatively, it can be stated in the protocol that if, on visual inspection of the normal probability plot, the data appears to be positively skewed, then the log transformation will be used to recover the normality of the data. These are examples of clearly defined algorithms leading to a well-defined method of analysis for calculating the *p*-value. Of course, *visual inspection* contains an element of subjectivity, but regulators can see a clear way through the decision-making process.

There is one possible area where pre-planning rules may be relaxed, which is in relation to orphan indications/small populations. Regulators recognise that many of these situations are very challenging. For example, even the primary endpoint could be difficult to define in a particular setting, and results for a range of endpoints, without the formal need to control for multiplicity, may be the only way to adequately demonstrate a treatment benefit.

#### **CHMP (2006): 'Guideline on clinical trials in small populations'**

*'The choice of the primary endpoint may pose considerable problems. In some cases, the "most appropriate" clinical endpoint may not be known or widely agreed or a validated clinical endpoint may not exist. . . . In such circumstances, the usual approach of pre-specifying the primary endpoint may be too conservative and more knowledge may be gained from collecting all sensible/possible endpoints and then presenting all the data in the final study report'.*

A further issue concerns new questions that arise during data analysis. These aspects should be clearly distinguished and will constitute only exploratory analyses.

#### **ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'**

*'Although the primary goal of the analysis of a clinical trial should be to answer the questions posed by its main objectives, new questions based on the observed data may well emerge during the unblinded analysis. Additional and perhaps complex statistical analysis may be the consequence. This additional work should be strictly distinguished in the report from work which was planned in the protocol'.*

### 21.3.8 Sensitivity and robustness

Statisticians and regulators alike, quite rightly, place great store on robustness and sensitivity analyses. All analyses are based on certain assumptions regarding the data, such as normality and constant variance or independent censoring in time-to-event data. Analyses could potentially be affected by the presence of single outlying data points or be sensitive to the definition of the full analysis set or the handling of missing data. It would be very unsatisfactory if the conclusions drawn from the data were driven by questionable assumptions or unduly influenced by different choices for dealing with specific aspects of the data. Throughout the statistical analysis, the sensitivity of the conclusions to assumptions of the kind mentioned should be evaluated. Section 8.4.2 contains a discussion on sensitivity analyses within the estimands framework. Such analyses are pre-planned. Other sensitivity analyses may also be conducted when certain aspects and issues with the data are unanticipated. The regulators mention these issues on several occasions and in relation to numerous aspects of analysis and interpretation:

- Missing data

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'  
*'An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial'.*

- Outliers

The following quote follows on from that on missing data:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'  
*'A similar approach should be adopted to exploring the influence of outliers . . . If no procedures for dealing with outliers was foreseen in the trial protocol, one analysis with the actual values and at least one other analysis eliminating or reducing the outlier effect should be performed and differences between the results discussed'.*

In certain cases, more specific guidance is given. The FDA discuss various sensitivity analyses concerning the analysis of progression-free survival, for example, in FDA (2018), *Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics*.

A general message arising from these issues is that where there are doubts regarding how to handle specific aspects of the data at the analysis stage, a range of different approaches (say, two or three) should be considered in the sensitivity analyses that hopefully support the robustness of the conclusions. However, if the conclusions are affected by this choice, this could undermine their validity.

## 21.4 The regulatory submission

The statistics group(s) involved in analysing and reporting each trial have a role in compiling the regulatory submission. Both the integrated summary of safety (ISS) and the integrated summary of efficacy (ISE) involve the analysis and

presentation of compiled results across the whole programme of trials. Formal meta-analysis may be employed or, alternatively, a pooling of data; the technical aspects of these methods are discussed in Chapter 18.

Once the regulatory submission has been made, questions and issues will inevitably come back from the regulators. There may be concerns about how the data have been handled from a statistical point of view. The regulators may request additional specific analyses to resolve uncertainty. And there may be open issues that the regulators are unhappy about and that may require a substantial amount of further analysis. In each of these cases, there will be a need for statistical consideration.

If additional analyses are specifically requested, providing them should be straightforward. However, if the questions are more general, the company may need to respond by providing a series of reanalyses to address the issues. In this case, the concept of pre-planning is irrelevant; those deciding what further analyses to present are unblinded to the data. This scenario of itself creates substantial difficulties. Of course, there is the temptation to reanalyse in several different ways but only present to the regulators those analyses that support the company's position, and the regulators are well aware of this. The best way to avoid potential criticism is to be open with the regulators and present a range of analyses that fit with the questions and issues raised.

In the US, the FDA request an electronic version of the database within the submission. This gives them the opportunity to not only reanalyse the data to confirm the results presented but also perform their own alternative analyses. This does not currently happen in Europe. Therefore, the process following submission is somewhat different in the US, and FDA statisticians take care of much of the interchange in terms of requesting and supplying alternative analyses.

## 21.5 Publications and presentations

Outside of the clinical report and regulatory submission, the results of trials need to be published in the medical literature and presented at conferences and meetings.

In recent years, there has been a range of recommendations regarding the structure of publications: how they should be laid out and what they should contain. These recommendations have usually been in the form of checklists, which have been encapsulated within the *Consolidated Standards of Reporting Trials* (CONSORT) statement (Altman et al., 2001; Moher et al., 2001). The statement was revised further in 2010 (Schulz et al., 2010). Increasingly, many medical journals have adopted this guidance in terms of requiring their clinical trial publications to conform to it. There is a website that provides up-to-date information and helpful resources and examples: [www.consort-statement.org](http://www.consort-statement.org).

The guideline breaks down the content of each publication into a series of 22 items, ranging from

1. Title and Abstract
2. Introduction – Background
3. Methods – Participants
4. Methods – Interventions

through to

7. Methods – Sample Size
13. Results – Participant Flow

and

22. Discussion – Overall Evidence

The precise content is too detailed to provide complete coverage here, but to give an impression, we will consider two areas: the sample size calculation (item 7) and participant flow (item 13).

The sample size calculation should be detailed in the trial publication, indicating the estimated outcomes in each treatment group (which defines the clinically relevant difference to be detected), the type I error, the type II error/power and, for a continuous primary outcome variable in a parallel-group trial, the within-group standard deviation for that measure. For time-to-event data, details on clinically relevant differences are usually specified in terms of either the median event times or the proportions event-free at a certain time point.

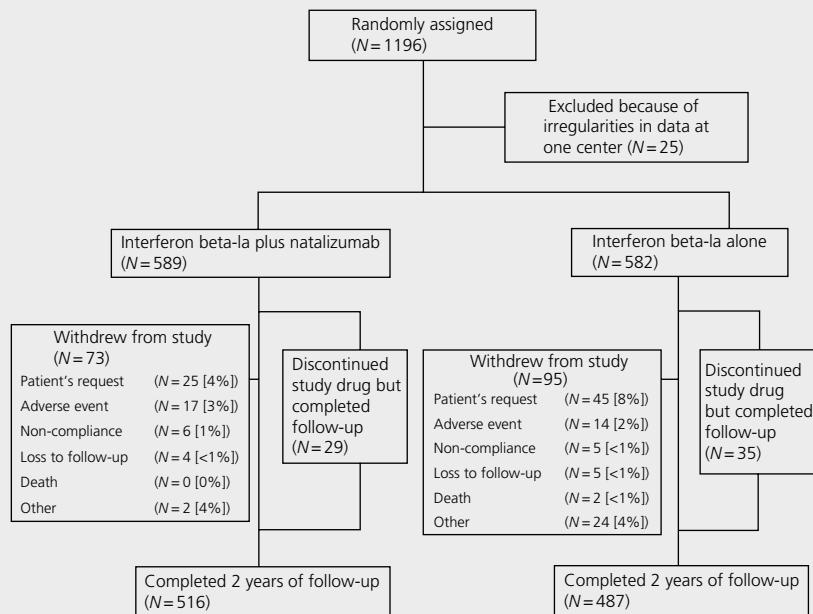
An important aspect of the reporting of any clinical trial is a clear indication of what happened to all patients randomised. CONSORT recommends that each publication contain a diagram showing the flow of participants through the trial, numbers randomised to each of the treatment groups receiving intended treatment, protocol deviations by treatment group classified by type of deviation and patient groups analysed for primary outcomes. Figure 21.1 in Example 21.1 shows an example.

The quality of publications is increasing, partly due to guidance of the type already described in this section. It is unfortunately the case, however, that many mistakes in design, analysis, reporting and interpretation are still made, even in leading journals, despite apparently rigorous refereeing procedures. Particular areas of statistics that seem to cause consistent difficulties include

- Assuming that a large clinical trial necessarily removes bias. Large studies have greater precision, but precision and bias are different things.
- Over-stratification. Stratification for important prognostic factors is important to ensure balance in the mix of patients. But stratifying by too many factors can have precisely the opposite effect.
- Incorrect interpretation of the odds ratio.
- Interpreting the  $p$ -value from the logrank test or the Cox model in terms of a comparison of the medians.
- Assuming that a non-significant  $p$ -value from a superiority comparison indicates that the treatments are similar in their effect.

**Example 21.1** Natalizumab plus interferon beta-1a for relapsing multiple sclerosis:  
The SENTINEL study

Figure 21.1 shows the participant flow of this placebo-controlled trial of natalizumab (Rudick et al., 2006). Reproduced by permission of Massachusetts Medical Society.



**Figure 21.1** Patient disposition in the SENTINEL trial. Source: Rudick RA, Stuart WH, Calabresi PA, Confavreux C, et al. for the SENTINEL Investigators (2006) Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *NEJM*, **354**, 911–923.

- The correct design, analysis, reporting and interpretation of non-inferiority trials. Remember, conventional *p*-values have no role.
- Selecting good and bad subgroups from, and general over interpretation of, a forest plot evaluating treatment effects in subgroups.
- Adjusting the analysis for baseline factors. With change from baseline as the outcome variable, including baseline as a covariate to avoid regression towards the mean is always needed. Note that it is incorrect, in general, to use covariates that are measured after randomisation.

More recently, we have seen several journals adopt a strict policy regarding the reporting of *p*-values. The proliferation of *p*-values in publications and consequent misinterpretation of those *p*-values without proper control of multiplicity has been a source of major concern to many statisticians over the years. The position

adopted by the *New England Journal of Medicine* (Harrington and D'Agostino, 2019) has already been reported in Section 10.3 but will be repeated here:

*'The new guidelines discuss many aspects of the reporting of studies in the Journal, including a requirement to replace P values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity'.*

Many other journals have adopted a similar position, which is to be applauded!

It is important to enlist the help of statistical colleagues when putting together publications, not only in terms of the actual analysis but for interpretation and reporting in the publication itself. Further, cast a critical eye over the statistical methodology in the papers you review, highlight possible problem areas and, again, request the help of your statistical colleagues. Do not automatically assume that just because a publication is in a leading journal, everything is correct from a statistical point of view.

Of course, presentations at conferences do not go through the same critical review process as publications. Even though abstracts are often submitted and reviewed in advance, mistakes and bad practices slip through. It is important to have statistical input when putting together these presentations. Errors in the statistics will invariably be picked up by some audience members, and the resulting bad press could be damaging. From the opposite perspective, look critically at other presentations that contain statistics, and challenge the presenters if you feel that inappropriate methods are being used.

# References

- Altman DG (1991) *Practical Statistics for Medical Research* London: Chapman & Hall.
- Altman DG (1998) 'Confidence intervals for the number needed to treat' *British Medical Journal*, **317**, 1309–1312.
- Altman DG and Bland JM (1994) 'Diagnostic tests 2: predictive values' *British Medical Journal*, **309**, 102.
- Altman DG, Schulz KF, Moher D, et al. (2001) 'The revised CONSORT statement for reporting randomized trials: explanation and elaboration' *Annals of Internal Medicine*, **134**, 663–694.
- Amit O, Heidberger RM and Lane PW (2008) 'Graphical approaches to the analysis of safety data from clinical trials' *Pharmaceutical Statistics*, **7**, 20–35.
- Anagelis A and Kavvounos P (2017) 'Multiple Criteria Decision Analysis (MCDA) for evaluating new medicines in Health Technology Assessment and beyond: The Advance Value Framework' *Social Science & Medicine*, **188**, 137–156.
- Arani RB, Soong S-J, Weiss HL, et al. (2001) 'Phase specific analysis of herpes zoster associated pain data: a new statistical approach' *Statistics in Medicine*, **20**, 2429–2439.
- Aroda VR, Saugstrup T, Buse JB, et al. (2019) 'Incorporating and interpreting regulatory guidance on estimands in diabetes clinical trials: The PIONEER 1 randomized clinical trial as an example' *Diabetes, Obesity and Metabolism*, **21**, 2203–2210.
- Austin PC and Mamdani MM (2006) 'A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use' *Statistics in Medicine*, **25**, 2084–2106.
- Austin PC and Stuart EA (2015) 'Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies' *Statistics in Medicine*, **34**, 3661–3679.
- Bate A, Lindquist M, Edwards IR, et al. (1998) 'A Bayesian neural network method for adverse drug reaction signal generation' *European Journal Clinical Pharmacology*, **54**, 315–21.
- Bauer P and Köhne K (1994) 'Evaluation of experiments with adaptive interim analyses' *Biometrics*, **50**, 1029–1041.
- Beigel JH, Tomashek KM, Dodd LE, et al. (2020) 'Remdesivir for the Treatment of Covid-19 – Final Report' *New England Journal of Medicine*, **383**, 1813–1826.
- Bellingan G, Maksimow M, Howell DC, et al. (2013) 'The effect of intravenous interferon-beta-1a (FP-1201) on lung CD73 expression and on acute respiratory distress syndrome mortality: an open-label study' *The Lancet Respiratory Medicine*, **2**, 98–107.
- Bland M (2004) 'Cluster randomised trials in the medical literature: two bibliometric surveys' *BMC Medical Research Methodology*, **4**, 21.
- Borenstein M, Hedges LV, Higgins JP and Rothstein HR (2010) 'A basic introduction to fixed-effect and random-effects models for meta-analysis' *Research Synthesis Methods*, **1**, 97–111.
- Bradburn MJ, Deeks JJ, Berlin JA and Localio AR (2007) 'Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events' *Statistics in Medicine*, **26**, 53–77.

- Breslow NE and Day NE (1994) *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies*, IARC Scientific Publications, No. 82. New York: Oxford University Press.
- Brodie MJ, Richens A and Yuen AW (1995) 'Double-blind comparison of lamotrigine and carbamazepine in newly diagnosed epilepsy. UK Lamotrigine/Carbamazepine Monotherapy Trial Group' *The Lancet*, **345**, 476–479.
- Brodie MJ and Whitehead J (2006) 'Active control comparisons: the ideal trial design' *Epilepsy Research*, **68**, 70–73.
- Campbell G (2021) 'The role of statistics in the design and analysis of companion diagnostic (CDx) studies' *Biostatistics & Epidemiology*, **5**, 218–231.
- Campbell MJ, Donner A and Klar N (2007) 'Developments in cluster randomised trials and Statistics in Medicine' *Statistics in Medicine*, **26**, 2–19.
- Cancer Therapy Evaluation Program (2006) *Common Terminology Criteria for Adverse Events Version 3.0*, DCTD, NCI, NIH, DHHS, 31 March 2003. [https://ctep.cancer.gov/protocoldevelopment/electronic\\_applications/docs/ctcaev3.pdf](https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcaev3.pdf), accessed on 12 October 2022.
- Cannon CP, Pratley R, Dagogo-Jack S et al. (2020) 'Cardiovascular Outcomes with Ertugliflozin in Type 2 Diabetes' *New England Journal of Medicine*, **383**, 1425–1435.
- Carpenter J and Kenward MG (2007) *Missing Data in Clinical Trials – a Practical Guide*. UK National Health Service, National Co-ordinating Centre for Research on Methodology.
- Chen JJ, Hsueh H-M and Liu J-P (2003) 'Simultaneous non-inferiority test of sensitivity and specificity for two diagnostic procedures in the presence of a gold standard' *Biometrical Journal*, **45**, 47–60.
- Chisholm O, Sharry P and Phillips L (2022) 'Multi-Criteria Decision Analysis for Benefit-Risk Analysis by National Regulatory Authorities' *Frontiers in Medicine*, **8**. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8790083/>, accessed on 12 October 2022.
- Chou R, Fu R, Huffman LH and Korthuis (2006) 'Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses' *The Lancet*, **368**, 1503–1515.
- Clamp AR, James EC, McNeish IA et al. (2019) 'Weekly dose-dense chemotherapy in first-line epithelial ovarian, fallopian tube, or primary peritoneal carcinoma treatment (ICON8): primary progression free survival analysis results from a GCIG phase 3 randomised controlled trial' *The Lancet*, **394**, 2084–2095.
- Cochran WG (1954) 'The combination of estimates from different experiments' *Biometrics*, **10**, 101–129.
- Cohen J (1960) 'A coefficient of agreement for nominal scales' *Educational and Psychological Measurement*, **20**, 37–46.
- Coronary Drug Project Research Group (1980) 'Influence and adherence to treatment and response of cholesterol on mortality in the coronary drug project' *New England Journal of Medicine*, **303**, 1038–1041.
- Cox DR (1972) 'Regression models and life tables (with discussion)' *Journal of the Royal Statistical Society, B*, **74**, 187–220.
- Crawford ED, Eisenberger MA, McLoed DG, et al. (1989) 'A controlled trial of leuprolide with and without flutamide in prostatic cancer' *New England Journal of Medicine*, **321**, 419–424.
- Cumming G (2009) 'Inference by eye: reading the overlap of independent confidence intervals' *Statistics in Medicine*, **28**, 205–220.
- D'Agostino RB (1998) 'Propensity score methods for bias reduction in the comparison of a treatment to a non-randomised control group' *Statistics in Medicine*, **17**, 2265–2281.
- de la Taille A, Irani J, Graefen M, et al. (2011) 'Clinical evaluation of the PCA3 assay in guiding initial biopsy decisions' *Journal of Urology*, **185**, 2119–2125.
- Deeks JJ, Dinnis J, D'Amico R, et al. (2003) 'Evaluating non-randomised intervention studies' *Health Technology Assessment*, **7**, iii–x, 1–173.

- DeLong ER, DeLong DM, and Clarke-Pearson DL (1988) 'Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach' *Biometrics*, **44**, 837–845.
- Doll R and Bradford Hill A (1950) 'Smoking and carcinoma of the lung' *British Medical Journal*, **2**, 739–748.
- Ebbutt AF and Frith L (1998) 'Practical issues in equivalence trials' *Statistics in Medicine*, **17**, 1691–1701.
- Egger M and Smith D (1998) 'Meta-analysis bias in location and selection of studies' *British Medical Journal*, **316**, 61–66.
- Egger M, Smith GD, Schneider M and Minder C (1997) 'Bias in meta-analysis detected by a simple, graphical test' *British Medical Journal*, **315**, 629–634.
- Ellenberg SS, Fleming TR and DeMets DL (2019) *Data Monitoring Committees in Clinical Trials: A Practical Perspective* (2nd edn). John Wiley & Sons, Inc.
- Emerson SS, Levin GP and Emerson SC (2011) 'Comments on adaptive increase in sample size when interim results are promising: a practical guide with examples' *Statistics in Medicine*, **30**, 3285–3301.
- Eriksson BI, Dahl OE, Rosenthaler N, et al. (2007) 'Dabigatran etexilate versus enoxaparin for prevention of venous thromboembolism after total hip replacement: a randomized, double-blind, non-inferiority trial' *The Lancet*, **370**, 949–956.
- Evans SJ, Waller PC and Davis S (2001) 'Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports' *Pharmacoepidemiol Drug Safety*, **10**, 483–486.
- Fleming TR and DeMets DL (1996) 'Surrogate end points in clinical trials: are we being misled?' *Annals of Internal Medicine*, **125**, 605–613.
- Friede T and Kieser M (2006) 'Sample size recalculation in internal pilot study designs: a review' *Biometrical Journal*, **48**, 537–555.
- Gardner MJ and Altman DG (1989) 'Estimation rather than hypothesis testing: confidence intervals rather than p-values' In: *Statistics with Confidence* (eds. MJ Gardner and DG Altman) London: *British Medical Journal*, 6–19.
- Gehan EA (1969) 'Estimating survival functions from the life table' *Journal of Chronic Diseases*, **21**, 629–644.
- Geyer CE, Forster J, Lindquist D, et al. (2006) 'Lapatinib plus capecitabine for HER-2-positive advanced breast cancer' *New England Journal of Medicine*, **355**, 2733–2743.
- Gibbons RD and Amatya AK (2016) *Statistical Methods for Drug Safety*. New York, CRC Press.
- Glasziou P, Simes RJ and Gelber RD (1990) 'Quality adjusted survival analysis' *Statistics in Medicine*, **9**, 1259–1276.
- Gray R (1988) 'A class of k-sample tests for comparing the cumulative incidence of a competing risk' *Annals of Statistics*, **16**, 1141–1154.
- Greenwood M (1926) 'The errors of sampling of the survivorship tables' *Reports on Public Health and Statistical Subjects*, No. 33, Appendix 1. London: HMSO.
- Grieve AP (2003) 'The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes?' *Pharmaceutical Statistics*, **2**, 87–102.
- Grimes DA and Schulz KF (2008) 'Making sense of odds and odds ratios' *American Journal of Obstetrics and Gynecology*, **111**, 423–426.
- Harrington D, D'Agostino B (2019) 'New guidelines for statistical reporting in the Journal' *New England Journal of Medicine*, **381**, 285–286.
- Haybittle JL (1971) 'Repeated assessment of results in clinical trials of cancer treatment' *British Journal of Radiology*, **44**, 793–797.
- Helsinki Declaration (2004) *Ethical Principles for Medical Research Involving Human Subjects* WMA General Assembly, Tokyo.

- Higgins JPT, Altman DG, Gotzsche PC, et al. (2011) 'The Cochrane Collaboration's tool for assessing risk of bias in randomised trials' *British Medical Journal*, **343**, d5928.
- Higgins JPT, Thomas J, Chandler J et al. (2019) 'Cochrane Handbook for Systematic Reviews of Interventions' Wiley Blackwell, 2nd Edition' Oxford: The Cochrane Collaboration and John Wiley & Sons, Ltd.
- Higgins JPT, Thompson SG, Deeks JJ and Altman DG (2003) 'Measuring inconsistency in metaanalyses' *British Medical Journal*, **327**, 557–560.
- Hochberg Y and Tamhane AC (1987) *Multiple Comparison Procedures* New York: John Wiley & Sons, Inc.
- Holm S (1979) 'A simple sequentially rejective multiple test procedure' *Scandinavian Journal of Statistics*, **6**, 65–70.
- Horman MS, Emmett L et al (2021) '[<sup>177</sup>Lu]Lu-PSMA-617 versus cabazitaxel in patients with metastatic castration-resistant prostate cancer (TheraP): a randomised, open-label, phase 2 trial' *The Lancet*, **397**, 797–804.
- Hussain M, Fizazi K, Saad F, et al. (2018) 'Enzalutamide in Men with Nonmetastatic, Castration-Resistant Prostate Cancer' *New England Journal of Medicine*, **378**, 2465–2474.
- The ICON and AGO Collaborators (2003) 'Paclitaxel plus platinum-based chemotherapy versus conventional platinum-based chemotherapy in women with relapsed ovarian cancer: the ICON4/AGO-OVAR-2.2 trial' *The Lancet*, **361**, 2099–2106.
- Jabbar-Lopez ZK, You ZZN, Ward V et al. (2017) 'Quantitative evaluation of biologic therapy options for psoriasis: A systematic review and network meta-analysis' *Journal of Investigative Dermatology*, **137**, 1646–1654.
- Jansen JP, Trikalinos T, Cappelleni JC et al. (2014) 'Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: An ISPOR-AMCP-NPC Good Practice Task Force Report' *Value in Health*, **17**, 157–173.
- Jenkins M, Stone A and Jennison C (2011) 'An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints' *Pharmaceutical Statistics*, **10**, 347–356.
- Jennison C and Turnbull BW (2000) *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, Boca Raton.
- Jensen MP, Karoly P, O'Riordan EF, et al. (1989) 'The subjective experience of pain. An assessment of the utility of 10 indices' *The Clinical Journal of Pain*, **5**, 153–159.
- Jones B, Jarvis P, Lewis JA and Ebbutt AF (1996) 'Trials to assess equivalence: the importance of rigorous methods' *British Medical Journal*, **313**, 36–39.
- Jonsson L, Sandin R, Ekman et al (2014) 'Analyzing overall survival in randomised controlled trials with crossover and implications for economic evaluation' *Value in Health*, **17**, 707–713.
- Jovanovic BD and Levy PS (1997) 'A look at the rule of three' *The American Statistician*, **51**, 137–139.
- Julious SA (2004) 'Using confidence intervals around individual means to assess statistical significance between two means' *Pharmaceutical Statistics*, **3**, 217–222.
- Julious SA and Mullee MA (1994) 'Confounding and Simpson's paradox' *British Medical Journal*, **309**, 1480–1481.
- Kaplan EL and Meier P (1958) 'Non-parametric estimation from incomplete observations' *Journal of the American Statistical Association*, **53**, 457–481.
- Kaul S and Diamond GA (2006) 'Good enough: a primer on the analysis and interpretation of non-inferiority trials' *Annals of Internal Medicine*, **145**, 62–69.
- Kay R (1995) 'Some fundamental statistical concepts in clinical trials and their application in herpes zoster' *Antiviral Chemistry and Chemotherapy*, **6**, 28–33.
- Kay R (2004) 'An explanation of the hazard ratio' *Pharmaceutical Statistics*, **3**, 295–297.

- Kay R (2006) Letter to the Editor on 'Phase specific analysis of herpes zoster associated pain data: a new statistical approach' *Statistics in Medicine*, **25**, 359–360.
- Kenward MG (2015) 'Controlled multiple imputation methods for sensitivity analyses in longitudinal clinical trials with dropout and protocol deviation' *Clinical Investigation*, **5**, 311–320.
- Lachin JM (2016) 'Fallacies of last observation carried forward' *Clinical Trials*, **13**, 161–168.
- Lan KKG and DeMets DL (1983) 'Discrete sequential boundaries for clinical trials' *Biometrika*, **70**, 659–663.
- Landis JR, Heyman ER and Koch GG (1978) 'Average partial association in three-way contingency tables: a review and discussion of alternative tests' *International Statistical Review*, **46**, 237–254.
- Landis JR and Koch GG (1977) 'The measurement of observer agreement for categorical data' *Biometrics*, **33**, 159–174.
- Latimer NR and Abrams KR (2014) 'Adjusting survival time estimates in the presence of treatment switching' NICE DSU Technical Support Document 16.
- Lee LL, McNeer JF, Stramer CF, et al. (1980) 'Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease' *Circulation*, **61**, 508–515.
- Lehmacher W and Wassmer G (1999) 'Adaptive sample size calculations in group sequential trials' *Biometrics*, **55**, 1286–1290.
- Lewis JA (2002) 'The European regulatory experience' *Statistics in Medicine*, **21**, 2931–2938.
- Lewis JA (2004) 'In defence of the dichotomy' *Pharmaceutical Statistics*, **3**, 77–79.
- Li Z, Chines AA and Meredith MP (2004) 'Statistical validation of surrogate endpoints: is bone density a valid surrogate for fracture?' *Journal of Musculoskeletal Neuron Interaction*, **4**, 64–74.
- Maca J, Bhattacharya S, Dragalin V, et al. (2006) 'Adaptive seamless phase II/III designs – background, operational aspects, and examples' *Drug Information Journal*, **40**, 463–473.
- Machin D, Campbell MJ, Tan SB, Tan SH (2011) *Sample Size Tables for Clinical Studies* (3rd edn), Chichester, John Wiley & Sons.
- Machin D, Cheug YB and Parmar MKB (2006) *Survival Analysis. A Practical Approach* (2nd edn), Chichester, Wiley & Sons.
- Mallinckrodt C, Molenberghs G, Lipkovich I and Ratitch B (2020) *Estimands, Estimators and Sensitivity Analysis in Clinical Trials*, CRC Press, Boca Raton.
- Mantel N and Haenszel W (1959) 'Statistical aspects of the analysis of data from retrospective studies of disease' *Journal of the National Cancer Institute*, **22**, 719–748.
- Marshall RJ and Chisholm EM (1985) 'Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer' *Statistics in Medicine*, **5**, 337–344.
- Martina R, Kay R, van Maaren R and Ridder A (2014) 'The analysis of incontinence episodes and other count data in patients with overactive bladder by Poisson and negative binomial regression' *Pharmaceutical Statistics*, **14**, 151–160.
- Matthews JNS, Altman DG, Campbell MJ and Royston P (1990) 'Analysis of serial measurements in medical research' *British Medical Journal*, **300**, 230–235.
- Mehta CR and Pocock SJ (2011) 'Adaptive increase in sample size when interim results are promising: a practical guide with examples' *Statistics in Medicine*, **30**, 3267–3284.
- Meier P (1978) 'The biggest public health experiment ever: the 1954 field trial of the Salk polio-myelitis vaccine' In: *Statistics: A Guide to the Unknown* (eds. J Tanur, F Mostellar, WH Kruskal, et al.) San Francisco: Holden Day.
- Miller DH, Khan OA, Sheremata WA, et al. (2003) 'A controlled trial of natalizumab for relapsing multiple sclerosis' *New England Journal of Medicine*, **348**, 15–23.
- Moher D, Schulz KF and Altman DG (2001) 'The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials' *Annals of Internal Medicine*, **134**, 657–694.
- Mulrow CD (1994) 'Rationale for systematic reviews' *British Medical Journal*, **309**, 597–599.

- Nakamura H, Arakawa K, Itakura H, *et al.* (2006) 'Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA study): a prospective randomised controlled trial' *The Lancet*, **368**, 1155–1163.
- Noseworthy PA, Yao X, Abraham NS, *et al.* (2016) 'Direct comparison of dabigatran, rivaroxaban, and apixaban for effectiveness and safety in nonvalvular atrial fibrillation' *Chest*, **150**, 1302–1312.
- O'Brien PC and Fleming TR (1979) 'A multiple testing procedure for clinical trials' *Biometrics*, **35**, 549–556.
- O'Brien RG and Castelloe J (2010) 'Sample-size analysis for traditional hypothesis testing: concepts and issues' In: *Pharmaceutical Statistics Using SAS: A Practical Guide* (eds. A Dmitrienko, C Chuang-Stein and R D'Agostino) Cary: SAS Institute Inc.
- Okwera A, Whalen C, Byekwaso F, *et al.* (1994) 'Randomised trial of thiacetazone and rifampicin-containing regimens for pulmonary tuberculosis in HIV-infected Ugandans. The Makerere University-Case Western University Research Collaboration' *The Lancet*, **344**, 1323–1328.
- Packer M, Coats AJS, Fowler MB, *et al.* (2001) 'Effect of carvedilol on survival in severe chronic heart failure' *New England Journal of Medicine*, **344**, 1651–1658.
- Page MJ, McKenzie JE, Bossuyt PM, *et al.* (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.' *British Medical Journal*; **372**:n71. doi: 10.1136/bmj.n71
- Pak K, Uno H *et al.* (2017) 'Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio' *Journal of the American Medical Association Oncology*, **3**, 1692–1696.
- Parmar MK, Ledermann JA *et al.* (2003) 'Paclitaxel plus platinum-based chemotherapy versus conventional platinum-based chemotherapy in women with relapsed ovarian cancer: the ICON4/AGO-OVAR-2.2 trial' *Lancet*, **361**, 2099–2106.
- Peto R and Peto J (1972) 'Asymptotically efficient rank invariant procedures' *Journal of the Royal Statistical Society, A*, **135**, 185–207.
- Peto R, Pike MC, Armitage P, *et al.* (1976) 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I. Introduction and design' *British Journal of Cancer*, **34**, 585–612.
- Piccart-Gebhart MJ, Procter M, Leyland-Jones B, *et al.* (2005) 'Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer' *New England Journal of Medicine*, **353**, 1659–1672.
- Pocock SJ (1977) 'Group sequential methods in the design and analysis of clinical trials' *Biometrika*, **64**, 191–199.
- Pocock SJ (1983) *Clinical Trials: A Practical Approach* New York: John Wiley & Sons, Ltd.
- Pocock SJ (2004) 'A major trial needs three statisticians: why, how and who?' *Statistics in Medicine*, **23**, 1535–1539.
- Pocock SJ, Clayton TC and Altman DG (2002) 'Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls' *The Lancet*, **359**, 1686–1689.
- Posch M, Koenig F, Branson M, *et al.* (2005) 'Testing and estimation in flexible group sequential designs with adaptive treatment selection' *Statistics in Medicine*, **24**, 3697–3714.
- Powderly WG, Saag MS, Cloud GA, *et al.* (1992) 'A controlled trial of fluconazole or amphotericin B to prevent relapse of cryptococcal meningitis in patients with the acquired immunodeficiency syndrome' *New England Journal of Medicine*, **326**, 793–798.
- Prochaska JJ and Hilton JF (2012) 'Risk of cardiovascular serious events associated with varenicline use for tobacco cessation: systematic review and meta-analysis' *British Medical Journal*, **344**, e2856.
- Proschan MA, Lan KKG and Wittes JT (2006) 'Statistical Monitoring of Clinical Trials: A Unified Approach' New York: Springer.

- Richardson PG, Ho VT, Giralt S, et al. (2012) 'Safety and efficacy of defibrotide for the treatment of severe hepatic veno-occlusive disease' *Therapeutic Advances in Hematology*, **3**, 253–265.
- Roes KCB (2004) 'Dynamic allocation as a balancing act' *Pharmaceutical Statistics*, **3**, 187–191.
- Rosenbaum PR and Rubin DB (1983) 'The central role of the propensity score in observational studies for causal effects' *Biometrika*, **70**, 41–55.
- Rosenthal R (1979) 'The "file drawer problem" and tolerance for null result' *Psychology Bulletin*, **48**, 638–641.
- Royston P, Altman DG and Sauerbrei W (2006) 'Dichotomizing continuous predictors in multiple regression: a bad idea' *Statistics in Medicine*, **25**, 127–141.
- Royston P and Parmar MKB (2013) 'Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome' *BMC Medical Research Methodology*, **13**, 152–165.
- Royston P and Parmar MKB (2020) 'A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome' *Trials*, **21**, 315–332.
- Rubin D.B. (1987) 'Multiple Imputation for Nonresponse in Surveys' New York: John Wiley & Sons, Inc.
- Rubin LJ, Badesch DB, Barst RJ, et al. (2002) 'Bosentan therapy for pulmonary arterial hypertension' *New England Journal of Medicine*, **346**, 869–903.
- Rudick RA, Stuart WH, Calabresi PA, et al. (2006) 'Natalizumab plus interferon beta-1a for relapsing multiple sclerosis' *New England Journal of Medicine*, **354**, 911–923.
- Sankoh AJ (1995) 'Interim analyses: an FDA reviewer's experience and perspective' *Drug Information Journal*, **29**, 729–737.
- Sargent DJ, Goldberg RM, Jacobson SD, et al. (2001) 'A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients' *New England Journal of Medicine*, **345**, 1091–1097.
- Schlumberger M, Tahara M, Wirth LJ, et al. (2015) 'Lenvatinib versus placebo in radioiodine-refractory thyroid cancer' *New England Journal of Medicine*, **372**, 621–630.
- Schuirmann DJ (1987) 'A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability' *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.
- Schulz KF, Altman DG, Moher D and CONSORT Group (2010) 'CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials' *British Medical Journal*, **340**, 698–702.
- Senn S (2002) *Cross-Over Trials in Clinical Research* (2nd edn) Chichester: John Wiley & Sons, Ltd.
- Senn S (2003) 'Disappointing dichotomies' *Pharmaceutical Statistics*, **2**, 239–240.
- Senn S (2007) *Statistical Issues in Drug Development* (2nd edn) Chichester: John Wiley & Sons, Ltd.
- Sherman DG, Albers GW, Bladin C, et al. (2007) 'The efficacy and safety of enoxaparin versus unfractionated heparin for the prevention of venous thromboembolism after acute ischemic stroke (PREVAIL study): an open-label randomised comparison' *The Lancet*, **369**, 1347–1355.
- Smith JG and Christophers AJ (1992) 'Phenoxy herbicides and chlorophenols: a case control study on soft tissue sarcoma and malignant lymphoma' *British Journal of Cancer*, **65**, 442–448.
- Solomon S, McMurray JJV, Pfeffer MA, et al. (2005) 'Cardiovascular risk associated with celecoxib in a clinical trial for colorectal adenoma prevention' *New England Journal of Medicine*, **352**, 1071–1080.
- Spiegelhalter DJ, Abrams KR and Myles JP (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* Chichester: John Wiley & Sons, Ltd.

- Sterne JAC, White IR, Carlin JB, *et al.* (2009) 'Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls' *British Medical Journal*, **339**, 157–160.
- Stokes ME, Davis CS and Koch GG (2000) *Categorical Data Analysis Using the SAS System* (2nd edn) Cary: SAS Institute Inc.
- Storosum JG, van Zwieten BJ, Vermeulen HDB, *et al.* (2001) 'Relapse and recurrence in major depression: a critical review of placebo-controlled efficacy studies with special emphasis on methodological issues' *European Psychiatry*, **16**, 327–335.
- Stutchfield P, Whitaker R and Russell I (2005) 'Antenatal betamethasone and incidence of neonatal respiratory distress after elective caesarean section: pragmatic randomised trial' *British Medical Journal*, **331**, 662–667.
- Szegedi A, Kohnen R, Dienel A and Kieser M (2005) 'Acute treatment of moderate to severe depression with hypericum extract WS 5570 (St John's wort): randomised controlled double blind non-inferiority trial versus paroxetine' *British Medical Journal*, **330**, 503–508.
- Tai BC, Wee J and Machin D (2011) 'Analysis and design of randomised clinical trials involving competing risks endpoints' *Trials*, **12**, 127–137.
- Tang BMP, Eslick GD, Nowson C, *et al.* (2007) 'Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis' *The Lancet*, **370**, 657–666.
- van Belle G, Fisher LD, Heagerty PJ and Lumley T (2004) *Biostatistics: A Methodology for the Health Sciences* (2nd edn) Hoboken: John Wiley & Sons, Inc.
- van Elteren PH (1960) 'On the combination of independent two-sample tests of Wilcoxon' *Bulletin of the International Statistical Institute*, **37**, 351–361.
- van Leth F, Phanuphak P, Ruxrungtham K, *et al.* (2004) 'Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN study' *The Lancet*, **363**, 1253–1263.
- Viswanathan M, Ansari MT, Berkman ND, *et al.* (2013) 'Assessing the risk of bias in individual studies in systematic reviews of health care interventions' In: *Methods Guide of Effectiveness Reviews*, AHRQ Publication No 10(13)-EHC063-EF. Rockville: Agency for Healthcare Research and Quality.
- von Elm E, Altman DG, Egger M, *et al.* (2007) 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies' *British Medical Journal*, **335**, 806–808.
- Wallentin L, Becker RC, Budaj A, *et al.* (2009) 'Ticagrelor versus clopidogrel in patients with acute coronary syndromes' *New England Journal of Medicine*, **361**, 1045–1057.
- Wang J, Zhao Z, Barber B, *et al.* (2011) 'A Q-TWiST analysis comparing panitumumab plus best supportive care (BSC) with BSC alone in patients with wild-type KRAS metastatic colorectal cancer' *British Journal of Cancer*, **104**, 1848–1853.
- Wang S-J, Hung HMJ and O'Neill R (2011) 'Adaptive design clinical trials and trial logistics models in CNS drug development' *European Neuropsychopharmacology*, **21**, 159–166.
- Westlake WJ (1981) 'Bioequivalence testing – a need to rethink (Reader Reaction Response)' *Biometrics*, **37**, 589–594.
- Wijeyesundara DN, Austin PC, Hux JE, *et al.* (2008) 'Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials' *Journal of Clinical Epidemiology*, **62**, 13–21.
- Woo SJ, Veith M, Hamouz J *et al.* (2020) 'Efficacy and safety of a proposed ranibizumab biosimilar product vs a reference ranibizumab product for patients with neovascular age-related macular degeneration: a randomized clinical trial' *JAMA Ophthalmology*, **139**, 68–76.
- The Xamoterol in Severe Heart Failure Study Group (1990) 'Xamoterol in severe heart failure' *The Lancet*, **336**, 1–6.

- The Young Infants Clinical Signs Study Group (2008) 'Clinical signs that predict severe illness in children under age 2 months: a multicenter study' *The Lancet*, **371**, 135–142.
- Zhang Y, Hedo R, Rivera A (2019) 'Post hoc power analysis: is it an informative and meaningful analysis?' *General Psychiatry*, **32**, e100069.
- Zhou XH, Obuchowski NA and McClish DK (2002) *Statistical Methods in Diagnostic Medicine* Hoboken: John Wiley & Sons, Inc.

## Regulatory Guidelines

### ICH Guidelines

- ICH E1 (1995) 'Population Exposure: The Extent of Population Exposure to Assess Clinical Safety'
- ICH E2E (2005) 'Note for Guidance on Planning Pharmacovigilance Activities'
- ICH E3 (1995) 'Note for Guidance on Structure and Content of Clinical Study Reports'
- ICH E9 (1998) 'Note for Guidance on Statistical Principles for Clinical Trials'
- ICH E9 (R1) (2019) 'Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials'
- ICH E10 (2001) 'Note for Guidance on Choice of Control Group in Clinical Trials'
- ICH E14 (2005) 'Note for Guidance on the Clinical Evaluation of QT/QTC Interval and Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs'

### FDA Guidelines

- FDA (1992) 'Points to Consider on Clinical Development and Labeling of Anti-Infective Drug Products'
- FDA (1998) 'Developing Antimicrobial Drugs – General Considerations for Clinical Trials'
- FDA (2001) 'Statistical Approaches to Establishing Bioequivalence'
- FDA (2004) 'Critical Path Initiative'
- FDA (2018) 'Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics'
- FDA (2006) 'Critical Path Opportunities List'
- FDA (2006) 'Establishment and Operation of Clinical Trial Data Monitoring Committees'
- FDA (2010) 'Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials'
- FDA (2007) 'Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation'
- FDA (2010) 'Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials'
- FDA (2012) 'Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products. Guidance for Industry'
- FDA (2014) 'Reference Product Exclusivity for Biological Products Filed Under Section 351(a) of the PHS Act'
- FDA (2016) 'Non-Inferiority Clinical Trials to Establish Effectiveness'
- FDA (2016) 'Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product: Draft Guidance for Industry and Food and Drug Administration Staff'
- FDA (2017) 'Multiple Endpoints in Clinical Trials. (Draft) Guidance for Industry'
- FDA (2018) 'Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making. Draft PDUFA VI Implementation Plan (FY 2018-2022)'
- FDA (2018) 'Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics'

- FDA (2019) 'Adaptive Design Clinical Trials for Drugs and Biologics. Guidance for Industry'
- FDA (2020) 'Type 2 Diabetes Mellitus: Evaluating the Safety of New Drugs for Improving Glycemic Control'

## **European (CHMP/EMA) Guidelines**

- CPMP (1999) 'Note for guidance on clinical evaluation of new vaccines'
- CPMP (2000) 'Points to consider on switching between superiority and non-inferiority'
- CPMP (2001) 'Note for guidance on clinical investigation of medicinal products for the treatment of acute stroke'
- CPMP (2001) 'Note for guidance on the investigation of bioavailability and bioequivalence'
- CPMP (2001) 'Points to consider on applications with 1. meta analyses; 2. one pivotal study'
- CPMP (2001) 'Points to consider on missing data'
- CPMP (2003) 'Note for guidance on evaluation of medicinal products indicated for treatments of bacterial infections'
- CPMP (2003) 'Points to consider on adjustment for baseline covariates'
- CPMP (2003) 'Points to consider on the clinical development of fibrinolytic medicinal products in the treatment of patients with ST segment elevation acute myocardial infarction (STEMI)'
- CHMP (2005) 'Guidance on the choice of non-inferiority margin'
- CHMP (2005) 'Guideline on data monitoring committees'
- CHMP (2005) 'Note for guidance on clinical evaluation of new vaccines'
- CHMP (2006) 'Guideline on clinical trials in small populations'
- CHMP (2007) 'Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan'
- CHMP (2008) 'Guideline on medicinal products for the treatment of Alzheimer's disease and other dementias'
- CHMP (2008) 'Reflection paper on benefit-risk assessment methods in the context of the evaluation of marketing authorisation application of medicinal products for human use'
- CHMP (2009) 'Guideline on clinical evaluation of diagnostic agents'
- CHMP (2010) 'Points to consider on missing data in confirmatory clinical trials'
- CHMP (2012) 'Guideline on the evaluation of medicinal products indicated for treatment of bacterial infections'
- CHMP (2012) 'Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (COPD)'
- CHMP (2013) 'Draft qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of phase II dose finding studies under model uncertainty'
- CHMP (2014) 'Guideline on similar biological medicinal products'
- CHMP (2015) 'Guideline on adjustment for baseline covariates in clinical trials' (updated from the 2003 version)
- CHMP (2016) 'Draft guideline on medicinal products in the treatment of Alzheimer's disease and other dementias'
- CHMP (2017) 'Guideline on multiplicity issues in clinical trials' (Draft)
- CHMP (2018) 'Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus' (Draft)
- CHMP (2019) 'Guideline on the investigation of subgroups in confirmatory clinical trials'
- EMA (2008) 'Guideline on the use of statistical signal detection methods in the EudraVigilance data analysis system'
- EMA (2010) 'Benefit-Risk Methodology Project: Work Package 2'

- EMA (2011a) 'Benefit-Risk Methodology Project: Work Package 1'
- EMA (2011b) 'Benefit-Risk Methodology Project: Work Package 3'
- EMA (2012) 'Benefit-Risk Methodology Project: Work Package 4'
- EMA (2018) 'Question and answer on adjustment for cross-over in estimating effects in oncology trials'

# Index

Please note: entries in *italics* refer to figures, entries in **bold** refer to tables.

- 0.05 problem, *p*-values, 161–163  
2NN study, 272–273  
10 Mark banknote, 32  
10 per cent rule, cure rates, 208–209  
>80 per cent power, 154  
95% confidence interval, 40–42
- absolute risk reduction (ARR), 68, 71  
accelerated failure time model, 232–233  
across trial summaries, safety evaluation, 339–340  
acute coronary syndromes, 174–175  
adaptive designs, 266–282  
    advantages and drawbacks, 266–267  
    bias minimisation, 268–273  
    estimation, 270–271  
    methodological issues, 271–272  
    operational, 272–273  
    type I error control, 268–271  
enrichment, 279–280  
flexible adaptations, 268  
inferentially seamless, 276  
multiplicity, 276–277  
operational bias, 272–273  
operationally seamless, 276  
*p*-values, 273  
placebo removal in non-inferiority  
    study, 280–281  
power, 279–280  
primary endpoints changes, 278–279  
regulatory issues, 267–268, 275–276, 278–279, 281–282  
restricted adaptations, 267–268  
seamless phase II/III studies, 275–278  
    framework, 275–276, 275, 276  
    logistical challenges, 278  
    multiplicity, 276–277  
    phase II data incorporation, 277–278  
sub-population, 278
- unblinded sample size re-estimation, 273–275  
weighting, 274  
adjusted analysis, 80–91  
    adjusted treatment difference, 81–82  
ANCOVA, 102–103  
baseline covariates, 84, 86–87, 90–91  
baseline imbalance, 83  
binary, categorical and ordinal endpoints, 88  
centre adjustment, 89  
centre combination, 90–91  
continuous endpoints, 81–84  
homogeneity, 85–88  
least squares means, 84–85  
main effects model, 100, 102  
multi-centre trials, 89–91  
null hypothesis, 82–83  
objectives, 80  
quantitative and qualitative interactions, 87–88  
standard deviation, 83  
treatment-by-centre interactions, 89–90  
treatment-by-factor interactions, 85–87  
two-way analysis of variance, 82
- adjusted significance level, 166  
    avoidance, 171–176  
Bonferroni correction, 166  
Hochberg correction, 169–170  
Holm correction, 168–169  
interim analysis, 170–171  
adjustment, sample sizes, 154–155  
administrative censoring, 233  
administrative interim analysis, 249  
adverse events (AE), 329–330, 332–335  
AE *see* adverse events  
age-related macular degeneration, 215–216  
alpha-spending functions, 244  
alternative hypothesis, 48, 53  
alternative inhaler propellant, 197  
amphotericin B, 200  
analysis of covariance (ANCOVA)

- accelerated failure time model, 233  
 and ANOVA, 110  
 baseline covariates, 109, 110–113  
 baseline imbalances, 111  
 baseline testing, 109–110  
 binary endpoints, 97–98, 105–106  
 continuous outcomes, 98–105
  - and adjusted analyses, 102–103
  - advantages, 103–104
  - disadvantages, 104
  - least squares means, 104–105, *105*
  - main effects models, 100, 102
  - random element, 105
  - single model, 102
  - treatment-by-covariate interactions, 100–102
 count endpoints, 108–109  
 regression towards the mean, 112  
 regulatory aspects, 110–113  
 selection bias adjustment, 293  
 analysis sets
  - equivalence and non-inferiority trials, 204–205
  - see also* full analysis set; per-protocol set
 analysis of variance (ANOVA)
  - accelerated failure time model, 232–233
  - one-way, 77–78
  - two-way, 82–83
 ANCOVA *see* analysis of covariance  
 ANOVA *see* analysis of variance  
 area under the curve (AUC), ROC, 361–362  
 arithmetic mean, 188  
 ARR *see* absolute risk reduction  
 assay sensitivity, 202–204  
 asthma, 44  
 attributes
  - case-control studies, 303
  - estimands, 134–136
 attrition bias, 285, **318**  
 AUC *see* area under the curve  
 avoidance
  - adjusted significance levels, 171–176
  - missing data, 125–126
 balance statistics, 302  
 baseline covariates, 84, 86–87, 90–91, 109, 110–113  
 baseline factors, 92  
 baseline imbalances, 82, 111, 291–292, 298–302  
 baseline observation carried forward (BOCF), 123  
 baseline testing, 109–110  
 Bayes theorem, 258–259  
 Bayesian neural networks, 353–354  
 Bayesian rank analysis, 326–327  
 Bayesian statistics, 255–265
  - Bayes theorem, 257–259
  - case study, 262–263
  - credible intervals, 261
  - empirical prior distributions, 264
  - frequentist methods, 259
 inference, 259–261  
 posterior beliefs, 256–257  
 posterior distributions, 257  
 posterior probabilities, 260–261  
 prior beliefs, 255–256  
 prior distribution, 255–256  
 prior probabilities, 255–256  
 regulatory aspects, 263–265  
 sceptical prior, 256  
 vague priors, 256  
 benefit-risk assessment, 341–350
  - balance, 276, 341–343, 346–348
  - favourable effects, 343, 346–348, 350
  - MCDA, 342–348
  - Q-TWiST, 348–350
  - unfavourable effects, 342, 343, 345–348
  - utility, 344–345, 348
 best-case classification, 124  
 beta distribution, 190  
 betamethasone, 107  
 between-patient designs, 12–13  
 between-subject variation, 109–110  
 bias
  - adaptive designs, 268–273
  - attrition, 285, **318**
  - detection, 286, **318**
  - funnel plots, 318–319, 319
  - meta-analysis, 317–319
  - minimisation, 269–273
  - observational studies, 285–289, 291–300
  - operational, 249, 272–273
  - performance, 286, **318**
  - and precision, 10–12
  - publication, 318–319, **318**
  - risk of, 317, **318**
  - selection, 285–289, 291–302, **318**
  - sources, 285–286
 binary endpoints, 18, 78–79
  - logistic regression, 97–98
  - sample sizes, 153
 biocreep, 210–211  
 bioequivalence, 14, 17, 201–202, 206  
 biosimilars, 206, 215–216  
 blind review, 374  
 blinding/masking, 3, 11, 110, 250, 285  
 block randomisation, 4–5  
 BOCF *see* baseline observation carried forward  
 bone mineral density, 22, 315–316  
 Bonferroni correction, 166, 167–168  
 bootstrapping, 47  
 bosentan therapy, 123  
 box plots, 29–30  
 Breslow–Day test, 88  
 calcium supplementation, 315–316  
 caliper, propensity scoring, 296  
 capecitabine plus lapatinib, 247–248

- cardiovascular disease  
 carvediol, 219, 231–232  
 prevastatin, 179–180
- carry-over effect, 13–14
- carvedilol, 219, 231–232
- case report forms (CRFs), 341, 371, 376
- case–control studies, 284, 302–305
- categorical endpoints, 18, 73–75, 78–79
- categorisation, 23–24
- cause-specific hazard function, 238
- censoring, 217, 228  
 administrative, 233  
 independent, 233–234  
 non-informative, 233–234
- central limit theorem, 30–31
- central randomisation, 8
- centre adjustment, 89
- chi-square distribution on one df, 65
- chi-square test  
 binary data, 63–66, 78–79  
 categorical data, 73–75, 78–79  
 Mantel–Haenszel chi-square test, 75–76, 79  
 multiple treatment groups, 78–79  
 ordered categorical data, 75–76, 78–79  
 Pearson, 63–66  
 sample sizes, 150  
 standard error, 66
- choice, endpoints, 20–24
- cholesterol lowering study, 33, 146–148
- chronic HIV-1 infection, 272–273
- CI *see* confidence intervals
- CIFs *see* cumulative incidence functions
- classical methods, 255
- classification, missing data, 126–127
- clinical importance  
 confidence intervals, 159–160  
 internal consistency, 163  
*p*-values, 157–159, 160–163
- Clinical Study Reports (CSR), 155–156
- clinical trial design  
 assay sensitivity, 202–204  
 between-and within-patient designs, 12–13  
 bias and precision, 10–12  
 confirmatory and exploratory trials, 16–17  
 control groups, 2  
 crossover trials, 13–14  
 endpoints, 18–24  
 noise, 15, 15  
 placebos and blinding, 3, 11  
 randomisation, 3–10  
 signal, 14–15  
 signal-to-noise ratio, 16  
 superiority, equivalence and  
 non-inferiority trials, 17–18
- clinically relevant difference (crd), 149, 153, 212, 380–381
- clofibrate, 117
- clopidogrel, 174–175
- cluster randomisation, 10
- CMH *see* Cochran–Mantel–Haenszel tests
- co-primary endpoints, 171–172
- Cochrane Collaboration, 307–308
- Cochran–Mantel–Haenszel (CMH) tests, 88
- Cochran’s Q statistic, 313–314
- coherence, 326
- combination, meta-analysis, 309–310
- combining centres, 90–91
- companion diagnostics, 367–369
- competing risks, 238–241
- complete cases analysis/completer analysis, 122
- components, estimands, 134–136
- composite endpoints, 23, 172–173, 237–241  
 cause-specific hazard function, 238  
 competing risks, 238–241  
 cumulative incidence functions, 237–238  
 Gray’s test, 238  
 regulatory aspects, 238–241
- composite strategy, 137
- composite time-to-event endpoints, 237–241
- composite variables (endpoints), 23, 237–241
- concealment, 3
- concordance, 365
- concurrent cohort studies, 284, 287–288
- conditional power, 246–247
- confidence coefficients, 42, 158
- confidence intervals (CI), 40–47  
 95 per cent, 40–42  
 bootstrap, 47  
 clinical importance, 158–160  
 difference between two means, 44–45  
 equivalence, 198, 198  
 general case, 46  
 interim analysis, 249  
 meta-analysis, 310, 323–324  
 multiplying constants, 42–43, 43  
 network meta-analysis, 323–324  
 non-inferiority, 199–200  
 odds ratio, 69–70  
 one-sided, 199–200  
*p*-values, 157–159  
 for proportions, 45–46  
 single mean, 40–44  
 standard error, 43–44
- confirmatory trials, 16–17
- Consolidated Standards of Reporting Trials (CONSORT), 155–156, 379–380
- CONSORT *see* Consolidated Standards of Reporting Trials
- constancy, 211
- constant hazard ratio, 225–226
- continuous endpoints, 18, 77–78

- adjusted analysis, ANOVA, 81–84  
 ANCOVA, 98–105  
 sample sizes, 153  
 control groups, 2, 152–153, 371  
 copy reference, 128  
 correlation, 113–114  
 correlation coefficients, 113–114  
   Pearson, 114  
   Spearman, 114  
 count endpoints, 19, 77  
   negative binomial regression, 97,  
     108–109  
   offset term, 98, 108–109  
 counterfactual, 140  
 covariates  
   baseline, 109, 110–113  
   baseline imbalances, 111  
   pre-planning, 110  
   regression towards the mean, 112  
   regulatory aspects, 110–113  
   time-dependent, 111  
 COVID-19 infections, 137  
 crd *see* clinically relevant difference  
 credible intervals, 261  
 CRFs *see* case report forms  
 critical value, 53, 66  
 cross-sectional studies, 284–285  
 crossover, oncology, 234–237  
 crossover trials, 13–14, 132  
 cryptococcal meningitis, 200  
 CSR *see* Clinical Study Reports  
 cumulative incidence functions (CIFs),  
   237–238  
 cure rates  
   10 per cent rule, 208–209  
   non-inferiority, 212  
 cut-offs, diagnosis, 360
- Data Monitoring Boards (DMBs), 250–251  
 Data Monitoring Committees (DMCs)  
   case report forms, 341  
   charter, 252–253  
   data tables, 252  
   establishment, 251  
   guidelines, 251–252  
   inadvertent unblinding, 249  
   long-term trials, 251  
   meetings and recommendations, 253–254  
   participants, 251–252  
   risk/benefit assessment, 250  
   roles and responsibilities, 250  
   safety monitoring, 250  
   sponsors, 251–252  
   statisticians, 252  
   structure and process, 252–253  
   tables and graphs, 340–341
- data validation plans, 373–374  
 degrees of freedom, 42–43  
 delta method, 128  
 depression, 270–271  
 detection bias, 286, 318  
 diabetes, 137–139  
 diagnosis, 356–370  
   companion diagnostics, 367–369  
   concordance, 365  
   cutoffs, 360  
   false positive/negative rates, 358  
   inter/intra-rater agreement, 367  
   kappa statistic, 365–367  
   likelihood ratio, 359  
   positive and negative predictive value, 358  
   predictive accuracy, 359–360  
   prevalence, 358–359  
   receiver operating characteristic curves,  
     361–363  
   regression models, 362–364  
   sensitivity and specificity, 357  
   surrogate standard of truth, 357  
   trial design, 364–365  
   weighted kappa, 367  
 difference between two means  
   confidence intervals, 44–45  
   standard error, 38–39  
 dirty data, 374  
 DMBs *see* Data Monitoring Boards  
 DMCs *see* Data Monitoring Committees  
 dose-ranging studies, 79  
 dropouts  
   regulatory aspects, 154  
   replacement, 132–133  
 drug-induced liver injury, 331  
 dry runs, 374  
 dynamic allocation and minimisation, 9–10
- ECG data, routine safety evaluation, 338–339  
 ECS trial *see* European carotid surgery trial  
 effect modifiers, 324–325  
 efficacy, interim analysis, 245–246  
 endpoints, 18–24  
   binary, 18, 63–66, 78–79, 97–98, 153  
   categorical, 18, 73–75, 78–79  
   categorisation, 23–24  
   choice, 20–24  
   co-primary, 171–172  
   composite, 23, 172–173, 237–241  
   continuous, 18, 77–78, 98–105, 153  
   count, 19, 77, 97–98  
   global assessment, 22  
   ordered categorical, 18, 75–76, 78–79  
   ordinal, 18  
   primary, 20–21  
   scores, 19

- endpoints (*cont'd*)  
 secondary, 21  
 surrogate, 21–22, 22  
 time-to-event, 19, 129–131, 237–241  
 types, 18–20
- enrichment, 279–280
- enzalutamide, 176–177
- equivalence margins, 198
- equivalence trials, 17–18, 196–216  
 analysis sets, 204–205  
 assay sensitivity, 202–204  
 biocreep, 210–211  
 bioequivalence, 201–202, 206  
 biosimilars, 206, 215–216  
 choice of  $\Delta$ , 205–210  
   10 per cent rule, 208–209  
   bioequivalence, 206  
   synthesis method, 209–210  
   therapeutic equivalence, 206
- confidence intervals, 198–200
- constancy, 211
- Ebbutt and Frith case study, 197
- p*-value approach, 201–202
- sample size calculations, 210–212
- similarity, 196–197  
 statistical principles, 198–199  
 synthesis method, 209–210
- establishment, Data Monitoring Committees, 251
- estimands, 134–143  
 attributes, 134–136  
 composite strategy, 137  
 counterfactual, 140  
 hypothetical strategy, 137  
 intercurrent event, 136  
 main estimators, 141  
 population, 135  
 principle stratum policy, 139–141  
 sensitivity, 142–143  
 strategies, 136–141  
 summary statistics, 136  
 supplementary analysis, 143  
 treatment policy strategy, 137  
 variables, 135  
 while on treatment, 139
- estimation, adaptive design, 271–272
- European carotid surgery (ECS) trial,  
 286–287, 296–298
- event rates  
 control groups, 152–153  
 relative risk, 220–221
- event times, median, 221–222
- expected frequencies, 64
- exploratory trials, 16–17
- F-test, 184–185
- failure classification, missing data, 123
- fallback procedure, 176–178
- false negative rate (FNR), 358, 360  
 false negatives, 145  
 false positive rate (FPR), 358, 360  
 false positives, 164  
 family-wise error rate (FWER), 166  
 FAS *see* full analysis set  
 favourable effects, 342  
 Fisher's combination test, 269–270  
 Fisher's exact test, 72–73  
 fixed effects model, 310, 313  
 fixed margin approach, 207–208  
 flexible adaptations, 268  
 fluconazole, 200  
 FNR *see* false negative rate  
 forest plots, 310–312, 311  
 FPR *see* false positive rate  
 fracture risk, 22, 315–316  
 frequentist methods, 255  
   Bayesian, 259–260  
 fudge factor, 189, 314  
 full analysis set (FAS), 119–120, 142, 182  
 funnel plots, 318–319, 319  
 futility, 246–247  
 FWER *see* family-wise error rate
- gamma distribution, 190
- Gaussian distribution, 30–33
- Gehan–Wilcoxon test, 223–224, 230–231, 379
- geometric mean, 188
- German 10 Mark banknote, 32, 32
- global assessment endpoints, 22
- graphical methods  
 adjusted analysis, 85, 87, 88  
 meta-analysis, 310–311  
 non-parametric, 185–186
- Gray's test, 238
- guidelines, Data Monitoring Committees, 251–252
- HAMA *see* Hamilton Anxiety Scale  
 HAMD *see* Hamilton Depression Scale  
 Hamilton Anxiety (HAMA) Scale, 19  
 Hamilton Depression (HAMD) Scale, 19, 62, 63, 130
- Haybittle & Peto method, 243
- hazard rate, 225
- hazard ratio, 225–229  
 adjusted for covariates, 231–232  
 constant, 225–226  
 hazard rate, 224–225, 224  
 KM curve calculation, 228–229  
 meta-analysis, 312, 323–324  
 network meta-analysis, 312, 323–324  
 non-constant, 226–227  
 proportional hazards, 227  
 survival curves link, 227–228
- HER2-positive advanced breast cancer,  
 45–46, 247–248

- heterogeneity, 312–313  
 hierarchical testing, 173–175  
 hip replacement, 207  
 histograms, 26–27  
 historical cohort studies, 284, 288–289  
 HIV-1 infection, 2NN study, 272–273  
 Hochberg correction, 169–170  
 Holm correction, 168–169  
 homogeneity of variance, 184–185  
 hypericum extracts, 270–271  
 hypothesis testing, 47–57
  - alternative hypothesis, 48, 53
  - complex situations, 52–55
  - null distribution, 54
  - null hypothesis, 48, 53
  - one-sided and two-sided tests, 56–57
 p-values
  - calculation, 49–56
  - interpretation, 48–49
  - randomness, 53–55
  - statistical significance, 56
 t-distribution, 54  
 test statistic, 53  
 hypothetical strategy, 137  
 Hy's law, 331, 340
- $I^2$  statistic, 312–313  
 IC *see* information component  
 ICE *see* intercurrent event  
 imputation models, 123–124, 127  
 inadvertent unblinding, 249  
 incoherence, 326  
 independent censoring, 233–234  
 indirect treatment comparisons, 323–326  
 individual patient data, meta-analysis, 314  
 inferential statistics *see* sampling and inferential statistics  
 inferentially seamless adaptive designs, 276  
 information component (IC), 354  
 integrated summary of efficacy (ISE), 378–379  
 integrated summary of safety (ISS), 378–379  
 intention-to-treat (ITT), 115–133
  - crossover trials, 132
  - dropout replacement, 132–133
  - full analysis set, 119–120
  - imputation models, 123–124, 127
  - jump to reference, 128
  - missing data, 121–128
    - avoidance, 125–126
    - baseline observation carried forward, 123
    - classification, 126–127
    - complete cases analysis, 122
    - last observation carried forward, 123
    - multiple imputation, 127–128
- sensitivity, 124–125  
 success/failure classification, 123  
 worst-case/best-case classification, 124  
 modified ITT populations, 119–120  
 per-protocol set, 120–121  
 pre-specification, 129  
 principle, 115–121  
 randomisation, 118  
 safety set, 131  
 sensitivity, 120–121  
 superiority trials, 120  
 time-to-event data, 129–131  
 inter/intra-rater agreement, 367  
 Interactive Voice Response System (IVRS), 8  
 Interactive Web Response System (IWRS), 8  
 intercurrent event (ICE), 136  
 interferon beta-1a, 381  
 interim analysis, 170–171, 243
  - administrative, 249
  - alpha-spending functions, 244
  - completion of recruitment, 248
  - conditional power, 246
  - confidence intervals, 249
  - efficacy, 245–246
  - futility, 246–247
  - Haybittle & Peto method, 243
  - inadvertent unblinding, 249
  - Lan–DeMets alpha spending function, 244
  - O'Brien and Fleming method, 243
  - operational bias, 249
  - overrun, 248
  - Pocock method, 243
  - point estimates, 249
  - safety monitoring, 249–250
  - stopping rules, 243–245
  - time-to-event endpoints, 244–245
  - type I error, 248
- internal consistency, 163  
 international stroke (IS) trial, 286–287, 296–298  
 inverse propensity score weighting (IPSW), 300–302  
 IPSW *see* inverse propensity score weighting  
 IS trial *see* international stroke trial  
 ISE *see* integrated summary of efficacy  
 ISS *see* integrated summary of safety  
 ITT *see* intention-to-treat  
 IVRS *see* Interactive Voice Response System  
 IWRS *see* Interactive Web Response System
- jump to reference, 128
- Kaplan–Meier (KM) curves
  - calculation, 228–229
  - cumulative incidence functions, 237–238
  - event rates and relative risk, 220
  - hazard rates, 228–229
  - median event times, 221–222

- Kaplan–Meier (KM) curves (*cont'd*)  
 plotting, 218–220  
 rank preserving structural failure time model, 235–236  
 risk set, 220  
 kappa statistic, 365–367  
 kidney stones removal, 306–307  
 KM *see* Kaplan–Meier curves
- laboratory data, routine safety evaluation, 335–338
- Lan–DeMets alpha spending function, 244
- lapatinib plus capecitabine, HER2-positive advanced breast cancer, 247–248
- last observation carried forward (LOCF), 123, 371
- late baselines, 111
- learn vs. confirm paradigm, 267, 275
- least squares, 93
- least squares means, 84–85, 104–105, 105
- lenvatinib, 236, 237
- likelihood ratio (LR), 359
- linear regression  
 least squares, 93  
 Pearson correlation coefficients, 114  
 simple, 92–95
- LOCF *see* last observation carried forward
- log normal distribution, 190
- log transformation, 188–189, 188
- logistic regression, 97–98, 105–107, 294, 297
- logistical challenges, seamless phase II/III studies, 278
- logit, 97, 189
- logrank test, 222–223, 226, 230–231, 232  
 Gray's test, 238  
 weighted, 225
- lower confidence limit, 40
- LR *see* likelihood ratio
- lung cancer, radiotherapy vs. surgery, 115–117
- main effects model, 100, 102
- main estimators, 141
- Mann–Whitney U-test, 190–191, 191
- Mantel–Haenszel (MH) chi-square test, 75–76, 79
- MAR *see* missing at random
- margins  
 equivalence trials, 205–210  
 non-inferiority trials, 207–208
- Markov Chain Monte Carlo (MCMC) method, 262
- matching, propensity scoring, 296–297
- MCAR *see* missing completely at random
- MCDA *see* multi-criteria decision analysis
- MCMC *see* Markov Chain Monte Carlo
- mean, 27  
 arithmetic/geometric, 188
- regression towards the mean, 112
- standard error for difference, 38–39
- standard error, 33–36
- mean absolute deviation, 28
- MedDRA *see* Medical Dictionary for Regulatory Activities
- Activities
- median, 27
- median event times, 221–222
- Medical Dictionary for Regulatory Activities (MedDRA) system, 331, 373
- meta-analysis, 306–327  
 adverse event rates, 307–308  
 bias, 317–319  
 case studies, 314–316, 322–323  
 combination methods, 309–310  
 confidence intervals, 310, 323–324  
 definition, 306–307  
 fixed effects model, 310  
 funnel plots, 318–319, 319  
 graphical methods, 310–311  
 heterogeneity, 312–313  
 individual patient data, 314  
 methodology, 309–314  
 objectives, 307–309  
 planning, 316–317  
 pooling, 306–308  
 PRISMA, 320  
 publication bias, 318–319, 318  
 random effects model, 310  
 rare events, 313–314  
 regulatory aspects, 320–321  
 risk of bias, 317, 318  
 robustness, 313  
 scientific validity, 316–320  
 serious adverse events, 313–314  
 systematic reviews, 306  
*see also* network meta-analysis
- MH *see* Mantel–Haenszel chi-square test
- MI *see* multiple imputation; myocardial infarctions
- minimum effective dose, 178
- mirabegron, 108–109
- misinterpretation, *p*-values, 160–161
- missing at random (MAR), 126–127, 128
- missing completely at random (MCAR), 126–127
- missing data, 121–128  
 avoidance, 125–126  
 baseline observation carried forward, 123  
 classification, 126–127  
 complete cases analysis, 122  
 copy reference, 128  
 dropout replacement, 132–133
- imputation models, 123–124, 127
- last observation carried forward, 123
- missing at random, 126–127, 128
- missing completely at random, 126–127
- missing not at random, 126, 128
- multiple imputation, 121, 127–128
- multiple testing, 182

- sensitivity, 124–125, 378  
success/failure classification, 123  
worst-case/best-case classification, 124  
missing not at random (MNAR), 126, 128  
mITT *see* modified intention to treat population  
Mixed Model for Repeated Measures  
(MMRM), 109–110, 127  
MMRM *see* Mixed Model for Repeated Measures  
MNAR *see* missing not at random  
modified intention to treat population  
(mITT), 119–120  
multi-centre trials  
centre adjusting, 89  
combining centres, 90–91  
treatment-by-centre interactions, 89–90  
multi-criteria decision analysis (MCDA),  
342–348, 343–347  
multiple comparisons, 177–178  
multiple imputation (MI), 127–128  
delta method, 128  
jump to reference, 128  
multiple regression, 95–97, 100  
multiple sclerosis, 381  
multiple testing *see* multiplicity  
multiplicity, 164–183  
alternate statistical tests, 161–162  
analysis sets, 182  
Bonferroni correction, 166, 167–168  
causes, 165–166  
co-primary endpoints, 171–172  
composite endpoints, 237–241  
fallback procedure, 176–178  
false positives, 164  
hierarchical testing, 173–175  
Hochberg correction, 169–170  
Holm correction, 168–169  
interim analysis, 170–171  
missing data, 182  
multiple comparisons, 177–178  
nominal significance, 183  
pre-planning, 182–183  
regulatory aspects, 166–167  
seamless phase II/III trials, 276–277  
subgroup testing, 178–181  
type I error inflation, 164–165  
multiplicity-adjusted *p*-values, 277  
multiplying constants, confidence intervals,  
42–43, 43  
myocardial infarctions (MI), 117, 174–175
- natalizumab, 192, 381  
natalizumab plus interferon beta-1a, 381  
negative binomial model, 77  
count endpoints, 97–98, 108–109  
offset term, 98, 108–109  
negative predictive value (NPV), 359, 360  
neonatal respiratory distress syndrome, 107  
network maps, 322–323, 323
- network meta-analysis (NMA), 321–327  
Bayesian rank analysis, 326–327  
case study, 322–323  
coherence, 326  
critiques, 325–326  
cross-trial calculations, 323–324  
effect modifiers, 324–325  
indirect treatment comparisons, 323–326  
randomisation, 325–326  
rank probabilities, 326–327  
surface under the cumulative ranking  
curve, 326–327  
transitivity, 323–326  
NMA *see* network meta-analysis  
NNH *see* number needed to harm  
NNT *see* number needed to treat  
noise, 15, 15  
nominal significance, 183  
non-constant hazard ratio, 226–227  
non-inferiority trials, 17–18  
analysis set, 204–205  
assay sensitivity, 202–204  
biosimilars, 206, 215–216  
choice of  $\Delta$ , 207–209  
confidence intervals, 199–200  
cure rates, 208–209, 212  
fixed margin approach, 207–208  
one-sided equivalence, 197  
placebo arm removal, 280–281  
primary comparison, 197  
sample sizes, 211–213  
superiority, 213–215  
synthesis method, 209–210  
non-informative censoring, 233–234  
non-normality and transformations  
arithmetic and geometric means, 188  
beta and gamma distribution, 190  
log transformation, 188–189, 188  
*p*-value, 188  
paired t-test setting, 189  
Poisson distribution, 190  
quantile–quantile plots, 189, 189  
square root and logit transformations, 189  
non-parametric and related methods  
advantages and disadvantages, 194  
assumptions, t-tests and extensions, 184  
homogeneity of variance, 184–185  
Mann–Whitney U-test, 190–192  
non-normality and transformations, 187–190  
normal distribution, *p*-value calculations, 193  
normality  
positively skewed distributions, 185, 186  
quantile–quantile plot, 185, 186  
Shapiro–Wilks test, 186  
outliers, 195  
van Elteren test, 194  
Wilcoxon signed rank test, 192–193  
normal distribution, 30–33

- normality  
 positively skewed distributions, 185, 186  
 quantile–quantile plot, 185, 186  
 Shapiro–Wilks test, 186
- notation, sampling/inferential statistics, 28–29
- NPV *see* negative predictive value
- null distribution, 54, 66
- null hypothesis, 48, 53, 56, 82–83
- number needed to harm (NNH), 69
- number needed to treat (NNT), 69, 76
- objective response rates (ORR), 343–345
- objectives, meta-analysis, 307–309
- O’Brien and Fleming method, 243
- OBS *see* overactive bladder syndrome
- observational studies, 283–305  
 baseline balance, 291–292  
 bias, 285–289, 291–302  
 case-control studies, 284, 302–305  
 concurrent cohort studies, 284  
 cross-sectional studies, 284–285  
 historical cohort studies, 284  
 historically controlled, 284  
 inverse propensity score weighting, 300–302  
 non-randomised comparisons, 283  
 propensity scoring, 293–297, 300–302  
 regulatory guidance, 290–291  
 selection bias, 285–289, 291–302  
 STROBE, 291  
 study types, 284–285
- observed frequencies, 64
- odds ratio (OR)  
 case-control studies, 304–305  
 confidence interval, 69–70  
 description, 67  
 meta-analysis, 310–312, 323–324  
 network meta-analysis, 323–324  
 and relative risk, 71
- offset term, 98, 108–109
- oncology, crossover, 234–237
- one-sided confidence intervals, 199–200
- one-sided equivalence, 197
- one-sided *p*-values, 56–57
- one-way analysis of variance (one-way ANOVA), 77–78
- operable lung cancer, surgery vs.  
 radiotherapy, 115–117
- operational bias, 249, 272–273
- operationally seamless adaptive designs, 276
- OR *see* odds ratio
- ordered categorical (ordinal) endpoints, 18, 75–76, 78–79
- ORR *see* objective response rates
- osteoporosis, fracture risk, 22
- outcomes *see* endpoints
- outcomes, case-control studies, 303
- outliers, 195, 378
- over stratification, 8
- overactive bladder syndrome (OBS), 108–109
- overrun, 248
- p*-values, 48–56  
 0.05 problem, 161–163  
 adaptive designs, 273, 277  
 adjusted analysis, 82  
 bioequivalence, 201–202  
 calculation, 50–56, 72–73  
 and confidence intervals, 157–159  
 Fisher’s exact test, 72–73  
 interpretation, 48–49  
 misinterpretation, 160–161  
 multiplicity, 341  
 multiplicity adjusted, 277  
 non-significant, 328  
 publication, 381–382  
 similarity, 160–161  
 statistically significant, 328, 334–335
- PA *see* predictive accuracy
- packaging, treatments, 4
- paired design, 13
- parallel-group cholesterol lowering study, 146–148
- parallel-group design, 12–13
- parametric methods, non-normality, 187–190
- Pearson chi-square test, 63–66
- Pearson correlation coefficients, 114
- per-protocol sets (PPS), 120–121, 142, 154, 182, 205
- performance bias, 286, 318
- period effects, 61
- permanent treatment effect, 223–224
- PFS *see* progression-free survival
- pharmacovigilance  
 Bayesian neural networks, 353–354  
 post-approval safety monitoring, 350–351  
 proportional reporting ratio, 351–353
- phase II/III trials  
 adaptive designs, 267–268, 275–278  
 seamless, 275–278, 276  
 incorporation of phase II data, 277–278  
 inferentially/operationally, 276  
 logistical challenges, 278
- placebo effect, 38
- placebos, 3
- planning  
 meta-analysis, 316–317  
*see also* pre-planning
- Pocock method, 170
- point estimates, 249
- Poisson distribution, 190
- polychotomous logistic model, 98
- pooling, 339, 379  
 meta-analysis, 306–308
- population distribution, 26–27
- populations, estimands, 135
- positive predictive value (PPV), 358, 359, 360

- post hoc power, 156–157  
 post-approval safety monitoring, 350–351  
 posterior beliefs, 256–257  
 posterior distributions, 257  
 posterior probabilities, 260–261  
 power, 145–157  
   >80 per cent, 154  
   clinically relevant difference, 149, 153  
   conditional, 246–247  
   definition, 145  
   event rates in control group, 152–153  
   per-protocol set, 154  
   post hoc, 156–157  
   regulatory aspects, 153–155  
   sample size, 148–153, 153  
   standard deviation, 152  
   treatment differences detection, 145–148,  
     **146, 148**  
   type I/II errors, 144–145, 154  
   unpaired t-test, 146
- PPS *see* per-protocol sets  
 PPV *see* positive predictive value  
 pre-planning  
   multiplicity, 182–183  
   Statistical Analysis Plan, 373–374  
 precision, and bias, 10–12  
 predictive accuracy (PA), 359–360, 363, 364  
 Preferred Reporting Items for Systematic Reviews  
   and Meta-Analyses (PRISMA)  
   statement, 320  
 presentations, 379–382  
 prevalence, 358–359  
 prevastatin, 179–180  
 primary comparison, 197  
 primary endpoints, 20–21  
   changing, 278–279  
 principle stratum policies, 139–141  
 prior beliefs, 255, 256  
 PRISMA statement *see* Preferred Reporting Items for  
   Systematic Reviews and Meta-Analyses  
 probability  
   normal distribution, 32, **32**  
   *p*-values, 48–56  
   *see also* Bayesian...; *p*-values  
 progression-free survival (PFS), 343  
 propensity scoring, 293–297, 300–302  
   balance statistics, 302  
   caliper, 296  
   definition, 293–295  
   inverse weighting, 300–302  
   matching, 296–297  
   regression, 296  
   stratification, 295–296  
 proportional hazards assumption, 227  
 proportional hazards model, 231–232  
 proportional reporting ratios (PRR), 350, 351–353  
 proportions, confidence intervals, 45–46  
 prostate cancer, enzalutamide, 176–177
- PRR *see* proportional reporting ratios  
 publication bias, 318–319, **318**
- Q-TWiST *see* quality-adjusted time without  
   symptoms or toxicity  
 qualitative interaction, 88  
 quality-adjusted time without symptoms or toxicity  
   (Q-TWiST), 348–350  
 quantile–quantile plots, 189  
 quantitative interaction, 87–88
- radiotherapy vs. surgery, operable lung  
   cancer, 115–117  
 random effects model, 310, 313, 315  
 random element, 105  
 randomisation, 3–10  
   block, 4–5  
   central, 8  
   cluster, 10  
   description, 3  
   dynamic allocation and minimisation, 9–10  
   intention-to-treat, 118  
   meta-analysis, 307  
   network meta-analysis, 325–326  
   packaging, 4  
   publication, 380  
   sample sizes, 151  
   stratified, 6–8  
   unequal, 6  
   unrestricted/simple, 4  
 randomness, hypothesis testing, 53–55  
 rank preserving structural failure time model  
   (RPSFTM), 234–237  
 rank probabilities, 326–327  
 rate ratio, 77, 323–324  
 re-estimation, sample sizes, 273–275  
 receiver operating characteristic (ROC)  
   curves, 360–362  
 recurrent events, 77  
 regression  
   baseline factors, 92  
   and correlation, 113–114  
   diagnostic models, 362–364  
   least squares, 93, 104–105  
   linear, 92–95, 114  
   logistic, 97–98  
   multiple, 95–97  
   multiple imputation, 127  
   Pearson correlation coefficients, 114  
   propensity scoring, 296  
   proportional hazards, 231–232  
   selection bias adjustment, 292–293  
 regression towards the mean, 112  
 regulatory aspects  
   adaptive designs, 267–268, 275–276, 278–279,  
     281–282  
   Bayesian analysis, 263–265  
   composite endpoints, 238–241

- regulatory aspects (*cont'd*)  
 covariate usage, 110–113  
 crossover, 236–237  
 meta-analysis, 320–321  
 multiplicity, 166–167  
 observational studies, 290–291  
 power, 154–155  
 sample size adjustment, 154–155  
 relative risk reduction (RRR), 68, 70, 71, 76  
 relative risk (RR), 67–68, 70, 71, 76, 88, 220–221  
   case-control studies, 304–305  
   network meta-analysis, 323–324  
 repeated measures, 109–110  
 replacement, dropouts, 132–133  
 reporting  
   sample size calculations, 155  
   statistical analysis, 375  
 residual variance, 105  
 residuals, 184  
 respiratory distress, neonatal, 107  
 responder analysis, 23, 122  
 restricted adaptations, 267–268  
 restricted mean survival time (RMST), 229–230  
 risk of bias, 317, **318**  
 risk ratio, 68  
 risk sets, 220  
 RMST *see* restricted mean survival time  
 robustness, 313, 378  
 ROC curves *see* receiver operating characteristic  
 routine evaluation  
   safety, 330–340  
     across trial summaries, 339–340  
     adverse events, 332–335  
     ECG data, 338–339  
     laboratory data, 335–338  
     vital signs, 339  
 RPSFTM *see* rank preserving structural failure time model  
 RR *see* relative risk  
 RRR *see* relative risk reduction  
 Rubin's rules, 127  
 rule of three, 329–330
- SAEs *see* serious adverse events  
 safety analysis and monitoring, 328–355  
   adverse events, 329–330, 332–335  
   Bayesian neural networks, 353–354  
   benefit-risk assessment, 341–350  
     current approaches, 341–342  
     multi-criteria decision analysis, 342–348, 343–347  
   Q-TWIST, 348–350  
   quality-adjusted time without symptoms or toxicity, 348–350  
   utility, 344–345, 348  
 Data Monitoring Committees, 340–341  
 electrocardiography, 332  
 pharmacovigilance, 350–354
- Bayesian neural networks, 353–354  
 post-approval safety monitoring, 350–351  
 proportional reporting ratios, 351–353  
 QT interval, 332  
 routine evaluation, 330–340  
   adverse events, 332–335  
   ECG data, 338–339  
   laboratory data, 335–338  
   safety set, 330  
   safety summary across trials, 339–340  
   serious adverse events, 331  
   vital signs, 332, 339  
 rule of three, 329–330  
 safety study, 340  
 safety monitoring, interim analysis, 249–250  
 safety set, 131, 330  
 Salk Polio Vaccine trial, 1–2  
 sample distribution, 26–27  
 sample sizes, 148–157  
   adaptive designs, 273–275  
   adjustment, 154–155  
   calculation, 148–151, 155–156, 211–213, 241–242  
   clinically relevant difference, 149, 153  
   event rates in control group, 152–153  
   event-driven studies, 241  
   non-inferiority trials, 211–213  
   post hoc power, 156–157  
   randomisation, 150–151, 151  
   re-estimation, 273–275  
   regulatory aspects, 154–155  
   reporting calculations, 155–156  
   standard deviation, 152  
   survival data, 241–242  
   time-to-event endpoints, 241–242  
   unpaired t-test, 148–149
- sampling and inferential statistics, 25–39  
   box plots, 29–30  
   median and mean, 27  
   normal/Gaussian distribution, 30–33  
   notation, 28–29  
   population distributions, 26–27  
   sample distributions, 25–26  
   standard deviation, 27–28  
   standard error, 36–39  
   statistics and population parameters, 26–30  
 sampling variation, 35  
 SAP *see* statistical analysis plan  
 sceptical priors, 256  
 scientific method, 49  
 scientific validity, meta analysis, 316–320  
   funnel plots, 319–320, 320  
   planning, 316–317  
   PRISMA statement, 320  
   publication bias, 318–319  
   risk of bias, 317, **318**  
 scores, endpoints, 19  
 sd *see* standard deviation

- se *see* standard error  
 seamless phase II/III trials, 275–278  
   incorporation of phase II data, 277–278  
   inferentially, 276  
   logistical challenges, 278  
   multiplicity, 276–277  
   operationally, 276  
 secondary endpoints, 21  
 selection bias, 285–289, 291–300, **318**  
   baseline imbalances, 291–292, 298–302  
   correction methods comparison, 297–300  
   inverse propensity score weighting, 300–302  
   propensity scoring, 293–297, 300–302  
   regression, 292–293  
   stratification, 292–293, 295–296  
 sensitivity  
   diagnosis, 357  
   estimands, 142–143  
   missing data, 124–125, 378  
   statistical analysis, 378  
 SENTINEL trial, 381  
 serious adverse events (SAEs), 45–46, 63–69, 249, 252–253, 313–314  
 severe heart failure, 219, 231–232  
 Shapiro–Wilks test, 186  
 signal-to-noise ratio, 16  
 significance level, 56  
 similarity, *p*-values, 160–161  
 simple linear regression, 92–95  
   dependence of time to disease recurrence, 92–93, 93  
   least squares, 93  
   least-squares regression line, 93, 95  
 simple randomisation, 4  
 simple treatment comparisons  
   chi-square test, 63–66  
   confidence intervals, 69–70  
   Fisher's exact test, 72–73  
   interpretation, 71  
   multiple treatment groups, 77–79  
   number needed to treat, 69  
   odds ratio, 67  
   relative risk, 67–68  
   relative risk reduction, 68  
 t-test  
   interpretation, 62–63  
   paired, 59–61  
   unpaired, 58–59  
 Simpson's paradox, 307  
 single mean confidence intervals  
   95 per cent, 40–42  
   confidence coefficients, 42  
   lower/upper limit, 40  
   multiplying constants, 42–43, **43**  
   standard error, 43–44  
 single pivotal trial and 0.05, 161–163  
 solifenacin, 108–109  
 Spearman (rank) correlation coefficients, 114  
 specificity, diagnosis, 357  
 sponsors, Data Monitoring Committees, 251–252  
 square root transformation, 189  
 standard deviation (sd), 27–28, 83, 152  
 standard error (se)  
   bootstrapping, 47  
   chi-square test, 66  
   computer simulation, 34  
   confidence intervals, 43–44  
   difference between two means, 38–39  
   general setting, 39  
   indirect treatment comparisons, 323–324  
   of the mean, 33–36  
   normal distribution, 33, 33  
   for proportions, 39  
   sampling distributions, 36  
   sampling variation, 35  
 standard normal, 191  
 statistical analysis plan (SAP), 373, 374  
 statistical models, 105  
 statistical significance, 56  
   confidence intervals, 157–160  
   correlation coefficients, 114  
   multiplicity, 164–183  
   nominal, 183  
   *p*-values, 157–159, 160–163  
 statistics and statisticians, 370–382  
   blind review, 374  
   Data Monitoring Committees, 252  
   data validation plans, 373–374  
   importance of statistical thinking, 370–371  
   methodological requirements, 372–378  
   pre-planning, 375–377  
   publication and presentations, 379–382  
   regulatory guidelines, 371–372, 375  
   regulatory submission, 378–379  
   reporting, 375  
   robustness, 378  
   sensitivity, 378  
   statistical analysis plans, 373, 374  
   trial and development plans, 371  
 step functions, 219  
 stopping rules, interim analysis, 243–245  
 stratification  
   over-stratification, 8  
   propensity scoring, 295–296  
   selection bias adjustment, 292–293, 295–296  
   survival data, 230–231  
 stratified analysis, 82  
 stratified randomisation, 6–8, 80  
 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), 291  
 stroke, 174–175  
 subgroup testing, 178–181, 229–230  
 success classification, missing data, 123  
 SUCRA *see* surface under the cumulative ranking curve  
 summary statistics, estimands, 136

- superiority trials, 17–18  
 intention-to-treat, 120  
 non-inferiority trials, 213–215
- supplementary analysis, estimands, 143
- surface under the cumulative ranking curve (SUCRA), 326–327
- surgery vs. radiotherapy, operable lung cancer, 115–117
- surrogate endpoints, 21–22, **22**
- survival data, 217–242  
 accelerated failure time model, 232–233  
 adjusted analyses, 230–233  
 censoring, 217, 228, 233–234  
 composite time-to-event endpoints, 237–241  
 crossover, 234–237  
 cumulative incidence functions, 237–238  
 Gehan–Wilcoxon test, 223–224  
 hazard ratio, 225–229  
 independent, non-informative censoring, 233–234  
 intention-to-treat, 129–131  
 KM curves, 218–222, 228–229, 235–236  
 logrank test, 222–223  
 long-term effects, 223–224  
 proportional hazards regression, 231–232  
 rank preserving structural failure time model, 234–237  
 restricted mean survival time, 229–230  
 sample size calculations, 241–242  
 stratified methods, 230–231  
 survival curve patterns, 223  
 time-to-event endpoints, 237–241  
 treatment comparisons, 222–225  
*see also* time-to-event outcomes
- synthesis method, 209–210
- systematic reviews, 306
- t-test**  
 binary, categorical and ordered categorical endpoints, 78–79  
 interpretation, 62–63  
 multiple treatment groups, 77–79  
 paired, 59–61, 62, 78  
 unpaired, 53, 58–59, 61, 62, 78, 148–149  
 test statistic, 54
- therapeutic equivalence, 206
- thyroid cancer, 236, 237
- ticagrelor, 174–175
- time-dependent covariates, 111
- time-to-event endpoints, 19, 129–131  
 composite, 237–241  
 interim analysis, 244–245  
 sample sizes, 241–242
- transitivity, 323–326
- trastuzumab, 45–46
- treatment differences, 145–148, **147, 148**
- treatment effect, 38  
 permanent, 223–224
- treatment policy strategy, 137
- treatment-by-centre interactions, 89–90  
 treatment-by-covariate interactions, 100–102  
 treatment-by-factor interactions, 85–87  
 treatments, packaging, 4
- trial design, diagnostic agents, 364–365
- two-sample t-test *see* unpaired t-test  
 two-sided *p*-values, 56–57  
 two-trial rule, 161–162  
 two-way analysis of variance (two-way ANOVA), 82–83
- type 2 diabetes, 137–139
- type I errors, 144–145  
 adaptive designs, 268–271  
 inflation, 164–165  
 multiplicity, 164–183
- type II errors, 144–145, 154
- unblinding, inadvertent, 249
- unequal randomisation, 6
- unintended drug effects, pharmacovigilance, 350–355
- unpaired t-test, 53, 58–59, 61, 62, 78, 81
- unrestricted randomisation, 4
- upper confidence limit, 40
- vague priors, 256
- validity  
 meta-analysis, 316–320  
 funnel plots, 319–320, 320  
 planning, 316–317  
 PRISMA statement, 320  
 publication bias, 318–319  
 risk of bias, 317, **318**
- van Elteren test, 194
- variables  
 estimands, 135  
*see also* endpoints  
 variance, 28  
 residual, 105  
 VAS *see* visual analogue scale  
 venous thromboembolism prevention, 207  
 visual analogue scale (VAS), 23  
 vital signs, 339  
 vitamin supplementation, 315–316
- weighted kappa, 367
- weighted logrank test, 225
- weighting  
 adaptive designs, 274  
 inverse, propensity scoring, 300–302
- Welch's approximation, 185
- Wilcoxon rank sum test, 190–191
- Wilcoxon signed rank test, 192–193
- within-patient designs, 12–13
- within-subject variation, 109–110
- worst-case classification, 124
- xamoterol, 155–156

# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.