Review Article | Experimental and Others

http://dx.doi.org/10.3348/kjr.2016.17.3.339
pISSN 1229-6929 · eISSN 2005-8330
Korean J Radiol 2016;17(3):339-350

Korean Journal of Radiology
KJR

How to Develop, Validate, and Compare Clinical Prediction Models Involving Radiological Parameters: Study Design and Statistical Methods

Kyunghwa Han, PhD[1], Kijun Song, PhD[2], Byoung Wook Choi, MD, PhD[1]

[1]Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, Korea; [2]Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, Seoul 03722, Korea

Clinical prediction models are developed to calculate estimates of the probability of the presence/occurrence or future course of a particular prognostic or diagnostic outcome from multiple clinical or non-clinical parameters. Radiologic imaging techniques are being developed for accurate detection and early diagnosis of disease, which will eventually affect patient outcomes. Hence, results obtained by radiological means, especially diagnostic imaging, are frequently incorporated into a clinical prediction model as important predictive parameters, and the performance of the prediction model may improve in both diagnostic and prognostic settings. This article explains in a conceptual manner the overall process of developing and validating a clinical prediction model involving radiological parameters in relation to the study design and statistical methods. Collection of a raw dataset; selection of an appropriate statistical model; predictor selection; evaluation of model performance using a calibration plot, Hosmer-Lemeshow test and c-index; internal and external validation; comparison of different models using c-index, net reclassification improvement, and integrated discrimination improvement; and a method to create an easy-to-use prediction score system will be addressed. This article may serve as a practical methodological reference for clinical researchers.

Index terms: *Prediction model; Prognosis; Diagnosis; Patient outcome*

INTRODUCTION

In medicine, improving prognosis is the ultimate goal of all diagnostic and therapeutic decisions. Prognosis commonly relates to the probability or risk of an individual developing a particular state of health. Health care providers need to make decisions to order tests for diagnostic work-up or decisions relating to therapeutics such as starting or delaying treatments, surgical decision-making, and changing or modifying the intensity of treatments. A clinical prediction model (often referred to as the clinical prediction rule, prognostic model, or risk score) is developed to calculate estimates of the probability of the presence/occurrence or future course of a particular patient outcome from multiple clinical or non-clinical predictors in order to help individualize diagnostic and therapeutic decision-making in healthcare practice (1).

For example, the Framingham risk score is well known in public health as a prediction model used to estimate the probability of the occurrence of a cardiovascular disease in an individual within 10 years. It is calculated using traditional risk factors such as age, sex, systolic blood pressure, hypertension treatment, total and high-density lipoprotein cholesterol levels, smoking, and

Received September 24, 2015; accepted after revision January 14, 2016.

Corresponding author: Kijun Song, PhD, Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
• Tel: (822) 2228-2491 • Fax: (822) 364-8037
• E-mail: biostat@yuhs.ac

diabetes (2). In clinical practice and medical research, baseline characteristics and laboratory values have been commonly used to predict patient outcome. For instance, risk estimation for the hepatocellular carcinoma in chronic hepatitis B score was developed using sex, age, serum alanine aminotransferase concentration, hepatitis B e antigen status, and serum hepatitis B virus DNA level (3).

In radiology practice and research, diagnostic imaging tests have traditionally been viewed and evaluated by assessing their performance to diagnose a particular target disease state. For example, ultrasound features of thyroid nodules have been used in risk stratification of thyroid nodules (4). Coronary computed tomography (CT) angiography (CCTA) has been evaluated for the selection of coronary artery bypass graft candidates (5). Recently, it appears that imaging findings are also often used as predictive parameters, either for standalone prediction or as an addition to the traditional clinical prediction models. For example, cardiac CT, including coronary artery calcium scoring and CCTA, provides prognostic information regarding mortality or disease recurrence and is expected to improve risk stratification of coronary artery disease (CAD) beyond clinical risk factors (6-9).

In evaluating diagnostic imaging tests, patient health outcomes can measure higher levels of efficacy than diagnostic accuracy (10). Radiologic imaging techniques are being developed for accurate detection and early diagnosis, which will eventually affect patient outcomes. Hence, with results attained through radiological means, especially diagnostic imaging results, being incorporated into a clinical prediction model, the predictive ability of the model may improve in both diagnostic and prognostic settings. In this review, we aim to explain conceptually the process for development and validation of a clinical prediction model involving radiological parameters with regards to the study design and statistical methods.

Diagnosis and Prognosis

There are several similarities between diagnostic and prognostic prediction models (11). The type of outcome is often binary; either the disease is present or absent (in diagnosis) or the future event occurs or does not occur (in prognosis). The key interest is in generating the probability of the outcome occurring for an individual. Estimates of probabilities are rarely based on a single predictor, and combinations of multiple predictors are used; therefore, prediction is inherently multivariable. The same challenges and measures exist for developing and assessing the prediction model. This model can be developed for either ill or healthy individuals.

The main difference is the time at which outcomes are evaluated. In diagnosis, outcome is evaluated at the same time of prediction for individuals with a suspected disease. In prognosis, models estimate the probability of a particular outcome, such as mortality, disease recurrence, complications, or therapy response, occurring in a certain period in the future; therefore, collecting follow-up data is more important. Consequently, diagnostic modeling studies involve a cross-sectional relationship, whereas prognostic modeling studies involve a longitudinal relationship.

Study Design for Raw Dataset

The performance of a statistical model depends on the study design and the quality of the analyzed data. Various study designs can be used to develop and validate a prediction model. Diagnostic accuracy studies are often designed as cross-sectional studies in which, typically, the diagnostic test results and the results of a reference test are compared to establish the ground truth regarding the presence or absence of the target disease performed for a group of subjects in a short duration. Prospective studies using a pre-specified protocol for systematic diagnostic work-up and reference standard testing in a well-defined clinical cohort, i.e., a cohort-type diagnostic accuracy study, are preferred to retrospective studies so as to minimize incomplete test results and/or assessment bias. Case-control-type accuracy studies can also be applied in prediction model studies (12). However, patient sampling for the presence or absence of the target disease and including non-consecutive patients leads to selection bias and loss of generalizability.

The best design for prognostic studies is a cohort study. Healthy or symptomatic participants are enrolled in the cohort at a certain time interval and are followed over time in anticipation of the outcome or event of interest. Such studies can be prospective or retrospective; the preferred design is a prospective longitudinal cohort study because it is efficient to optimally control and measure all predictors and outcomes and minimize the number of missing values and those lost to follow-up. Alternatively, retrospective cohort studies are often performed with existing databases such as hospital records systems or registries. In these cases, it is possible to have longer follow-up times, and obtaining records incurs a relatively lower cost than in the prospective design. However, some information is often missing or incorrect, leading to a selection bias.

Randomized clinical trials are a subtype of the prospective cohort design, and thus can also be used for prognostic models. However, the main disadvantages may be in the selection of patients. Participants are enrolled strictly according to inclusion/exclusion criteria and receipt of informed consent, decreasing the generalizability of this model.

Candidate predictors include patient demographics, disease type and severity, history characteristics, physical examination, biomarkers, and tests results. Predictors should be reliably measured and well defined by any observer. Inter-observer variability caused by subjective interpretation is a specific concern when imaging test results are involved as predictors. Predictor assessors should always be blinded to the outcome and vice versa, particularly if the predictors involve subjective interpretation of an imaging test or pathology results, in order to prevent potential bias in estimation of the association between predictors and outcome.

The outcomes of a prediction model are preferably focused on patient-relevant outcomes. In diagnostic modeling, the outcome is the presence or absence of a specific target disease state (i.e., reference standard information) at the time of the diagnostic test. For prognostic models, common outcomes relate to mortality (e.g., all-cause or cause-specific), non-fatal events (e.g., disease progression, tumor growth, and cardiovascular events), and patient-centered outcomes (e.g., symptoms, and health-related quality of life).

Development of Prediction Model

Sample Size for Model Derivation

It is important to have an adequate sample size when developing a prediction model; however, what constitutes an "adequate" sample size is unclear. In medical research, a larger sample size will yield more findings of high reliability. When the number of predictors is much larger than the number of outcome events, there is a risk of overestimating/overfitting the predictive performance of the model. In principle, the sample size could be estimated regarding a precision of metrics of prediction model performance (R^2 or c-index to be discussed later). Generally, the smaller number of the binary outcome (events or non-events) dictates the effective sample size in prediction studies. From some empirical simulation studies, a rule of thumb for sample size has been suggested (13, 14) in which at least 10 events (or non-events depending on which side is smaller) are required per candidate predictor, although other investigations have found the value of 10 to be too strict (15) or, conversely, too lax (16). For example, when the number of events is 400 out of 1000, we can consider as many as 40 (= 400/10) variables as candidate predictors in developing the prediction model. It is often the case that a data set may already be readily available from a large cohort or registry; it would make sense to use the entire data set for maximum power and generalizability of the results, regardless of whether it meets specific sample size calculations.

Statistical Model Selection

As mentioned previously, prognostic studies are inherently multivariable. Hence, the most frequently used approach is multivariable regression modeling. A linear regression can be applied to predict a continuous outcome; a logistic regression model is commonly used to predict a binary endpoint; and a Cox proportional hazards regression model is used for time-to-event outcomes. The logistic regression is usually used in diagnostic models or short-term prognostic events (e.g., 30-day mortality), and the Cox regression is used for long-term prognostic outcomes (e.g., 10-year cardiovascular disease risk).

Considerations in Selecting Predictors

Predictor selection can be performed before and during modeling. Before modeling, a set of variables could be selected by considering clinical reasoning (i.e., commonly applied in clinical practice), costs or burden of measuring, relevant literature (e.g., a systematic review of the literature), and/or knowledge of experts in the field. For continuous variables, categorization (e.g., dichotomization) is commonly used for user (e.g., clinician) convenience or easier presentation of prediction models. However, information loss or various results from different cut points are non-ignorable. A data-driven cut point (e.g., mean or median of the predictor as optimal cut point) may produce a biased regression coefficient (17). Thus, researchers should carefully use the categorization of predictor values and should explain the rationale for any categorization.

After preselection of predictors, candidate variables can be selected based on univariable (unadjusted) association with the outcome. However, this approach may reject important predictors due to confounding by other predictors in the model. Non-significance does not imply that there is evidence for a zero effect of a predictor; in other words, absence of evidence is not evidence of absence (18). Automated variable selection methods, such as forward selection, backward elimination, or stepwise selection, can be implemented in some statistical packages. Usually, automated variable selection methods make a model smaller. However, different criteria for predictors to be included or excluded or different statistics of model fit (i.e., F statistic or Akaike information criterion) are used in the different statistical packages and, along with the random variability that exists in the data, means that the selection of predictors may be unstable (19). Alternatively, repeating the automated variable selection process by bootstrapping, i.e., most frequently selected (e.g., at least 50% or 75% of the bootstrap samples) variables within bootstrap samples being included in the prediction model, may identify true predictors (20). Austin (21) showed that the bootstrapping-based method tended to have similar variable selection to backward variable elimination with a significance level of 0.05 for variables retention.

Multicollinearity is a relevant issue in regression modeling because it affects the reliable estimation of each predictor's estimate. However, this is not relevant for adequate reliability of the prediction model as a whole. The purpose of multivariable prediction modeling is to predict outcome with consideration of the joint effects of predictors that are correlated with each other. Prediction is about estimation rather than risk factor testing, and thus it is quite reasonable to include all clinically relevant predictors in the prediction model despite non-significant univariable association or multicollinearity (1).

Missing values, for either predictors or outcomes, often occur in clinical prediction research. There are several approaches to deal with missing values, such as complete case analysis or imputation. Traditional "complete case" or "available case" analyses lead to selection bias of subjects and statistically inefficient results. A more effective method can be used, the so-called imputation approach, in which the missing value is replaced by a mean or median, or replaced by a predicted value from the observed data. It is related to observed variables, and thus assumes a missing at random mechanism. Multiple imputations can be performed to incorporate uncertainty in the imputed values. For example, in Rubin's method (22), multiple simulated with imputation data sets are analyzed by standard methods and then the results (e.g., regression coefficients and predictive performance measures) are combined to produce overall estimates.

Assumption for Model Development

Unlike linear regression, logistic regression does not require many assumptions such as normality, homogeneity of variance, and linearity between predictors and outcome. When modeling a Cox proportional hazard regression, a key assumption is the proportionality of hazard, i.e., the predictor effects being constant over time. There are a number of approaches for testing proportional hazard–for instance, Kaplan-Meier curves for categorical predictors with few levels, testing for time-dependent covariates, or using the scaled Schoenfeld residuals. When the assumptions were not met, other approaches can be considered, such as transformation (e.g., log transformation) for the predictors, nonlinear modeling (e.g., restricted cubic spline), or stratified analysis according to the variables that did not satisfy the assumptions.

Data Example

We illustrate the process of prediction model development with artificial data from a retrospective cohort study aiming to evaluate the findings of CCTA and related patient outcomes (e.g., all-cause mortality).

The cohort consists of 960 consecutive retrospectively-identified patients from a CT registry database who underwent CCTA over a period of two years (2008–2009), and follow-up data after CCTA were collected up to December 2011. Baseline characteristics such as age, sex, hypertension, diabetes, and hyperlipidemia were obtained from electronic medical records, and the CCTA findings were evaluated in consensus by two experienced radiologists blinded to patients' clinical findings. Through CCTA, observers determined the presence of significant stenosis and the number of segments with significant stenosis, finally determining whether each patient had "significant CAD". Data about death status was obtained for all patients from the National Statistics database.

To estimate the absolute risk of all-cause mortality, the logistic regression models employed cardiac risk factors including age as a continuous predictor, and sex, hypertension, diabetes, hyperlipidemia, and CCTA finding (significant CAD) as dichotomous predictors. Although the

Cox proportional hazard regression could be implemented to account for time-to-death, we refer only to status of death for the purpose of simple illustration in this article.

Table 1 shows the baseline characteristics and CCTA findings in the cohort of 960 patients (referred to as a derivation cohort) and an external cohort of additional patients (referred to as a validation cohort, to be discussed further later). There were 9.6% (= 92/960) deaths in the derivation cohort. Although two variables (hypertension and hyperlipidemia) were not statistically significant in univariable comparison, we chose to include these factors to build the multivariable prediction model because they are well-known cardiac risk factors. We developed two models, an old model that includes the five cardiac risk factors excluding the CCTA finding, and a new model that includes all six variables including the CCTA finding. The results of the two multivariable logistic regression models in the derivative cohort are shown in Table 2. The adjusted odds ratio of the five baseline characteristics were deemed similar between the two models, and the risk of death in the patients with significant CAD calculated in the new model is relatively higher (adjusted OR = 4.669, 95% confidence interval [CI], 2.789–7.816).

Evaluation of Model Performance

The measures of association are not directly linked to a predictor's ability to classify a participant (23). In other words, the prediction models are focused on absolute risks, not on relative risks such as odds ratios or hazard ratios. To evaluate the strength of the predictions from the model, measures of performance (not of association) are needed.

Traditional overall performance measures can be quantified by the distance between the predicted and actual outcome. The coefficient of determination (denoted as R^2 in linear regression models) of the percentage of total variance of outcomes explained by the prediction model is used for continuous outcomes. For binary or time-to-event outcomes, the newly proposed R^2 or Brier score can be used to present overall performance measures (24-26). However, the common types of outcome in the prediction model in medicine are the binary or time-to-event outcomes, and in such cases, the most important aspects of model performance can be assessed in terms of calibration and discrimination.

Calibration

Calibration is related to goodness-of-fit, which reflects the agreement between observed outcomes and predictions. A calibration plot has the predicted probabilities for groups defined by ranges of individual predicted probabilities (e.g., 10 groups of equal size) on the x-axis, and the mean observed outcome on the y-axis, as shown in Figure 1. Perfect calibration should lie on or around a 45° line of

Table 1. Baseline Characteristics and CCTA Findings

Variables	Derivative Cohort		External Validation Cohort	
	Death (n = 92)	Survivor (n = 868)	Death (n = 15)	Survivor (n = 221)
Age, years (mean ± SD)	75.5 ± 4.3	74.3 ± 4.0	75.9 ± 5.6	73.9 ± 3.6
Sex, male (%)	68 (73.9)	354 (40.8)	11 (73.3)	80 (36.2)
Hypertension (%)	71 (77.2)	590 (68.0)	14 (93.3)	129 (58.4)
Diabetes (%)	45 (48.9)	212 (24.4)	13 (86.7)	35 (15.9)
Hyperlipidemia (%)	15 (16.3)	169 (19.5)	1 (6.7)	34 (15.4)
Significant CAD at CCTA (%)	70 (76.1)	296 (34.1)	11 (73.3)	75 (33.9)

CAD = coronary artery disease, CCTA = coronary computed tomographic angiography, SD = standard deviation

Table 2. Multivariable Logistic Regression Analysis in Derivative Cohort

	Old Model			New Model		
	Adjusted OR	95% CI	*P*	Adjusted OR	95% CI	*P*
Age, years	1.073	1.021–1.128	0.005	1.059	1.005–1.115	0.031
Sex, male	3.899	2.381–6.385	< 0.001	3.311	1.996–5.492	< 0.001
Hypertension	1.458	0.861–2.468	0.161	1.282	0.745–2.206	0.369
Diabetes	2.755	1.750–4.338	< 0.001	2.407	1.504–3.852	< 0.001
Hyperlipidemia	0.838	0.457–1.538	0.569	0.754	0.403–1.413	0.379
Significant CAD at CCTA				4.669	2.789–7.816	< 0.001

CAD = coronary artery disease, CCTA = coronary computed tomographic angiography, CI = confidence interval, OR = odds ratio the plot. This plot is a graphical illustration of the Hosmer-Lemeshow goodness-of-fit test (27) for binary outcomes, or the counterpart tests for survival outcomes including the Nam-D'Agostino test (28). A p value < 0.05 for these calibration tests would indicate poor calibration of the model. In our example, the Hosmer-Lemeshow test showed that the calibration of the two models is adequate with p > 0.05 (p = 0.648 in the old model, p = 0.113 in the new model). The Hosmer-Lemeshow test has some drawbacks in that it is often non-significant for small sample sizes but nearly always significant for large sample sizes, and has limited power to assess poor calibration (29).

Discrimination

Various statistics can be used to evaluate discriminative ability. Discrimination refers to the ability of a prediction model to differentiate between two outcome classes. The well-known statistical measures used to evaluate discrimination (classification) performance of diagnostic tests, particularly in radiological research, are true-positive rates, false-positive rates, and receiver operating characteristic (ROC) curves with the area under the ROC curve (AUC) (30). The concordance statistic (c-index), which is mathematically identical to the AUC for a binary outcome, is the most widely used measure to indicate discriminatory ability. The c-index can be interpreted as the probability that a subject with an outcome is given a higher probability of the outcome by the model than a randomly chosen subject without the outcome (31). A value of 0.5 indicates that the model has no discriminatory ability, and a value of 1.0 indicates that the model has perfect discrimination. In our example, the c-index was 0.801 (95% CI, 0.749–0.852) for the new model.

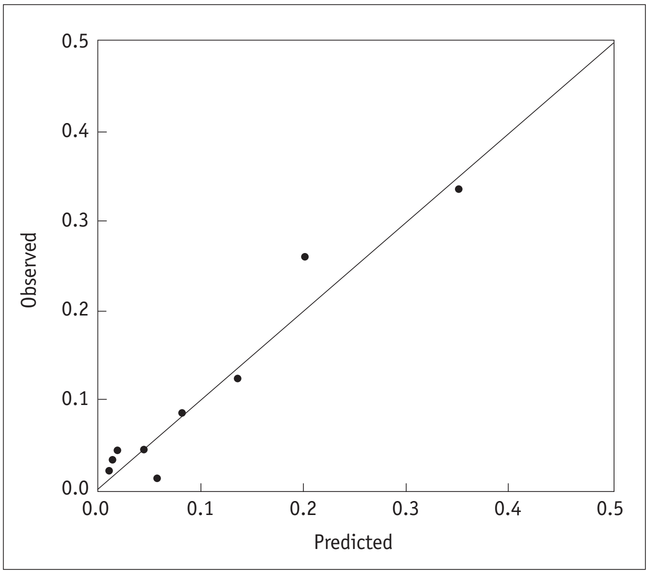

Fig. 1. Calibration plot.

For time-to-event (often referred to as "survival" data), Harrell's c-statistic is an analogous measure of the proportion of all subject pairs that can be ordered such that the subject with the higher predicted survival is the one who survived longer (31). A number of different approaches to the c-index for survival models have been proposed, and researchers should carefully state which measure is being used (32). Extensions of the c-index have been proposed for polytomous outcome (33), clustered time-to-event data (34), and competing risks (35).

Although ROC curves are widely used, there are some criticisms. First, AUC interpretation is not directly clinically relevant. Second, the predicted risk can differ substantially from the actual observed risk, even with perfect discrimination. Therefore, calibration is also very important in prediction model evaluation.

Validation of Prediction Model

The purpose of a prediction model is to provide valid prognoses for new patients; hence, validation is an important aspect of the predictive modeling process. Internal validation is a necessary part of model development. It determines the reproducibility of a developed prediction model for the derivative sample; and prevents over-interpretation of current data. Resampling techniques such as cross-validation and bootstrapping can be performed; bootstrap validation, in particular, appears most attractive for obtaining stable optimism-corrected estimates (1). The optimism is the decrease between model performance (e.g., c-index) in the bootstrap sample and in the original sample, which can adjust the developed model for over-fitting. To obtain these estimates, we first develop the prediction model in the development cohort (n = 960 in our example), and then generate a bootstrap sample by sampling n individuals with replacement from the original sample. After generating at least 100 bootstrap samples, the optimism-corrected model performance can be obtained by subtracting the estimated mean of the optimism estimate value from the c-index in the original sample. In our example study, with 500 bootstrap replications, the estimated optimism is 0.005, and the optimism-corrected c-index of 0.796 (= 0.801 - 0.005) showed good discrimination. In addition, statistical shrinkage techniques for adjusting regression coefficients can be used to recalibrate the model (1).

Another aspect is external validity. External validation is essential to support generalizability (i.e., general applicability) of a prediction model for patients other than those in the derivative cohort. External validation can be achieved by evaluating the model performance (e.g., c-index) in data other than that used for the model development. Therefore, it is performed after developing a prediction model. There are several types of external validation, such as validation in more recent patients (temporal validation), in other places (geographic validation), or by other investigators at other sites (fully independent validation). Formal sample size calculations based on statistical power considerations are not well investigated for external validation studies. However, a substantial sample size is required to validate the prediction model to achieve adequate model performance in the validation set. The number of events and predictors has an effect in determining the sample size for external validation data for the development of the prediction model. Simulation studies suggested that a minimum of 100 events and/or 100 non-events is required for external validation of the prediction model (36, 37), and a systematic review found that small external validation studies are unreliable and inaccurate (38). In our example study, we had a new dataset including 236 patients from another hospital in January–June 2013. External validation using these data showed that our model (new model) discriminates well (c-index = 0.893, 95% CI, 0.816–0.969).

Comparison of Prediction Models

To compare different prediction models, the improvement in discrimination can be assessed by quantifying an incremental value such as the change in the c-index. The statistical significance of the difference between the two models can be tested by the method used by DeLong et al. (39), which was designed to compare two correlated ROC curves. In our example, the c-index values were 0.748 (95% CI, 0.696–0.800) for the old model, and 0.801 (95% CI, 0.749–0.852) for the new model, and the difference of 0.053 (95% CI, 0.021–0.085) between the two models was statistically significant (p = 0.001). It can be interpreted that discriminatory ability improved significantly when CCTA finding (significant CAD) was added to the old model. Figure 2 illustrates ROC curves for the two models.

However, when the new predictor(s) adds to an existing clinical prediction model as in our example, the c-index is often conservative in model comparisons (40). A simulation study showed that statistical testing on the difference of the AUC (c-index), such as the method from DeLong et al. (39), is not recommended when the test of the added predictor is not significant (41). Ware (42) showed that the measure of association between the risk factor and predicted outcome (e.g., odds ratio or hazard ratio) does not reflect the predictive (discriminative) ability. Thus, very large associations are required to significantly increase the prognostic performance.

To overcome these drawbacks, Pencina et al. (43) proposed two indexes–net reclassification improvement (NRI) and integrated discrimination improvement (IDI)– to quantify the amount of overall reclassification and to evaluate the incremental value of predictors. These were originally proposed for comparing nested models, but they can also be used for comparing two non-nested models.

The NRI is the net proportion of events reclassified correctly plus the net proportion of nonevents reclassified correctly, for a pre-specified set of cut-off points. If D denotes the disease status, the NRI is defined as

NRI = (P [up|D = 1] - P [down|D = 1]) - (P [up|D = 0] - P [down|D = 0])

where an upward movement (up) refers to a change of the predicted risk into a higher category based on the new model and a downward movement (down) refers to a change in the opposite direction. Thus, the NRI is an index that combines four proportions and can have a range from -2 to 2. The above NRI has two components–event NRI and non-

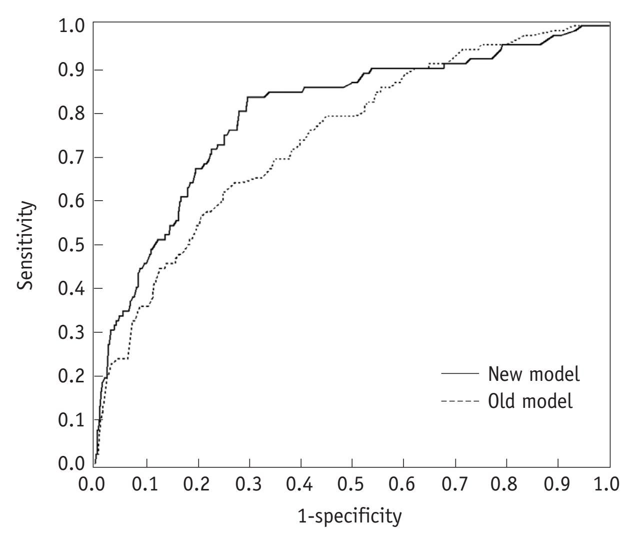

Fig. 2. ROC curves for two prediction models. ROC = receiver operating characteristic

event NRI–the net percentage of persons with (without) the event of interest correctly classified upward (downward). In our example, we measured the incremental prognostic value of the CCTA finding by calculating the NRI and the IDI. We chose risk categories based on frequently used cut-off values in the cardiovascular disease field corresponding to 10% to 20% (7). The results regarding reclassification by the model with or without CCTA finding including calculation details, are shown in Table 3. The categorical NRI was 0.266 (95% CI, 0.131–0.400), with 24.0% and 2.6% of patients who died and survived, respectively, correctly reclassified by the model with CCTA finding added. As the weights for event and nonevent groups are equal in the overall NRI, this may be problematic in cases of low disease occurrence. Therefore, it is recommended that authors report separately both the event and non-event components of the NRI separately, along with the overall NRI (44-46).

Category-based NRI is highly sensitive to the number of categories, and higher numbers of categories lead to increased movement of subjects, thus inflating the NRI value. In several diseases, there are no established risk categories, and the selection of thresholds defining the risk categories can influence the NRI amount. In this situation, two or more NRIs can be employed according to the different risk categories (47). Additionally, the continuous (category-free) NRI suggested by Pencina et al. (48) was presented, which considers any change (increase or decrease) in predicted risk for each individual, and it is not affected by the category cut-off. Presenting *p* values from statistical testing for NRI significance is discouraged because it has been proved mathematically that it is equivalent to the testing for adjusted effects of a newly added factor controlling for existing factors; instead, only confidence intervals for the NRI should be provided (45, 49, 50).

Table 3. Reclassification Tables

Model without CCTA Finding	Model with CCTA Finding		
	< 10%	10–20%	≥ 20%
Death (n = 92)			
< 10%	17 (18.5)	13 (14.1)	0 (0.0)
≥ 10% and < 20%	5 (5.4)	4 (4.4)	19 (20.7)
≥ 20%	0 (0.0)	5 (5.4)	29 (31.5)
Survivor (n = 868)			
< 10%	525 (60.5)	70 (8.1)	0 (0.0)
≥ 10% and < 20%	104 (12.0)	25 (2.9)	58 (6.7)
≥ 20%	12 (1.4)	35 (4.0)	39 (4.5)

Values are numbers (percentages). Event NRI = (13 + 19 + 0) / 92 - (5 + 5 + 0) / 92 = (14.1% + 20.7%) - (5.4% + 5.4%) = 24.0%, Non-event NRI = (104 + 35 + 12) / 868 - (70 + 58 + 0) / 868 = (12.0% + 4.0% + 1.4%) - (8.1% + 6.7% + 0.0%) = 2.6%, Category-based NRI = 0.240 + 0.026 = 0.266 (95% CI, 0.131–0.400), Category-free NRI = 0.840 (95% CI, 0.654–1.025).
CCTA = coronary computed tomographic angiography, CI = confidence interval, NRI = net reclassification improvement

The IDI is the difference in predicted probabilities between in those who do and do not have the outcome. It estimates the magnitude of the probability improvements or worsening between two models (nested or not) over all possible probability thresholds. The IDI can be interpreted as equivalent to the difference in mean predicted probability in subjects without and with the outcome. Thus, the estimation of IDI can be expressed as follows:

$$\widehat{\text{IDI}} = (\bar{\hat{p}}_{\text{new, events}} - \bar{\hat{p}}_{\text{old, events}}) - (\bar{\hat{p}}_{\text{new, nonevents}} - \bar{\hat{p}}_{\text{old, nonevents}})$$

where $\hat{p}_{\text{new, events}}$ is the mean predicted probabilities of an event in the new model for event group, $\hat{p}_{\text{old, events}}$ is the corresponding quantity in the old model, $\hat{p}_{\text{new, nonevents}}$ is the mean predicted probabilities of a non-event in the new model for nonevent group and $\hat{p}_{\text{old, nonevents}}$ is the corresponding quantity based on the old model. Although statistical testing regarding IDI was proposed by the original authors (43), other investigators demonstrated that the *p* value for the IDI may be not be valid even in large samples (51). Therefore, the bootstrap confidence interval for the IDI would be more appropriate. In our example, the estimated IDI was 0.057 (= 0.051 - [-0.005], 95% CI, 0.043–0.071).

Several statisticians have concluded that increases in AUC, IDI, and NRI offer complementary information. They therefore recommend reporting all three values together as measures that characterize the performance of the final model (52).

How to Present Prediction Model?

Regression Formula

To allow individualized predictions, the estimated regression models can be represented regarding the predictive probability of the outcome occurring. In logistic regression, the predicted probability of the outcome event is

$$\text{Probability} = \frac{\exp\ (\beta_0 + \beta_1 X_1 + \cdots + \beta_\kappa X_\kappa)}{1 + \exp\ (\beta_0 + \beta_1 X_1 + \cdots + \beta_\kappa X_\kappa)}$$

where, β_0: intercept in model, $\beta_1, \cdots \beta_\kappa$: regression coefficient (= log odds ratio) for each predictor ($X_1, \cdots X_\kappa$).

In our example, the predicted probability for a patient to death based on the new model can be expressed as

$$p = \frac{\text{exp (-8.527 + 0.057 age + 1.197 male + 0.249 hypertension}}{\text{1 + exp (-8.527 + 0.057 age + 1.197 male + 0.249 hypertension}}$$

$$\frac{\text{+ 0.878 diabetes - 0.282 hyperlipidemia + 1.541 significant CAD)}}{\text{+ 0.878 diabetes - 0.282 hyperlipidemia + 1.541 significant CAD)}}$$

For example, the predicted probability for a 77-year-old man with both hypertension and diabetes and significant CAD on CCTA is estimated as

$$p = \frac{\text{exp (-8.527 + 0.057 x 77 + 1.197}}{\text{1 + exp (-8.527 + 0.057 x 77 + 1.197}}$$

$$\frac{\text{+ 0.249 + 0.878 + 1.541)}}{\text{+ 0.249 + 0.878 + 1.541)}} = 43.22\%.$$

Scoring System

Simplified scoring is a useful method to present the predicted probability of an outcome that is easy to use in practice. It is developed in several ways, based on converting the regression coefficient or relative risk (odds ratio or hazard ratio) for each predictor to integers. Motivated by the Framingham heart study, Sullivan et al. (53) developed a so-called points (scoring) system that can simply compute the risk estimates without a calculator or computer. In many studies, this approach was used to create risk scoring systems (54, 55). We developed a scoring system for presenting the prediction model in our example. The algorithm for developing the scoring system using our example is shown in Table 4. Each step with details is as follows:

1) *Estimate the regression coefficients (β) of the multivariable model*

2) *Organize the risk factors into categories and determine the baseline category and reference values for each variable*

In our example, we consider a 70–74 year old, non-hypertensive, non-diabetic, and non-hyperlipidemic female with non-significant CAD on CCTA as the referent profile. Reference values (W_{REF}) for continuous variables such as age are determined as mid-point values of each category. For categorical variables, assign 0 point to the reference category and 1 point to the other category in the scoring system.

3) *Determine how far each category is from the reference category in regression units*

The quantities for each category determined using β (W - W_{REF}).

4) *Set the base constant* (*constant B*)

It means the number of regression units that reflects one point in the point scoring system. Generally, the smallest regression coefficient in the model can be used. In our example, the constant *B* was determined

Table 4. Scoring System to Calculate Point Values for Risk Score

Variables	β (1)	Categories (2)	Reference Value (W) (2)	β (W - W_{REF}) (3)	Points$_i$ = β (W - W_{REF}) / B (4, 5)
Age	0.057	70–74*	72 (W_{REF})	0	0
		75–79	77	0.285	1
		80–84	82	0.570	2
		85–92	88.5	0.941	3
Sex	1.197	Female*	0 (W_{REF})	0	0
		Male	1	1.197	4
Hypertension	0.249	No*	0 (W_{REF})	0	0
		Yes	1	0.249	1
Diabetes	0.878	No*	0 (W_{REF})	0	0
		Yes	1	0.878	3
Hyperlipidemia	-0.282	No*	0 (W_{REF})	0	0
		Yes	1	-0.282	-1
Significant CAD	1.541	No*	0 (W_{REF})	0	0
		Yes	1	1.541	5

*Reference category
1) Estimate the regression coefficients (β) of the multivariable model
2) Organize the risk factors into categories, determine the reference category, and reference values for each variable
3) Determine how far each category is from the reference category in regression units
4) Set the base constant (constant B)
5) Determine the number of points for each of the categories of each variable
CAD = coronary artery disease

in terms of the increase in risk associated with a 5-year increase in age based on the work from the Framingham study. The constant *B* in our example was set 0.057 x 5 = 0.285.

5) *Determine the number of points for each of the categories of each variable*
Points for each of the categories of each variable are computed by β (W - W_{REF}) / B. The final points are rounded to the nearest integer. As a result, a single point was meant to represent the increase in all-cause mortality associated with a 5-year increase in age. For example, a 77-year-old man with both hypertension and diabetes and assessed has significant CAD on CCTA would have a score of 15 (= 1 + 4 + 1 + 3 + 5).

6) *Create risk categories according to the total score*
In our example, the maximum total score is 16 points. For simple interpretation in a clinical setting, risk categories are often suggested. For example, the patients can be classified according to their total score into three categories: < 6 points, low-risk; 6–10 points, intermediate-risk; > 10 points, high-risk group. Table 5 indicates these three risk groups within derivation and validation cohorts.

Caution and Further Considerations

Prediction models can be used for diagnostic or prognostic purposes. Such models will be more generalizable when the properties including range or usability of predictors and outcome in the new population for application are similar to those seen in the development population. In the previous example study, the developed prediction model was designed primarily for use with elderly patients, so this model cannot be generalized to young or middle-aged adults. Rapidly changing predictors–for example, continuing developments in diagnostic tests or imaging modalities in radiology–can limit the application of the developed prediction model.

Various prediction models have been published with or without validation results. In cases where a researcher applied such an existing model to their own data, the model performance using the data was often revealed to be poor, indicating that the published prediction model requires "updating". For example, re-calibration and re-estimation of some regression coefficients including new predictors can be performed (56).

Optimal study design and statistical methods have been simultaneously developed. However, these cannot remedy any limitations in the collection of raw data and/or missing or misclassified information. Therefore, the quality of the raw data is emphasized. Efforts should be made to minimize missing values, especially required patient characteristics and/or follow-up information.

To overcome small sample size and reduced generalizability in a single-center study, individual participant data sharing in multicenter collaborative studies or "big data" from national or worldwide surveys or registries are being used to derive and/or validate the prediction model.

Reporting guidelines for various study types including randomized trial or observational studies have been published. Radiology researchers may be familiar with the standards for the reporting of diagnostic accuracy studies (STARD) statement (57). Recently, an international group of prediction model researchers developed the reporting guidelines for prediction models–transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (11). In addition to the TRIPOD statements, an accompanying explanation and elaboration document with more than 500 references from statistics, epidemiology, and clinical decision-making are also available (58). Each checklist item of the TRIPOD is explained, and corresponding examples of good reporting from published articles are provided.

SUMMARY

· Diagnostic imaging results are often combined with other clinical factors to improve the predictive ability of a clinical prediction model in both diagnostic and prognostic settings.

· The model-based prediction is inherently multivariable;

Table 5. Risk Groups within Derivation and Validation Cohort

Risk Group	Score*	Derivation Cohort	Validation Cohort
Low	1–5	13/529 (2.5)	1/135 (0.7)
Intermediate	6–10	36/305 (11.8)	6/82 (7.3)
High	10–16	43/126 (34.1)	8/19 (42.1)

*Sum of scores for each variable as shown in Table 4.

and, therefore, the most frequently used approach is multivariable regression modeling.

- Predictors should be selected using both clinical knowledge and statistical reasoning.
- The model performance should be evaluated in terms of both calibration and discrimination.
- The validation, especially external validation, is an important aspect of establishing a predictive model.
- Performance of different predictive models can be compared using c-index, NRI, and IDI.
- A predictive model may be presented in the form of a regression equation or can be converted into a scoring system for an easier use in practice.

REFERENCES

1. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer, 2009
2. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117:743-753
3. Yang HI, Yuen MF, Chan HL, Han KH, Chen PJ, Kim DY, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. *Lancet Oncol* 2011;12:568-574
4. Kwak JY, Jung I, Baek JH, Baek SM, Choi N, Choi YJ, et al. Image reporting and characterization system for ultrasound features of thyroid nodules: multicentric Korean retrospective study. *Korean J Radiol* 2013;14:110-117
5. Kim SY, Lee HJ, Kim YJ, Hur J, Hong YJ, Yoo KJ, et al. Coronary computed tomography angiography for selecting coronary artery bypass graft surgery candidates. *Ann Thorac Surg* 2013;95:1340-1346
6. Yoon YE, Lim TH. Current roles and future applications of cardiac CT: risk stratification of coronary artery disease. *Korean J Radiol* 2014;15:4-11
7. Shaw LJ, Giambrone AE, Blaha MJ, Knapper JT, Berman DS, Bellam N, et al. Long-term prognosis after coronary artery calcification testing in asymptomatic patients: a cohort study. *Ann Intern Med* 2015;163:14-21
8. Lee K, Hur J, Hong SR, Suh YJ, Im DJ, Kim YJ, et al. Predictors of recurrent stroke in patients with ischemic stroke: comparison study between transesophageal echocardiography and cardiac CT. *Radiology* 2015;276:381-389
9. Suh YJ, Hong YJ, Lee HJ, Hur J, Kim YJ, Lee HS, et al. Prognostic value of SYNTAX score based on coronary computed tomography angiography. *Int J Cardiol* 2015;199:460-466
10. Sunshine JH, Applegate KE. Technology assessment for radiologists. *Radiology* 2004;230:309-314
11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55-63
12. Bossuyt PM, Leeflang MM. *Chapter 6: Developing Criteria for Including Studies*. In: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [updated September 2008]. Oxford: The Cochrane Collaboration, 2008
13. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373-1379
14. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;48:1503-1510
15. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007;165:710-718
16. 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;54:774-781
17. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:127-141
18. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485
19. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 2004;57:1138-1146
20. Austin PC, Tu JV. Bootstrap methods for developing predictive models. *Am Stat* 2004;58:131-137
21. Austin PC. Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. *J Clin Epidemiol* 2008;61:1009-1017.e1
22. Little RJA, Rubin DB. *Statistical analysis with missing data*, 2nd ed. New York: John Wiley & Sons, 2014
23. 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;159:882-890
24. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika* 1991;78:691-692
25. Tjur T. Coefficients of determination in logistic regression models—A new proposal: the coefficient of discrimination. *Am Stat* 2009;63;366-372
26. Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol* 2010;63:938-939; author reply 939
27. Hosmer Jr DW, Lemeshow S. *Applied logistic regression*. New York: John Wiley & Sons, 2004
28. D'Agostino R, Nam, BH. *Evaluation of the performance of survival analysis models: discrimination and calibration*

measures. In: Balakrishnan N, Rao CO, eds. *Handbook of statistics: advances in survival analysis*. Vol 23. Amsterdam: Elsevier, 2004:1-25
29. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16:965-980
30. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol* 2004;5:11-18
31. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001
32. Pencina MJ, D'Agostino RB Sr, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Stat Med* 2012;31:1543-1553
33. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat Med* 2012;31:2610-2626
34. Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. *Stat Med* 2010;29:3160-3171
35. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics* 2014;15:526-539
36. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-483
37. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35:214-226
38. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837-845
40. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928-935
41. Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med* 2012;31:2577-2587
42. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med* 2006;355:2615-2617
43. 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;27:157-172; discussion 207-212
44. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol* 2011;173:1327-1335
45. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med* 2014;160:122-131
46. Pepe MS, Janes H. Commentary: reporting standards are needed for evaluations of risk reclassification. *Int J Epidemiol* 2011;40:1106-1108
47. Widera C, Pencina MJ, Bobadilla M, Reimann I, Guba-Quint A, Marquardt I, et al. Incremental prognostic value of biomarkers beyond the GRACE (Global Registry of Acute Coronary Events) score and high-sensitivity cardiac troponin T in non-ST-elevation acute coronary syndrome. *Clin Chem* 2013;59:1497-1505
48. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;30:11-21
49. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med* 2013;32:1467-1482
50. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? *J Natl Cancer Inst* 2014;106:dju041
51. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am J Epidemiol* 2011;174:364-374
52. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 2012;176:473-481
53. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004;23:1631-1660
54. Imperiale TF, Monahan PO, Stump TE, Glowinski EA, Ransohoff DF. Derivation and Validation of a Scoring System to Stratify Risk for Advanced Colorectal Neoplasia in Asymptomatic Adults: A Cross-sectional Study. *Ann Intern Med* 2015;163:339-346
55. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB Sr, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 2009;373:739-745
56. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76-86
57. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology* 2015;277:826-832
58. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-W73