MANCHESTER 1824

The University of Manchester

Validation of Prediction Models

Dr Glen Martin

glen.martin@manchester.ac.uk

Intended Learning Objectives

The University of Manchester

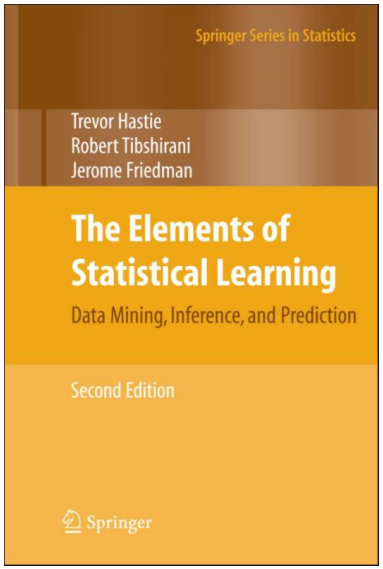
By the end of this session, you will be able to:

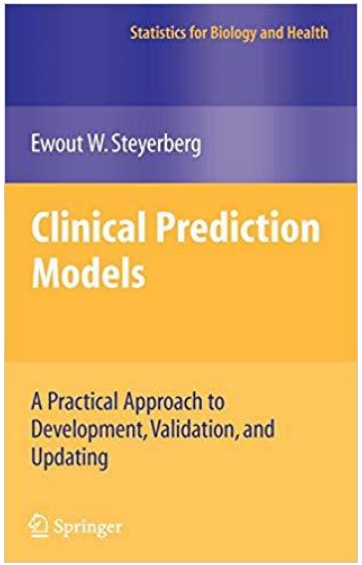
1) Describe the processes/steps required to validate a clinical prediction model, including the difference between internal and external validation

2) Understand the key metrics used to evaluate the predictive performance of a clinical prediction model

The concepts covered in this lecture will be applied in Practical Session II.

Reading





This lecture draws on the material covered in Chapter 7 of "The Elements of Statistical Learning" by Hastie et al., and chapters 15 and 17 of "Clinical Prediction Models" by Ewout Steyerberg

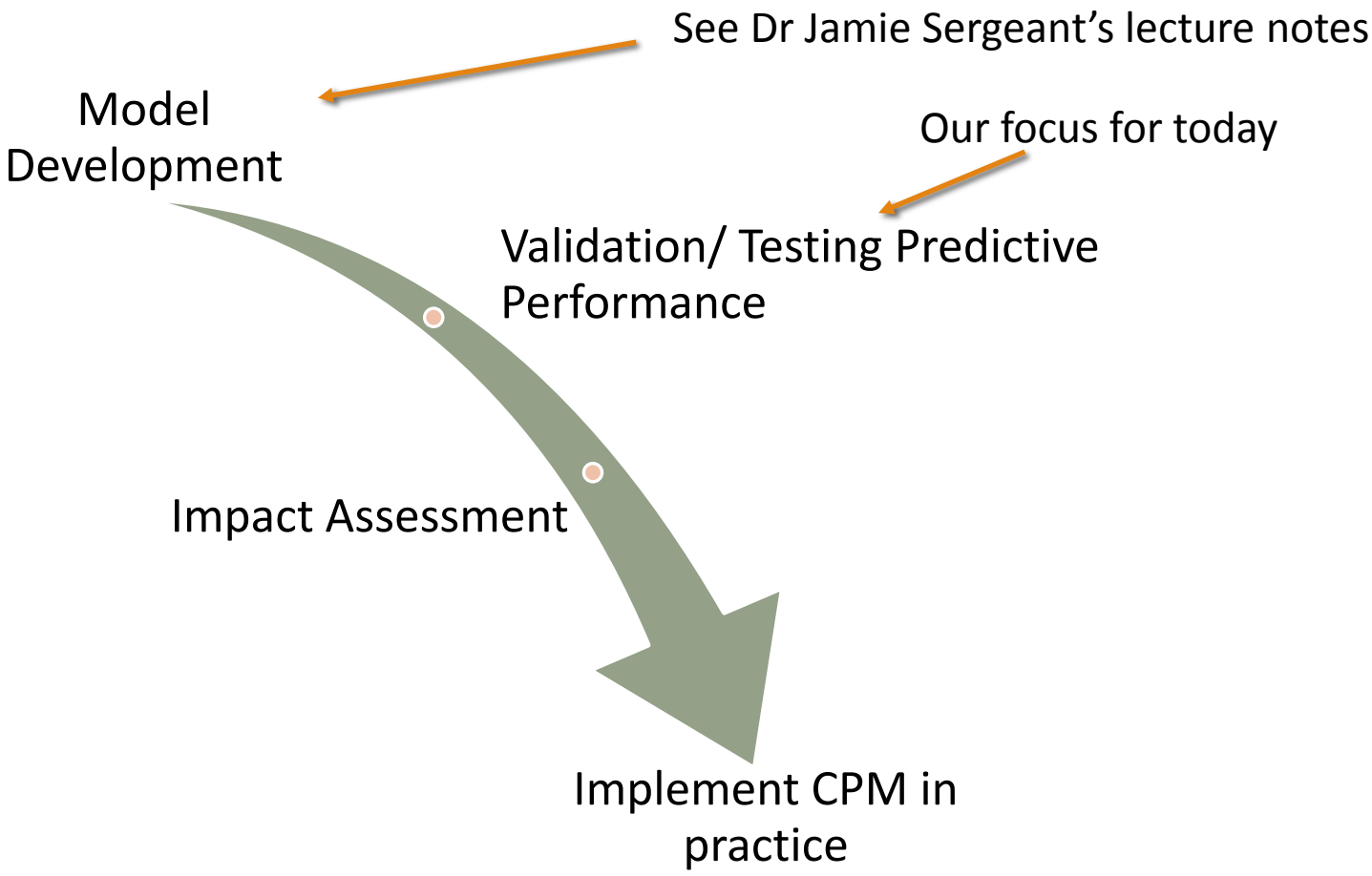
MANCHESTER 1824

The University of Manchester

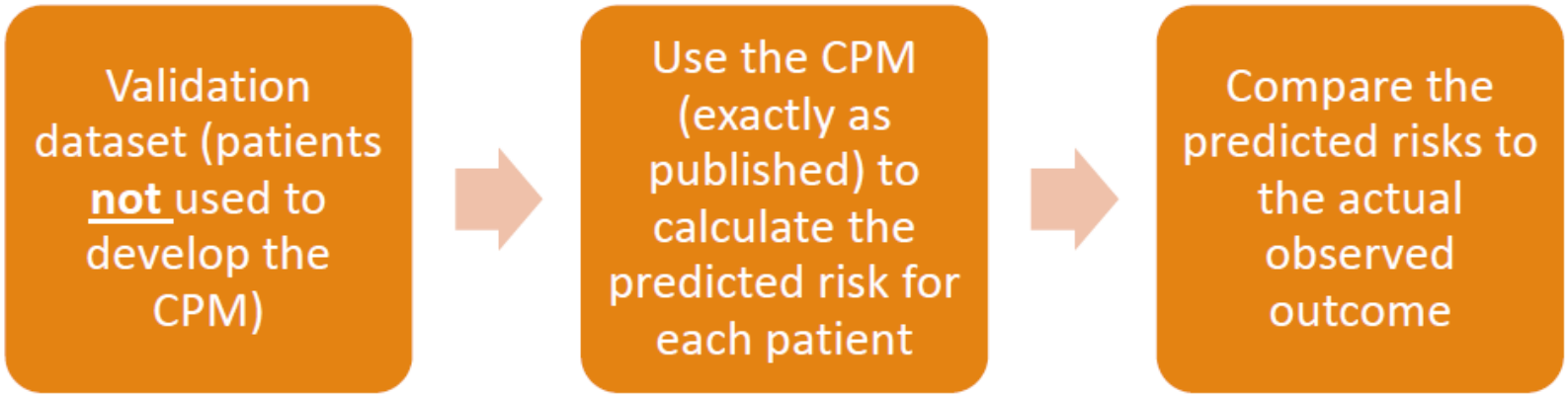
Validation of Prediction Models

DR GLEN MARTIN, UNIVERSITY OF MANCHESTER

The CPM Pipeline



General Process of Validating a Prediction Model



Validation

Before we delve into this, it is important to differentiate two separate goals that we might have in mind at this stage

Model selection: testing and comparing the performance of different models to choose the "best" (e.g. variable selection)

Model validation: having selected a model, estimate how well it performs at making predictions in new data

We are focussing on the latter here. Model selection is more akin to steps taken during model development stage of the CPM pipeline

Why new data?

…in the previous slide, we emphasised that the data used to validate the CPM must not be the same as used to develop/fit/derive it – why do you think this is?

- That is, why cant we just estimate predictive performance in the training data?

Obtaining completely new data will not always be possible – do you have any thoughts, at this stage, about how we might overcome this and still estimate performance? We will return to this later!

Internal and External Validation

The University of Manchester

Internal Validation

- This is well-defined: an examination of the performance of a model in the same population or setting that the model was developed in – how well the model replicates

External Validation

- This is less well-defined, but in general terms it means capturing some notion of a model's ability to generalise to new settings or populations that differ from the setting or population in which the model was developed.
  - How people interpret this differs greatly; we will return to this point later

We will now explore each of these in more detail, in turn.

Internal Validation

This is often evaluated at the same time as the model is developed, and using the same data as used to derive the model

This, therefore, presents us with a challenge: how do we estimate predictive performance while also dealing with in-sample optimism?

We will answer both of these questions over the next few slides…

Bias-Variance Trade off

Given a training set $\mathrm{T} = \{(x_1, y_1),\ldots,(x_N, y_N)\}$, drawn at random from a population, we develop a prediction model $f(X)$. Denote $L\big(Y, f(X)\big)$ as the loss function for errors between Y and $f(X)$.

Training error is the average loss over the training sample: $\frac{1}{N}\sum_1^N L\big(y_i, f(x_i)\big)$

Now, given new observations $(\tilde{x}_1, \tilde{y}_1)$ (test data), test error is defined as $E\big[L\big(\tilde{y}_1, f(\tilde{x}_1)\big)|\mathrm{T}\big]$.

Ideally, our goal when validating a model is to estimate test error.

Here, T is fixed – so test error refers to the expected error for a specific training test

…this conditional error is hard to estimate given only the information in the given training set

Bias-Variance Trade off

In general, we will be able to estimate prediction error (expected test error) using observed data:

$$\text{Err} = E\left[L\left(\tilde{y}_1, f\left(\tilde{x}_1\right)\right)\right]$$

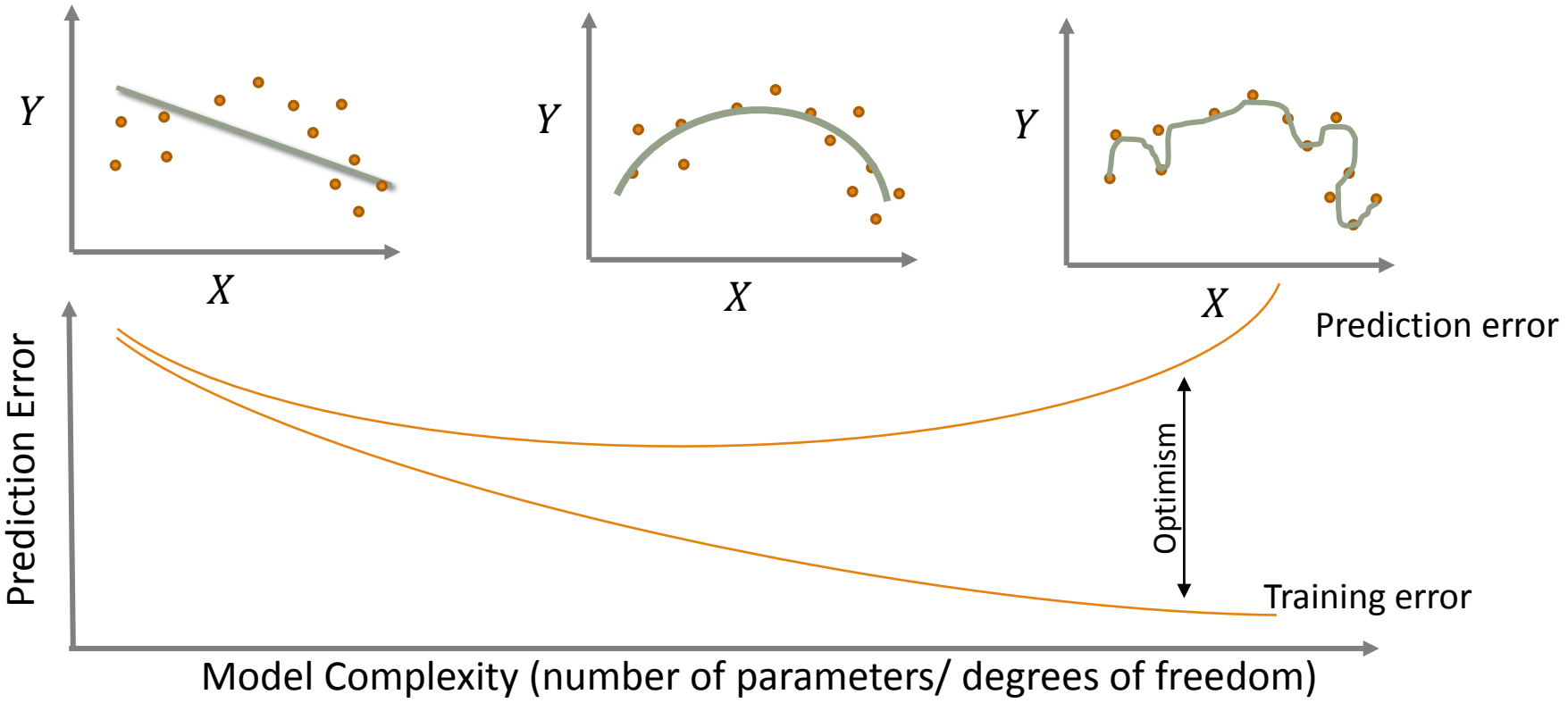
Note that this is taking the expectation over everything, **including the randomness in $\mathbf{T}$**!

Informally, how well the model predicts in the test data.

But how?

Can we just use the training data to estimate prediction error?...

Training error is a poor estimate of test error



Training error decreases with model complexity. The more complex a model is, the more likely it is to be overfit to the training data, which leads to large test error. The difference between training and test error is **in-sample optimism**

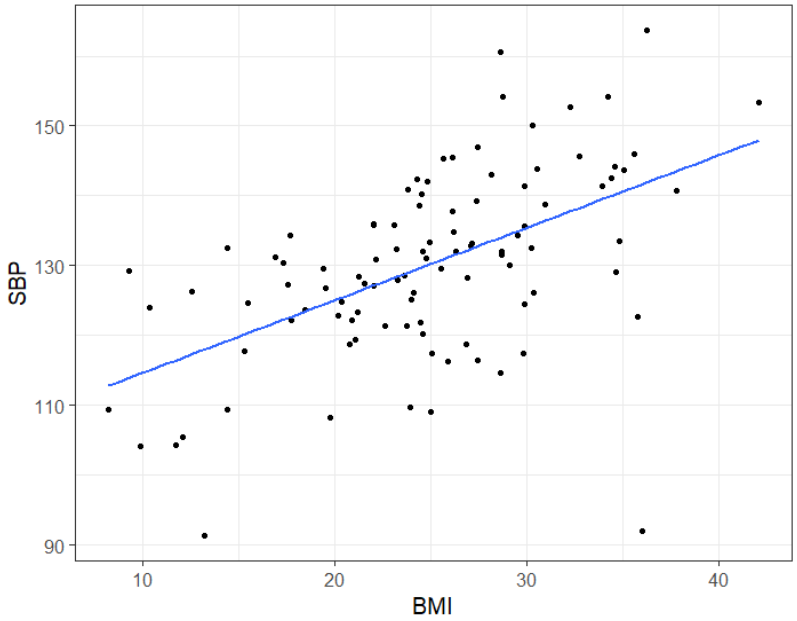
In-sample optimism

To illustrate the concept of in-sample optimism, lets return to an example from yesterday – the concept applies to any model type.

100 observations of SBP (outcome) and BMI; obtained MLEs of the intercept and slope using least squares – as shown.

Informally speaking, any fitted model "adapts to" the training data.

Therefore, the apparent error/performance (training error) will be an overly optimistic estimate of the error/performance in new observations from the same population (test error).



In-sample optimism

Lets explore this formally….

Using the 100 training samples (x_1, y_1),…,(x_{100}, y_{100}) drawn at random from a population (x=BMI, y=SBP), we obtained

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Now suppose we have some test data $(\tilde{x}_1, \tilde{y}_1)$, …, $(\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data.

Theory:

If $R_{tr}(\hat{\alpha}, \hat{\beta}) = \frac{1}{100}\sum_{i=1}^{100}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ (i.e. sum of squared residuals) and $R_{test}(\hat{\alpha}, \hat{\beta}) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\alpha} - \hat{\beta}\tilde{x}_1)^2$, then $E[R_{tr}(\hat{\alpha}, \hat{\beta})] \leq E[R_{test}(\hat{\alpha}, \hat{\beta})]$

We leave the proof of this as an exercise for those interested. The difference is optimism

Methods to estimate test error/ predictive performance

The University of Manchester

During internal validation, we will (usually) only have access to one dataset of random samples from a population.

We therefore need to use this data to both develop the model (see yesterday's lecture) and internally validate it.

If we are in a very (very) data rich situation one possibility is an approach called split-sample.

Internal Validation: Split-sample

The University of Manchester

The split-sample approach to internal validation is where we divide the data randomly into two groups

- Fit the data in one group and validate it in the other.

Development sample (e.g. 75% of data)	⇒ Develop the CPM
Validation sample (e.g. 25% of data)	⇒ Test/Validate the CPM

What are your thoughts on this approach? Is it the most optimal use of the data?

Avoid split-sample

In general, we should avoid using the single split-sample approach.

Data is precious, and so we need to maximise the amount we use to develop our models while also handling in-sample optimism to estimate predictive performance

Splitting a portion away for validation is inefficient use of data

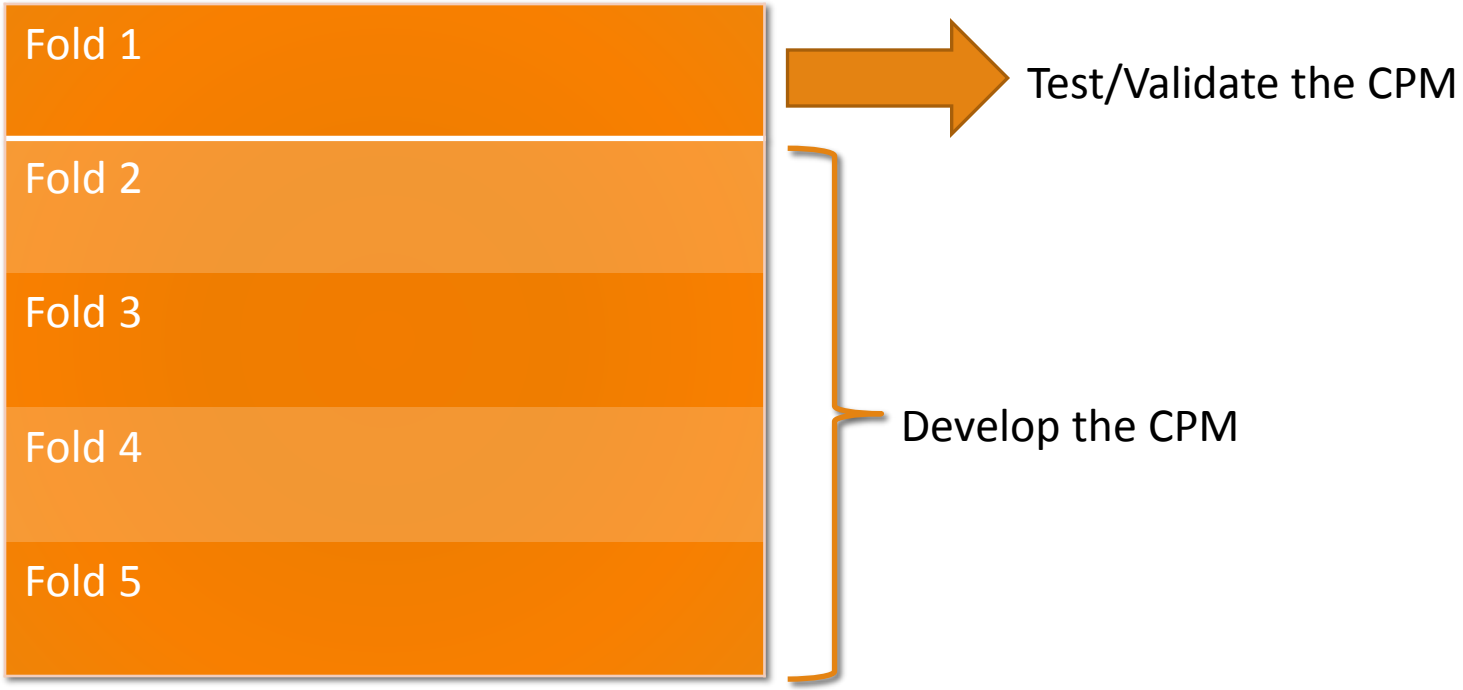
Using resampling methods of cross-validation or bootstrapping are preferred.

These methods directly target estimates of prediction error (i.e. the average error/predictive performance when the CPM is applied to an independent test set drawn from the same population as developed in)

Internal Validation: Cross-Validation

The University of Manchester

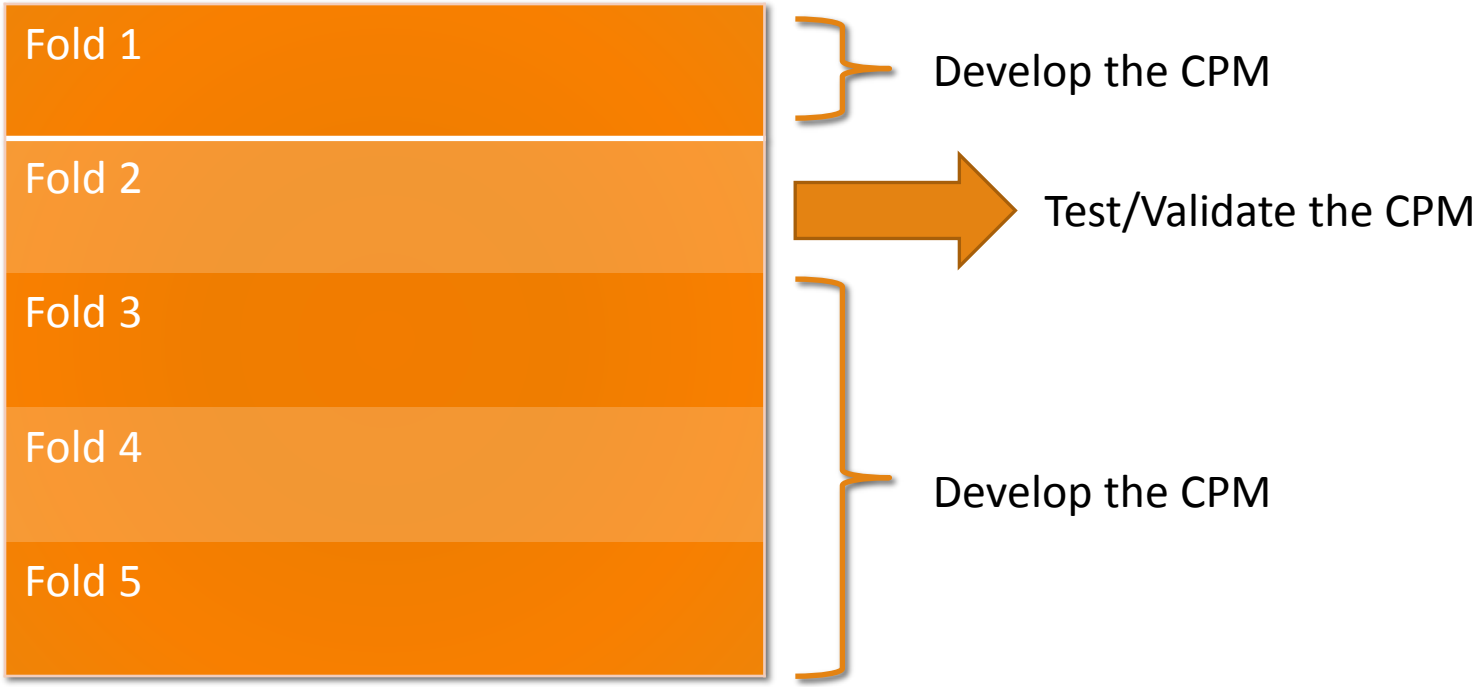
For example 5-fold cross validation:



Internal Validation: Cross-Validation

The University of Manchester

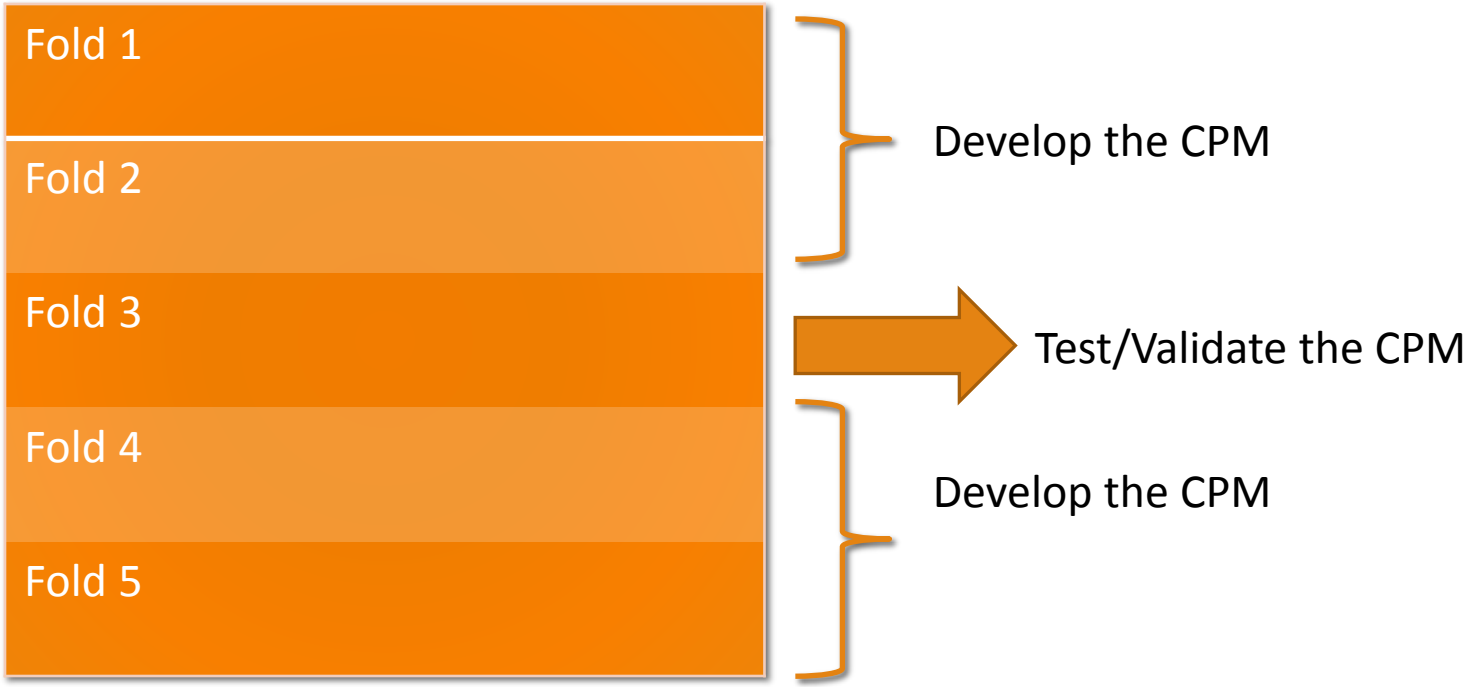
For example 5-fold cross validation:



Internal Validation: Cross-Validation

The University of Manchester

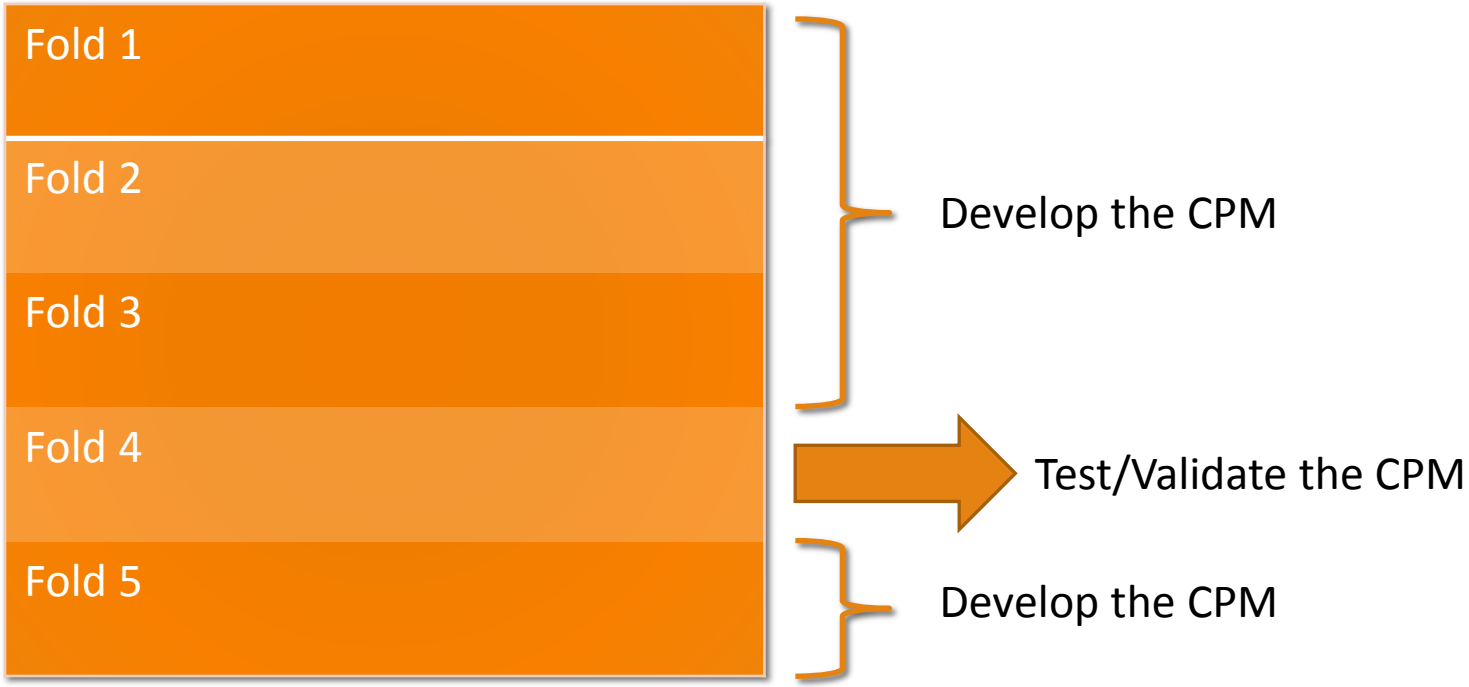
For example 5-fold cross validation:



Internal Validation: Cross-Validation

The University of Manchester

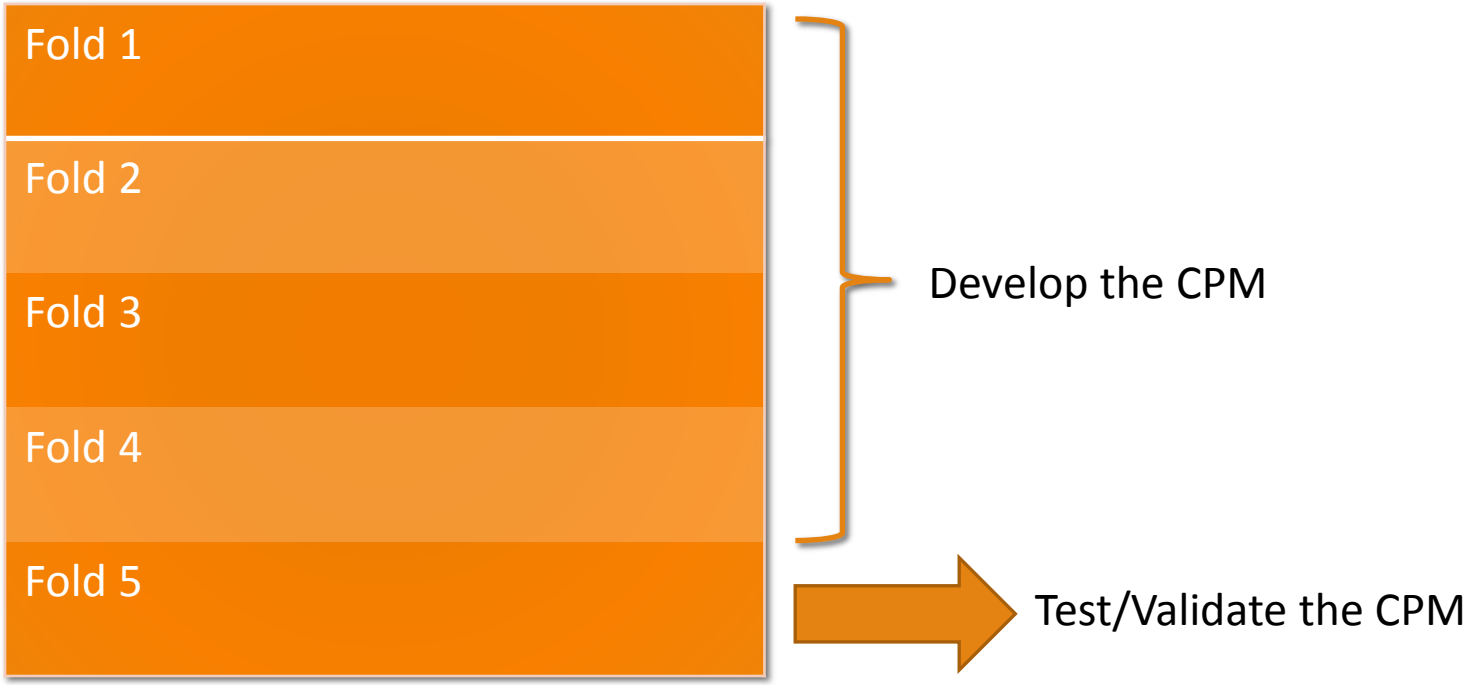
For example 5-fold cross validation:



Internal Validation: Cross-Validation

The University of Manchester

For example 5-fold cross validation:



Internal Validation: Cross-Validation

The University of Manchester

For example 5-fold cross validation:

Predictive Performance from validation in Fold 1	**Predictive Performance from validation in Fold 2**	**Predictive Performance from validation in Fold 3**	**Predictive Performance from validation in Fold 4**	**Predictive Performance from validation in Fold 5**

After we have repeated this process across all folds, we take the average of the performance/validation results from each fold.

Internal Validation: Cross-Validation – choice of k

The University of Manchester

Usual choices are either 5-fold or 10-fold cross-validation.

An extreme is $k = N$ cross-validation; a.k.a. leave-one-out cross-validation, where the model is fitted to all observations bar one, and error is assessed on the left-out observation.

That is $\frac{1}{N}\sum_{i=1}^{N} L(y_i, f^{-i}(x_i))$

Where $f^{-i}(x_i)$ is the model fitted to all data except observation i, evaluated using the ith value(s) of x

What are your thoughts on the choice of k? What would be the strengths and limitations of $k = N$ compared with $k = 5$?

Internal Validation: Bootstrapping

MANCHESTER 1824
The University of Manchester

Bootstrapping simply means sampling rows, with replacement, from an observed dataset to create a 'mock' dataset with the same number of rows.

Internal Validation: Bootstrapping

The University of Manchester

Bootstrap Validation Involves:

1. Derive the CPM on the full (raw) dataset
2. Sample rows (with replacement) from the raw data to create a bootstrap dataset
3. Derive a CPM in the bootstrap data (using identical analysis steps to 1)
4. Validate the model from step 3 within the bootstrap data
5. Validate the model from step 3 in the raw data
6. Subtract the performance metrics from step 4 with those from step 5 (i.e. calculate optimism)
7. Repeat steps 2-6 lots of times (e.g. >500 times), storing the optimism estimates
8. Average the optimism estimates
9. Validate the model from step 1 within the raw data and 'subtract' the average optimism estimate from step 8

The University of Manchester

Activity (5 mins)

In your tables (or small groups), discuss the relative strengths and weaknesses of each of the above approaches to internal validation

We will then discuss this as a full group.

MANCHESTER 1824

The University of Manchester

External Validation

Recall that in general terms we define EV to mean capturing some notion of a model's ability to generalise to new settings or populations.

How people interpret this varies in three main ways:

1. Defines external validation as an examination of performance in some different population or setting. One might develop a model using one dataset, and then externally validate the model in a second dataset (representing a different population or context). This second dataset may be arbitrary or chosen for convenience
2. Defines external validation as testing performance in a new dataset that is carefully chosen to match the intended use of the model
3. Defines external validation as a broad concept of showing that a model generalises and/or transports well

While option 1 is commonly done, 2 and 3 should be the focus as they are more informative

The University of Manchester

Approach 2

Image we wish to develop a model, $f(x)$, intended to predict length of stay in Intensive Care Unit at Manchester Royal Infirmary (MRI). We develop $f(x)$ using data collected from MRI.

We might test internal performance using bootstrapping

For external validation, the most sensible choice is to obtain more/new data from MRI (e.g. at a later timepoint) and evaluate prediction error/performance.

Note: testing performance in another (arbitrary) hospital would not make sense for this intended use-case!

Approach 3

Image we wish to develop a model, $f(x)$, intended to predict risk of sudden cardiac death after myocardial infarction to support international guidelines. We develop $f(x)$ using data collected from multiple countries around the world.

For external validation, the most sensible choice is to obtain more/new data internationally and evaluate heterogeneity in prediction error/performance – i.e. testing how well $f(x)$ generalises/ transports.

Note: evaluating the model's performance in a single population/ country would not be sufficient in this case, unlike the previous example.

Internal-External Cross-Validation (IECV)

Approach 3 is a common requirement of EV – but how can we do this in practice?

IECV is a incredibly powerful and robust option to test generalisability/ transportability of a CPM

This process involves leaving out one of the datasets (e.g., countries), with the remaining datasets used to develop a CPM.

The left-out dataset is then used to validate the model.

This process cycles through leaving out each dataset in-turn, resulting in sets of predictive performance estimates, one per dataset, which are then combined by random effects meta-analysis.

Meta-analysis of predictive performance

The University of Manchester

Imagine we have J countries. Using IECV, we obtain J predictive performance estimates, θ_j, for all $j \in [1, J]$. To summarise these and quantify heterogeneity in performance, random-effects meta-analysis assumes that

$$\theta_j \sim N(\mu_j, S_j^2)$$

$$\mu_j \sim N(\mu, \tau^2)$$

where μ_j is the "true" performance for dataset j with known variance S_j^2, and that these performance statistics are assumed to be drawn from an overarching normal across datasets with mean performance μ and between-study variance τ^2.

Prediction intervals can then be used to summarise the expected model performance in a new but similar dataset to those included in the meta-analysis (i.e. the actual information of interest under approach 3 to EV!). A $100(1-\alpha)\%$ prediction interval can be calculated as

$$\hat{\mu} \pm t_{\alpha,k-2} \sqrt{\hat{\tau}^2 + \frac{1}{\sum_j \frac{1}{\hat{S}_j^2 + \hat{\tau}^2}}}$$

where $t_{\alpha,k-2}$ is the $100\left(1-\frac{\alpha}{2}\right)\%$ percentile of the t-distribution for k-2 degrees of freedom

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(11) 3505–3522

DOI: 10.1177/0962280217705678
journals.sagepub.com/home/smm
SAGE

Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?

Kym IE Snell,[1] Joie Ensor,[1] Thomas PA Debray,[2,3] Karel GM Moons[2,3] and Richard D Riley[1]

RESEARCH METHODS AND REPORTING

External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges

Richard D Riley,[1] Joie Ensor,[1] Kym I E Snell,[2] Thomas P A Debray,[3,4] Doug G Altman,[5] Karel G M Moons,[3,4] Gary S Collins[5]

MANCHESTER 1824

The University of Manchester

Metrics to Evaluate Predictive Performance

Predictive Performance Measures

We have so far covered what data/methods we should use to validate a CPM. We will now cover different metrics used to quantify the predictive performance of a CPM, which should be estimated through IV and/or EV.

We will be covering:

1) Overall predictive performance (R^2, Brier Score)

2) Discrimination (AUC, C-statistic)

3) Calibration (Calibration plot, observed:expected ratio)

Overall Predictive Performance - R^2

MANCHESTER 1824
The University of Manchester

R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s) – it is sometimes given the name of "**explained variation**"

More formally, given a set of observed continuous outcomes, $y_1, \dots, y_N$, to which we have fit a (linear regression) model to obtain predictions $\hat{y}_1, \dots, \hat{y}_N$, then the **residuals are $\boldsymbol{\epsilon_i = y_i - \hat{y}_i}$** for $i = 1, \dots, N$.

The total sum of squares is $SS_{tot} = \sum_{i=1}^{N}(y_i - \bar{y})^2$, where $\bar{y}$ is the mean of the observed outcomes.

Similarly, recall that the residual sum of squares is $SS_{res} = \sum_{i=1}^{N} \epsilon_i^2$.

Then, $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$.

There are **various variants of this for other outcome types such as binary outcomes and survival outcomes**.

Overall Predictive Performance – Brier Score

The University of Manchester

For binary outcomes, an alternative is to use the Brier Score

It is a quadratic scoring rule

Given a prediction model that estimates the probability of a binary outcome, y_i, as p_i (for $i = 1, \dots, N$), then the Brier Score is

$$\sum_{i=1}^{N}(y_i - p_i)^2$$

Here, the Brier Score will be minimised when the prediction is closer to 0 for individuals with $y_i = 0$ and when its closer to 1 for individuals with $y_i = 1$.

Calibration & Discrimination

The overall performance measures above combine two important aspects: calibration and discrimination.

Calibration refers to the agreement between the observed risk and that estimated by the model.

Discrimination the extent to which the model can differentiate between individuals who have the event and those that do not. i.e. are those that ultimately go onto have the outcome assigned a higher predicted risk than those who do not have the outcome?

Discrimination

Several measures can quantify the discriminative ability of a model.

The concordance (C-)statistic is the most commonly used performance measure to indicate the discriminative ability of generalized linear regression models.

For a binary outcome, the C-statistic is identical to the area under the receiver operating characteristic (ROC) curve (AUC).

Sensitivity and Specificity

The University of Manchester

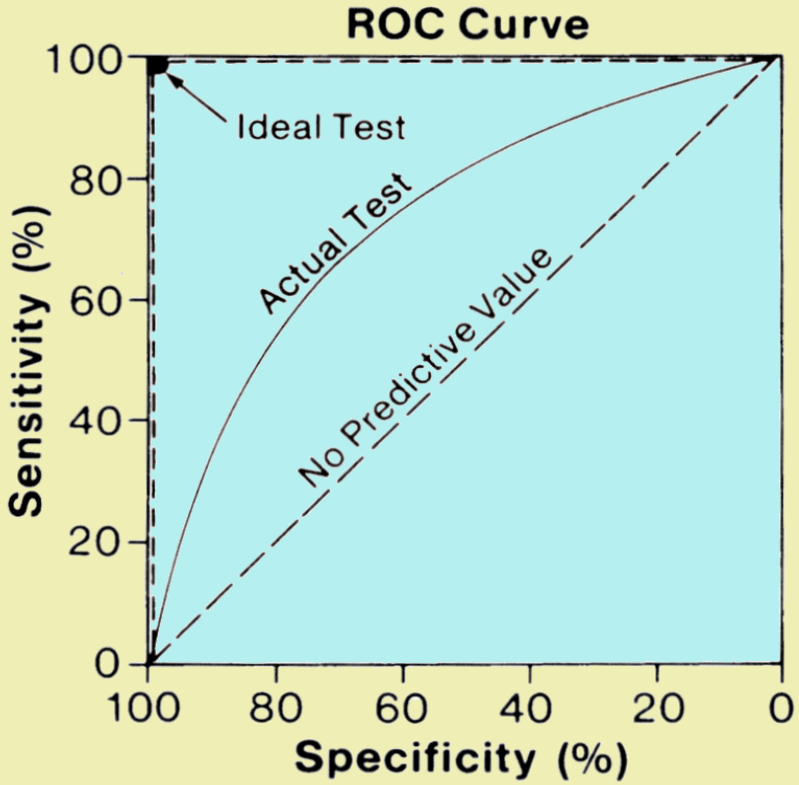
Sensitivity: (a.k.a. true positive rate), measures the proportion of patients who had the outcome and were correctly identified as such by the CPM

Specificity: (a.k.a. true negative rate), measures the proportion of patients who did not have the outcome and were correctly identified as such by the CPM

We can calculate these metrics from a prediction model, for a given cut-off of the predicted risks from the model (e.g. from a logistic regression model):

	Outcome +'ve	Outcome –'ve
Probabilty $\geq$ cut-off	TP	FP
Probabilty $<$ cut-off	FN	TN
	Total with outcome	Total without outcome

ROC Curve



By varying the cut-off value across multiple values between 0% and 100%, we can get corresponding sensitivity and specificity estimates

If we plot these on a plot of sensitivity against 1-specificity, we get a ROC curve.

This can be summarised with the area under the curve (AUC):

- 0.5 implies no predictive value, while 1 implies perfect discrimination

Calibration

Recall that is the agreement between the observed outcomes and those predicted by the model, across the full risk range.

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 74 (2016) 167–176

Journal of Clinical Epidemiology

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster[a,b,*], Daan Nieboer[b], Yvonne Vergouwe[b], Bavo De Cock[a], Michael J. Pencina[c,d], Ewout W. Steyerberg[b]

[a]KU Leuven, Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Bel
[b]Department of Public Health, Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Nethe
[c]Duke Clinical Research Institute, Duke University, 2400 Pratt Street, Durham, NC 27705, U
[d]Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC

Accepted 23 December 2015; Published online 6 January 2016

Abstract

Objective: Calibrated risk models are vital for valid decision support. We define four levels of calibration
model development and external validation of predictions.

Study Design and Setting: We present results based on simulated data sets.

Results: A common definition of calibration is "having an event rate of *R*% among patients with a prec
refer to as "moderate calibration." Weaker forms of calibration only require the average predicted risk (mea
prediction effects (weak calibration) to be correct. "Strong calibration" requires that the event rate equals th
variate pattern. This implies that the model is fully correct for the validation setting. We argue that this is unrea
incorrect, the linear predictor is only asymptotically unbiased, and all nonlinear and interaction effects sho
addition, we prove that moderate calibration guarantees nonharmful decision making. Finally, results indica
of calibration in small validation data sets is problematic.

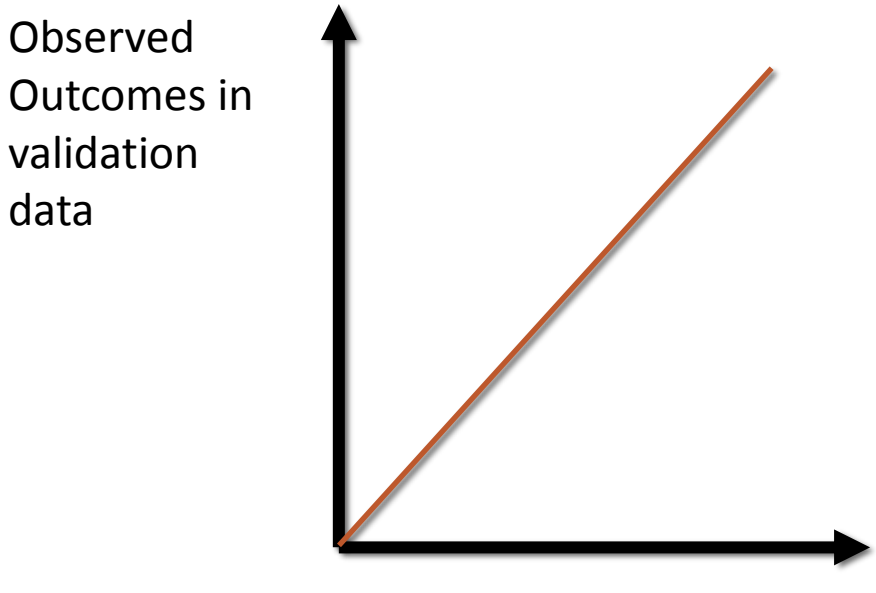
Conclusion: Strong calibration is desirable for individualized decision support but unrealistic and counter pr
development of overly complex models. Model development and external validation should focus on moderate calib

Defines a hierarchy for assessing calibration, from weak to strong assessment

The University of Manchester

Calibration Plots

Calibration plots can be a powerful way of (visually) assessing calibration of a model.

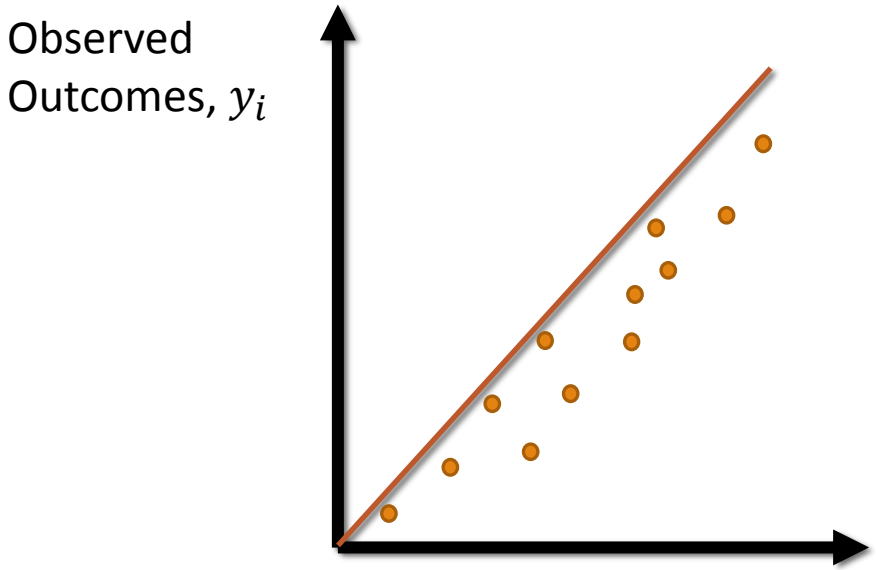


Any deviation from the 45 degree line indicates miscalibration

Calibration Plots

Imagine we have a risk model for a continuous outcome, where we have observed outcomes y_i and predicted outcomes $\hat{y}_i$ (in a validation set).

Then, we have a set of points: $(y_1, \hat{y}_1)$, $(y_2, \hat{y}_2)$, …, $(y_N, \hat{y}_N)$, which we can plot



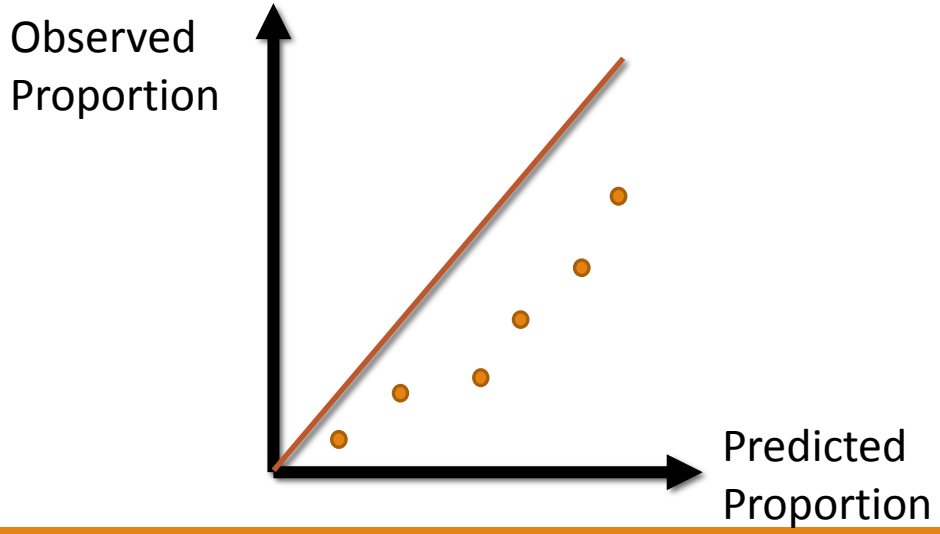
Any deviation from the 45 degree line indicates miscalibration

Calibration Plots

For binary outcomes, this is slightly more involved…

We now plot observed proportion of events against expected outcome proportions

- A common way to estimate these is to estimate observed proportions in groups based on the predicted risks;
- e.g. split into ten groups based on bands of predicted risk and calculate observed proportion who had event, and mean predicted risk



calibration-in-the-large & calibration slope

The University of Manchester

The calibration-in-the-large and calibration slope are two useful metrics to calculate. We will illustrate these here assuming we have a logistic regression CPM that we wish to validate – the concepts apply to any model type.

calibration-in-the-large & calibration slope

The University of Manchester

The calibration-in-the-large and calibration slope are estimated within a validation dataset by fitting a logistic regression model to the observed outcomes in the validation set with each patient's linear predictor as the only covariate:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha_0 + \alpha_1 LP$$

Here, the intercept α_0 indicates the extent that predictions from the CPM are systematically too low or too high (so-called calibration-in-the-large), where zero is the reference. Values of α_0 less than zero indicate that the CPM systematically over-predicts risk and *vice versa*.

The calibration slope, α_1 should be one for a well calibrated CPM, with values less than one indicate overfitting

Similar techniques can be used for other types of models.

Calibration

Other methods are also possible, but we do not have time to cover them here.

See the paper below for more information

ELSEVIER
CrossMark
Journal of Clinical Epidemiology 74 (2016) 167–176
Journal of Clinical Epidemiology

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster[a,b,*], Daan Nieboer[b], Yvonne Vergouwe[b], Bavo De Cock[a], Michael J. Pencina[c,d], Ewout W. Steyerberg[b]

[a]KU Leuven, Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Belgium
[b]Department of Public Health, Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands
[c]Duke Clinical Research Institute, Duke University, 2400 Pratt Street, Durham, NC 27705, USA
[d]Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27719, USA

Accepted 23 December 2015; Published online 6 January 2016

Abstract

Objective: Calibrated risk models are vital for valid decision support. We define four levels of calibration and describe implications for model development and external validation of predictions.

Study Design and Setting: We present results based on simulated data sets.

Results: A common definition of calibration is "having an event rate of *R*% among patients with a predicted risk of *R*%," which we refer to as "moderate calibration." Weaker forms of calibration only require the average predicted risk (mean calibration) or the average prediction effects (weak calibration) to be correct. "Strong calibration" requires that the event rate equals the predicted risk for every covariate pattern. This implies that the model is fully correct for the validation setting. We argue that this is unrealistic: the model type may be incorrect, the linear predictor is only asymptotically unbiased, and all nonlinear and interaction effects should be correctly modeled. In addition, we prove that moderate calibration guarantees nonharmful decision making. Finally, results indicate that a flexible assessment of calibration in small validation data sets is problematic.

Conclusion: Strong calibration is desirable for individualized decision support but unrealistic and counter productive by stimulating the development of overly complex models. Model development and external validation should focus on moderate calibration.

A final note on sample size

The University of Manchester

Statistics in Medicine

RESEARCH ARTICLE | Open Access

Minimum sample size for external validation of a clinical prediction model with a continuous outcome

Lucinda Archer, Kym I. E. Snell, Joie Ensor, Mohammed T. Hudda, Gary S. Collins, Richard D. Riley

First published: 04 November 2020 | https://doi.o

Statistics in Medicine

RESEARCH ARTICLE | Open Access

Minimum sample size for external validation of a clinical prediction model with a binary outcome

Richard D. Riley, Thomas P. A. Debray, Gary S. Collins, Lucinda Archer, Joie Ensor, Maarten van Smeden, Kym I. E. Snell

J Clin Epidemiol. 2021 Jul;135:79-89. doi: 10.1016/j.jclinepi.2021.02.011. Epub 2021 Feb 14. .9025 | Citations: 3

External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb

Kym I E Snell [1], Lucinda Archer [2], Joie Ensor [2], Laura J Bonnett [3], Thomas P A Debray [4], Bob Phillips [5], Gary S Collins [6], Richard D Riley [2]

MANCHESTER 1824

The University of Manchester

A quick quiz…

Go to www.menti.com

MANCHESTER 1824
The University of Manchester

Take-Home Messages

1. Robust validation (internal and external settings) is fundamental to establish the predictive performance of a model
2. For internal validation, methods such as cross-validation or bootstrapping should be used to obtain estimates of prediction error, avoiding in-sample optimism
3. For external validation, the approach and data used for the assessment should be based on the intended use of the model – IECV is a strong way to assess transportability of a model
4. Calibration and Discrimination should be assessed and **both** are equally important in assessing predictive performance

Additional Reading

This is a dense topic and a very active research field.