JAMA | Users' Guides to the Medical Literature

Discrimination and Calibration of Clinical Prediction Models
Users' Guides to the Medical Literature

Ana Carolina Alba, MD, PhD; Thomas Agoritsas, MD, PhD; Michael Walsh, MD, PhD; Steven Hanna, PhD; Alfonso Iorio, MD; P. J. Devereaux, MD, PhD; Thomas McGinn, MD; Gordon Guyatt, MD, MSc

Accurate information regarding prognosis is fundamental to optimal clinical care. The best approach to assess patient prognosis relies on prediction models that simultaneously consider a number of prognostic factors and provide an estimate of patients' absolute risk of an event. Such prediction models should be characterized by adequately discriminating between patients who will have an event and those who will not and by adequate calibration ensuring accurate prediction of absolute risk. This Users' Guide will help clinicians understand the available metrics for assessing discrimination, calibration, and the relative performance of different prediction models. This article complements existing Users' Guides that address the development and validation of prediction models. Together, these guides will help clinicians to make optimal use of existing prediction models.

JAMA. 2017;318(14):1377-1384. doi:10.1001/jama.2017.12126

Supplemental content

CME Quiz at jamanetwork.com/learning

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Ana Carolina Alba, MD, PhD, Toronto General Hospital, 585 University Ave, 6EN-246, Toronto, ON M5G 2N2, Canada (carolina.alba@uhn.ca).

Clinical Scenario

You are a general internist seeing an ambulatory consult. This new patient is a 54-year-old male with a history of hypertension treated with calcium channel blockers. He smokes and has a sedentary lifestyle but has not had any previous cardiovascular events. What is the risk of a cardiovascular event for this patient? Recent laboratory results show normal levels of total cholesterol (198 mg/dL) and low-density lipoprotein cholesterol (138 mg/dL), but a decreased level of high-density lipoprotein cholesterol (39 mg/dL). On physical examination, his systolic and diastolic blood pressure is 130 mm Hg and 80 mm Hg, respectively. Based on this information and using an online tool (Pooled Cohort Equations [modified from the Framingham risk score, which is recommended by the American Heart Association]), you calculate his risk of a cardiovascular event (myocardial infarction, stroke, or death due to coronary artery disease) to be 12.4% at 10 years.[1] Given this risk, current US, European, and Canadian guidelines recommend smoking cessation, regular physical activity, and initiation of statin therapy for primary prevention.[2]

When presented with the risk estimate, the patient is concerned, and although open to trying to stop smoking and exercise more, is reluctant to begin medication. Questioning the risk estimate, he points out that based on his family history of long survival without heart disease, the instrument has probably overestimated his risk. You are aware that another laboratory assay, N-terminal pro-B-type natriuretic peptide (NT-proBNP), may help to further differentiate patients at varying risk and consider whether this would help resolve the current dilemma. Before suggesting the test to the patient, you review the evidence regarding whether NT-proBNP will help to better categorize your patient's risk.

Why Clinicians Measure Prognosis

Accurate prognostic information is vitally important for patients and physicians to make optimal health-related and life decisions. For example, if a patient is at low risk of a future adverse event, the absolute benefit offered by an effective therapy may be small in relation to the potential harm, burden, and cost. Among higher-risk patients, the same treatment may offer substantial benefits. Thus, accurate prognostic assessment assists patients and physicians in the shared decision-making process, preventing testing in low-risk situations, and avoiding delays in treatment when there is a high probability of a favorable net benefit.

How Can Clinicians Estimate Prognosis?

There are several ways to estimate a patient's prognosis. First, it may sometimes be adequate for patients and clinicians to use their intuition. However, use of intuition is often limited; for example, patients with heart failure tend to overestimate their life expectancy,[3] particularly if they are younger and more symptomatic. In contrast, primary care physicians overestimate mortality risk in patients with heart failure, particularly in patients who are stable and mildly symptomatic.[4]

Second, clinicians might consider estimates of the mean prognosis from observational studies or randomized clinical trials of broad populations, or from meta-analyses of such studies. For example, a report from the registry of the International Society for Heart and Lung Transplantation described, in approximately 23 000 adults receiving their first heart transplant between 2006 and 2012, a 1- and 5-year mortality risk of 14% and 27%, respectively.[5] Such risk estimates may provide a starting point, but are limited by failure to consider factors (eg, age and the presence of comorbidities) that are likely to define the individual risk more accurately than the population average.

Box. Users' Guide to Assessing the Risk Prediction or Diagnostic Capability of a Prediction Model in a Validation Cohort

Do the authors provide information about both the discrimination and calibration of the model? If not, skepticism is warranted.

What is the discriminatory ability of the model? Does the model adequately discriminate between patients at varying risk? If a model does not adequately discriminate, do not use it.

How well calibrated is the model? Does a visual representation suggest a consistent good fit between predicted and observed outcomes in patients at different risk categories?

When comparing 2 models or a modification of an existing model (ie, addition of a new variable), does 1 of the models predict risk more accurately?

Third, clinicians may rely on studies that address such prediction factors. However, interpretation of the relative effects from these prognostic studies depends on accurate estimates of baseline risk. Many prognostic studies report only relative measures of association such as a relative risk or hazard ratio. For example, a large US cohort study reported that among patients with an implantable cardioverter-defibrillator, receiving a shock was associated with a doubling of mortality.[6] This study failed to report the absolute risk of dying. If the baseline 1-year mortality risk in patients who have an implantable cardioverter-defibrillator without shocks is 2%, receiving a shock increases the risk to 4% (an absolute increase of 2%), whereas if the baseline risk is 20%, receiving a shock increases the risk to 40% (an absolute increase of 20%). If physician-judged estimates of baseline risk differ substantially from the actual risk, the application of the relative effects associated with patient characteristics will be misleading.

Fourth, a more informative approach relies on models that simultaneously consider a number of prognostic factors (eg, age, sex, symptoms, signs, laboratory results) and provide an estimate of patients' absolute risk of an event (eg, prediction or prognostic rules, guides, or models to assess the risk of stroke[7] or death[8,9]) or the likelihood of a diagnosis (eg, diagnostic rules, guides, or models to aid in the diagnosis of Alzheimer disease[10] or pulmonary embolism[11] or to address the possibility of having a concerning intraabdominal diagnosis in patients visiting emergency departments[12]). Investigators developing such models use sophisticated statistical techniques to identify a manageable number of variables that provide estimates of a patient's likelihood of the target outcome or diagnosis of interest.

Although there are a wide variety of clinical prediction rules or models available, many are seldom used. Reasons for this limited uptake may relate to the absence of guidelines recommending their use and consequent physician unawareness,[13] complexity of the underlying clinical problem,[13] impracticality of a model or difficulty accessing the model,[14] or uncertainty of the effect on clinical management.[15]

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement[16] provides guidelines for researchers reporting on studies presenting a new model or validating an existing model. Previous Users' Guides[17,18] have (1) reviewed the steps in the development and validation of prediction models (eFigure in the Supplement); (2) provided an approach for judging the credibility of the prediction models (eBox in Supplement); and (3) focused on the available study designs, issues in implementation of the designs, and the appropriate inferences given the design, implementation, and results. Among the issues the guides have addressed is the optimal conduct of the studies in which authors developed their models. Clinicians particularly interested in this issue can, in addition to previous Users' Guides, find useful discussion elsewhere.[19]

Investigators can address the prognostic capability of their model at the time they first develop it or at the stage of initial or more robust validation. An important issue that previous Users' Guides have left largely unaddressed is evaluating the prediction capability of the model (ie, judging how well the model predicts the outcome). The purpose of this Users' Guide is to introduce clinicians to the available approaches for assessing the prognostic capability of prediction rules, outline the strengths and limitations of these approaches, and provide guidance on how to interpret the results (**Box**).

What Makes a Good Model?

An ideal but probably unrealistic model will correctly identify every patient who will develop an event vs those who will not, and will not misclassify any patients. The extent to which any model comes close to achieving this goal can be characterized by the 2 related properties of discrimination and calibration.

Discrimination refers to how well the model differentiates those at higher risk of having an event from those at lower risk. Discrimination depends on the distribution of patient characteristics in the population in which the model is being used. A model could well discriminate patients with events from those without events in a heterogenous population with widely different values of predictors included in the model (eg, age, sex, laboratory values); however, the same model could fail to discriminate patients in a more homogenous population. An extreme example is a model including age as a predictor that may work well in a population with a flat age distribution between 20 and 90 years, but fail in another population in which all patients are between the ages of 50 and 60 years.

However, discrimination alone is insufficient to assess a model's prediction capability. Consider a model that accurately characterizes patient 2 at double the risk of patient 1, and patient 3 at 3 times the risk of patient 1. Such a relative characterization may be correct, but will nevertheless be misleading if it predicts that patients 1, 2, and 3 have respective absolute risks of 1%, 2%, and 3% when in fact their true risks are 10%, 20%, and 30%.

Although discrimination alone is insufficient, clinicians should not use a model that fails to differentiate between those with higher risk and those with lower risk. When discrimination is poor, there is no need to further assess other model characteristics.

The second necessary property of a model informs clinicians how similar the predicted absolute risk is to the true (observed) risk in groups of patients classified in different risk strata. Calibration refers to the accuracy of absolute risk estimates. To the extent that the estimates are accurate, the model is well calibrated.

Calibration may be excellent in some patients, but not as good in others. For example, a model can be accurate at estimating risk for individuals in the range 0% to 20%, but overestimates risks in

individuals at higher risk. Such poor calibration among patients with higher risk may or may not be a problem. For instance, if the threshold for decision making is 10% (ie, an intervention is warranted if the risk is ≥10%), a model that overestimates risk by more than 20% would still be clinically applicable. The overestimates among patients at higher risk would be irrelevant. A useful model should have both satisfactory discrimination and calibration, but this example suggests that a superficial examination may provide a misleading inference about what is satisfactory.

Measuring Discrimination

Measures of discrimination address the extent to which a model predicts a higher probability of having an event among patients who will vs those who will not have an event. There are a number ways to assess discrimination. For binary (eg, dead or alive) outcomes, discrimination is typically characterized using the receiver operating characteristic (ROC) curve or C statistic.[20] After taking all possible pairs of patients and comparing the predicted probabilities, if the model cannot discriminate between which patients will have vs will not have an event, the C statistic (or ROC) is 0.5 (ie, no better than chance). If the model always produces a higher probability for patients having events vs those not having events, the C statistic is 1.0.[21-23]

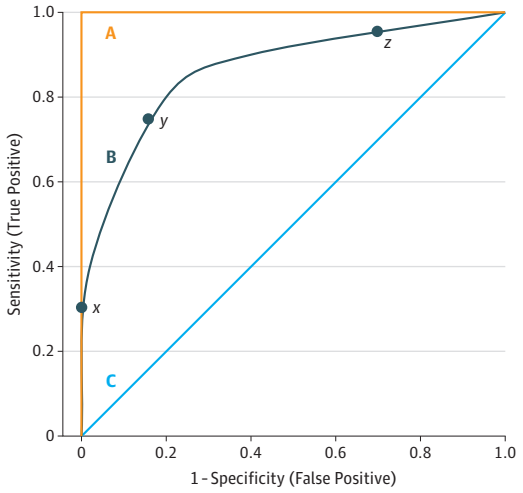
The true-positive and false-positive rates must be balanced against each other by choosing a cut point or threshold of the test result (or model score) for deciding whether a patient is at increased or decreased risk of experiencing an event. Choosing a cut point that increases the true-positive rate will increase the false-positive rate (and vice versa).

Figure 1 depicts the hypothetical relationship between a model's sensitivity and false-positive rate, which is called an ROC curve. This example illustrates the performance of a score derived from a regression model in which different weights (β coefficients) given to each predictor are added (ie, in a model with 3 predictors: predictor 1 × β_1 + predictor 2 × β_2 + predictor 3 × β_3). Looking at Figure 1, consider first curve B. Using a high threshold of a test result (ie, requiring a high risk score before predicting the patient will have an event), there will be no false-positive results; however, only 30% of those destined to have the event will be correctly identified (point *x* on curve B in Figure 1). Choosing a lower threshold increases sensitivity, but at the cost of more false-positive results (point *y*, sensitivity of 75%, 20% are false-positive results). The cost of a high sensitivity with an even lower threshold limits the model's ability to identify patients who will not experience an event (point *z*, sensitivity of 95%, 70% are false-positive results).

The greater the area under the ROC curve (the less is given up in false-positive results as sensitivity increases), the better the prediction model. A perfect ROC curve would have an area under the curve of 100%. The model will accurately classify all patients with events as higher risk relative to those without events (curve A in Figure 1). An ROC curve no better than chance would have an area under the curve of 50%. The model would have equal probability of assigning patients with vs those without events to a risk stratum (curve C in Figure 1).

Both the ROC curve and the C statistic represent the probability that the score or risk prediction is higher for a random patient with an event than for a random patient without an event,[24,25] which is a characterization that unfortunately fails to provide an intuitive sense of the value of the prediction guide. A generally accepted approach suggests that an area under the ROC curve or C statistic of less than 0.60 reflects poor discrimination; 0.60 to 0.75, possibly helpful discrimination; and more than 0.75, clearly useful discrimination.[21,22]

Figure 1. Receiving Operating Characteristic Curve and Area Under the Curve (AUC)



Three hypothetical curves are shown. Curve A represents the discriminatory capacity of the ideal test. The test or model correctly identifies all patients experiencing an event without misclassifying any patients who do not experience an event (AUC = 1.0). Curve B shows a more typical curve of a potentially useful test. The model correctly classifies more patients with events than misclassifies patients without events (AUC = 0.8). Curve C shows a test that it is no better than chance. The model has similar chances of correctly classifying patients with vs patients without events (AUC = 0.5). Point *x* on curve B represents a model with a high threshold for a positive result (ie, requiring a high-risk score before predicting the patient will have an event) and no false-positive results; however, only 30% of those destined to have the event will be correctly identified. Point *y* represents an example of a model with increased sensitivity but at the cost of more false-positive results (sensitivity of 75%, 20% are false-positive results). Point *z* represents a model with a high sensitivity (95%) but with limited ability to identify patients who will not experience an event (70% are false-positive results).

Such thresholds provide some notion of the prediction capability for existing models (a number of which are listed in the eTable in the Supplement) but are arbitrary to some extent. This is particularly true because the clinical implications of the probability of patients correctly vs incorrectly classified are specific to the clinical context. For example, it may (or may not) be more problematic to misclassify a patient with pulmonary embolism as being thrombosis-free than misclassifying a patient as having a pulmonary embolism.

A model can have excellent discrimination (and thus a high C statistic) and still provide misleading absolute risks. Correctly identifying a 5-fold gradient in risk is impressive, but if the model predicts risks of 1% and 5% in the 2 groups and the real risks are 10% and 50%, it is not useful. Such a model is referred to as being poorly calibrated.

For example, the European System for Cardiac Operative Risk Evaluation Score II (EuroSCORE II), a model to predict mortality associated with cardiac surgery, showed excellent discrimination (C static of 0.80) in a validation cohort.[26] However, it substantially

overestimated mortality in higher-risk patients. For instance, an individual undergoing elective surgery with a EuroSCORE II predicted mortality risk of 50% would have an actual risk of dying after surgery of 25%. This falsely high mortality risk may lead a patient to choose not to undergo surgery, whereas understanding the true risk may lead to the opposite decision. Thus, models need more than discriminatory power to function optimally, and clinicians need to know more than discrimination to judge their usefulness.

Critics of the ROC curve and the C statistic have further noted that they are insensitive to changes in absolute risk estimates.[27] The addition of a new predictor may contribute important prognostic and discriminatory power to a model; however, the C statistic may not increase substantially, particularly if the initial C statistic model is already high.

Measuring Calibration

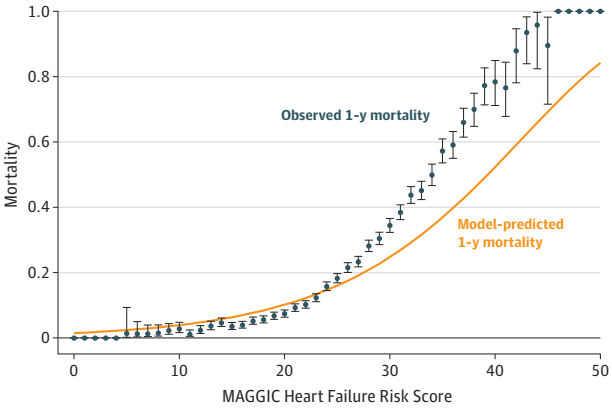
Calibration or goodness of fit is often considered the most important property of a model, and reflects the extent to which a model correctly estimates the absolute risk (ie, if the values predicted by the model agree with the observed values). Poorly calibrated models will underestimate or overestimate the outcome of interest.

Assessing calibration involves comparing predicted and observed risk at different levels[28] in (1) the whole population (mean calibration); (2) different groups of patients based on predicted risk; or (3) different groups of patients based on combinations of predictors (covariates). An excellent model will show strong calibration for different groups of patients with different characteristics; however, simulation exercises suggest that accurate calibration for different populations may be unrealistic.[28] Models that are well calibrated to predict average population risk could have poor discrimination and thus no clinical value. A model with adequate calibration by predicted risk strata will provide useful information for clinical decision making.

Consider the following example. Based on registry data, patients can expect a survival rate of approximately 85% at 1 year after heart transplantation.[5] Patients with advanced heart failure should have a predicted mortality of more than 15% prior to heart transplantation to provide a mortality benefit over the course of 1 year. The Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) risk score has been developed to predict mortality. A validation study showed good overall discrimination using the MAGGIC risk score among patients with heart failure as assessed by the C statistic (0.77).[29] However, among patients with an actual mortality of greater than 30%, the model underestimated mortality risk by 10% or greater (**Figure 2**). Some individuals in whom the model predicted mortality risk of 30% might still decide to continue with medical therapy and delay heart transplant (such individuals had an actual mortality of approximately 50%). Therefore, use of the model would lead some patients to inappropriately defer transplant.

Although the visual representation of the relationship between predicted and observed (Figure 2) is the best way to evaluate calibration, an alternative is to evaluate the difference between the predicted and the observed values using statistical tests (eg, the Hosmer-Lemeshow test) to determine whether chance can explain the difference between the predicted and the observed event rate. However, the Hosmer-Lemeshow test has limitations. A statistically significant result suggests that there is a difference between the predicted and the observed event rate that cannot be explained by chance, but gives no indication of the magnitude of the difference or whether (as in Figure 2) it varies among patients with low vs high risk.

Figure 2. Observed and Predicted 1-Year Mortality



The graph represents the relationship between observed (the data markers represent the mean and the error bars represent the 95% CI) and predicted 1-year mortality using the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) risk score (orange line). The model underestimated mortality in patients with a predicted mortality greater than 30%. Adapted from Sartipy et al.[29]

The Hosmer-Lemeshow test may be also misleading when the sample size is large. In this case, a clinically trivial difference between the predicted and the observed risk could lead to a statistically significant Hosmer-Lemeshow test result.

Discrimination and calibration are both important characteristics in the evaluation of model performance; however, they remain underreported in the published medical literature. A systematic review addressing prediction models of cardiovascular outcomes noted that only 63% reported on discrimination and only 36% reported on calibration.[30]

Comparison of 2 Models

The effectiveness of risk prediction models can be assessed in various ways. For example, the Akaike index criterion and the Bayesian index criterion provide statistics that rate model calibration as better or worse (a lower value using the Akaike or Bayesian index criterion suggests a better model fit or calibration), but provide no indication of how close the best model predicts the actual observed risk. Similarly, comparing C statistics does not provide insight into the relative merits of alternative models for predicting risk.

One popular method for assessing how much better a new model is at predicting risk than another is called net reclassification. Net reclassification involves classifying patients in risk categories and determining how a new model reclassifies patients into various risk categories compared with a previous model. Risk differences are classified based on the actual outcome patients experienced (ie, having vs not having the event of interest such as death or a stroke).[31]

Figure 3. Hypothetical Examples Illustrating Risk Reclassification Analysis

	Patients with events	Patients without events
Correct reclassification	$x^+ = a^+ + b^+ + c^+$	$x^- = d^- + e^- + f^-$
Incorrect reclassification	$y^+ = d^+ + e^+ + f^+$	$y^- = a^- + b^- + c^-$
Net reclassification	$z^+ = x^+ - y^+$	$z^- = x^- - y^-$

No reclassification
Correct reclassification
Incorrect reclassification

$$\text{Additive NRI} = \left(\frac{z^+}{\text{Total No. of patients with event}} \times 100\right) + \left(\frac{z^-}{\text{Total No. of patients without event}} \times 100\right)$$

$$\text{Absolute NRI} = \frac{(z^+ + z^-)}{\text{Total No. of patients}} \times 100$$

A Example 1

Patients with events (n=10 000, 50% of total)

OLD MODEL \ NEW MODEL	Low risk	Moderate risk	High risk	Totals, old model
Low risk	2500	1200 a^+	400 b^+	4100
Moderate risk	300 d^+	2500	600 c^+	3400
High risk	50 e^+	350 f^+	2100	2500
Totals, new model	2850	4050	3100	**10 000**

Patients without events (n=10 000, 50% of total)

OLD MODEL \ NEW MODEL	Low risk	Moderate risk	High risk	Totals, old model
Low risk	4000	300 a^-	100 b^-	4400
Moderate risk	100 d^-	4000	150 c^-	4250
High risk	50 e^-	100 f^-	1200	1350
Totals, new model	4150	4400	1450	**10 000**

	Patients with events, No.	Patients without events, No.		
Correct reclassification	2200	250	Additive NRI	12
Incorrect reclassification	700	550	Absolute NRI	6%
Net reclassification	1500	-300		

B Example 2

Patients with events (n=1000, 9% of total)

OLD MODEL \ NEW MODEL	Low risk	Moderate risk	High risk	Totals, old model
Low risk	110	120 a^+	40 b^+	270
Moderate risk	30 d^+	300	60 c^+	390
High risk	5 e^+	35 f^+	300	340
Totals, new model	145	455	400	**1000**

Patients without events (n=10 000, 91% of total)

OLD MODEL \ NEW MODEL	Low risk	Moderate risk	High risk	Totals, old model
Low risk	4000	300 a^-	100 b^-	4400
Moderate risk	100 d^-	3000	150 c^-	3250
High risk	50 e^-	100 f^-	2200	2350
Totals, new model	4150	3400	2450	**10 000**

	Patients with events, No.	Patients without events, No.		
Correct reclassification	220	250	Additive NRI	12
Incorrect reclassification	70	550	Absolute NRI	-1.36%
Net reclassification	150	-300		

Example 1 shows a hypothetical new model with an additive net reclassification index (NRI) of 12 and an absolute NRI of 6% based on a sample of patients who have equal proportions of having and not having events. Example 2 shows a hypothetical new model with an additive NRI of 12 and an absolute NRI of −1.36% based on better classification of patients with events but worse reclassification of patients without events. The gray cells on the diagonal represent patients retaining the same risk category in both the old and new model; the pink cells show patients incorrectly reclassified by the new model; and the green cells represent patients correctly reclassified by the new model.

The relative merit of the 2 models may be summarized by calculating (1) the number of patients who had the event of interest to whom the new model assigns a higher risk category than the prior model (a desirable reclassification), and (2) the number of patients who did not have the event to whom it assigns a lower risk category (also a desirable reclassification).[32]

There are different ways of summarizing the extent of reclassification. One of the most widely used summary statistics has been labeled as the net reclassification index (NRI). Calculating the NRI requires adding the percentage of patients having an event correctly reclassified (No. with the event who have a greater risk in the new model compared with the old model − No. with the event and

Figure 4. Risk Reclassification of Patients Undergoing Noncardiac Surgery Based on the Addition of Coronary Computed Tomographic Angiography (CCTA) to a Model That Predicts Postoperative Cardiovascular Death or Myocardial Infarction

Patients with myocardial infarction (n=74, 7.7% of total)

CCTA + RCRI

RCRI	Low risk	Moderate risk	High risk	Totals, old model
Low risk (<5%)	5	10 [a+]	0 [b+]	15
Moderate risk (5%-15%)	0 [d+]	41	7 [c+]	48
High risk (>15%)	0 [e+]	1 [f+]	10	11
Totals, new model	5	52	17	74

Patients without myocardial infarction (n=881, 92.3% of total)

CCTA + RCRI

RCRI	Low risk	Moderate risk	High risk	Totals, old model
Low risk (<5%)	191	114 [a-]	0 [b-]	305
Moderate risk (5%-15%)	47 [d-]	453	37 [c-]	537
High risk (>15%)	0 [e-]	10 [f-]	29	39
Totals, new model	238	577	66	881

	Patients with myocardial infarction, No.	Patients without myocardial infarction, No.		
Correct reclassification	17	57	Additive NRI	11
Incorrect reclassification	1	151	Absolute NRI	-8%
Net reclassification	16	-94		

The addition of CCTA to a risk prediction model results in an additive net reclassification index (NRI) of 11 and an absolute NRI of −8%. There is better classification of patients with events but worse reclassification of patients without events.[33] The gray cells on the diagonal represent risk stratification that is unchanged by the addition of CCTA to the model. The pink cells represent patients incorrectly reclassified by the addition of CCTA, whereas the green cells represent patients correctly reclassified by CCTA. RCRI indicates Revised Cardiac Risk Index.

a lower risk in the new vs to old model/total No. of patients with the event) to the percentage of patients who do not experience an event who are correctly reclassified (No. without the event and who have a lower predicted risk in the new model relative to the old model − No. without the event and a greater risk in the new vs the old model/total No. of patients without the event). This is the most commonly used method and is called the additive NRI (**Figure 3**).

The absolute NRI calculates the absolute number of patients correctly reclassified: net reclassification of patients with the event + net reclassification of patients without the event/total No. of patients.

The additive NRI can range from 200 (all patients with events had higher risk prediction and all patients without events had lower risk prediction with the new model) to −200 (the opposite) and its value does not represent a proportion, whereas the absolute NRI can range from −100% to 100% representing the proportion of patients incorrectly or correctly reclassified. The major limitation of the additive NRI is that it does not consider the prevalence of the events and nonevents in the population. Therefore, if a prediction model does better among patients who had an event in a sample in which the number of patients with events is small, but worse in a sample in which the number of patients without events is large, the results could be misleading. The absolute NRI avoids this problem.[33] The following examples illustrate this advantage.

Consider the hypothetical scenarios presented in Figure 3A in which there is an equal proportion of patients with and without events. In this scenario, the additive NRI will always be 2-fold the absolute NRI (12 and 6%, respectively). The absolute NRI represents the actual proportion reclassified better and is 6% in this case.

In Figure 3B, the proportion of patients with and without the event is 9% and 91%, respectively. The amount of reclassification is the same as in Figure 3A. There is an identical additive NRI of 12, favoring the new model. The new model does considerably better in the patients with events, but these patients now represent a far smaller group. The model still incorrectly classifies patients without events, but these patients are now a much larger group. The absolute NRI results in −150/11 000 or −1.34% of patients. When the incidence of events is low, the additive NRI can be misleading, whereas the absolute NRI better reflects the model's limitations.

The use of NRI and the pitfalls associated with additive NRIs can be illustrated from the results of a study[33] evaluating the benefit of adding the coronary computed tomographic angiography (CCTA) score to a perioperative score (Revised Cardiac Risk Index) that predicts the postoperative risk of myocardial infarction or cardiovascular death. In this study, 955 patients underwent noncardiac surgery and 74 patients (7.7%) experienced cardiovascular death or myocardial infarction.[33] **Figure 4** shows the risk reclassification of patients with and without events using the Revised Cardiac Risk Index and also when CCTA was added. The addition of CCTA improved classification of patients with events; however, it worsened the classification of patients without events. The additive NRI of 11 suggests improved prediction when CCTA was added to the Revised Cardiac Risk Index score. Because there were more patients (92.3% vs 7.7%) without a myocardial infarction or cardiovascular death, the absolute NRI is −78 of 955 or approximately −8%. The overall reclassification by adding CCTA to the Revised Cardiac Risk Index worsened the risk prediction in 8% of the patients compared with using the Revised Cardiac Risk Index alone.

The absolute NRI assumes that the consequences of having an event are the same as the consequences of not having an event. This may not always be the case. A refinement of NRI analysis accounts for this.[34] If the misclassification of patients with the event leads to more serious consequences than the misclassification of those without the event, one might value the model that does better classifying patients with the event, and place less weight on the classification of patients without the event.

For instance, consider the example in Figure 4. Of the 955 patients, 74 experienced myocardial infarction or cardiovascular death

and 16 patients were better classified by the model including CCTA. Of the 881 who did not have an event, 94 were better classified without CCTA. For CCTA to be judged equally good overall, a correct risk classification in a patient destined to have a myocardial infarction or die would have to be judged approximately 6 times more important than misclassifying a patient destined not to experience an event. For CCTA to be judged superior, the deleterious consequences associated with a false-negative result would have to be more than 6 times as important as those associated with a false-positive result.

There are some metrics (eg, relative utility, standardized net benefit) that facilitate the quantification of the relative clinical effect of false-positive and false-negative misclassifications.[35] These metrics are not yet widely used.

In summary, the total number of patients misclassified is likely to be informative, but clinicians should also consider the degree to which misclassification of patients with and without events is problematic.

The reclassification analysis further informs the value of the additional test. The additive NRI was 9.4; however, the additional use of NT-proBNP resulted in incorrect reclassification in approximately 146 patients without events (3.3%), which was the larger group of patients, and correct reclassification in approximately 40 patients (12.7%) who had an event. Thus, the absolute NRI was approximately −2.4%. These results demonstrate that the measurement of NT-proBNP will, relative to using only Pooled Cohort Equations, increase rather than decrease the absolute number of patients misclassified.

Does this necessarily mean a physician should not use the additional NT-proBNP results in predicting cardiovascular events? The answer is probably, but not necessarily. If a physician puts a much higher value on a correct prediction among patients destined to have events (adding NT-proBNP will result in fewer false-negative results) than among those destined not to have events (adding NT-proBNP will result in more false-positive results), then he or she might still consider using the additional information from the test.

Resolving the Clinical Scenario

Returning to our clinical scenario, a study[36] that enrolled individuals free of cardiovascular disease reported that NT-proBNP independently predicted cardiovascular risk vs the Pooled Cohort Equations. The C statistic generated using Pooled Cohort Equations was 0.753 and increased to 0.757 after NT-proBNP was added to the model, suggesting slight improvement in discrimination. The study did not report calibration.

Conclusions

Considering the limited valuable prognostic information provided by NT-proBNP and the patient inclination to avoid unnecessary statin therapy, the final decision is against additional testing (relatively more value in avoiding false-positive results). You accept your patient's preferences and values as reflected in his reluctance to start statin therapy, and strongly recommend that he quit smoking and engage in regular exercise.

ARTICLE INFORMATION

Accepted for Publication: August 15, 2017.

Author Affiliations: Heart Failure and Transplant Program, Toronto General Hospital, University Health Network, Toronto, Ontario, Canada (Alba); Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada (Agoritsas, Walsh, Hanna, Iorio, Devereaux, Guyatt); Divisions of Clinical Epidemiology and General Internal Medicine, University Hospitals of Geneva, Geneva, Switzerland (Agoritsas); Feinstein Institute for Medical Research, Northwell School of Medicine, Hofstra University, Hempstead, New York (McGinn).

Author Contributions: Dr Alba had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Alba, Walsh, Hanna, Iorio, Devereaux, McGinn, Guyatt.
Acquisition, analysis, or interpretation of data: Alba, Agoritsas, Iorio.
Drafting of the manuscript: Alba, Agoritsas, McGinn.
Critical revision of the manuscript for important intellectual content: Alba, Walsh, Hanna, Iorio, Devereaux, Guyatt.
Obtained funding: Alba.
Administrative, technical, or material support: Alba, McGinn.
Supervision: Guyatt.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Devereaux reported receiving grant funding from Abbott Diagnostics, Boehringer Ingelheim, Covidien, Octapharma, Roche Dignostics, and Stryker. No other disclosures were reported.

REFERENCES

1. Stone NJ, Robinson JG, Lichtenstein AH, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014;129(25)(suppl 2):S1-S45.

2. Nayor M, Vasan RS. Recent update to the US cholesterol treatment guidelines: a comparison with international guidelines. *Circulation*. 2016;133(18):1795-1806.

3. Allen LA, Yager JE, Funk MJ, et al. Discordance between patient-predicted and model-predicted life expectancy among ambulatory patients with heart failure. *JAMA*. 2008;299(21):2533-2542.

4. Muntwyler J, Abetel G, Gruner C, Follath F. One-year mortality among unselected outpatients with heart failure. *Eur Heart J*. 2002;23(23):1861-1866.

5. Lund LH, Edwards LB, Kucheryavaya AY, et al. The registry of the International Society for Heart and Lung Transplantation: thirty-second official adult heart transplantation report—2015; focus theme: early graft failure. *J Heart Lung Transplant*. 2015;34(10):1244-1254.

6. Saxon LA, Hayes DL, Gilliam FR, et al. Long-term outcome after ICD and CRT implantation and influence of remote device follow-up: the ALTITUDE survival study. *Circulation*. 2010;122(23):2359-2367.

7. Oldgren J, Hijazi Z, Lindbäck J, et al; RE-LY and ARISTOTLE Investigators. Performance and validation of a novel biomarker-based stroke risk score for atrial fibrillation. *Circulation*. 2016;134(22):1697-1707.

8. Klein KB, Stafinski TD, Menon D. Predicting survival after liver transplantation based on pre-transplant MELD score: a systematic review of the literature. *PLoS One*. 2013;8(12):e80661.

9. Regoli F, Scopigni F, Leyva F, et al; Collaborative Study Group. Validation of Seattle Heart Failure Model for mortality risk prediction in patients treated with cardiac resynchronization therapy. *Eur J Heart Fail*. 2013;15(2):211-220.

10. Sala I, Illán-Gala I, Alcolea D, et al. Diagnostic and prognostic value of the combination of two measures of verbal memory in mild cognitive impairment due to Alzheimer's disease. *J Alzheimers Dis*. 2017;58(3):909-918.

11. Shen J-H, Chen H-L, Chen J-R, Xing J-L, Gu P, Zhu B-F. Comparison of the Wells score with the revised Geneva score for assessing suspected pulmonary embolism: a systematic review and meta-analysis. *J Thromb Thrombolysis*. 2016;41(3):482-492.

12. Aaronson EL, Chang Y, Borczuk P. A prediction model to identify patients without a concerning intraabdominal diagnosis. *Am J Emerg Med*. 2016;34(8):1354-1358.

13. Plüddemann A, Wallace E, Bankhead C, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. *Br J Gen Pract*. 2014;64(621):e233-e242.

14. McGinn T. Putting meaning into meaningful use: a roadmap to successful integration of evidence at the point of care. *JMIR Med Inform*. 2016;4(2):e16.

15. Wallace E, Uijen MJ, Clyne B, et al. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. *BMJ Open*. 2016;6(3):e009957.

16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594

17. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS; Evidence-Based Medicine Working Group. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. *JAMA*. 2000;284(1):79-84.

18. McGinn T, Wyer PC, McCullagh L, et al. Clinical prediction rules. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature*. New York, NY: McGraw-Hill; 2014.

19. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.

20. Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. *JAMA*. 2015;314(10):1063-1064.

21. Hosmer DW, Lemeshow S. Assessing the fit of the model. In: Hosmer DW, Lemeshow S, eds. *Applied Logistic Regression*. 2nd ed. New York, NY: John Wiley & Sons; 2000:143-202.

22. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. *Handbook of Statistics v23: Advances in Survival Analysis*. Amsterdam, the Netherlands: Elsevier; 2004.

23. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J*. 2008;50(4):457-479.

24. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

25. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005;61(1):92-105.

26. Barili F, Pacini D, Rosato F, et al. In-hospital mortality risk assessment in elective and non-elective cardiac surgery: a comparison between EuroSCORE II and age, creatinine, ejection fraction score. *Eur J Cardiothorac Surg*. 2014;46(1):44-48.

27. 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.

28. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176.

29. Sartipy U, Dahlström U, Edner M, Lund LH. Predicting survival in heart failure: validation of the MAGGIC heart failure risk score in 51,043 patients from the Swedish Heart Failure Registry. *Eur J Heart Fail*. 2014;16(2):173-179.

30. Wessler BS, Lai Yh L, Kramer W, et al. Clinical prediction models for cardiovascular disease: Tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circ Cardiovasc Qual Outcomes*. 2015;8(4):368-375.

31. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.

32. 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):157-172.

33. Sheth T, Chan M, Butler C, et al; Coronary Computed Tomographic Angiography and Vascular Events in Noncardiac Surgery Patients Cohort Evaluation Study Investigators. Prognostic capabilities of coronary computed tomographic angiography before non-cardiac surgery: prospective cohort study. *BMJ*. 2015;350:h1907.

34. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? *Ann Intern Med*. 2014;160(2):136-137.

35. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.

36. Daniels LB, Clopton P, deFilippi CR, et al. Serial measurement of N-terminal pro-B-type natriuretic peptide and cardiac troponin T for cardiovascular disease risk assessment in the Multi-Ethnic Study of Atherosclerosis (MESA). *Am Heart J*. 2015;170(6):1170-1183.