Published in final edited form as:
Epidemiology. 2010 January ; 21(1): 128–138. doi:10.1097/EDE.0b013e3181c30fb2.

Assessing the performance of prediction models: a framework for some traditional and novel measures

Ewout W. Steyerberg[1], **Andrew J. Vickers**[2], **Nancy R. Cook**[3], **Thomas Gerds**[4], **Mithat Gonen**[2], **Nancy Obuchowski**[5], **Michael J. Pencina**[6], and **Michael W. Kattan**[5]

[1]Dept of Public Health, Erasmus MC, Rotterdam, the Netherlands [2]Dept of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA [3]Brigham and Women's Hospital, Harvard Medical School, Boston MA, USA [4]Institute of Public Health, University of Copenhagen, Denmark [5]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland OH, USA [6]Department of Mathematics and Statistics, Boston University, Boston MA, USA

Abstract

The performance of prediction models can be assessed using a variety of different methods and metrics. Traditional measures for binary and survival outcomes include the Brier score to indicate overall model performance, the concordance (or *c*) statistic for discriminative ability (or area under the receiver operating characteristic (ROC) curve), and goodness-of-fit statistics for calibration.

Several new measures have recently been proposed that can be seen as refinements of discrimination measures, including variants of the *c* statistic for survival, reclassification tables, net reclassification improvement (NRI), and integrated discrimination improvement (IDI). Moreover, decision–analytic measures have been proposed, including decision curves to plot the net benefit achieved by making decisions based on model predictions.

We aimed to define the role of these relatively novel approaches in the evaluation of the performance of prediction models. For illustration we present a case study of predicting the presence of residual tumor versus benign tissue in patients with testicular cancer (n=544 for model development, n=273 for external validation).

We suggest that reporting discrimination and calibration will always be important for a prediction model. Decision-analytic measures should be reported if the predictive model is to be used for making clinical decisions. Other measures of performance may be warranted in specific applications, such as reclassification metrics to gain insight into the value of adding a novel predictor to an established model.

1. Introduction

From a research perspective, diagnosis and prognosis constitute a similar challenge: the clinician has some information and wants to know how this relates to the true patient state, whether this can be known currently (diagnosis) or only at some point in the future (prognosis). This information can take various forms, including a diagnostic test, a marker value, or a statistical model including several predictor variables. For most medical applications, the outcome is our interest is binary and the information can be expressed as

Correspondence: E.W. Steyerberg, PhD, Dept of Public Health, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands, 31 - 10 - 703 8470 (tel); e.steyerberg@erasmusmc.nl.

probabilistic predictions [1]. Predictions are hence absolute risks, which go beyond assessments of relative risks, such as regression coefficients, odds ratios or hazard ratios [2].

There are various ways to assess the performance of a statistical prediction model. The traditional statistical approach is to quantify how close predictions are to the actual outcome, using measures such as explained variation (e.g. using R^2 statistics) and the Brier score [3]. Performance can further be quantified in terms of calibration (do close to *x* of 100 patients with a risk prediction of *x*% have the outcome?), using e.g. the Hosmer-Lemeshow "goodness-of-fit" test [4]. Furthermore, discrimination is essential (do patients who have the outcome have higher risk predictions than those who do not?), which can be quantified with measures such as sensitivity, specificity, and the area under the receiver operating characteristic curve (or concordance statistic, *c*) [15].

Recently, several new measures have been proposed to assess performance of a prediction model. These include variants of the *c* statistic for survival [67], reclassification tables [8], net reclassification improvement (NRI), and integrated discrimination improvement (IDI) [9], which are refinements of discrimination measures. The concept of risk reclassification has caused substantial discussion in the methodological and clinical literature [1011121314]. Moreover, decision–analytic measures have been proposed, including 'decision curves' to plot the net benefit achieved by making decisions based on model predictions [15]. These measures have not yet widely been used in practice, which may partly be due to their novelty to applied researchers [16]. In this paper, we aim to clarify the role of these relatively novel approaches in the evaluation of the performance of prediction models.

We first briefly discuss prediction models in medicine. Next, we review the properties of a number of traditional and relatively novel measures for the assessment of the performance of an existing prediction model, or extensions to a model. For illustration we present a case study of predicting the presence of residual tumor versus benign tissue in patients with testicular cancer.

2. Prediction models in medicine

Developing valid prediction models

We consider prediction models that provide predictions for a dichotomous outcome, since these are most relevant in medical applications. The outcome can be either an underlying diagnosis (e.g. presence of benign or malignant histology in a residual mass after cancer treatment), an outcome occurring within a relatively short time after making the prediction (e.g. 30-day mortality), or a long-term outcome (e.g. 10-year incidence of coronary artery disease, with censored follow-up of some patients).

At model development we aim for at least internally valid predictions, i.e. predictions that are valid for subjects from the underlying population [17]. Preferably, the predictions are also generalizable to 'plausibly related' populations [18]. Various epidemiologic and statistical issues need to be considered in a modeling strategy for empirical data [11920]. When a model is developed, it is obvious that we want some quantification of its performance, such that we can judge whether the model is adequate for its purpose, or better than an existing model.

Model extension with a marker

We recognize that a key interest in contemporary medical research is whether a marker (e.g. molecular, genetic, imaging) adds to an existing model. Often, new markers are selected from a large set based on strength of association in a particular study. This poses a high risk of overoptimistic expectations of the marker's performance [2122]. Moreover, we are only interested in the incremental value of a marker, on top of predictors that are readily

accessible. Validation in fully independent, external data is the best way to compare the performance a model with and without a new marker [2123].

Usefulness of prediction models

Prediction models can be useful for several purposes, such as for inclusion criteria or covariate adjustment in a randomized controlled trial [242526]. In observational studies, a prediction model may be used for confounder adjustment or case-mix adjustment in comparing outcome between centers [27]. We concentrate on the usefulness of a prediction model for medical practice, including public health (e.g. screening for disease) and patient care (diagnosing patients, giving prognostic estimates, decision support).

An important role of prediction models is to inform patients on their prognosis, for example after a cancer diagnosis has been made [28]. A natural requirement to a model for this situation is that predictions are well calibrated (or 'reliable') [2930].

A specific situation may be that only limited resources are available, which hence need to be targeted to those with the highest expected benefit, such as those at highest risk. This situation calls for a well discriminating model which separates those at high risk from those at low risk.

Decision support is another important area, including decisions on the need for further diagnostic testing (tests may be burdensome or costly to a patient), and therapy (e.g. surgery with risks of morbidity and mortality) [31]. Such decisions are typically binary and require the definition of clinically relevant decision thresholds.

3. Traditional performance measures

We briefly consider some of the more traditionally used performance measures in medicine, without intending to be comprehensive (Table 1).

Overall performance measures

The distance between the predicted outcome and actual outcome is central to quantify overall model performance from a statistical modeler's perspective [32]. The distance is $Y - \hat{Y}$ for continuous outcomes. For binary outcomes, with Y defined 0 – 1, $\hat{Y}$ is equal to the predicted probability p, and for survival outcomes it is the predicted event probability at a given time (or as a function of time). These distances between observed and predicted outcomes are related to the concept of 'goodness-of-fit' of a model, with better models having smaller distances between predicted and observed outcomes. The main difference between goodness-of-fit and predictive performance is that the former is usually evaluated in the same data while assessment of the latter requires either new data or cross-validation.

Explained variation (R^2) is the most common performance measure for continuous outcomes. For generalized linear models, Nagelkerke's R^2 is often used [133]. This is a logarithmic scoring rule. For binary outcomes Y, we score a model with the logarithm of predictions p: $Y*\log(p) + (Y-1)*(\log(1-p))$. Nagelkerke's R^2 can also be calculated for survival outcomes, based on the difference in −2 log likelihood of a model without and a model with one or more predictors.

The Brier score is a quadratic scoring rule, where the squared differences between actual binary outcomes Y and predictions p are calculated: $(Y - p)^2$[234]. We can also write this similar to the logarithmic score: $Y*(1-p)^2 + (1-Y)*p^2$. The Brier score for a model can range from 0 for a perfect model to 0.25 for a non-informative model with a 50% incidence of the outcome. When the outcome incidence is lower, the maximum score for a non-

informative model is lower, e.g. for 10%: $0.1*(1–0.1)^2 + (1–0.1)*0.1^2 = 0.090$. Similar to Nagelkerke's approach to the LR statistic, we could scale Brier by its maximum score under a non-informative model: $Brier_{scaled} = 1 – Brier / Brier_{max}$, where $Brier_{max} = mean(p)*(1 – mean(p))$, to let it range between 0% and 100%. This scaled Brier score happens to be very similar to Pearson's R^2 statistic [35].

Calculation of the Brier score for survival outcomes is possible with a weight function, which considers the conditional probability of being uncensored during time [36373]. We can then calculate the Brier score at fixed time points, and create a time-dependent curve. It is useful to use a benchmark curve, based on the Brier score for the overall Kaplan-Meier estimator, which does not consider any predictive information [3]. It turns out that overall performance measures compose of two important characteristics of a prediction model, discrimination and calibration, each of which can be assessed separately.

Discrimination

Accurate predictions discriminate between those with and those without the outcome. Several measures can be used to indicate how well we classify patients in a binary prediction problem. The concordance (*c*) statistic is the most commonly used performance measure to indicate the discriminative ability of generalized linear regression models. For a binary outcome, *c* is identical to the area under the Receiver Operating Characteristic (ROC) curve, which plots the sensitivity (true positive rate) against 1 – (false positive rate) for consecutive cutoffs for the probability of an outcome.

The *c* statistic is a rank order statistic for predictions against true outcomes, related to Somers' D statistic [1]. As a rank order statistic, it is insensitive to systematic errors in calibration such as differences in average outcome. A popular extension of the *c* statistic with censored data can be obtained by ignoring the pairs that cannot be ordered [1]. It turns out that this results in a statistic that depends on the censoring pattern. Gonen and Heller have proposed a method to estimate a variant of the *c* statistic which is independent of censoring, but holds only in the context of a Cox proportional hazards model [7]. Furthermore, time-dependent *c* statistics have been proposed [638].

In addition to the *c* statistic, the discrimination slope can be used as a simple measure for how well subjects with and without the outcome are separated [39]. It is calculated as the absolute difference in average predictions for those with and without the outcome. Visualization is readily possible with a box plot or a histogram, which will show less overlap between those with and those without the outcome for a better discriminating model. Extensions of the discrimination slope have not yet been made to the survival context.

Calibration

Calibration refers to the agreement between observed outcomes and predictions [29]. For example, if we predict a 20% risk of residual tumor for a testicular cancer patient, the observed frequency of tumor should be approximately 20 out of 100 patients with such a prediction. A graphical assessment of calibration is possible with predictions on the x-axis, and the outcome on the y-axis. Perfect predictions should be on the 45° line. For linear regression, the calibration plot is a simple scatter plot. For binary outcomes, the plot contains only 0 and 1 values for the y-axis. Smoothing techniques can be used to estimate the observed probabilities of the outcome (p(y=1)) in relation to the predicted probabilities, e.g. using the loess algorithm [1]. We may however expect that the specific type of smoothing may affect the graphical impression, especially in smaller data sets. We can also plot results for subjects with similar probabilities, and thus compare the mean predicted probability to the mean observed outcome. For example, we can plot observed outcome by decile of

predictions, which makes the plot a graphical illustration of the Hosmer-Lemeshow goodness-of-fit test. A better discriminating model has more spread between such deciles than a poorly discriminating model. We note however that such grouping, though common, is arbitrary and imprecise.

The calibration plot can be characterized by an intercept *a*, which indicates the extent that predictions are systematically too low or too high ('calibration-in-the-large'), and a calibration slope *b*, which should be 1 [40]. Such a recalibration framework was already proposed by Cox [41]. At model development, $a=0$ and $b=1$ for regression models. At validation, calibration-in-the-large problems are common, as well as *b* smaller than 1, reflecting overfitting of a model [1]. A value of *b* smaller than 1 can also be interpreted as reflecting a need for shrinkage of regression coefficients in a prediction model [4243].

4. Novel performance measures

We now discuss some relatively novel performance measures, again without pretending to be comprehensive.

Novel measures related to reclassification

Cook proposed to make a 'reclassification table' to show how many subjects are reclassified by adding a marker to a model [8]. For example, a model with traditional risk factors for cardiovascular disease was extended with the predictors 'parental history of myocardial infarction' and 'CRP'. The increase in *c* statistic was minimal (from 0.805 to 0.808). However, when they classified the predicted risks into four categories (0–5, 5–10, 10–20, >20 per cent 10-year CVD risk), about 30% of individuals changed category when comparing the extended model with the traditional one. Change in risk categories, however, is insufficient to evaluate improvement in risk stratification; the changes must be appropriate. One way to evaluate this is to compare the observed incidence of events in the cells of the reclassification table to the predicted probability from the original model. Cook proposed a reclassification test as a variant of the Hosmer-Lemeshow statistic within the reclassified categories, leading to a chi-square statistic [44].

Pencina et al extended the reclassification idea by conditioning on the outcome: reclassi cation of subjects with and without the outcome should be considered separately [9]. Any 'upward' movement in categories for subjects with the outcome implies improved classi cation, and any 'downward movement' indicates worse reclassi cation. The interpretation is opposite for subjects without the outcome. The improvement in reclassi cation was quantified as the sum of differences in proportions of individuals moving up minus the proportion moving down for those with the outcome, and the proportion of individuals moving down minus the proportion moving up for those without the outcome. This sum was labeled the Net Reclassification Improvement (NRI). Also, a measure that integrates the NRI over all possible cut-offs for the probability of the outcome was proposed (integrated discrimination improvement, IDI) [9]. The IDI is equivalent to the difference in discrimination slopes of 2 models, and to the difference in Pearson R^2 measures [45], or the difference is scaled Brier scores.

Novel measures related to clinical usefulness

Some performance measures imply that false negative and false positive classifications are equally harmful. For example, the calculation of error rates is usually made by classifying subjects as positive when their predicted probability of the outcome exceeds 50%, and as negative otherwise. This implies an equal weighting of false-positive and false-negative classifications.

In the calculation of the NRI, the improvement in sensitivity and the improvement in specificity are summed. This implies relatively more weight for positive outcomes if a positive outcome was less common, and less weight if a positive outcome was more common than a negative outcome. The weight is equal to the non-events odds: (1-mean(p)) / mean(p), where mean(p) is the average probability of a positive outcome. Accordingly, although weighting in not equal, it is not explicitly based on clinical consequences. Defining the best diagnostic test as the one closest to the top left hand corner of the ROC curve – that is, the test with the highest sum of sensitivity and specificity (the Youden index: Se + Sp – 1, [46]) – similarly implies weighting by the non-events odds.

Vickers et al proposed decision curve analysis as a simple approach to quantify the clinical usefulness of a prediction model (or an extension to a model) [15]. For a formal decision analysis, harms and benefits need to be quantified, leading to an optimal decision threshold [47]. It may however often be difficult to define this threshold [15]. Difficulties may lie at the population level, i.e. that we do not have sufficient data on harms and benefits. Moreover, the relative weight of harms and benefits may differ from patient to patient, necessitating individual thresholds. Hence, we may consider a range of thresholds for the probability of the outcome, similar to ROC curves that consider the full range of cut-offs rather than a single cut-off for a sensitivity/specificity pair.

The key aspect of decision curve analysis is that a single probability threshold can be used both to categorize patients as positive or negative and to weight false positive and false negative classifications [48]. If we assume that the harm of unnecessary treatment (a false-positive decision) is relatively limited – such as antibiotics for infection - the cut-off should be low. In contrast, if overtreatment is quite harmful, such as extensive surgery, we should use a higher cut-off before a treatment decision is made. The harm to benefit ratio hence defines the relative weight w of false-positive decisions to true-positive decisions. For example, a cut-off of 10% implies that FP decisions are valued at 1/9th of a TP decision, and $w = 0.11$. The performance of a prediction model can then be summarized as a Net Benefit: $NB = (TP - w\,FP) / N$, where TP is the number of true positive decisions, FP the number of false positive decisions, N is the total number of patients and w is a weight equal to the odds of the cut-off ($p_t/(1-p_t)$, or the ratio of harm to benefit [48]. Documentation and software for decision curve analysis is publicly available (www.decisioncurveanalysis.org).

Validation graphs as summary tools

We may extent the calibration graph to a validation graph [20]. This entails that the distribution of predictions in those with and without the outcome is plotted at the bottom of the graph, capturing information on discrimination, similar to what is shown in a box plot. Moreover, it is important to have 95% confidence intervals around deciles (or other quantiles) of predicted risk to indicate uncertainty in the assessment of validity. From the validation graph we can learn the discriminative ability of a model (e.g. study the spread in observed outcomes by deciles of predicted risks), the calibration (closeness of observed outcomes to the 45 degree line), and the clinical usefulness (how many predictions are above or below clinically relevant thresholds).

5. Application to testicular cancer case study

Patients

Men with metastatic non-seminomatous testicular cancer can often be cured nowadays by cisplatin based chemotherapy. After chemotherapy, surgical resection is a generally accepted treatment to remove remnants of the initial metastases, since residual tumor may still be present. In the absence of tumor, resection has no therapeutic benefits, while it is associated with hospital admission, and risks of permanent morbidity and mortality. Logistic

regression models were developed to predict the presence of residual tumor, combining well-known predictors, such as the histology of the primary tumor, pre-chemotherapy levels of tumor markers, and (reduction in) residual mass size [49].

We first consider a data set with 544 patients to develop a prediction model that includes 5 predictors (Table 2). We then extend this model with the pre-chemotherapy level of the tumor marker lactate dehydrogenase (LDH). This illustrates ways to assess the incremental value of a marker. LDH values were log transformed, after standardizing by dividing by the local upper levels of normal values, after examination of nonlinearity with restricted cubic spline functions [50]. In a later study, we externally validated the 5 predictor model in 273 patients from a tertiary referral center, where LDH was not recorded [51]. This illustrates ways to assess the usefulness of a model in a new setting.

A clinically relevant cut-off for the risk of tumor was based on a decision analysis, where estimates from literature and from experts in the field were used to formally weigh the harms of missing tumor against the benefits of resection in those with tumor [52]. This analysis indicated that a risk threshold of 20% would be clinically reasonable.

Incremental value of a marker

Adding LDH to the 5 predictor model increased the model chi-square from 187 to 212 (LR statistic 25, $p<0.001$) in the development data set. LDH hence had statistically significant additional predictive value. Overall performance improved: Nagelkerke's R^2 increased from 39% to 43%, and the Brier score decreased from 0.17 to 0.16 (Table 3). The discriminative ability showed a small increase (*c* rose from 0.82 to 0.84, Fig 1). Similarly, the discrimination slope increased from 0.30 to 0.34 (Fig 2). The IDI hence was 4%.

Using a cut-off of 20% for the risk of tumor led to classification of 465 and 469 patients as at high risk for residual tumor with the original and extended models respectively (Table 4). The extended model reclassified 19 of the 465 patients as low risk (4%). On the other hand, 23 of 79 were reclassified as high risk while initially classified as low risk (29%). The total reclassification was hence 7.7% (42/544). Based on the observed proportions, those who were reclassified were placed into more appropriate categories. Cook's reclassification test was statistically significant (p=0.030), comparing predictions from the original model with observed outcomes in the 4 cells of Table 4. A more detailed assessment of the reclassification is obtained by a scatter plot with symbols by outcome (tumor or necrosis, Fig 3). We note especially that some patients with necrosis have higher predicted risks according to the model without LDH than according to the model with LDH (circles in right lower corner of the graph). The improvement in reclassification for those with tumor was 1.7% ((8-3)/299), and for those with necrosis 0.4% ((16–15)/245). The NRI hence was 2.1% [95% CI −2.9 to +7.0%], which is a much lower percentage than the 7.7% for all reclassified patients. The IDI was already estimated from Fig 1 as 4%.

A cut-off of 20% implies a relative weight of 1:4 for false-positive decisions against true-positive decisions. For the model without LDH, the Net Benefit was (TP – w*FP)/N = (284 – 0.25*(465-284))/544=0.439. If we would do resection in all, the NB would however be similar: (299 – 0.25*(544-299))/544=0.437. The model with LDH has a better NB: (289 0.25*(469-289))/544=0.449. Hence, at this particular cut-off, the model with LDH would be expected to lead to 1 more mass with tumor being resected per 100 patients at the same number of unnecessary resections of necrosis. The decision curve shows that the NB would be much larger for higher threshold values (Fig 4), i.e. patients accepting higher risks of residual tumor.

External validation

Overall model performance in the new cohort of 273 patients (197 with with residual tumor) was less than at development, according to R^2 and scaled Brier scores (25% instead of 39% and 20% instead of 30% respectively). Also, the *c* statistic and discrimination slope were poorer. Calibration was on average correct (calibration-in-the-large coefficient close to zero), but the effects of predictors were on average smaller in the new setting (calibration slope 0.74). The Hosmer-Lemeshow test was of borderline significance. The Net Benefit was close to zero, which was explained by the fact that very few patients had predicted risks below 20% and that calibration was imperfect around this threshold (Figs 2 and 5).

Software

All analyses were done in R version 2.8.1 (R Foundation for Statistical Computing, Vienna, Austria), using the Design library. The syntax is provided in the Appendix.

6. Discussion

This paper provided a framework for a number of traditional and relatively novel measures to assess the performance of an existing prediction model, or extensions to a model. Some measures relate to the evaluation of the quality of predictions, including overall performance measures such as explained variation and the Brier score, and measures for discrimination and calibration. Other measures quantify the quality of decisions, including decision-analytic measures such as the Net Benefit and decision curves, and measures related to reclassification tables (NRI, IDI).

Having a well discriminating model will commonly be most relevant for research purposes, such as covariate adjustment in a RCT. But a well discriminating model (e.g. *c* 0.8) may be useless if the decision threshold for clinical decisions is outside the range of predictions provided by the model. And a poorly discriminating model (e.g. *c* 0.6), may be clinically useful if the clinical decision is close to a “toss up” [53]. This implies that the threshold is right in the middle of the distribution of predicted risks, which is for example the case for models in fertility medicine [54]. For clinical practice, providing insight beyond the *c* statistic has been a motivation for some recent measures, especially in the context of extension of a prediction model with additional predictive information, e.g. from a biomarker [8945]. Many measures provide numerical summaries that may be difficult to interpret (see e.g. Table 3).

Evaluation of calibration is important if model predictions are used to inform patients or physicians to make decisions. The widely used Hosmer-Lemeshow test has a number of drawbacks, including limited power and poor interpretability [155]. Instead, the recalibration parameters as proposed by Cox (intercept and calibration slope) are more informative [41]. Validation plots with the distribution of risks for those with and without the outcome provide a useful graphical depiction, in line with previous proposals [45].

The net benefit, with visualization in a decision curve, is a simple summary measure to quantify clinical usefulness when decisions are to be supported by a prediction model [15]. We recognize however that other measures may give additional insights instead of providing a single summary measure. If a threshold is clinically well accepted, such as the 10% and 20% 10-year risks thresholds for cardiovascular events, reclassification tables and its associated measures may be particularly useful. For example, Table 4 clearly illustrates that LDH makes that a few more subjects with tumor are in the high risk category (289/299=97% instead of 284/299=95%) and one less subject without tumor is in the high risk category (180/245=73%. instead of 181/245=74%). This illustrated that key information for comparing performances of two models is contained in the margins of the reclassification tables [12].

In sum, we suggest that reporting discrimination and calibration will always be important for a prediction model. Decision-analytic measures should be reported if the predictive model is to be used for making clinical decisions. Other measures of performance may be warranted in specific applications, such as reclassification metrics to gain insight into the value of adding a novel predictor to an established model

A key issue in the evaluation of the quality of decisions is that false-positive and false-negative decisions will usually have quite different weight in medicine. Using equal weights for false-positive and false-negative decisions is 'absurd' in many medical applications [56]. Several measures of clinical usefulness have been proposed before which are consistent with decision-analytic considerations [483157585960].

We recognize that binary decisions can fully be evaluated in a ROC plot. The plot may however be obsolete unless the predicted probabilities at the operating points are indicated. Optimal thresholds can be defined by the tangent line to the curve, defined by the incidence of the outcome and the relative weight of false-positive and false-negative decisions [58]. If a prediction model is perfectly calibrated, the optimal threshold in the curve corresponds to the threshold probability in the Net Benefit analysis. The tangent is a 45 degree line if the outcome incidence is 50% and false-positive and false-negative decisions are weighted equally. We consider the Net Benefit and related decision curves preferable to graphical ROC curve assessment in the context of prediction models, although these approaches are obviously related [59].

Most performance measures can also be calculated for survival outcomes, which pose the challenge of dealing with censoring observations. Naïve calculation of ROC curves for censored observations can be misleading, since some of the censored observation would have had events if follow-up were longer. Also, the weight of false-positive and false-negative decisions may change with the follow-up time considered. Another issue is to consider competing risks in survival analyses of non-fatal outcomes, such as failure of heart valves [61], or mortality due to different causes [62]. Disregarding competing risks often leads to overestimation of absolute risk [63].

Any performance measure should be estimated with correction for optimism, as can e.g. be achieved with cross-validation or bootstrap resampling. To determine generalizability to other, plausibly related, settings, an external validation data set of sufficient size is required [18]. Some statistical updating may then be necessary for parameters in the model [64]. After repeated validation under different circumstances, an analysis of the impact of using a model for decision support should follow, which requires formulation of a model as a simple decision rule [65].

We have tried to sketch a framework for performance evaluation of predictions and decisions based on prediction models, both for newly developed or existing models, and for the situation of assessing the incremental value of a predictor such as a biomarker. Many more measures are available than discussed in this paper, which may have specific value in specific circumstances. The novel measures on reclassification and clinical usefulness can provide valuable additional insight on the value of prediction models and extensions to models, which goes beyond traditional measures of calibration and discrimination.

Acknowledgments

This paper was based on discussions at an international symposium "Measuring the accuracy of prediction models" (Cleveland, OH, Sept 29, 2008, http://www.bio.ri.ccf.org/html/symposium.html), which was supported by the Cleveland Clinic Department of Quantitative Health Sciences and the Page Foundation. We thank Dr Margaret

Pepe and Jessie Gu (University of Washington, Seattle, WA) for their critical review and helpful comments, as well as two anonymous reviewers.

References

1. Harrell, FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
2. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004; 159(9): 882–90. [PubMed: 15105181]
3. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008; 50(4): 457–79. [PubMed: 18663757]
4. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997; 16(9):965–80. [PubMed: 9160492]
5. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003; 229(1):3–8. [PubMed: 14519861]
6. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005; 61:92–105. [PubMed: 15737082]
7. Gonen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005; 92(4):965–970.
8. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115(7):928–35. [PubMed: 17309939]
9. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008; 27(2):157–72. discussion 207–12. [PubMed: 17569110]
10. Pepe MS, Janes H, Gu JW. Letter by Pepe et al regarding article, "Use and misuse of the receiver operating characteristic curve in risk prediction". Circulation. 2007; 116(6):e132. author reply e134. [PubMed: 17679623]
11. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Comments on 'Integrated discrimination and net reclassification improvements-Practical advice'. Stat Med. 2008; 27(2): 207–12.
12. Janes H, Pepe MS, Gu W. Assessing the Value of Risk Predictions by Using Risk Stratification Tables. Ann Intern Med. 2008; 149(10):751–760. [PubMed: 19017593]
13. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008; 168(21):2304–10. [PubMed: 19029492]
14. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009; 150(11):795–802. [PubMed: 19487714]
15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006; 26(6):565–74. [PubMed: 17099194]
16. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. Med Decis Making. 2008; 28(1):146–9. [PubMed: 18263565]
17. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000; 19(4):453–73. [PubMed: 10694730]
18. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999; 130(6):515–24. [PubMed: 10075620]
19. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54(8):774–81. [PubMed: 11470385]
20. Steyerberg, EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
21. Simon R. A checklist for evaluating reports of expression profiling for treatment selection. Clin Adv Hematol Oncol. 2006; 4(3):219–24. [PubMed: 16728933]

22. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008; 19(5):640–8. [PubMed: 18633328]
23. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics. 2007; 23(14):1768–74. [PubMed: 17485430]
24. Vickers AJ, Kramer BS, Baker SG. Selecting patients for randomized trials: a systematic approach based on risk group. Trials. 2006; 7:30. [PubMed: 17022818]
25. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol. 2004; 57(5):454–60. [PubMed: 15196615]
26. Hernandez AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann Epidemiol. 2006; 16(1):41–8. [PubMed: 16275011]
27. Iezzoni, LI. Risk adjustment for measuring health care outcomes. 3. Chicago: Health Administration Press; 2003.
28. Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst. 2003; 95(9):634–5. [PubMed: 12734304]
29. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med. 1978; 17(4):227–37. [PubMed: 366335]
30. Hand DJ. Statistical methods in diagnosis. Stat Methods Med Res. 1992; 1(1):49–67. [PubMed: 1341652]
31. Habbema JD, Hilden J. The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. Methods Inf Med. 1981; 20(2):80–96. [PubMed: 7017344]
32. Vittinghoff, E. Statistics for biology and health. New York: Springer; 2005. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models.
33. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika. 1991; (78):691–692.
34. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev. 1950; 78:1–3.
35. Hu B, Palta M, Shao J. Properties of R(2) statistics for logistic regression. Stat Med. 2006; 25(8): 1383–95. [PubMed: 16059870]
36. Schumacher M, Graf E, Gerds T. How to assess prognostic models for survival data: a case study in oncology. Methods Inf Med. 2003; 42(5):564–71. [PubMed: 14654892]
37. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J. 2006; 48(6):1029–40. [PubMed: 17240660]
38. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med. 2006; 25(20):3474–86. [PubMed: 16220486]
39. Yates JF. External correspondence: decomposition of the mean probability score. Org Beh Hum Perf. 1982; 30:132–156.
40. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making. 1993; 13(1):49–58. [PubMed: 8433637]
41. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958; 45:562–565.
42. Copas JB. Regression, prediction and shrinkage. J R Stat Soc, Ser B. 1983; 45(3):311–354.
43. van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med. 1990; 9(11): 1303–25. [PubMed: 2277880]
44. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008; 54(1):17–23. [PubMed: 18024533]
45. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008; 167(3): 362–8. [PubMed: 17982157]
46. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3(1):32–5. [PubMed: 15405679]

47. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980; 302(20):1109–17. [PubMed: 7366635]
48. Peirce CS. The numerical measure of success of predictions. Science. 1884; 4:453–454.
49. Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Toner GC, Schraffordt Koops H, Mulders PF, Messemer JE, Ney K, Donohue JP, et al. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol. 1995; 13(5):1177–87. [PubMed: 7537801]
50. Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JD. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Stat Med. 2001; 20(24):3847–59. [PubMed: 11782038]
51. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol. 2001; 165(1):84–8. [PubMed: 11125370]
52. Steyerberg EW, Marshall PB, Keizer HJ, Habbema JD. Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. Cancer. 1999; 85(6):1331–41. [PubMed: 10189139]
53. Pauker SG, Kassirer JP. The toss-up. N Engl J Med. 1981; 305(24):1467–9. [PubMed: 7300866]
54. Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. Hum Reprod. 2004; 19(9):2019–26. [PubMed: 15192070]
55. Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. J Clin Epidemiol. 2007; 60(5): 491–501. [PubMed: 17419960]
56. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). Stat Med. 2008; 27(2):199–206. [PubMed: 17729377]
57. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol. 2002; 20(2):96–107. [PubMed: 12012295]
58. McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med. 1975; 293(5):211–5. [PubMed: 806804]
59. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991; 11(2):95–101. [PubMed: 1865785]
60. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics. 2005; 6(2): 227–39. [PubMed: 15772102]
61. Grunkemeier GL, Jin R, Eijkemans MJ, Takkenberg JJ. Actual and actuarial probabilities of competing risks: apples and lemons. Ann Thorac Surg. 2007; 83(5):1586–92. [PubMed: 17462363]
62. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. JASA. 1999; 94:496–509.
63. Gail M. A review and critique of some models used in competing risk analysis. Biometrics. 1975; 31(1):209–22. [PubMed: 1164533]
64. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004; 23(16):2567–86. [PubMed: 15287085]
65. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006; 144(3):201–9. [PubMed: 16461965]

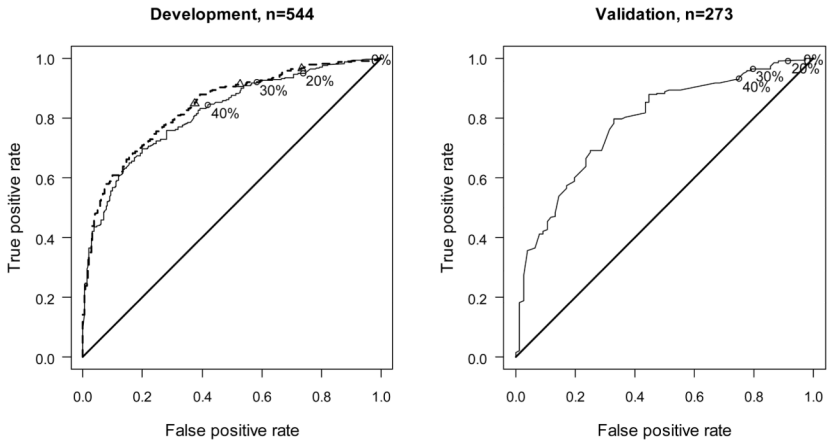

Fig 1.

Receiver operating characteristic (ROC) curves for the predicted probabilities without (solid line) and with the tumor marker LDH (dashed line) in the development data set (left) and for the predicted probabilities without the tumor marker LDH from the development data set in the validation data set (right). Threshold probabilities are indicated.

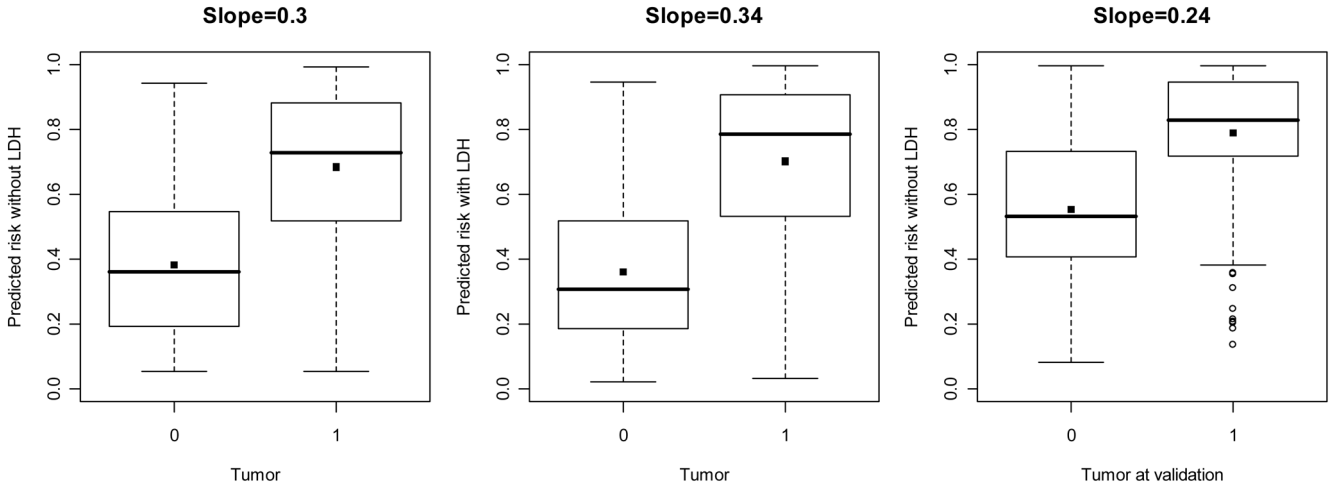

Fig 2.

Box plots of predicted probabilities without and with the tumor marker LDH. The discrimination slope is calculated as the difference between the mean predicted probability with and without residual tumor (solid dots indicate means). The difference between discrimination slopes is equivalent to integrated discrimination index (IDI=0.04).

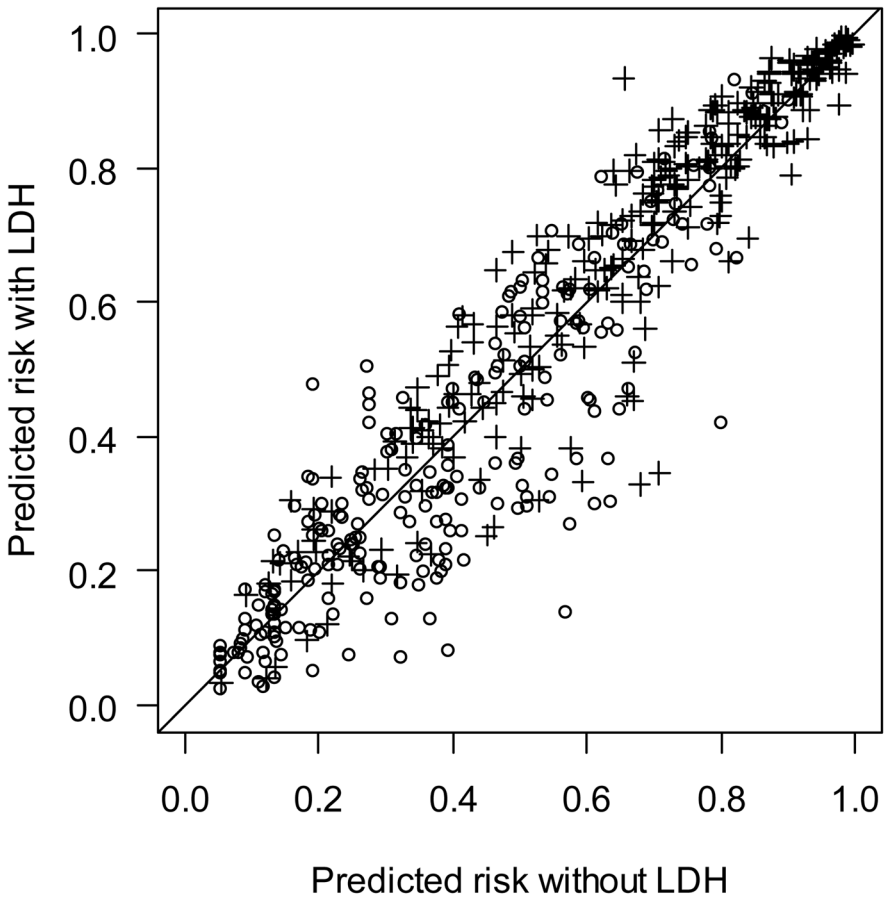

Fig 3.

Scatter plot of predicted probabilities without and with the tumor marker LDH (+: tumor; o: necrosis). Some patients with necrosis have higher predicted risks of tumor according to the model without LDH than according to the model with LDH (circles in right lower corner of the graph). For example, we note a patient with necrosis and an original prediction of nearly 60%, who is reclassified as less than 20% risk.

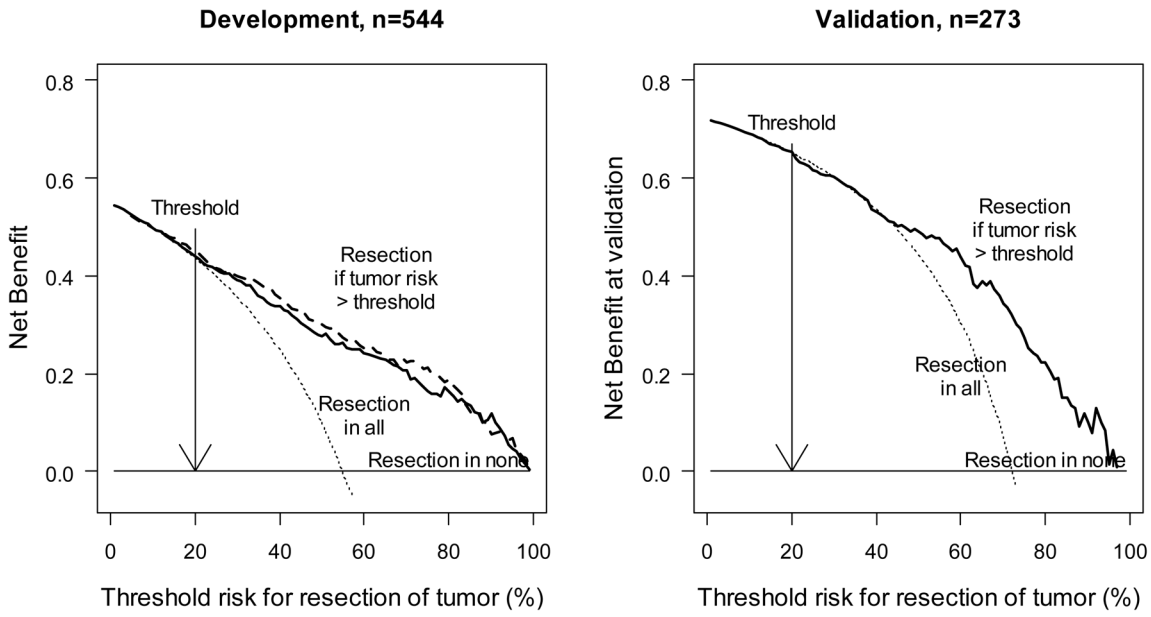

Fig 4.

Decision curves for the predicted probabilities without (solid line) and with the tumor marker LDH (dashed line) in the development data set (left) and for the predicted probabilities without the tumor marker LDH from the development data set in the validation data set (right).

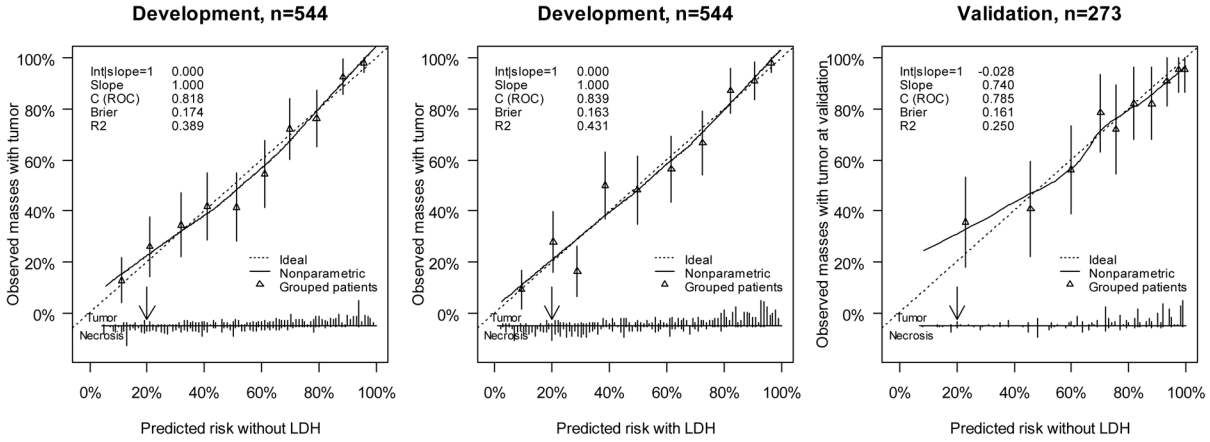

Fig 5.

Validation plots of prediction models for residual masses in patients with testicular cancer without and with the tumor marker LDH. The arrow indicates the decision threshold of 20% risk of residual tumor.

Table 1

Characteristics of some traditional and novel performance measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2 Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects.
Discrimination	C statistic	ROC curve	Rank order statistic; Interpretation for a pair of patients with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; Easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean(y) versus mean($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation related to 'shrinkage' of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification calibration		Compare observed and predicted within cross-classified categories
	Net Reclassification Index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right correction
	Integrated Discrimination Index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net Benefit (NB)	Cross-table	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)
	Decision curve analysis (DCA)	Decision curve	

Table 2

Logistic regression models in testicular cancer data set (n=544), without and with the tumor marker LDH. The outcome was residual tumor at postchemotherapy resection (299/544, 55%).

Characteristic	Without LDH	With LDH
Primary tumor teratoma-positive?	2.7 [1.8 – 4.0]	2.5 [1.6 – 3.8]
Prechemotherapy AFP elevated?	2.4 [1.5 – 3.7]	2.5 [1.6 – 3.9]
Prechemotherapy HCG elevated?	1.7 [1.1 – 2.7]	2.2 [1.4 – 3.4]
Square root of postchemotherapy mass size (mm)	1.08 [0.95 – 1.23]	1.34 [1.14 – 1.57]
Reduction in mass size per 10%	0.77 [0.70 – 0.85]	0.85 [0.77 – 0.95]
Prechemotherapy LDH (log(LDH/upper limit of local normal value))	-	0.37 [0.25 – 0.56]

Values are odds ratios with 95% confidence intervals. Continuous predictors were first studied with restricted cubic spline functions, and then simplified to simple parametric forms.

Table 3

Performance of testicular cancer models with or without the tumor maker LDH

	Development		External validation
Performance measure	**Without LDH**	**With LDH**	**Without LDH**
Overall			
Brier	0.174	0.163	0.161
$Brier_{scaled}$	29.8%	34.0%	20.0%
R^2 (Nagelkerke)	38.9%	43.1%	25.0%
Discrimination			
C stat	0.818 [0.78 – 0.85]	0.839 [0.81 – 0.87]	0.785 [0.73 – 0.84]
Discrimination slope	0.301	0.340	0.237
Calibration			
Calibration-in-the-large	0	0	−0.03
Calibration slope	1	1	0.74
H-L test	Chi-square 6.2, p=0.63	Chi-square 12.0, p=0.15	Chi-square 15.9, p=0.07
Clinical usefulness			
Net Benefit at threshold 20%*	0.2%	1.2%	0.1%

* compare to resect all

Table 4

Reclassification for the predicted probabilities without and with the tumor marker LDH in the development data set

		With LDH		
		Risk <=20%	**Risk > 20%**	**Total**
Without LDH	Risk <=20%	56	23	79
		7 tumor (12%)	8 tumor (35%)	15 tumor (19%)
	Risk > 20%	19	446	465
		3 tumor (16%)	281 tumor (63%)	284 tumor (61%)
	Total	75	469	544
		10 tumor (13%)	289 tumor (62%)	299 tumor (55%)