Chapter 16
Translational Risk Models

Donna Pauler Ankerst, Vanadin Seifert-Klauss, and Marion Kiechle

With rapid progression of computing and other technological advances, the practice of modern medicine has moved from primarily anecdotal to largely quantitative. With due credit to the Internet and the new cyber-society, individuals have taken a more active role in the decision-making process concerning their health, from deciding whether or not to get screened for a disease to which treatment is best for their specific clinical profile. Treating physicians are more connected with latest medical breakthroughs through vast dissemination via the Internet. Statistical prediction models assembled on large well-designed cohorts, multiply validated and easily accessible through online calculators play a role in translating basic science results to implementation in the community for public health benefit. This chapter describes the risk model building process that forms the basis of modern medical decision-making, from statistical estimation to validation and implementation on the Internet. The early diagnosis of cancer is used as the context to illustrate principles, though the concepts immediately transcend to other disciplines as concluding examples in forestry and finance will show.

Keywords Logistic regression · Calibration · Discrimination · Prediction · Validation

D.P. Ankerst (✉)
Biostatistics, Center for Mathematical Sciences, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: ankerst@tum.de

D.P. Ankerst
Health Science Center at San Antonio, University of Texas, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA

V. Seifert-Klauss
Gynaecology, Department of Medicine, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Str. 22, 81675 Munich, Germany

M. Kiechle
Chair of Gynaecology, Department of Medicine, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Str. 22, 81675 Munich, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_16,

The Facts

- Logistic regression for predicting disease from multiple risk factors, or more generally a dichotomous outcome from covariates, will be covered.
- Latest developments in external validation, including its distinct components of discrimination, calibration and net benefit, will be reviewed.
- How risk prediction tools get converted to online risk calculators will be discussed.

1 Introduction

Risks provide the currency by which doctors, patients and individual members of the general population communicate and make informed decisions regarding health. Examples of daily media encounters with risk include claims that diets rich in fruits and vegetables reduce the risk of heart disease, smoking increases the risk of lung cancer, or one glass of red wine per day reduces blood pressure, just to name a few. Increasingly it has been recognized that risks agglomerate from a multitude of factors rather than being the product of any single factor acting in isolation, for example, that both diet and exercise work more effectively in combination to reduce the risk of cardiovascular disease. Experience has also revealed that there is uncertainty underpinning estimates of risk, in other words, that one study may undo a previous study report on risk. Most people obtain information concerning health risk either passively through the media, or more actively, through the Internet. These sources in turn obtain their information from peer-reviewed published scientific studies and hence often serve as the translators of basic science to public use. The scientific studies have typically involved observation of a cohort or group of voluntary participants under the relevant controlled or uncontrolled environments of a clinical trial or observational study, respectively, followed by subsequent observation of outcome and an observed statistically significant association between risk factors, interventions and outcomes. Statisticians, epidemiologists and other quantitative scientists scrutinize the findings from such cohorts, determining which biases may have been at play that could ultimately limit validity of the findings or generalizability to populations beyond that on which the studies have been performed. For example, a risk model constructed primarily from people of one ethnicity may not apply to people of other ethnicities. They build risk models when applicable, and validate them in other cohorts, a process sometimes taking years beyond the already many years invested in conducting the original study collecting the data and unfortunately, sometimes resulting in failure to validate, thus limiting the scientific impact of the original study reporting a positive finding.

This chapter describes the process beginning with the end of a published study reporting a significant association between risk factors and outcomes and ending with implementation of a risk prediction model for public use via the Internet. The general concept of building a risk model applies to a vast variety of applications,

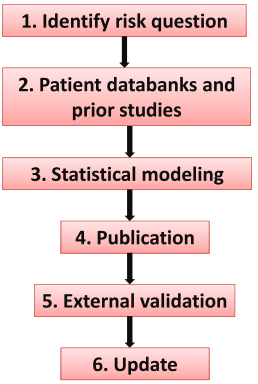

Fig. 1 The risk building paradigm

data and statistical models, the details of which cannot be delivered in a single chapter or even a single textbook. Hence, to concretely illustrate the general concepts, a specific application of the online Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) for detection of prostate cancer will be used throughout. In later sections of the chapter, an overview of the most prominently used risk models in prostate, breast, lung, colorectal, and ovarian cancer will be reviewed, followed by generalizations to other diseases and other disciplines, such as ecology, business and finance.

2 Building Risk Models

The fundamental paradigm for building risk models is shown in Fig. 1 and proceeds as follows. First a population- or clinically-relevant risk question is identified. Second, the appropriate data sources are located that could be used to address the risk question. For example, existing data repositories could be mined retrospectively to see what risk factors led to development of a disease or outcome, or a new study could be designed to prospectively follow individuals. Once the data are collected, the third step requires statistical analysis, entailing data cleaning, fitting of an appropriate model, selection of covariates to include in the model, adjustments for potential biases in the data collection and then internal testing of model performance. Once this is complete, the model is published in order to disseminate the results, for the media, for further validation, which is the fifth step, and hopefully ultimately for use by the public. Technologies and risk factors change over time and new biomarkers of disease (biological entities that can be measured in the blood or urine) or new risk factors are continually discovered. Therefore the last step of the process is the continual task of keeping a risk model contemporary. To give an example of the process a brief overview of the PCPTRC from its conception to ongoing efforts to update as new biomarkers for prostate cancer are discovered is illustrated. Details for the specific steps of Fig. 1 will be more thoroughly described in subsequent sections of the chapter.

Example 2.1 We illustrate the risk building paradigm through the PCPTRC.

(1) (*Identify risk question*) Prostate cancer has the highest incidence of all cancers affecting U.S. men and is the second leading cause of cancer-related death behind lung cancer [10]. Prostate-specific antigen (PSA) is the leading blood test used for the early detection of prostate cancer and it is now common for men over 55 years of age to undergo routine screening for prostate cancer. Of concern to older men is given their PSA values and other clinical test results, what is their risk of prostate cancer? If sufficiently high, then they might be advised to undergo prostate biopsy, a more invasive diagnostic procedure.

(2) (*Patient databanks and prior studies*) The PCPTRC was developed based on analysis of data from 5519 placebo arm participants who had undergone annual PSA and digital rectal examination (DRE) screening as part of the 7-year Prostate Cancer Prevention Trial (PCPT) [24]. All PCPT participants were requested to undergo prostate biopsy, both during the trial when prompted by a PSA value exceeding 4 ng/mL or abnormal digital rectal exam (DRE) result and at the end of the trial regardless of PSA and DRE findings. The latter aspect made the PCPT cohort unique in the world in having prostate cancer status ascertained by biopsy even among men who did not meet the clinical criteria for recommendation to biopsy.

(3) (*Statistical modeling*) For predicting prostate cancer outcome all potential risk factors measured on participants during the trial were identified, including age, family history of prostate cancer in a first degree relative, whether or not a prior prostate biopsy had been performed that was negative for prostate cancer, race, ethnicity, and PSA and DRE outcomes within one year prior to the biopsy result used in the analysis. Participants could have multiple biopsies up until either a positive cancer diagnosis or the end-of-study required biopsy; only the last biopsy of each participant was used. Logistic regression was used to statistically model the association between the multiple risk factors to the outcome, prostate cancer or not, on biopsy. A separate logistic regression was performed for the association of risk factors to high grade prostate cancer, defined as prostate cancer with Gleason grade ≥ 7. High grade cancer is a particularly aggressive form of cancer that is more often associated with mortality.

(4) (*Publication*) The PCPTRC appeared online as soon as the algorithm for the PCPTRC appeared by [24].

(5) (*External validation*) Accuracy of the PCPTRC has been validated in a range of external populations, from healthy populations undergoing annual screening with men referred to prostate biopsy for elevated PSA or abnormal DRE, similar in art to the PCPT [16], to clinical populations where men underwent biopsy based on clinical symptoms [5, 7, 8, 15].

(6) (*Update*) The PCPT was enhanced in 2008 to include a new urine marker for prostate cancer, PCA3 [2] and due to the online posting of the updated calculator, soon thereafter externally validated [18]. It has recently been updated to include the biomarkers percent free PSA and [−2]proPSA (two recently discovered relatives of PSA that are also measurable in the blood) and externally

validated [3]. Minor updates to the PCPTRC were made to tailor for men currently taking finasteride [25] and to incorporate body mass index [12].

3 Statistical Models

Risk calculators can predict a range of outcome types, such as the probability of having prostate cancer on biopsy, of developing breast cancer in the next 5 years, or surviving past 10 years post-treatment for a disease. Accordingly they are built on the appropriate statistical model for the outcome of interest. For example, Cox's proportional hazards model is a popular model for predicting times to an event with possible censoring (end of follow-up time preceding occurrence of the event) because it incorporates risk factors and makes minimal assumptions on the baseline event hazard rate. Logistic regression is commonly used to predict binary events, and will be used to illustrate the principles throughout this chapter. The principles of model selection and testing applied to logistic regression easily extend to other statistical models for other outcome types.

Logistic regression is a method for relating multiple risk factors $X_1, \ldots, X_p$ assembled in a vector $X = (X_1, \ldots, X_p)$ to a dichotomous outcome Y which will be assumed to take the value of 1 for a bad outcome, such as disease and 0 for the opposing good outcome, such as no disease. Specifically it relates the log odds of the bad outcome ($Y = 1$) to the risk factors X through the relationship:

$$\log \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \alpha + \beta' X, \tag{1}$$

where in the formula, log denotes the natural logarithm (base e), α is an intercept, and β a vector of log odds ratios, one for each risk factor assembled in X.

To understand why β defines log odds ratios, it is helpful to consider the simple scenario of just one dichotomous risk factor, X, taking the value 1 for an unfavorable risk factor value versus 0 for a favorable risk factor value. Based on the logistic model, the odds of the bad outcome based on the single risk factor X is:

$$\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \exp\{\alpha + \beta X\},$$

where exp denotes the exponential function. This equation implies that for the individual with risk factor X, the probability of the bad outcome is a multiple, $\exp\{\alpha + \beta X\}$, times the probability of the good outcome. The odds of the bad outcome for individuals with the unfavorable risk factor ($X = 1$) and favorable risk factor ($X = 0$) are given by:

$$\frac{P(Y = 1|X = 1)}{1 - P(Y = 1|X = 1)} = \exp\{\alpha + \beta\}, \qquad \frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)} = \exp\{\alpha\},$$

respectively. From these expressions one might expect β to be greater than 0 since individuals with the unfavorable risk factor should have a higher probability, and hence odds, of the bad outcome than individuals with the favorable outcome. The

ratio of odds for individuals with the unfavorable ($X = 1$) to favorable ($X = 0$) risk factor describes the magnitude by which the odds of the bad outcome accordingly changes:

$$\frac{\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}}{\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)}} = \frac{\exp\{\alpha+\beta\}}{\exp\{\alpha\}} = \exp\{\beta\}.$$

The simplified expression, $\exp\{\beta\}$, is the odds ratio (OR) for individuals with the unfavorable compared to favorable risk factor.

For the case of a single predictor X that is continuous rather than dichotomous, the OR gives the ratio of odds of outcome Y for a unit-increase in X (to see this compute the OR for $X = x + 1$ compared to $X = x$). An OR > 1 implies that an increase in the risk factor increases the odds of the bad outcome, OR < 1 means it decreases the odds and OR $= 1$ means it has no impact. From the relationship above, $\beta = \log\{\exp\{\beta\}\}$ is the log odds ratio (log OR), and values of $\beta > 0$, <0, and $=0$ have the same interpretations as for OR > 1, <1 and $=1$, respectively.

If X were a categorical risk factor with more than two levels, such as race with levels African American, Caucasian, and Other, the logistic model can still be fit by choosing one level as a reference (say Caucasian) and then returning two odds ratios, one for the comparison of each of the remaining levels, African Americans and Other, to the reference. In many aspects such as this logistic regression operates similarly as for linear regression. In the general case of multiple risk factors (1), $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of respective log OR's for each of the multiple risk factors comprising $X = (X_1, X_2, \ldots, X_p)$:

$$\log \frac{P(Y=1|X)}{1-P(Y=1|X)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

The interpretation of each parameter β_i is the log OR corresponding to a unit increase in the respective risk variable X_i, with all other risk variables in the model held constant.

Statistical packages return estimates of log odds ratios (β's), their standard errors, and p-values for tests of the null hypotheses that they equal 0 (no effect) versus two-sided alternative hypotheses that they do not equal 0. From these approximate 95 percent confidence intervals for log ORs can be constructed as (estimated log OR) $\pm$ 1.96 $\times$ (standard error); to obtain estimates and confidence intervals for the OR's, take the exponent of the estimates and 95 percent confidence interval bounds, respectively.

Typically many risk factors or individual characteristics, including demographic, environmental or other variables are available for potential inclusion in the model and additionally more complicated relationships between the variables and outcomes can be modeled using transformations and interactions. Therefore, a variety of model selection techniques are available in statistical packages, many of which automatically sort through large numbers of models. Some of the most commonly used model selection techniques are based on finding the model with the lowest Akaike's information criterion, AIC $= -2 \times$ maximized log likelihood $+ 2 \times$ number of parameters, or the lowest Bayesian Information criterion,

BIC $= -2 \times$ maximized log likelihood $+$ log(sample size) $\times$ number of parameters [1, 19]. The first terms of both criteria seek to find the model maximizing goodness of fit to the developmental data set, while the second terms penalize for over-parameterization, with the BIC tending to penalize more, and hence selecting smaller models with fewer parameters than the AIC on average.

As a preliminary indication of how the model may validate, internal validation can be performed by splitting the dataset into a training and test set or into a group of equally sized subsets that alternatively serve as training and test sets. A logistic regression model is fit from scratch on the training set and then evaluated on the test set using any one of the metrics to be defined later for external validation. For multiple splits of the dataset, test set performances are simply averaged. Bootstrapping, repeated random sampling with replacement of test and training sets, can also be used.

Example 3.1 The BIC and average cross-validated area underneath the receiver operating characteristic curve (AUC) were used to find the optimal multivariable logistic regression model relating potential risk factors to prostate cancer outcome on biopsy on data from the 5519 PCPT placebo arm participants used to develop the PCPTRC [24]. The AUC is a rank-based measure of how well a risk model discriminates the bad outcomes it aims to predict from the good outcomes and will be further defined in later sections. For the PCPT, 4-fold internal cross-validation was implemented, whereby the developmental dataset of 5519 observations was randomly partitioned into four subsets, three of size 1380 and one of size 1379, with randomization stratified to keep the proportion of prostate cancer cases between 20 % and 23 % in each subset. Over 50 models, some including two-way interactions, were evaluated by a combination of forward, backward and stepwise selection and subjective measures, such as including only statistically significant effects at the 0.05 level. BIC and cross-validated AUC values were tabulated for each of the models. The model with the lowest BIC value contained only main effects and no interactions among risk factors and was also one of the models with lowest AUC values. Therefore this model was selected to form the PCPTRC. The final selected logistic regression model contained four risk factors: PSA (OR $= 2.34$ for logPSA), DRE (OR $= 2.47$), family history of prostate cancer (OR $= 1.31$) and history of a prior negative prostate biopsy (OR $= 0.64$). All were statistically significant with a p-value less than 0.001 except for family history with a p-value of 0.002 [24].

4 External Validation

Once a risk model has been constructed, it is critical to evaluate its performance on a population independent to that on which it was developed. Internal validation, evaluating the model on the same population as on which the model was developed, even though it has been split into separate training and test sets, is not enough, since unmeasurable cohort effects will still favorably bias the performance of the

model compared to what might be achieved in a completely distinct cohort collected elsewhere. A variety of evaluation methods for risk models have been proposed in the literature, and these can be grouped into those that measure discrimination, calibration, or both. Recent reviews detangle the different objectives of the many metrics currently employed to evaluate risk prediction models [20, 21]. All of these metrics require an external validation cohort or data set, whereby all individuals in the cohort have all risk factors X required for evaluation of the risk prediction model and the true outcome Y.

For missing covariates X, [9] showed by simulation that imputation results in less biased estimates of validation metrics than other currently used practices of either excluding the entire patient from the analysis or throwing the covariate out of a model. The current state of the art in imputation for X is based on specification of full conditional distributions for missing covariates and termed Multivariate Imputation by Chained Equations (MICE) and implementable in the R statistical package [26]. For missing outcomes Y in logistic regression, verification bias algorithms, which repeatedly impute the missing Y values using the assumed logistic regression form can be used if the missing data mechanism is assumed to be missing-at-random (MAR), meaning that the reason for missing data does not depend on the missing outcome value [4, 23].

4.1 Discrimination

Discrimination metrics focus on how well risk prediction models perform if used as the basis for making binary decisions as to whether individuals will have bad or good outcomes, sometimes referred to as hard classifications. Moving from a risk prediction, varying from 0 % to 100 %, to a positive versus negative decision on the bad outcome requires selection of a threshold r such that a risk above r corresponds to a positive test and below r, a negative test. The misclassification rate, or number of wrong test results made, is calculated separately for the subpopulations with bad and good outcomes.

How successfully the risk prediction predicts bad outcomes is termed sensitivity and on the external validation set is estimated by the percent positive tests among the bad outcomes:

$$\text{Sensitivity}(r) = \frac{\text{Number of bad outcomes with risk} > r}{\text{Number bad outcomes}},$$

where sensitivity is indexed by r as a reminder that it depends on the user-selected threshold r. How successfully the risk prediction tests negative for the good outcomes is termed specificity and is accordingly estimated by:

$$\text{Specificity}(r) = \frac{\text{Number of good outcomes with risk} \leq r}{\text{Number good outcomes}}.$$

The higher the sensitivity and specificity at any threshold r the better the risk prediction tool is. The problem is that as r increases from 0 % to 100 % specificity

increases from 0 % to 100 % while sensitivity decreases from 100 % to 0 %, so that finding the threshold r that simultaneously optimizes sensitivity and specificity is difficult to achieve in practice. Sensitivity is often referred to as the true positive rate (TPR), one minus the sensitivity as the false negative rate (FNR) and one minus the specificity as the false positive rate (FPR).

The receiver operating characteristic (ROC) curve provides a summary of sensitivity and specificity for all choices of r ranging from 0 % to 100 %; it typically displays sensitivity on the y-axis and false positive rates on the x-axis, with both axes ranging from 0 % to 100 % [22]. The higher the ROC curve, the better its capacity for distinguishing bad from good outcomes. An appealing feature of ROC curves is that they are invariant with respect to measurement scales, for example, risks and the logits of risks (1) will yield the same ROC curve. This makes ROC curves particularly useful when comparing tests on completely different measurement scales, for example for directly comparing risk predictions from a model to the leading risk factor or covariate in the risk prediction model. Finally, as rank-based measures, ROC curves are by definition independent of disease prevalence in external validation set and hence can be applied to the case-control study situation in addition to prospective studies. An interesting single summary of the ROC curve is the area under the ROC curve (AUC), which in addition, conveniently holds the intuitive definition as the probability that a randomly chosen individual with a bad outcome has a higher risk prediction than a randomly chosen individual with a good outcome. The AUC ranges from a minimum at 50 %, implying predictive power of the risk prediction tool no better than flipping a coin to a maximum of 100 % for a perfectly discriminating risk prediction tool.

As seen by their formulas sensitivities and specificities for each threshold r can be calculated by just computing the appropriate sample proportions in the external validation set. The AUC is equivalent to the non-parametric Mann-Whitney or Wilcoxon rank sum statistic for comparing two populations and is hence easily computable using standard statistical software. The Wilcoxon test can be used for testing the null hypothesis that the AUC equals 0.5 versus the alternative that it exceeds 0.5. External packages can be imported into the statistical package R for computing the AUC and for performing various statistical tests for comparing AUCs of multiple tests.

Example 4.1 In 2009 the generalizability of the PCPTRC, which had been developed on a cohort of primarily Caucasian, healthy and elderly men, for potential applicability to other populations was investigated. The Early Detection Research Network (EDRN) clinical cohort comprised 645 men, some younger than members of the PCPT cohort, who had been referred to multiple urology practices across 5 states in the northeastern U.S. and had received a prostate biopsy due to some clinical indication, including persistent elevated PSA or abnormal DRE [7]. PCPTRC risks were calculated for each member of the EDRN cohort and compared to the actual clinical outcome on biopsy using sensitivities, specificities and the AUC. The PCPTRC demonstrated statistically significant superior discrimination for detecting prostate cancer cases compared to PSA (AUC = 69.1 % compared to 65.5 %,

respectively, p-value = 0.009), and the ROC curve for the PCPTRC consistently fell at or above that for PSA for all false positive rates, with the greatest difference for false positive rates less than 25 %. For example, the thresholds of the PCPT Risk Calculator and PSA which obtained a false positive rate of 20 % were 48.4 % and 6.9 ng/mL, respectively (Table 2 of [7]). One can view these as two alternative tests for referral to further intensive diagnostic testing by prostate biopsy, each with equal specificities: the PCPTRC refers a patient to prostate biopsy if his PCPT risk exceeds approximately 50 % and the PSA test if his PSA exceeds 6.9 ng/mL. If these two diagnostic tests had been implemented in the EDRN population to "rule in" patients who should undergo prostate biopsy and "rule out" patients who should not, the PCPTRC would have correctly referred 47.1 % of the prostate cancer cases (sensitivity) and the PSA test 35.4 % of the prostate cancer cases. Although better than PSA, the PCPTRC would still have missed 50 % of the prostate cancer cases! Insisting that 80 % of prostate cancer cases get caught for both tests would have meant that the thresholds for referral would have had to be lowered, to 38.0 % and 4.0 ng/mL, for the PCPTRC and PSA test, respectively (Table 3 of [7]). But this would have approximately halved the specificity of both tests, to 40.3 % for the PCPT Risk Calculator and 44.1 % for PSA. In other words, approximately 60 % of the men who did not have prostate cancer would have been referred to a prostate biopsy unnecessarily (false positive rate), an error rate unacceptable from a public screening perspective.

4.2 Calibration

Calibration concerns itself with how close predicted risks from a model are to actual risks observed in an external validation population. Observed risks should match predicted risks among homogenous groups defined by the same risk profile. However, obtaining an external validation set large enough to have enough individuals with the same risk factors in order to make comparisons quickly becomes infeasible as the number of risk factors increases; hence approximations are made by further grouping.

One of the most commonly used calibration tests is based on an approximation to Pearson's chi-square goodness-of-fit test recommended by Lemeshow and Hosmer [11]:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)},$$

where the validation set has been partitioned into k equally-sized groups (typically $k = 10$ with groups defined by deciles of the predicted risks in the validation set), O_i are the observed numbers of bad outcomes in the groups, n_i the observed numbers of participants in the groups, and π_i the mean risks of the groups. Under the null hypothesis, observed outcomes O_i are close to expected outcomes $n_i\pi_i$, hence X^2

should be small. Asymptotically under the null hypothesis, the X^2 statistic follows a chi-square distribution with k degrees of freedom.

An alternative measure of calibration, which measures reliability, was proposed by [6] and elaborated upon in [13]. The approach requires logistic regression of the outcomes ($Y_i = 0$ good outcome, $Y_i = 1$ bad outcome) on the logit of the predicted risks (π_i) as covariates for the $i = 1, \ldots, N$ individuals in the validation set:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \alpha + \beta \log \frac{\pi_i}{1 - \pi_i}.$$

A perfect match of predicted to actual risks would occur when $\alpha = 0$ and $\beta = 1$. Therefore, a test of the composite null hypothesis $H_0 : \alpha = 0, \beta = 1$ provides an overall reliability test for the predictions. More specifically, the intercept α controls the calibration of the model, which is most clearly seen when $\beta = 1$. When $\beta = 1$, $\alpha < 0$ implies the predicted risks are too high and $\alpha > 0$, too low. When $\beta \neq 1$, noting that for $\pi_i = 0.5$:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \alpha,$$

one can interpret the intercept α as a calibration measure at $\pi_i = 0.5$. The slope parameter β is referred to as the refinement parameter: $\beta > 1$ implies the predicted risks do not vary enough, $0 < \beta < 1$ they vary too much, and $\beta < 0$ they show the wrong direction. Therefore, additional tests of calibration given appropriate refinement, $H_0 : \alpha = 0 | \beta = 1$, and of refinement given appropriate calibration, $H_0 : \beta = 1 | \alpha = 0$, can be performed.

Example 4.2 In [7] it was reported that the average PCPTRC risk over all 645 men of the EDRN cohort was 45.1 %, which is fairly high in keeping with the nature of the cohort as elicited from multiple Urology practices. As a first indication of calibration the average PCPTRC risk among the cohort should correspond to the actual percent of the cohort that did have prostate cancer on biopsy. The percentage of the 645 men in the EDRN cohort diagnosed with prostate cancer was 43.4 %, fairly close to the average PCPTRC risk, providing a crude indication of calibration.

As an exploratory and primarily descriptive analysis of calibration among specific risk groups, Table 4 of [7] assessed the degree to which the PCPTRC calibrated to actual risks for specific subgroups, such as for Caucasians, African Americans, men with a positive family history and men with PSA less than 4.0 ng/mL. Across all subgroups the average PCPTRC risk never varied by more than approximately 5 or 6 percentage points from the observed risk but there were some subgroups where PCPTRC risks were better calibrated to actual risks than others. For example, among the 47 African American participants in the cohort, 51.1 % had prostate cancer but the average PCPTRC risk among these men was only 45.4 %. Application of the Lemeshow and Hosmer test of calibration yielded a p-value of 0.10, not rejecting the null hypothesis of a good fit at the 0.05 level of statistical significance. Cox's logistic regression of observed prostate cancer status on logits of predicted PCPTRC risks was also performed. The composite hypothesis test of reliability was not performed; however, the intercept from the logistic regression was estimated as -0.014

with standard error 0.091 and the slope by 1.291 with standard error 0.159. Separate 95 % confidence intervals for these estimates overlapped with 0 and 1, respectively, indicating that predicted PCPTRC risks were reliable estimates of observed risks in the EDRN population.

4.3 Net Benefit

Discrimination and calibration metrics objectively summarize accuracy but do not provide information as to which thresholds of a prediction model might be useful for basing clinical decisions. Towards this end, Vickers and Elkin [27] proposed a measure of net benefit justified through a layman's decision analysis framework that does not rely on user-specified costs associated with various outcomes as full-blown decision analyses typically do. As with the other accuracy measures, net benefit is evaluated on an external cohort to the one on which the risk model was developed as the expectation over the true and false positive counts:

$$\begin{aligned}\text{NetBenefit}(\text{Cohort}, r) = &\frac{\text{True Positive Count}(\text{Cohort}, r)}{\text{Sample Size}(\text{Cohort})} \\ &- \frac{\text{False Positive Count}(\text{Cohort}, r)}{\text{Sample Size}(\text{Cohort})}\left(\frac{r}{1-r}\right),\end{aligned}$$

where for emphasis dependencies on the chosen cohort and user-selected cutoff r are included in the definitions. The expression for the net benefit can be rewritten to show that it is also a function of the discrimination measures sensitivity and 1-specificity, $\text{TPR}(\text{Cohort}, r)$ and $\text{FPR}(\text{Cohort}, r)$, respectively, evaluated on the cohort and weighted by the proportions of bad outcomes (% Bad Outcomes(Cohort)) and good outcomes (% Good Outcomes(Cohort)) in the cohort:

$$\begin{aligned}\text{NetBenefit}(\text{Cohort}, r) = &\text{TPR}(\text{Cohort}, r) \times \%\ \text{Bad Outcomes}(\text{Cohort}) \\ &- \text{FPR}(\text{Cohort}, r) \times \%\ \text{Good Outcomes}(\text{Cohort}) \\ &\times \left(\frac{r}{1-r}\right).\end{aligned}$$

This expression illustrates further the dependence of the net benefit on the evaluation cohort. The discrimination metrics TPR and FPR already tend to depend on the cohort, net benefit further relies on how prevalent the disease is in the evaluation cohort. In other words, for two cohorts with the same operating characteristics of a prediction model, the cohort with a higher disease prevalence will demonstrate higher net benefit for using the prediction tool for clinical decisions.

Vickers and Elkin suggested evaluating the net benefit over all possible thresholds r of the prediction model ranging from 0 to 1. The specific numbers obtained for the net benefit can be difficult to interpret in isolation so they also recommended overlaid decision curves for the strategies of referring no patients to action or all patient's to action regardless of the threshold r selected. For these curves the last expression $[r/(1-r)]$ remains the same but the TPR and FPR are calculated based

on the test rule that assigns no patients positive (in other words, $r > 1$) and all patients positive (in other words, $r < 0$). For taking no action, the TPR and FPR are identically 0 so the net benefit curve for taking no action is the horizontal line at 0 across all thresholds r. For taking action on all patients the TPR and FPR are 1 and the net benefit curve is % Disease(Cohort) − % NotDisease(Cohort)$[r/(1-r)]$, which seemingly ironically, still depends on the threshold r, but that is an artifact from the derivation of relative values of false positive results used in the derivation of the net benefit.

4.4 Overall Performance Measures

There are overall measures that combine discrimination and calibration that have been proposed for evaluating risk models but these have not gained widespread use largely for two reasons. Firstly, they have awkward properties because of the dichotomous nature of the outcome predicted by a continuous measure and secondly, they do not have an intuitive clinical interpretation.

The ubiquitous R^2 measure of proportion of variability explained by a linear regression of a continuous outcome Y on a series of variables has been extended to the case of generalized linear models, including logistic regression, where Y is dichotomous in the form of Nagelkerke's R^2 [14]:

$$R^2 = 1 - \exp\left[-\frac{2}{n}\left\{l(\hat{\beta}) - l(0)\right\}\right] = 1 - \left[\frac{L(0)}{L(\hat{\beta})}\right]^{2/n},$$

where $L(\cdot)$ and $l(\cdot)$ are the likelihood and loglikelihood functions, respectively, defined at the maximized values of β, the logistic regression log odds ratios, and for a null model with no covariates ($\beta = 0$). The problem with this measure is that for dichotomous outcomes it has a maximum less than 1, so is not as easy to judge as for continuous outcomes, where the maximum of R^2 is 1. Modifications by the max obtainable R^2 have been proposed but these are awkward to implement in practice. Therefore the criterion has not become widely used outside the case of linear regression with Normal outcomes.

A similar metric extended for dichotomous outcomes that has not found widespread use is the Brier score, which is simply the squared difference between the 0–1 Y outcomes and predictions from the model:

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{p}_i)^2.$$

Predictions are good if the Brier score is small but squared Euclidean distance between a dichotomous outcome Y and a continuous predictor p is not intuitive and will give coarse results for small sample sizes n. The Brier score also obtains a maximum less than one and similarly, attempts to correct it are awkward [21].

4.5 Integrated Discrimination Index

Noting shortcomings in the AUC for comparing risk prediction tools, Pencina and colleagues [17] proposed the integrated discrimination index (IDI) for comparing risk predictions from a new model to risk predictions from an old model that is simply the difference in discrimination slopes between the new and old predictions as proposed by Yates [28]:

$$IDI = \left(\frac{1}{n_{events}} \sum_{i=1}^{n_{events}} p_{new,i} - \frac{1}{n_{nonevents}} \sum_{i=1}^{n_{nonevents}} p_{new,i} \right) - \left(\frac{1}{n_{events}} \sum_{i=1}^{n_{events}} p_{old,i} - \frac{1}{n_{nonevents}} \sum_{i=1}^{n_{nonevents}} p_{old,i} \right),$$

where n_{events} are the number of events, or bad outcomes, and $n_{nonevents}$ are the number of non-events, or good outcomes, and the summations sum over the predicted probabilities from the new and old models as subscripted on the n's. The logic of the IDI is clear, a good prediction model should provide higher estimated risks among the bad outcomes in the validation set compared to the good outcomes, how good is determined by the discrimination slopes of the models. A positive IDI would indicate a new model has better discrimination slope than the old.

Example 4.3 Ankerst and colleagues [30] have developed an extension of the PCPTRC to incorporate the novel prostate cancer markers % freePSA and [−2]proPSA, which are both obtainable by blood tests. The methodology for updating the PCPTRC for new markers is based on Bayes algorithm for updating the prior odds of prostate cancer, which in this case are based on PCPTRC risks, via likelihood ratios of the distributions of the new marker among prostate cancer cases and controls to obtain posterior odds and hence updated posterior risks for prostate cancer; for more details see [29]. The updated PCPTRC is now available online at the same location as the PCPTRC. A developmental dataset of 474 participants in the San Antonio Biomarkers of Risk (SABOR) study were used to build the updated PCPTRC and the model was validated on an external EDRN dataset comprising 575 men. The IDI for comparing the new updated PCPTRC incorporating the two new markers to the standard PCPTRC evaluated on the EDRN validation set was 6.3 % (95 % CI 3.0 to 9.6 %), indicating a statistically significant positive improvement to using the updated model. Figure 2 compares the net benefit curves of the updated PCPTRC model (called posterior PCPTRC risks), the original PCPTRC, and the strategies of simply referring all men or no men to prostate biopsy irrespective of any risk prediction model. The benefit curves indicated benefit of using both the PCPTRC and updated PCPTRC for situations where risk thresholds exceeding 20 % for both of these rules would be used for referral to biopsy over the blanket rule of referring all men in the EDRN cohort to prostate biopsy, but no clear benefit of the more complicated updated PCPTRC to the standard PCPTRC in this region. Both the standard and updated PCPTRC provided benefit over the rule referring no patients to biopsy.

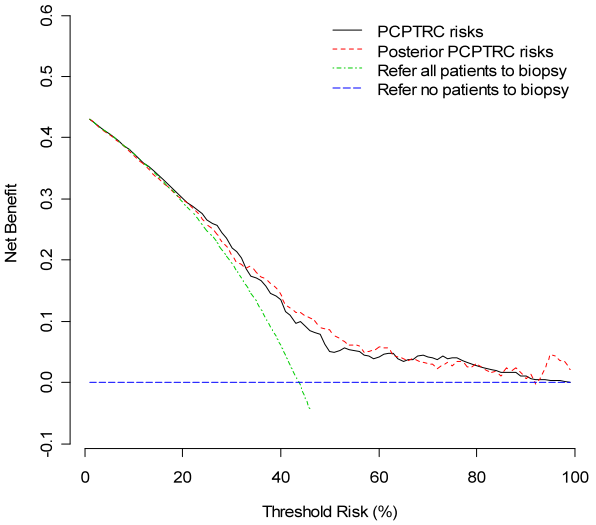

Fig. 2 Net benefit curves

5 Food for Thought

To illustrate fundamental principles this chapter has focused on risk models for cancer diagnosis, but similar principles apply for all aspects of cancer treatment and follow-up care, and easily extend to prediction problems in other disciplines outside of medicine.

For projecting risks over expanded time periods, such as the 10-year risk of heart attacks or other cardiovascular events in elderly people, models incorporating risks of death from other causes, referred to in the medical literature as competing risks, need to be implemented. Prognostic models refer to models used to predict treatment outcomes, such as how long a patient can expect to live after a given a treatment is administered. They may rely on other methodologies than logistic regression, such as the Cox proportional hazards models which are appropriate for handling the commonly occurring censored survival times. A censored survival time refers to the case where the exact date of death of a patient is not observed; it is only known that the patient has lived a certain number of years, such as up until the end of the clinical trial. Projections from such models can be used as a basis for making treatment decisions, by favoring treatments that have the longest survival period for specific patient clinical characteristics. The picture is not unidimensional as benefits in survival might be offset by loss in quality of life due to side effects. More complicated decision functions incorporating multiple outcomes are required for weighing the cost-benefits of competing treatment options. Increasingly, models addressing multiple long-term effects of preventative or curative treatments for cancer, such as higher incidences of ovarian and endometrial cancer in women taking tamoxifen, are being implemented in order to provide a unified picture of the pros and cons of various actions, providing many avenues for research in risk prediction for the future of medicine.

The principles of model building and online prediction outlined in this chapter also directly apply to other disciplines, including forestry, ecology, informatics and finance. For forest management, [32] have developed an online tool called SILVA that projects growth of trees over expanded time periods. Their model accounts for man-induced thinning, mortality, and other natural- and human-induced impacts. Using an expanded database of 40000 trees observed at 5-year intervals in Bavaria, Boeck and colleagues [31] implemented logistic regression to update the mortality simulator module of the SILVA program. Similar collaborations with ecologists are working towards prediction of microhabitats on trees representing biological diversity, a concept of interest to forest conservationalists. Informatic scientists are using the techniques to develop online predictions of project margins for large and complex software development portfolios. One can foresee similar applications for online predictions of financial success indicators based on the types of advanced time series models used in that field.

6 Summary

This chapter has detailed the step-by-step program by which risk prediction models are built, using as one illustration construction of the PCPTRC, one of the currently most widely used prostate cancer risk calculators. The importance of external validation across multiple cohorts pushing the envelope in terms of generalizability of the risk tool has been emphasized, as well as the separate components of validation which address discrimination, calibration, and net benefit. As risk prediction tools are typically founded on once-in-a-lifetime large well-designed studies, methodologies are needed for updating them based on new data and risk factor discoveries based on smaller more recent studies. This chapter has discussed the need for comparing existing to updated risk models, using the integrated discrimination index as one possible measure. To end a summary on risk prediction tools in current use was provided along with extensions to other outcomes in medicine and applications in other disciplines such as forestry and finance.

References

Selected Bibliography

1. H. Akaike, A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)
2. D.P. Ankerst, J. Groskopf, J.R. Day et al., Predicting prostate cancer risk through incorporation of prostate cancer gene 3. J. Urol. **180**, 1303–1308 (2008)
3. D.P. Ankerst, T. Koniarski, Y. Liang et al., Updating risk prediction tools: a case study in prostate cancer. Biom. J. **54**, 127–142 (2012)

4. C.B. Begg, R.A. Greenes, Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics **39**, 207–215 (1983)
5. V. Cavadas, L. Osório, F. Sabell, F. Teves, F. Branco, M. Silva-Ramos, Prostate cancer prevention trial and European randomized study of screening for prostate cancer risk calculators: a performance comparison in a contemporary screened cohort. Eur. Urol. **58**, 551–558 (2010)
6. D.R. Cox, Two further applications of a model for binary regression. Biometrika **45**, 562–565 (1958)
7. S.J. Eyre, D.P. Ankerst, J.T. Wei et al., Validation in a multiple urology practice setting of the prostate cancer prevention trial calculator for predicting prostate cancer detection. J. Urol. **182**, 2653–2658 (2009)
8. D.J. Hernandez, M. Han, E.B. Humphreys et al., Predicting the outcome of prostate biopsy: comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. BJU Int. **103**, 609–614 (2009)
9. K.J.M. Janssen, A.R.T. Donders, F.E. Harrell Jr. et al., Missing covariate data in medical research: to impute is better than to ignore. J. Clin. Epidemiol. **63**, 721–727 (2010)
10. A. Jemal, R. Siegel, J. Xu, E. Ward, Cancer statistics, 2010. CA Cancer J. Clin. **60**, 277–300 (2010)
11. S. Lemeshow, D.W. Hosmer Jr., A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. **115**, 92–106 (1982)
12. Y. Liang, D.P. Ankerst, M. Sanchez, R.J. Leach, I.M. Thompson, Body mass index adjusted prostate-specific antigen and its application for prostate cancer screening. Urology **76**, 1268.e1–1268.e6 (2010)
13. M.E. Mille, S.L. Hui, W.M. Tierney, Validation techniques for logistic regression models. Stat. Med. **10**, 1213–1226 (1991)
14. N.J. Nagelkerke, A note on a general definition of the coefficient of determination. Biometrika **78**, 691–692 (1991)
15. C.T. Nguyen, C. Yu, A. Moussa, M.W. Kattan, J.S. Jones, Performance of prostate cancer prevention trial risk calculator in a contemporary cohort screened for prostate cancer and diagnosed by extended prostate biopsy. J. Urol. **183**, 529–533 (2010)
16. D.J. Parekh, D.P. Ankerst, B.A. Higgins et al., External validation of the prostate cancer prevention trial risk calculator in a screened population. Urology **68**, 1153–1155 (2006)
17. M.J. Pencina, R.B. D'Agostino Sr., R.B. D'Agostino Jr., R.S. Vasan, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat. Med. **27**, 157–172 (2008)
18. S. Perdonà, V. Cavadas, G.D. Lorenzo et al., Prostate cancer detection in the grey area of prostate-specific antigen below 10 ng/ml: head-to-head comparison of the updated PCPT calculator and Chun's nomogram, two risk estimators incorporating prostate cancer antigen 3. Eur. Urol. **59**, e1–e4 (2011)
19. G. Schwarz, Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
20. E.W. Steyerberg, *Clinical Prediction Models* (Springer, New York, 2010)
21. E.W. Steyerberg, A.J. Vickers, N.R. Cook et al., Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology **21**, 128–138 (2010)
22. J.A. Swets, R.M. Pickett, *Evaluation of Diagnostic Systems: methods from Signal Detection Theory* (Academic Press, New York, 1982)
23. I.M. Thompson, D.P. Ankerst, C. Chi et al., The operating characteristics of prostate-specific antigen in a population with initial PSA of 3.0 ng/ml or lower. JAMA **294**, 66–70 (2005)
24. I.M. Thompson, D.P. Ankerst, C. Chi et al., Assessing prostate cancer risk: results from the prostate cancer prevention trial. J. Natl. Cancer Inst. **98**, 529–534 (2006)
25. I.M. Thompson, D.P. Ankerst, C. Chi et al., Prediction of prostate cancer for patients receiving finasteride: results from the prostate cancer prevention trial. J. Clin. Oncol. **25**, 3076–3081 (2007)
26. S. van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. **16**, 219–242 (2007)

27. A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Mak. **26**, 565–574 (2006)
28. J.F. Yates, External correspondence: decomposition of the mean probability score. Organ. Behav. Hum. Perform. **30**, 132–156 (1982)

Additional Literature

29. D.P. Ankerst, J. Groskopf, J.R. Day et al., Predicting prostate cancer risk through incorporation of prostate cancer gene 3. J. Urol. **180**, 1303–1308 (2008)
30. D.P. Ankerst, T. Koniarski, Y. Liang et al., Updating risk prediction tools: a case study in prostate cancer. Biom. J. **54**, 127–142 (2012)
31. A. Boeck, J. Dieler, P. Biber, H. Pretzsch, D.P. Ankerst, Predicting tree mortality for European beech in southern Germany using spatially-explicit competition indices. For. Sci. (in press)
32. H. Pretzsch, P. Biber, J. Dursky, The single tree-based stand simulator SILVA: construction, application and evaluation. For. Ecol. Manag. **162**, 3–21 (2002)