Chapter 19
Patterns of External Validity

Background Generalizability depends on the quality of the prediction model as developed for the development setting (internal validity), and on characteristics of the population where the model is applied (validity of regression coefficients and distribution of predictor values). The general framework of validity of predictions was discussed in Chap. 17 (see in particular Fig. 17.1). Here, we first consider a number of typical situations that we may encounter when a prediction model is applied in an external setting. Theoretical relationships are illustrated with a large sample simulation and findings in some case studies. Approximate power calculations are given for tests of invalidity of a prediction model.

19.1 Determinants of External Validity

We concentrate on the external validity of predictions for a binary outcome Y. We consider a number of differences between populations that determine this external validity, related to case-mix and regression coefficients β (Table 19.1).

19.1.1 Case-Mix

With case-mix we refer to the distribution of predictors X that are included in the regression model $Y \sim X$, as well as the distribution of predictors that are not included in the model, either observable or unobservable. Predictors not included in the model are referred to as "missed predictors," despite the fact that some may in fact be observable. Since the linear predictor (lp) is a linear function of the predictors X, we will for simplicity consider one predictor "x" in the model $Y \sim x$. Here, x represents a linear combination of X. Similarly, the missed predictors Z are represented as one variable "z" in the regression model $Y \sim x + z$.

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_19, © Springer Science+Business Media, LLC 2009

Table 19.1 Differences between populations that determine external validity

Scenario	Characteristic	Differences	Example
Case-mix	Distribution of observed predictors ("*X*")	Different selection, e.g. more-severe patients are selected; or inclusion criteria smaller/wider	Validation in referral centre; validation in/outside RCT
	Distribution of missed predictors ("*Z*")	Different selection based on predictors not considered in the model	Validation in different setting
	Distribution of outcomes ("*Y*")	Retrospective sampling of cases and controls	Case-control design
Coefficients	Coefficients β smaller than expected	Overfitted model is validated	Validation of model from small development sample
	Coefficients β different	Truly different population	Validation in different setting

19.1.2 Differences in Case-Mix

A different case-mix may be encountered because the setting differs compared with the development situation; e.g. model development in secondary care and validation in a primary or tertiary care setting. Or a model was developed in patients participating in a randomized controlled trial (RCT) and is applied in a less selected population. Such situations make that the distribution of observed predictors *X* is different between development and validation setting. The distribution of missed predictors *Z* may also differ when we apply a model in a different setting; per definition, such differences cannot be excluded a priori. Missed predictors *Z* may be fully independent of *X*, or be correlated. Finally, the design of a study may cause differences in incidence of the outcome *Y*, and may hence influence case-mix indirectly. For example, a case-control design can be followed, where the ratio of cases to controls is different than in the population.

19.1.3 Differences in Regression Coefficients

Regression coefficients β can be different between settings because of true differences between populations. Various reasons can be thought of, including definitions of predictors, the definition of the outcome, and a different selection of patients.

In practice, the coefficients β are not known for the development setting, but only estimated from a finite sample size. The same holds for a validation sample from a validation setting. This makes it impossible that the same regression coefficients are found when a regression model is re-estimated in a validation sample. Even if the underlying true coefficients are identical, some of the re-estimated coefficients will be larger and some smaller than in the development sample.

Another problem is that regression coefficients may on average have been estimated too large because of overfitting in the development data set. Such overfitting is most likely for models developed in small data sets with a relatively large number of (candidate) predictors (see e.g. Chap. 5, 11, and 13). Shrinkage of coefficients at model development should have prevented overestimation of coefficients for predictive purposes, but this is not the case for many currently developed models.

19.2 Impact on Calibration, Discrimination, and Clinical Usefulness

In the following we will consider various scenarios for differences between populations (Table 19.1). We will study the impact of these differences on calibration, discrimination, and clinical usefulness of prediction models for binary outcomes. We simulate an outcome *y*, which depends on *x* and a missed predictor *z* (both with standard normal distribution). In the development population, we estimate a logistic regression model with an intercept α_0 and coefficient β_1 for *x*, while in fact the outcome *y* is determined by *x* and *z*. The missed predictor *z* and *x* are uncorrelated, weakly correlated, or moderately correlated (Pearson correlation coefficients *r*, 0, 0.33, 0.5, Table 19.2 and Fig. 19.1). Findings are summarized in Table 19.3.

Table 19.2 Design of simulations with predictor *x* and missed predictor *z*, for a logistic regression model *Y* ~ *x* + *z* (adjusted analysis) and *Y* ~ *x* (unadjusted analysis)

Correlation $x - z$	Adjusted coefficients	Unadjusted coefficient
Pearson $r=0$, $r^2=0\%$	2.05**x* + 1.5**z*	1.5**x*
Pearson $r=0.33$, $r^2=11\%$	1.5**x* + 1.5**z*	1.5**x*
Pearson $r=0.5$, $r^2=25\%$	1.18**x* + 1.5**z*	1.5**x*

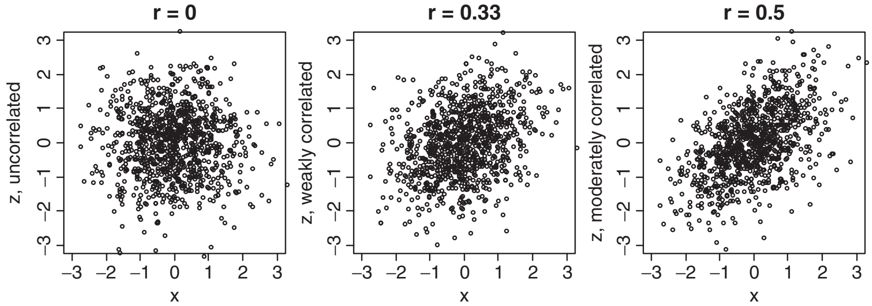

Fig. 19.1 Correlation between *x* (represented in the linear predictor) and *z* (a missed predictor), with or without correlation. Illustration for $n=1{,}000$; $n=500{,}000$ in further simulations

Table 19.3 Patterns of invalidity for a prediction model for binary outcomes in relation to differences between development and validation populations

Scenario	Characteristics	Differences	$a\|b=1$	b	c stat	NB
Development setting	$y = 1.5*x$ ($x \sim N(0,1)$)	–	0	1	0.81	0.055
Case-mix in validation setting	Distribution of observed predictors ("x")	More-severe patients	0	1	0.77	0.006
		Less-severe patients	0	1	0.77	0.104
		More heterogeneous	0	1	0.90	0.104
		Less heterogeneous	0	1	0.75	0.030
	Distribution of missed predictors ("z")	More-severe patients[a]	0.70	1	0.81	0.001
		Less-severe patients[a]	−0.70	1	0.81	0.109
		More heterogeneous[a]	0	1	0.83	0.062
		Less heterogeneous[a]	0	1	0.81	0.053
	Distribution of outcomes ("y")	2 times more cases	log(2)	1	0.81	NA
		2 times less cases	−log(2)	1	0.81	NA
Coefficients in validation setting	Coefficients β smaller than expected	Slope 0.8	0	0.8	0.77	0.037
		Slope 0.6	0	0.6	0.72	0.014
	Coefficients β different	X effects * 0.5 or 1.5	0	0.84	0.78	0.040
		X effects * 0.25 or 1.5	0	0.68	0.74	0.023

[a] For correlation $x - z$ of 0.33; detailed results in Figs. 19.5 and 19.6; $a|b=1$, intercept given that calibration slope is 1, indicating "calibration-in-the-large"; b, calibration slope; c stat, c statistic to indicate discriminative ability; NB, net benefit; NA, not applicable

19.2.1 Simulation Set-Up

We create a validation population where we apply the developed model. Various differences are simulated for the validation population compared with the development population. We first consider populations ($n=500{,}000$) and later samples of smaller size to illustrate sampling variability and statistical power. We consider a scenario inspired by the testicular cancer case study, with average incidence of tumor close to 50%, and a decision threshold for the probability of tumor of 30% (Chaps. 15 and 16). We consider a good discriminating model, with a c statistic of 0.81. This c statistic is achieved with a logistic regression model with a single predictor x, with x normally distributed and regression coefficient β, 1.5. We can hence define the linear predictor "lp" as $\text{lp}=1.5*x$.

We generate the outcome y with inclusion of the missed predictor z (uncorrelated or correlated). In the underlying model the lp is a function of x and z. With uncorrelated $x - z$, we define the lp as $\text{lp}=2.05*x + 1.5*z$. The adjusted regression coefficient for x is 2.05 rather than 1.5. This may be surprising, but is related to the "stratification" or "conditioning" effect in non-linear models such as logistic regression and Cox regression models. In such models, adjusted effects are more extreme than unadjusted effects when a covariate is considered that is related to the outcome, but uncorrelated to other covariates. This is well known in the analysis of randomized clinical trials (see Chaps. 2 and 22).[133,182,348,403] In unadjusted analysis, the coefficient for x is 1.5 (Table 19.2).

For moderately correlated $x - z$ data ($r=0.5$), we define the lp as lp=1.18*x + 1.5*z. Now the adjusted regression coefficient is 1.18 rather than 1.5, which is caused by the positive correlation between x and z. The confounding effect of this correlation was stronger than the stratification effect. In unadjusted analysis, the coefficient for x is again 1.5. An intermediate situation was identified by trial and error, where the correlation between x and z was 0.33, such that the negative effect of confounding and positive effect of stratification on z are exactly balanced in the adjusted analysis. The model is lp=1.5*x + 1.5*z.

In both development and validation settings, we study predictions only in relation to x, since z is a missed predictor. The observed relation is lp=1.5*x at development, with a c statistic of 0.81. At validation we hope to see y=0+1*lp in a logistic regression model (see Chap. 15 and 20 for more background on this calibration model).[86] This relation between y and lp may be influenced by changes in the distribution of the x, z, or y, or differences in the regression coefficients that determine the lp (see Table 19.1).

19.2.2 Performance Measures

In the following, we concentrate on a limited number of indicators of calibration, discrimination, and clinical usefulness, although many more performance measures can be considered for validation of predictions for binary outcomes (see Chaps. 15 and 16). For calibration we consider calibration-in-the-large (intercept given that slope b is set to 1, $a|b=1$) and the calibration slope (b). Both are determined in logistic regression models: y ~ lp. The linear predictor lp is entered as an offset variable ($a|b=1$), or as the only predictor (to estimate slope b) in a logistic regression model estimated in the validation data. The c statistic is used to indicate discriminative ability (Chap. 15). For clinical usefulness, we calculate the net benefit (NB), with the formula NB = (TP – w FP) / N, where TP is the number of true-positive classifications, FP the number of false-positive classifications, and w is a weight equal to the odds of the threshold (cutoff/(1 – cutoff)), or the ratio of harm to benefit (see Chap. 16). We compare the NB of the model with a cutoff at 30% with the strategy with the next best NB ("treat all" or "treat none"). With an incidence of the outcome at 50% and a threshold of 30%, "treat all" has the next best NB: for every 100 patients, 50 true-positive classifications are made, and 50 false-positive classifications (which are weighted as 3/7). The NB of treat all hence is 50 – 3/7 * 50 = 28.6 /100 patients. A clinically useful model should have an NB higher than this refe-rence value.

When the considered model is applied in the development population, the calibration is perfect ($a|b$=1 = 0; slope $b=1$) and discrimination good (c = 0.81, Fig. 19.2). The increase in NB by 0.055 means that 5.5 more true-positive classifications are obtained per 100 patients, at the same number of false-positive classifications (see Chap. 16). The model performance is identical whether uncorrelated or correlated missed predictors are present.

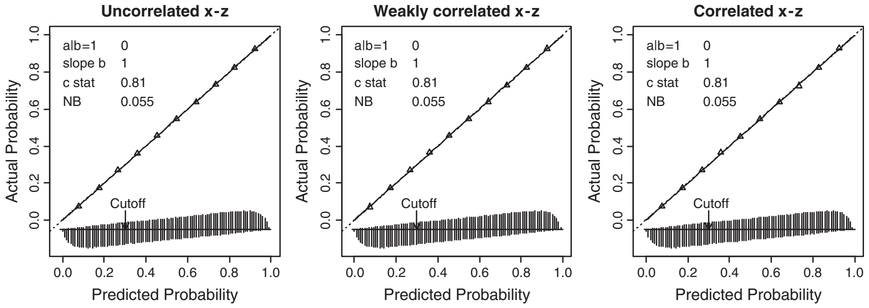

Fig. 19.2 Calibration, discrimination, and clinical usefulness when the prediction model is applied in a population with identical distribution of predictors *x* and missed predictor *z* (from left to right: $r=0$, 0.33, 0.5). $a|b=1$, intercept given slope *b* is 1; slope *b*, calibration slope in a model *y* ~ lp; *c* stat, *c* statistic indicating discriminative ability; NB, net benefit compared with "treat all." The value of 0.055 means that 5.5 more true positive decisions are taken per 100 patients, at the same number of false-positive decisions (see Chap. 16). Triangles represent deciles of patients grouped by similar predicted probability. The distribution of patients is indicated with spikes at the *bottom* of the graph, separately for those with and without the outcome

19.3 Distribution of Predictors

We consider various selection mechanisms based on observed predictors *X* and missed predictors *Z*. Such selection is an example of missing not at random (MNAR, Chap. 7). We already know that regression coefficients of a predictor *X* remain unbiased with an MNAR mechanism. Hence, calibration is expected to remain unaffected. Of interest is any influence on discrimination and clinical usefulness.

19.3.1 More- or Less-Severe Case-Mix According to X

Subjects may be more likely to be included in the validation setting because they have higher *X* values ("more suspect cases"). For example, we may assume that only the higher *X* values are represented (correlation with missingness, 0.62; R^2, 39%). This leads to a more-severe case-mix at validation.

```
n  <- 500000
x  <- rnorm(n)                           #standard normal x
xM <- ifelse(rnorm(n=n, sd=.8)<x,x,NA)   #50% missing, r=0.62
```

In our particular simulation, the more-severe case-mix is associated with somewhat less spread in predictions, and hence a lower *c* statistic (0.77 instead of 0.81, Fig. 19.3 left panel). Moreover, only few patients have predictions below the postulated threshold of 30%, reducing the NB to 0.006 instead of 0.055. The prediction model would be judged of very limited use in this validation setting. If the missingness was even more selective ($r>0.75$), the NB would become zero, meaning

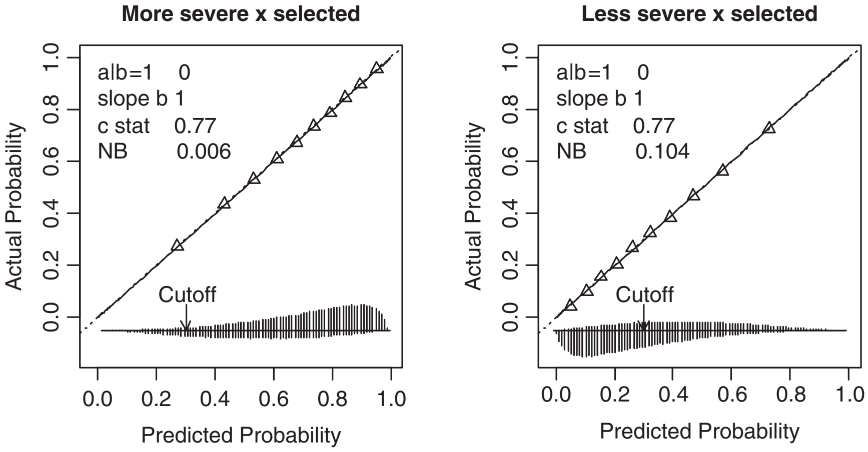

Fig. 19.3 Influence of selection of more- or less-severe cases according to observed predictor values ("*x*"). 50% of the subjects were selected, with higher or lower likelihood of selection with higher *x* values. Validation with a less-severe case-mix makes the prediction model clinically more useful (*right panel*)

that "treat all" would be as good a policy as using the model. In contrast, a less-severe case-mix led to a higher NB (NB, 0.104; Fig. 19.3, right panel). These patterns were identical with uncorrelated or correlated *z*.

*19.3.2 Example: Interpretation of Testicular Cancer Validation

These findings are important for the interpretation of the external validity of the testicular cancer example presented in Chaps. 15 and 16. When applied in more-severe patients treated at a tertiary referral centre (Indiana University Medical Center), we noted a decrease in clinical usefulness of the prediction model. But we have to realize that not all testicular cancers undergo surgical resection; there is "verification bias."[30] Typically a selection is made towards those with a suspicion of residual tumor (e.g. larger residual masses). If all testicular cancer patients were considered, the model would also indicate resection in some patients who are not candidates for resection under current policies. Clinical usefulness would hence be judged higher.

19.3.3 More or Less Heterogeneous Case-Mix According to X

Another situation is that a more heterogeneous setting is considered, which is fully represented by the *X* values. For example, inclusion criteria may be wider in surveys of patients with traumatic brain injury (TBI) compared with randomized controlled trials (RCTs). RCTs typically have a list of inclusion and exclusion criteria. If these criteria apply to predictors that are all considered in the prediction model, the distribution of X values will be more heterogeneous in surveys. Note that the selection on X values may lead to extrapolation of model predictions in the validation data beyond observed X values in the development data.

The heterogeneity in case-mix translates into a higher discriminative ability; we can distinguish more patients with very low or very high prediction (c statistic, 0.90 instead of 0.81, Fig. 19.4, left panel). More patients have predictions below the postulated threshold of 30%, doubling the NB (0.104 instead of 0.055). The prediction model would be judged quite useful in this more heterogeneous validation setting. The reverse is found for validation in a setting with less heterogeneity (lower c statistic, 0.75; lower NB, 0.03, Fig. 19.4, right panel). These patterns were identical with uncorrelated or correlated z.

19.3.4 More- or Less-Severe Case-Mix According to Z

Similar to distributions of observed predictors, distributions of missed predictors Z may also differ between development and validation settings. We will see that the correlation between observed predictors X and missed predictors Z is especially relevant for calibration.

The first situation is that a prediction model is applied in a setting of more- or less-severe cases, according to predictors that are not (or not fully) captured in the prediction model. A more-severe case-mix mainly causes a systematic miscalibration of predictions (Fig. 19.5, top row). The calibration-in-the-large ($a|b$=1) values are around 0.7, which reflects that approximately twice as many cases are found

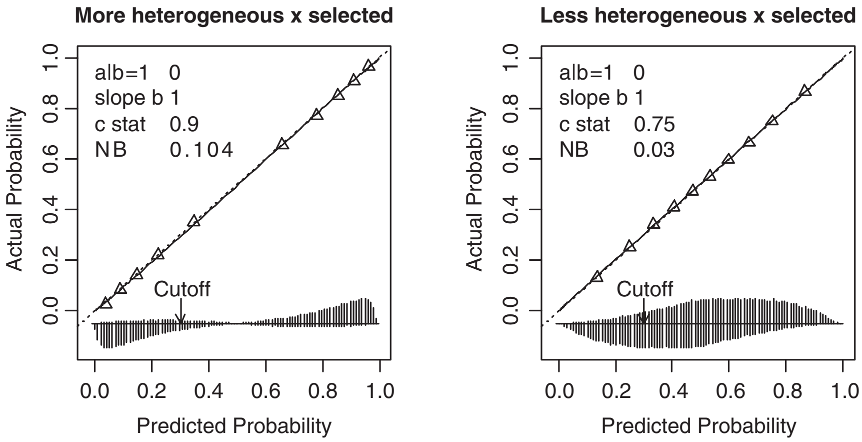

Fig. 19.4 Influence of selection of more or less heterogeneous cases according to observed predictor values for "x." Approximately 35% of the subjects were selected, with higher or lower likelihood of selection with more extreme x values. Validation with a more heterogeneous case-mix makes the prediction model more discriminatory and clinically more useful (*left panel*)

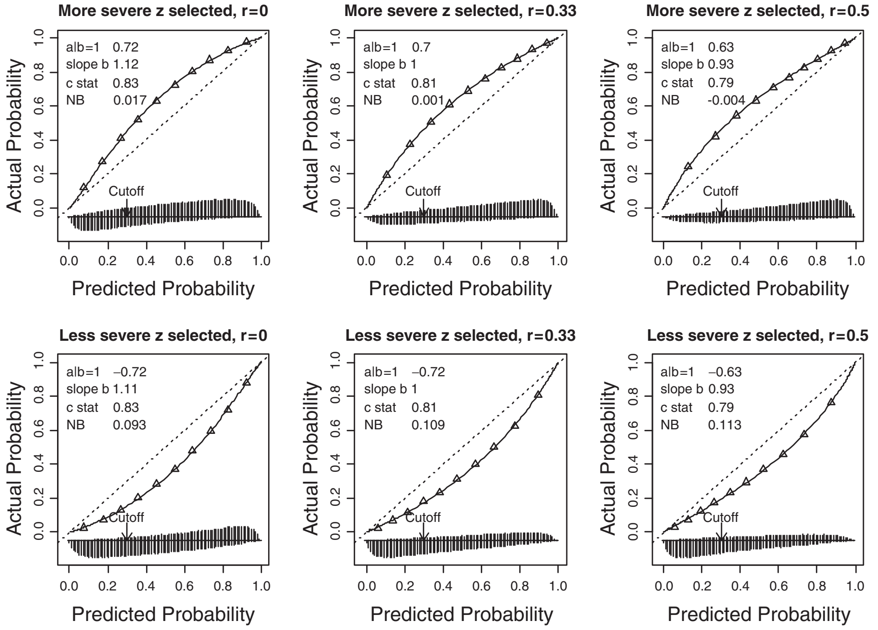

Fig. 19.5 Influence of selection of more- or less-severe cases according to a missed predictor ("*z*," *x* – *z* correlation, 0, 0.33, or 0.5). 50% of the subjects were selected, with higher or lower likelihood of selection with higher *z* values

than predicted (odds ratio, exp(0.7) = 2.0). The calibration slope is around 1. Without correlation between *x* and *z* ($r=0$, Fig. 19.5, upper left panel), the slope is 1.12, which is explained by the reduced stratification effect of *z* in the regression model. In the development setting, the stratification effect was such that the adjusted coefficient was 2.05 for an unadjusted coefficient of 1.5 for *x*; with less stratification, the unadjusted coefficient is 1.12*1.5 = 1.68. With moderate correlation (r=0.5, Fig. 19.5, upper right panel), the confouding effect is weaker, leading to an unadjusted coefficient of 0.93*1.5 = 1.4 for *x*.

The discrimination follows the same pattern as the calibration slope, with values around the original estimate of 0.81. The poor calibration causes the model to have at most small clinical usefulness. The NB of the model may even become negative (–0.004 in Fig. 19.5, upper right panel). This means that worse decisions are made with the model than the reference strategy of "treat all." This can be understood by realizing that the model assigns patients with a prediction under 30% to "no treatment," while predictions are systematically miscalibrated. Hence, many among those with a prediction under 30% have actual probabilities over 30% and should have been classified for "treat." On balance, the loss of inappropriately withholding treatment from those with actual probabilities over 30% was larger than the gain of reducing false-positive classifications (100% with a "treat all" strategy).

The reverse pattern is noted when selection is on less-severe patients according to some missed predictor (Fig. 19.5, second row). Calibration-in-the-large is the

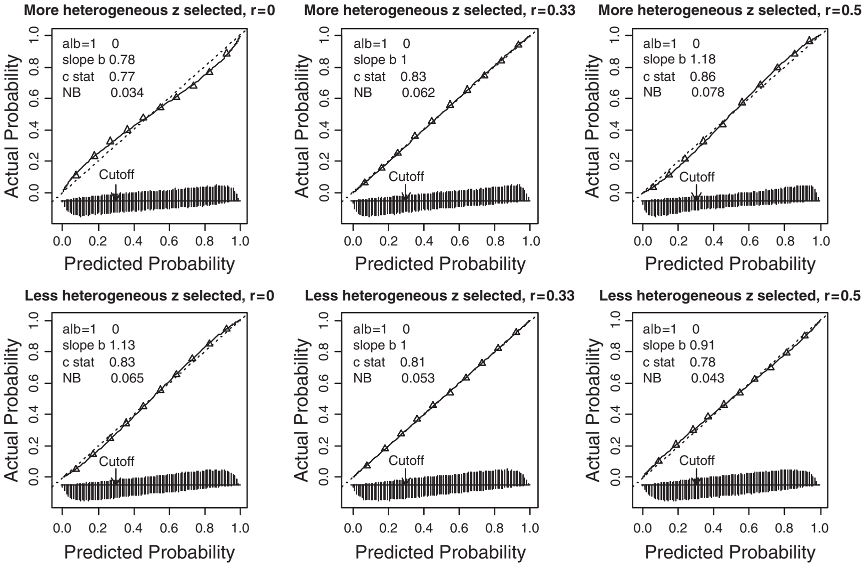

Fig. 19.6 Influence of selection of more or less heterogeneous cases according to a missed predictor ("*z*," *x* – *z* correlation, 0, 0.33 or 0.5). Approximately 35% of the subjects were selected, with higher or lower likelihood of selection with more extreme *z* values

main problem. But interestingly, the clinical usefulness is now increased, despite this miscalibration.

19.3.5 More or Less Heterogeneous Case-Mix According to Z

Similar to observed predictors, we can imagine that missed predictors may have a more or less heterogeneous distribution in a validation setting. Such distributional changes affect the calibration slope, but not calibration-in-the-large (Fig. 19.6). The specific patterns can again be explained by the magnitude of stratification and confounding effects. Discrimination and clinical usefulness were better with higher calibration slopes.

19.4 Distribution of Observed Outcomes *Y*

A case–control design allows for separate sampling of cases ($y=1$) and controls (y=0). Cases and controls should come from the same underlying populations as would be considered in a cohort study (Chap. 3). In our examples, the ratio of cases and controls was 1:1 (50% incidence of the outcome *Y*). The effect of manipulating the outcome incidence is reflected in calibration-in-the-large. With a ratio of 2 cases to 1 control, the odds ratio of the intercept is 2. Indeed, the coefficient is 0.69,

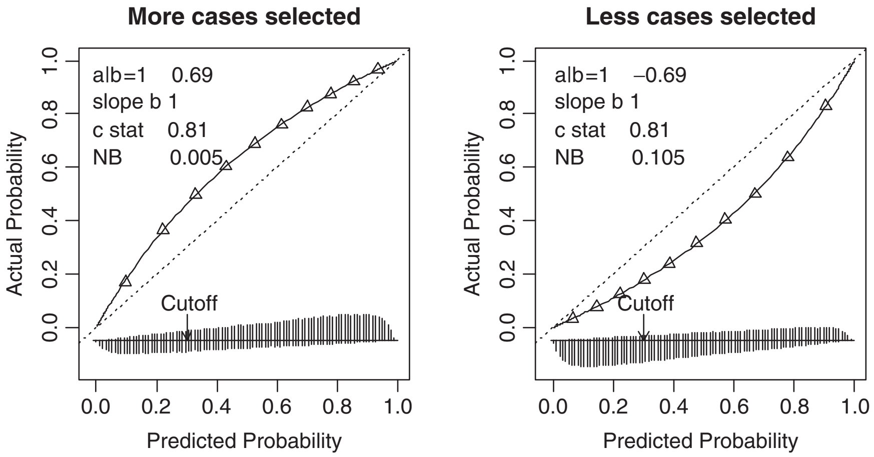

Fig. 19.7 Influence of a case–control design on the model intercept; calibration slope and discrimination remain unaffected. The ratio of cases to controls was set to 2:1 (*left*) and 1:2 (*right panel*)

or log(2) (Fig. 19.7, left panel). Conversely, a ratio of 1 case to 2 controls leads to an intercept of –0.69. With a proper case–control design, the effects of predictors remain identical (calibration slope = 1), as well as the *c* statistic (0.81). Calculation of clinical usefulness is only sensible after correction of the intercept, which can be seen as translating a case–control design back to clinical practice.

In a traditional case–control design, the number of controls is unknown. This makes it impossible to correctly adjust the intercept. In a *nested* case–control design, we sample the cases and controls from a defined underlying cohort. The number of controls is known in such a design, which makes it straightforward to adjust the intercept, for example by weighting the controls by the inverse of their sampling ratio.

19.5 Coefficients β

19.5.1 Coefficient of Linear Predictor < 1

Overfitting is a major problem of predictive modelling (Chaps. 4–18). At external validation, we may often find less predictive effect of the linear predictor lp. This reduced effect might have been detected already at internal validation, and might have led to incorporation of a shrinkage factor to compensate for overfitting. True differences in predictive effects may also play a role, for example caused by definition and selection issues.

A typical shrinkage factor found at internal validation is 0.8; more-severe overfitting might lead to a shrinkage factor of 0.6. At external validation, we find that such patterns of overfitting lead to a reduction in discriminative ability (*c*, 0.77 or 0.72 instead of 0.81) and a reduction in clinical usefulness (NB, 0.037 or 0.014 instead of 0.055, Fig. 19.8).

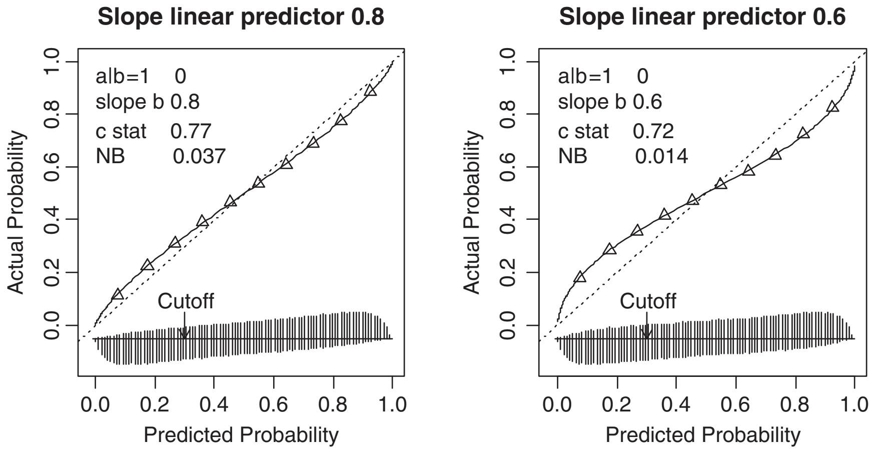

Fig. 19.8 Influence of overfitting in model development. The slope of the linear predictor is 0.8 or 0.6, with lower discriminative ability (c=0.77 or 0.72), and lower clinical usefulness (net benefit, 0.037 or 0.014)

19.5.2 Coefficients Different

In addition to being overestimated on average, regression coefficients may truly differ between development setting and validation setting. Various causes can be imagined, all related to the validation population not being "plausibly related" anymore to the development population.[222] Terrin et al. considered various scenarios of different effects of predictors in a validation setting. In simulation studies, they simulated weaker effects of predictors, motivated by clinical scenarios, and found reductions in c statistic from 0.75 to 0.72.[431]

In Chap. 5, we used an arbitrary example of differences in predictor effects, with half of the predictors having 0.5 and half having 1.5 times the effect of the development setting. We use this example here for illustration, and a more extreme situation, with half of the predictors having a very small effect at validation (0.25).

**19.5.3 R Code*

The programming code may help to understand how simulations were performed. First 10 x variables were created, with decreasing standard deviation:

```
n    <- 500000
x1   <- rnorm(n,sd=1)
x2   <- rnorm(n,sd=.9)
…
x10  <- rnorm(n,sd=.1)
```

For the development setting, we assume that each x has a coefficient of 1; but in the two validation settings these weights are different.

```
#development data
xsum   <-x1+x2+x3+x4+x5+x6+x7+x8+x9+x10
#validation data:2 scenarios
xval   <-.50*x1+1.5*x2+.50*x3+1.5*x4+…+.50*x9+1.5*x10
xval2  <-.25*x1+1.5*x2+.25*x3+1.5*x4+…+.25*x9+1.5*x10
```

Logistic regression models were constructed with the `xsum`, `xval`, and `xval2` variables. When the latter[2] variables are multiplied by 0.76, the *c* statistics are 0.81 for both models. We validate predictions from the model with `xsum` as predictor in settings where 0.76*`xval` or 0.76*`xval2` is the true linear predictor determining outcome.

19.5.4 Influence of Different Coefficients

Calibration-in-the-large may remain unaffected when predictive effects are different (Fig. 19.9). However, the calibration slope was smaller than 1. When effects in the validation setting remained close to the effects at development, the slope was 0.84, and discrimination was slightly decreased (0.78 instead of 0.81). When differences in coefficients were more substantial (Fig. 19.9, right panel), the calibration slope was 0.68, the *c* statistic 0.74, and clinical usefulness smaller (0.023 instead of 0.055). Hence, differences between effects in the development setting vs. the validation setting may seriously deteriorate model performance.

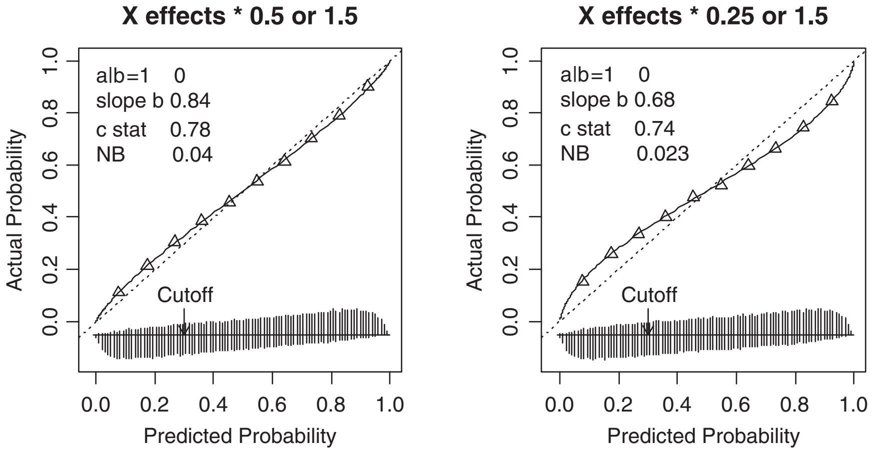

Fig. 19.9 Influence of differences in regression coefficients between development and validation setting. Regression coefficients were 0.5 or 1.5 times as large in the *left panel*, and 0.25 or 1.5 times as large in the right panel. In the right panel, miscalibration was severe (slope, 0.68), discriminative ability and clinical usefulness modest (*c* statistic, 0.74; net benefit, 0.023)

*19.5.5 Other Scenarios of Invalidity

Thus far, we considered one element at a time for differences between development and validation populations. All simulation results depend on the specific parameters chosen; with more extreme parameters, differences will be larger. We can consider other scenarios that are also plausible in medical research, where we combine differences in distribution of x, z, and regression coefficients (Table 19.4). Detailed results are provided at the book's Web site.

19.5.6 Summary of Patterns of Invalidity

- Calibration

  In the development setting, the calibration was perfect, the c statistic 0.81 and the NB of applying the model 0.055. Calibration remained perfect when the validation setting consisted of more- or less-severe patients according to predictor values, or more or less heterogeneous patients according to observed or missed predictor values. Calibration can be systematically disturbed by a more- or less-severe distribution of missed predictor values (z, e.g. intercept +0.70 or −0.70). A similar disturbance can be caused by a case–control design; however, the case:control ratio is under the influence of the researcher, while the distribution of a missed predictor usually is not. Calibration can also be affected by overfitting at model development (e.g. slope, 0.8 or 0.6), or truly differential predictive effects (coefficients of individual predictors 0.25 / 0.5 / 1.5 times as large).

- Discrimination

  Discriminative ability is related to the calibration slope, with a lower c statistic associated with a lower calibration slope. Another reason for a lower c statistic is a less heterogeneous case-mix (e.g., slope = 1, but $c = 0.75$ instead of 0.81, Fig. 19.4, right panel). A high c statistic such as 0.90 was found for a more heterogeneous

Table 19.4 Combinations of differences between development and validation populations and their impact on validity of a prediction model for binary outcomes

Scenario	x	z	Coefficients	$a\|b=1$	b	c stat	NB
Change of setting	–	More severe	X effects * 0.5 or 1.5	0.67	0.87	0.78	−0.008
	–	Less severe		−0.67	0.87	0.78	0.098
RCT vs. survey	More heterogeneous	More severe	X effects * 0.5 or 1.5	0.64	1.04	0.88	0.027
	Less heterogeneous	More severe		0.69	0.59	0.65	−0.036
	More heterogeneous	Less severe		−0.68	1.03	0.88	0.167
	Less heterogeneous	Less severe		−0.68	0.59	0.65	0.037

setting (Fig. 19.4, left panel). More heterogeneity in missed predictors had only small effects (Fig. 19.6). These examples illustrate that discrimination is determined by validity of estimated regression coefficients β and case-mix. Poor discrimination can hence result from both aspects (i.e. poor calibration and/or relatively homogeneous case-mix).

- Clinical usefulness

Tables 19.3 and 19.4 highlight the importance of calibration for clinical usefulness. A systematic miscalibration, e.g. caused by a more-severe case-mix according to a missed predictor z, may lead to a model without clinical usefulness. With incorrect calibration, we can even make systematically wrong decisions. This is not the case if predictions are well calibrated. Discrimination and calibration slope are linked, with a low calibration slope or low discrimination both associated with a low clinical usefulness.

Perfect calibration and good discrimination do not guarantee clinical usefulness. Discrimination is important; better discrimination may lead to better decision making. If a model has no discriminative ability, it cannot be clinically useful. Discrimination is hence a *necessary but not sufficient* condition for clinical usefulness.

When applying the model in more- or less-severe patients, the c statistic was 0.77 for both settings, but clinical usefulness was 0.006 for a more-severe setting and 0.104 for a less-severe setting. These findings are in line with the lack of clinical usefulness of the testicular cancer case study in Chap. 16, where we noted that few patients had a prediction above the threshold of 70% for the probability of benign tissue at external validation.

Case-mix is also very relevant. The case-mix in observed predictors (X) affects clinical usefulness through the distribution of predictions around the decision threshold, while leaving calibration largely intact. The case-mix in missed predictors (Z) may predominantly affect clinical usefulness through poor calibration-in-the-large.

19.6 Reference Values for Performance

The distribution of predictors X can be taken into account in the calculation of "reference values" for model performance. Reference values indicate a model's performance under the condition that the model predictions are valid in the validation sample. For a regression model this means that the regression coefficients for predictors X and the model intercept are fully correct for the validation setting. Such reference values may be very useful to obtain insight in what is happening at validation: Are there differences in case-mix or differences in regression coefficients compared with the development setting?

19.6.1 Calculation of Reference Values

For calibration, obvious reference values are 0 for calibration-in-the-large, and 1 for the calibration slope. For discrimination, we noted that the c statistic can vary with case-mix. Similarly, overall performance measures such as R^2 and Brier score depend on the case-mix in observed predictors *X*.[202]

A practical approach is to simulate the outcome *Y* for the observed case-mix in *X*, given that the prediction model is correct. This is simply obtained by first calculating the predictions for each subject in the validation data, and subsequently randomly assigning an outcome *Y* based on this prediction. With at least 100 repetitions for each patient, a stable estimate of the reference values is obtained. We illustrate the calculation below for 1,000 repetitions per patient in a logistic regression model.

*19.6.2 R Code

```
# fit in development data
fit          <- lrm(y~x1 + x2, data=dev.data)
# linear predictor for validation data
lp           <- predict(fit, newdata=val.data)
# External validation
val.prob(logit=lp, y=val.data$y, …)
# Start simulation of outcomes
n            <- nrow(val.data)
nsamples     <- 1000 # for stable results
perf.m       <- matrix(nrow=nsamples*n, ncol=2)
perf.m[,1]   <- rep(lp, nsamples) # repeat lp nsamples times
# Generate y for validation data
perf.m[,2]   <- ifelse(runif(length(perf.m[,1])) <=
                plogis(perf.m[,1]), 1, 0)
# Determine reference values
val.prob(logit=perf.m[,1],y = perf.m[,2], … )
```

19.6.3 Performance with Refitting

Another type of reference value is the performance obtained by refitting the model in the validation data. The regression coefficients are then optimal for the validation data, and hence provide an upper bound for the performance, which would be obtained if the coefficients from the development setting were exactly equal to those in the validation setting. However, this upper bound does not only depend on case-mix, but also on the effects of predictors in the validation setting. It is hence not simple to compare performance between development and validation settings: Differences may be attributable to both case-mix and/or coefficients.

Table 19.5 Examples of reference values for performance of two prediction models, developed in one setting and applied in another setting

Example	Measure	Apparent	Internally validated	Externally validated	Reference	Refitted
Testicular cancer	*c* stat	0.818	0.812	0.785	0.824	0.819
	R^2	38.9%	37.4%	26.7%	37.0%	34.2%
	Brier	0.174	0.178	0.161	0.144	0.147
Traumatic brain injury	*c* stat	0.767	0.765	0.816	0.804	0.819
	R^2	27.9%	27.3%	37.1%	35.3%	38.0%
	Brier	0.186	0.188	0.180	0.181	0.168

*19.6.4 Examples: Testicular Cancer and TBI

We apply the calculation of reference values to the testicular cancer and traumatic brain injury (TBI) case studies (Table 19.5). The apparent performance is calculated for n=544 testicular cancer patients (n=245 (45%) with benign histology) and n=2,036 TBI patients (n=798 (39%) with unfavorable 6-month outcome). The 544 testicular cancer patients are mostly from secondary care centres,[417] while validation was done in 273 patients from a tertiairy care centre (Indiana). A benign outcome was less frequent among these patients (n=76/273, 28%).[466] The 2,036 TBI patients were from the Tirilazad randomized controlled trials,[203] with validation in three largely unselected series (UK 4 centre study, European Brain Injury Consortium survey, Traumatic Coma Databank, n=2,090).[271] These patients more often had an unfavourable outcome at 6 months (n=1,249/2,090, 60%) compared with the development sample.

In the testicular cancer case study, the apparent *c* statistic was 0.818, with 0.006 optimism according to a bootstrap procedure. At external validation, the *c* statistic was 0.785, while 0.824 was expected based on the case-mix of the predictor variables ("reference," Table 19.5). When the model was refitted, the performance was slightly lower than this reference value (*c* statistic, 0.819 vs. 0.824). A similar pattern was noted for the R^2 and Brier statistics. We might test the statistical significance of these differences in performance, but concentrate here on the point estimates.

In the TBI case study, the apparent *c* statistic was 0.767, with negligible optimism. Surprisingly, the *c* statistic was higher at external validation (0.816), while 0.804 was expected based on the case-mix of the predictor variables. When the model was refitted, the performance was also higher than the reference (*c* 0.819 vs. 0.804). A similar pattern was noted for the R^2 and Brier statistics.

The interpretation of Table 19.5 is a follows:

1. Internal validation corrects for the statistical problem of overfitting in the development setting; case-mix is unchanged

2. External validation tests the model in a sample from a new setting, where both case-mix and coefficients may be different than in the development sample
3. The reference performance corrects for the new case-mix according to predictor values in the validation sample, while keeping the coefficients at the values from the development setting (that would ideally have been corrected with shrinkage to correct for overfitting)
4. The refitted performance corrects for the new case-mix and estimates optimal regression coefficients in the validation sample

The poorer external performance of the testicular cancer model is not explained by case-mix, at least not in the distribution of observed predictor values, since the reference performance was very similar to that in the development sample. The poorer external validity should hence be attributed to differences in regression coefficients between the settings. The refitted performance was similar to the reference performance, indicating that the predictors had similar predictiveness in both settings when refitted.

The better external performance of the TBI model is partly explained by case-mix, since the reference performance was higher than in the development sample. The surprisingly good external validity should further be attributed to differences in regression coefficients between the settings; predictive effects were overall stronger in the validation setting (calibration slope, 1.08), in line with the even better refitted performance (refitted *c*, 0.819, Table 19.5).

19.7 Estimation of Performance

Thus far, we examined theoretical patterns of invalidity with very large simulated samples, which can be considered as populations. The testicular cancer and TBI case studies considered more limited sample sizes for model development and validation; differences in model performance might at least partly be attributed to chance. Performance parameters such as model intercept ($a|b=1$), calibration slope (b), *c* statistic, and measures of clinical usefulness are subject to sampling error in real life.

19.7.1 Uncertainty in Validation of Performance

We illustrate the empirical behaviour of measures for calibration and discrimination of logistic regression models. The prediction model is the same as before, with a linear predictor defined by ten normally distributed *x* variables, each with a regression coefficient of 0.76. The model has a *c* statistic of 0.812. We consider small to large sample sizes for model development (N_{dev} = 100–10,000) and for model validation (N_{val} = 100–10,000), with outcome incidence 50% or 10%. Simulations are first performed under the Null hypothesis, i.e. that both samples originate from the same underlying population (Table 19.6). Case-mix and regression coefficients

Table 19.6 Estimation of calibration and discrimination of logistic regression models in small to large sample sizes for model development and for model validation

Scenario	Events/N_{dev}	Events/N_{val}	$a\|b=1$	slope b	c statistic
Incidence 50%					
Large sizes	5,000/10,000	5,000/10,000	0 ± 0.03	1.00 ± 0.03	0.81 ± 0.004
Small development samples	50/100	5,000/10,000	0 ± 0.28	0.64 ± 0.15	0.77 ± 0.017
	100/200		0 ± 0.17	0.82 ± 0.13	0.79 ± 0.010
	250/500		0 ± 0.12	0.92 ± 0.09	0.80 ± 0.006
	500/1000		0 ± 0.08	0.95 ± 0.07	0.81 ± 0.005
	1,000/2,000		0 ± 0.06	0.97 ± 0.05	0.81 ± 0.004
Small validation samples	5,000/10,000	50/100	0 ± 0.24	1.06 ± 0.24	0.82 ± 0.043
		100/200	0 ± 0.16	1.03 ± 0.17	0.81 ± 0.030
		250/500	0 ± 0.11	1.01 ± 0.10	0.80 ± 0.018
		500/1,000	0 ± 0.08	1.00 ± 0.07	0.81 ± 0.014
		1,000/2,000	0 ± 0.06	1.00 ± 0.05	0.81 ± 0.009
Small development samples, half size validation	50/100	25/50	0 ± 0.52	0.71 ± 0.31	0.77 ± 0.070
	100/200	50/100	0 ± 0.34	0.83 ± 0.25	0.79 ± 0.048
	250/500	100/200	0 ± 0.20	0.95 ± 0.18	0.80 ± 0.030
	500/1,000	250/500	0 ± 0.13	0.98 ± 0.11	0.81 ± 0.018
	1,000/2,000	500/1,000	0 ± 0.10	0.99 ± 0.09	0.81 ± 0.014
Small development samples and equal size validation samples	50/100	50/100	0 ± 0.44	0.66 ± 0.23	0.77 ± 0.051
	75/150	75/150	0 ± 0.32	0.77 ± 0.22	0.78 ± 0.039
	100/200	100/200	0 ± 0.27	0.82 ± 0.19	0.79 ± 0.033
	175/350	175/350	0 ± 0.19	0.89 ± 0.15	0.80 ± 0.023
	250/500	250/500	0 ± 0.15	0.93 ± 0.13	0.80 ± 0.019
	500/1,000	500/1,000	0 ± 0.11	0.97 ± 0.09	0.81 ± 0.014
	1,000/2,000	1,000/2,000	0 ± 0.08	0.99 ± 0.07	0.81 ± 0.010
Incidence 10%					
Large sizes	1,000/10,000	1,000/10,000	0 ± 0.05	1.00 ± 0.05	0.83 ± 0.007
Selected combinations of development and validation sample sizes	50/500	50/500	0 ± 0.25	0.85 ± 0.18	0.81 ± 0.033
		100/1,000	0 ± 0.23	0.85 ± 0.15	0.81 ± 0.021
		200/2,000	0 ± 0.19	0.86 ± 0.14	0.81 ± 0.018
		1,000/10,000	0 ± 0.18	0.86 ± 0.14	0.81 ± 0.010
	100/1,000	50/500	0 ± 0.22	0.93 ± 0.17	0.82 ± 0.032
		100/1,000	0 ± 0.18	0.93 ± 0.13	0.82 ± 0.021
		200/2,000	0 ± 0.15	0.93 ± 0.11	0.82 ± 0.015
		1,000/10,000	0 ± 0.13	0.93 ± 0.11	0.82 ± 0.008
	200/2,000	50/500	0 ± 0.19	0.95 ± 0.15	0.82 ± 0.031
		100/1,000	0 ± 0.18	0.96 ± 0.13	0.82 ± 0.022
		200/2,000	0 ± 0.11	0.97 ± 0.10	0.83 ± 0.017
		1,000/10,000	0 ± 0.09	0.96 ± 0.07	0.82 ± 0.007
	1,000/10,000	50/500	0 ± 0.17	1.01 ± 0.15	0.83 ± 0.030
		100/1,000	0 ± 0.13	0.99 ± 0.10	0.83 ± 0.021
		200/2,000	0 ± 0.09	1.00 ± 0.07	0.83 ± 0.015

Numbers are mean ± standard error, as observed in simulations (100–1,000 repetitions for sufficiently stable results)

were hence identical in both settings, and estimates may only vary because of finite sample sizes at development and/or validation.

With 50% incidence of the outcome in very large development and validation sizes ($N_{\text{dev}}=10{,}000$ and $N_{\text{val}}=10{,}000$), the standard errors (SEs) are small: The SE around the calibration-in-the-large and calibration slope *b* is 0.03, around the *c* statistic 0.004. With 10% incidence, $N_{\text{dev}}=10{,}000$ and $N_{\text{val}}=10{,}000$, the SEs are larger, corresponding to the lower number of events (1,000 instead of 5,000).

We find that the calibration-in-the-large is 0 on average in all scenarios; the SE depends on the size of the development sample and the size of the validation sample. With only 100 subjects for model development, the SE is 0.28 if validation is in 10,000 subjects; if validation is in 50 or 100 subjects, the SE is much larger (± 0.52 and ± 0.44 respectively). A quite low SE (± 0.06) is found with $n=2{,}000$ for model development and 10,000 for model validation, or with a reversal of this design (development n=10,000, validation $n=2{,}000$).

The calibration slope is below 1 when small samples are used for model development (e.g. slope b=0.65 with $N_{\text{dev}}=100$ and $N_{\text{val}}=10{,}000$, reflecting clear overfitting and a need for shrinkage of coefficients). In contrast, small validation samples lead to an upward bias for the slope (e.g. slope $b=1.08$ with $N_{\text{dev}}=10{,}000$ and $N_{\text{val}}=100$). The SE is somewhat larger with small validation samples than with small development samples (e.g. $N_{\text{dev}}=100$: SE ± 0.15; N_{val}= 100: SE ± 0.25).

The discriminative ability (*c* statistic) was 0.81 in the population, but smaller with small development samples (e.g. c=0.77 with $N_{\text{dev}}=100$, $N_{\text{val}}=10{,}000$). Again small validation samples led to an upward bias (e.g. c=0.82 with $N_{\text{dev}}=10{,}000$ and $N_{\text{val}}=100$). The SE was markedly higher with small validation samples (e.g. $N_{\text{dev}}=100$: SE ± 0.017; $N_{\text{val}}=100$: SE ± 0.043). Apparently, small development samples lead to poor discriminating models, which can reliably be quantified with large validation samples, but small validation samples lead anyway to uncertain estimates of discrimination.

*19.7.2 Estimating Standard Errors in Validation Studies

In Table 19.6, we calculate SEs empirically by studying the distribution of coefficients over samples. We can also use the asymptotic SE for the performance measures. The SE of calibration-in-the-large and calibration slope can be obtained from the variance estimates in logistic regression models. The SE of the *c* statistic can be calculated with standard formulas for rank order statistics.[172] We found that the asymptotic SEs agreed rather well with the empirical estimates.

19.7.3 Summary Points

- Variability is substantial with small development samples, but especially with small validation samples

- The effective sample size is largely determined by the number of events rather than the total sample size
- SEs can be estimated with asymptotic formulas or from simulations ("empirically")

19.8 Design of External Validation Studies

The variability in performance has implications for the design and power of validation studies (see references for validation of linear regression models).[392,336] We have seen in Chap. 17 that the bootstrap is generally preferable for internal validation purposes. Despite its inefficiency, some researchers may like a split-sample approach to convince their readership. This design was discouraged in Chap. 17. A common ratio in such a design is 2/3 of the sample for model development and 1/3 for validation. According to Table 19.6, a lower variability of performance is obtained with a half–half split-sample design; but this design has more optimism in calibration slope and discrimination. A 2:1 ratio may be a reasonable balance between optimizing bias and variability.

For external validation we may well choose a temporal validation design.[222] But we then face the same question on how to choose the size of the development data set vs. the size of the more recent validation set. With spatial validation, e.g. "leave-one-centre-out" cross-validation, the validation sets may be much smaller than the development set. The results in Table 19.6 show that this makes the performance quite uncertain in each validation part per se.

Another situation is that a model was published, and we simply wish to externally validate this model for our setting. We set up a fully independent external validation study, and wonder about a reasonable sample size, accepting the developed model as reasonable to test. This design requires some estimates of power to detect relevant differences in performance.

19.8.1 Power of External Validation Studies

Power calculations depend on various quantities: Statistical Type I and Type II error; the variability in the quantity we want to test, and the "clinically relevant" difference we do not want to miss. Type I error is conventionally set at 5%, and type II error at 20% (power 80%). The variability of performance measures is shown in Table 19.6. Note that these are empirically derived SEs for one specific logistic regression model (with ten normally distributed predictors). In practice, we may only know the asymptotic (i.e. estimated) SE of some measures such as the model intercept. Clinically relevant differences may be context-dependent. For logistic regression models we might consider a systematic over or underestimation by 1.5 times the odds of the outcome (intercept + or – ln(1.5)), a calibration slope less than

0.8 (difference 0.2 with ideal slope of 1), and a decrease in *c* statistic by more than 0.05 (given the same case-mix).

Some specific issues come up in power calculations for validation studies. The first is whether we should perform one-sample or two-sample tests. If we consider the prediction model as a system generating predictions, a one-sample test is reasonable to test whether the validation performance deviates from hypothesized values. For calibration, these values are obvious: 0 for calibration-in-the-large and 1 for calibration slope. For the *c* statistic, we may consider the reference value given the case-mix in the validation setting (see Sect. 19.11). For the *c* statistic we might also consider a two-sample test, including uncertainty in the estimate from the development setting. A further issue is whether we should perform one-sided or two-sided tests. Calibration-in-the-large asks for a two-sided test, since the incidence in the validation setting may be higher or lower than predicted. But for calibration slope we could test for slope<1, rather than slope <> 1. Similarly, only a decrease in discrimination is an interesting alternative hypothesis.

Finally, one might argue that we should consider assessment of validity as a non-inferiority design. This implies that we change the Null hypothesis to stating that the model is invalid, and test whether the model performance is within reasonable limits from the expected value. The reasonable limits may be context dependent, similar to defining "clinically relevant" differences in traditional sample-size calculations.

*19.8.2 Required Sample Sizes for Validation Studies

We first approximate the power given the SE under the null hypothesis, i.e. the model was actually valid in both development and validation setting. We consider SEs for model development with a large sample size in Table 19.6, such that the predominant source of variability is the validation sample size. For simplicity we use one-sample tests for all measures. For calibration-in-the-large, we use a two-sided test; for calibration slope, a one-sided test (slope<1); for the *c* statistic, a one-sided test ($c < c_{\text{reference}}$). The critical values[1] for power calculations are determined by Type I and Type II error, which we set at 5% (one-sided or two-sided) and 20% (one-sided). The critical value is 1.96+0.84=2.80 for two-sided tests, and 1.64+0.84=2.49 for one-sided tests. We multiply these critical values with the SE to obtain the minimum differences that can be detected with 80% power (Table 19.7).

As expected, small validation sizes only have 80% power to detect substantial invalidity. For example, if we validate a model in a sample with 50 events and 50 non-events, we only have enough power to detect a calibration-in-the-large problem with twice too high, or twice too low predictions (odds ratio, 1.96); a dramatically poor calibration slope (less than 0.4), and a decrease in *c* statistic over 0.1 (Table 19.7). To detect a more modest calibration-in-the-large problem, such as 1.5 times too low or too high predictions, we would need at least 100 events and 100 non-events

[1] Critical value: the value that a test statistic must exceed for the null hypothesis to be rejected.

Table 19.7 Required sample size for 80% power when validating a logistic regression model in a setting with 50% or 10% incidence of the outcome

Scenario	Events/N_{val}	$a\|b$=1 <> 1[a]	slope b < 1[b]	$c_{validation} < c_{reference}$[c]
Incidence 50%	50/100	±0.67, OR=1.96	<0.40	<–0.107
	100/200	±0.45, OR=1.57	<0.58	<–0.077
	250/500	±0.31, OR=1.36	<0.75	<–0.045
	500/1,000	±0.22, OR=1.25	<0.83	<–0.035
	1,000/2,000	±0.17, OR=1.18	<0.88	<–0.022
Incidence 10%	50/500	±0.45, OR=1.61	<0.63	<–0.075
	100/1000	±0.34, OR=1.44	<0.75	<–0.052
	200/2000	±0.25, OR=1.29	<0.83	<–0.037

OR, Odds ratio

[a] Asymptotic SE and minimum OR that can be detected with 80% power

[b] Minimum slope that can be detected with 80% power

[c] Minimum differences in c statistic that can be detected with 80% power

Table 19.8 Power for slope < 1 (true value, 0.84) and c statistic decrease (true decrease from, 0.821 to 0.778, −0.043)

Scenario	Events/N_{val}	slope b 0.84	c statistic −0.043
Incidence 50%	50/100	15%	11%
	100/200	25%	24%
	250/500	50%	57%
	500/1,000	78%	87%
	1,000/2,000	97%	99%

(total sample size > 200). This sample size would also have 80% power for a slope less than 0.58, and a decrease in c by 0.077. With more non-events (incidence of outcome, 10%), the picture is slightly better in terms of number of events required, but the total sample size should be at least 1,000 (100 events) for reasonable power.

In a secondary analysis, we simulate power in the case that the prediction model is invalid. We create a model with coefficients 0.76 for ten normally distributed predictors x1 to x10, and validate in a setting where the coefficients are 0.5 or 1.5 times as large (see Fig. 19.9). In the validation setting, calibration-in-the-large is fine (average, 0), but the slope is 0.84 instead of 1, and the c statistic is 0.778 instead of 0.821 in the development setting (decrease, 0.043). From Table 19.7, we expect that the power for detecting that the slope is lower than 0.84 is slightly below 80% with 500 events; indeed we find 78% power with this sample size (Table 19.8). For a decrease in c statistic by –0.043, we expect that more than 250 events and 250 non-events are required; indeed the power is 57% with these numbers, and 87% with 500 events.

19.8.3 Summary Points

- The variability of external validation assessments depends on the size of the development sample and the size of the validation sample

- For statistical testing, we may accept the prediction model as given, and hence perform one-sample tests in the validation data
- For such tests to have reasonable power, we need at least 100 events and at least 100 non-events in external validation studies, but preferably more (>250 events). With lower numbers the uncertainty in performance measures is large.

19.9 Concluding Remarks

The performance of a prediction model in a new setting ("generalizability" or "transportability") essentially depends on two aspects: the validity of the regression coefficients, and the case-mix in the validation setting. The validity of regression coefficients can be assessed by comparing regression coefficients between settings. Indeed we note that many validation studies report on the coefficients in their sample and compare these to the previous estimates. With relatively small development and validation samples it would be highly coincidental if coefficients agreed well. Even if the two samples came from exactly the same underlying population, chance processes will cause the coefficients in both samples to differ from each other to some extent, with some coefficients larger and some smaller than expected from the development sample.

Differences in case-mix between development and validation setting are usually considered informally, by comparing patient characteristic in a kind of "Table 1." One usually makes only informal comparisons to the case-mix in the development sample. Some statistical measures have previously been proposed for a more formal assessment of comparability, such as the "M statistic" to compare trauma populations.[53] With this approach, survival probabilities of trauma patients are grouped, for example as 0–25%, 26–50%, 51–75%, 76–90%, 91–95%, and 96–100%. The fraction of patients in these groups at validation is compared with the fraction at model development. The smaller of the two fractions is summed over all groups. This creates a number ranging from 0–1. *M* values close to 1 indicate a perfect match with the development case-mix, while 0 indicates a total discrepancy between the two samples. An arbitrary cutoff point of 0.88 has been suggested, and studies with *M* values below 0.88 should be "interpreted cautiously."[53]

We followed a more systematic approach to study the influence of differences in case-mix. Differences in predictor distributions ("*X*") do not affect calibration, and only discrimination aspects, as long as the model is correctly specified for the range of *X* values examined. If non-linearities and/or interactions had been missed at model development, we can imagine that shifting to another predictor distribution may impact on calibration as well. Furthermore, we may assume that a very different distribution in *X* implies that differences in missed predictors ("*Z*") are also likely. Differences in missed predictors between settings may severely invalidate a prediction model, both with respect to calibration (especially calibration-in-the-large) and discrimination. When predictions are systematically miscalibrated, we can make systematically wrong decisions based on the model. This may lead to a negative NB of using the model, compared with a default policy without using the model. It is therefore important to perform external validation studies.[40]

We also noted that the distribution of predictors can formally be taken into account in the calculation of reference values for model performance, given that the model is valid in the validation sample. This may be very useful to obtain insight in what is happening at validation: differences in case-mix or differences in regression coefficients.

Finally, we studied design issues of validation studies for predictive regression models. If a temporal split is made, a 2:1 ratio may be reasonable. This limits overfitting at development, and still gives reasonable power at validation. A validation data set should contain at least 100 events and 100 non-events for reasonable power.[401,465] For the detection of smaller but still quite relevant invalidity, higher sample sizes are advisable, e.g. 250 events and 250 non-events or 100 events and 900 non-events. [450,493]

Questions

19.1 Differences between populations (Table 19.1)
Consider a hypothetical model that is developed with logistic regression analysis in a sample of 100 patients in a clinical setting. The model is validated in a screening setting. What differences would you expect with respect to
(a) case-mix
(b) regression coefficients

19.2 Validity of a model
What would happen to the calibration and discrimination of a prediction model if
(a) units of measurement were wrong, e.g. mg/dl vs. mmol/L?
(b) a different measurement device was used, with random deviations compared with the measurements in the development setting
(c) a more heterogeneous case-mix was present in the validation setting
(d) a treatment that was very effective for all patients was used
(e) a treatment that was very effective for one subgroup was used

19.3 Influence of case-mix on clinical usefulness
A less-severe case-mix led to a higher net benefit than a more-severe case-mix (NB 0.104 vs. 0.006, Fig. 19.3). How do you explain this finding?

19.4 Disturbance of calibration (Sect. 19.8)
We found that calibration is not disturbed when the validation setting consists of (a) more-or less-severe patients according to predictor values, or (b) more or less heterogeneous patients according to observed or missed predictor values.
(a) What disturbs calibration-in-the-large?
(b) What disturbs the calibration slope?

19.5 Discrimination and clinical usefulness
Why is discrimination a *necessary but not sufficient* condition for clinical usefulness?

19.6 Reference values for performance (Sect. 19.6)
Reference values indicate a model's performance under the condition that the model predictions are valid in the validation sample. How is it possible that the reference value for performance can be better than the performance estimate in the development setting?

19.7 Power of validation studies (Table 19.7)
Suppose we wish to detect a possible deterioration in calibration-in-the-large of an odds ratio of 1.5, and a calibration slope < 0.8. What sample size would you recommend?

19.8 Study design: epidemiologic and statistical aspects
Suppose we can do a single centre study with 1,000 patients, where 200 (20%) will have the event of interest. Alternatively, we can do a multi-centre study with three centres, each contributing 300 patients. Among the 900, we expect 180 (20%) patients with the event of interest.
Which design would you prefer? Explain why, weighing epidemiological considerations (such as generalizability) and statistical considerations (such as standard error).