Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 **www.annals.org**
For author affiliations, see end of text.
For members of the TRIPOD Group, see the **Appendix**.

In medicine, numerous decisions are made by care providers, often in shared decision making, on the basis of an estimated probability that a specific disease or condition is present (diagnostic setting) or a specific event will occur in the future (prognostic setting) in an individual. In the diagnostic setting, the probability that a particular disease is present can be used, for example, to inform the referral of patients for further testing, to initiate treatment directly, or to reassure patients that a serious cause for their symptoms is unlikely. In the prognostic context, predictions can be used for planning lifestyle or therapeutic decisions on the basis of the risk for developing a particular outcome or state of health within a specific period (1–3). Such estimates of risk can also be used to risk-stratify participants in therapeutic intervention trials (4–7).

In both the diagnostic and prognostic setting, probability estimates are commonly based on combining information from multiple predictors observed or measured from an individual (1, 2, 8–10). Information from a single predictor is often insufficient to provide reliable estimates of diagnostic or prognostic probabilities or risks (8, 11). In virtually all medical domains, diagnostic and prognostic multivariable (risk) prediction models are being developed, validated, updated, and implemented with the aim to assist doctors and individuals in estimating probabilities and potentially influence their decision making.

A multivariable prediction model is a mathematical equation that relates multiple predictors for a particular individual to the probability of or risk for the presence (diagnosis) or future occurrence (prognosis) of a particular outcome (10, 12). Other names for a prediction model include *risk prediction model*, *predictive model*, *prognostic* (or *prediction*) *index* or *rule*, and *risk score* (9).

Predictors are also referred to as *covariates*, *risk indicators*, *prognostic factors*, *determinants*, *test results*, or—more statistically—*independent variables*. They may range from demographic characteristics (for example, age and sex), medical history–taking, and physical examination results to results from imaging, electrophysiology, blood and urine measurements, pathologic examinations, and disease stages or characteristics, or results from genomics, proteomics, transcriptomics, pharmacogenomics, metabolomics, and other new biological measurement platforms that continuously emerge.

Diagnostic and Prognostic Prediction Models

Multivariable prediction models fall into 2 broad categories: diagnostic and prognostic prediction models (Box A). In a diagnostic model, multiple—that is, 2 or more—predictors (often referred to as *diagnostic test results*) are combined to estimate the probability that a certain condition or disease is present (or absent) at the moment of prediction (Box B). They are developed from and to be used for individuals suspected of having that condition.

In a prognostic model, multiple predictors are combined to estimate the probability of a particular outcome or event (for example, mortality, disease recurrence, complication, or therapy response) occurring in a certain period in the future. This period may range from hours (for example, predicting postoperative

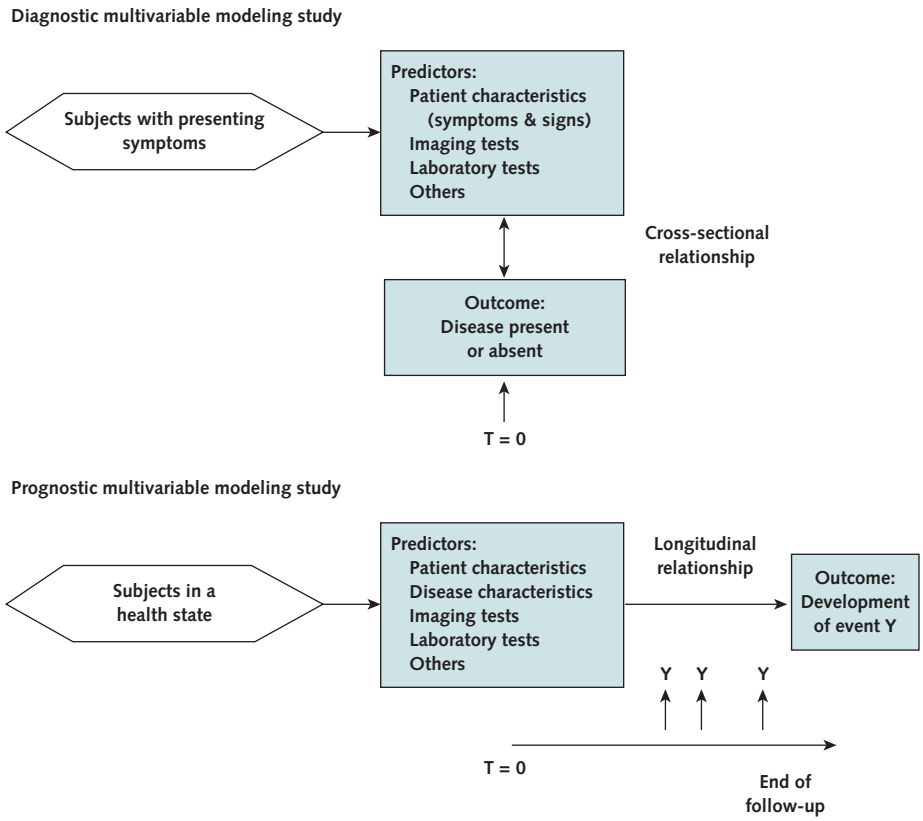
See also:

Related article 55
Editorial comment 73

© 2015 American College of Physicians

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box A. Schematic representation of diagnostic and prognostic prediction modeling studies.



The nature of the prediction in diagnosis is estimating the probability that a specific outcome or disease is present (or absent) within an individual, at this point in time—that is, the moment of prediction (T = 0). In prognosis, the prediction is about whether an individual will experience a specific event or outcome within a certain time period. In other words, in diagnostic prediction the interest is in principle a cross-sectional relationship, whereas prognostic prediction involves a longitudinal relationship. Nevertheless, in diagnostic modeling studies, for logistical reasons, a time window between predictor (index test) measurement and the reference standard is often necessary. Ideally, this interval should be as short as possible without starting any treatment within this period.

complications [13]) to weeks or months (for example, predicting 30-day mortality after cardiac surgery [14]) or years (for example, predicting the 5-year risk for developing type 2 diabetes [15]).

Prognostic models are developed and are to be used in individuals at risk for developing that outcome. They may be models for either ill or healthy individuals. For example, prognostic models include models to predict recurrence, complications, or death in a certain period after being diagnosed with a particular disease. But they may also include models for predicting the occurrence of an outcome in a certain period in individuals without a specific disease: for example, models to predict the risk for developing type 2 diabetes (16) or cardiovascular events in middle-aged nondiseased individuals (17), or the risk for preeclampsia in pregnant women (18). We thus use *prognostic* in the broad sense, referring to the prediction of an outcome in the future in individuals at risk for that outcome, rather than the narrower definition of predicting the course of patients who have a particular disease with or without treatment (1).

The main difference between a diagnostic and prognostic prediction model is the concept of *time*. Diagnostic modeling studies are usually cross-sectional, whereas prognostic modeling studies are usually longitudinal. In this document, we refer to both diagnostic and prognostic prediction models as "prediction models," highlighting issues that are specific to either type of model.

DEVELOPMENT, VALIDATION, AND UPDATING OF PREDICTION MODELS

Prediction model studies may address the development of a new prediction model (10), a model evaluation (often referred to as model *validation*) with or without updating of the model [19–21]), or a combination of these (**Box C** and **Figure 1**).

Model development studies aim to derive a prediction model by selecting predictors and combining them into a multivariable model. Logistic regression is commonly used for cross-sectional (diagnostic) and short-term (for example 30-day mortality) prognostic outcomes and Cox regression for long-term (for example, 10-year risk) prognostic outcomes. Studies may also focus on quantifying the incremental or added

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box B. Similarities and differences between diagnostic and prognostic prediction models.

Despite the different nature (timing) of the prediction, there are many similarities between diagnostic and prognostic prediction models, including:

- **Type of outcome is often binary: either disease of interest present versus absent (in diagnosis) or the future occurrence of an event yes or no (in prognosis).**
- **The key interest is to generate the probability of the outcome being present or occurring for an individual, given the values of 2 or more predictors, with the purpose of informing patients and guiding clinical decision making.**
- **The same challenges as when developing a multivariable prediction model, such as selection of the predictors, model-building strategies, and handling of continuous predictors and the danger of overfitting.**
- **The same measures for assessing model performance.**

Different terms for similar features between diagnostic and prognostic modeling studies are summarized below.

Diagnostic Prediction Modeling Study		Prognostic Prediction Modeling Study
	Explanatory variables, predictors, covariates (X variables)	
Diagnostic tests or index tests		Prognostic factors or indicators
	Outcome (Y variable)	
Target disease/disorder (presence vs. absence)		Event (future occurrence: yes or no)
Reference standard and disease verification		Event definition and event measurement
	Missing outcomes	
Partial verification		Loss to follow-up and censoring

predictive value of a specific predictor (for example, newly discovered) (22) to a prediction model.

Quantifying the predictive ability of a model on the same data from which the model was developed (often referred to as *apparent performance* [Figure 1]) will tend to give an optimistic estimate of performance, owing to overfitting (too few outcome events relative to the number of candidate predictors) and the use of predictor selection strategies (23–25). Studies developing new prediction models should therefore always include some form of internal validation to quantify any optimism in the predictive performance (for example, calibration and discrimination) of the developed model and adjust the model for overfitting. Internal validation techniques use only the original study sample and include such methods as bootstrapping or cross-validation. Internal validation is a necessary part of model development (2).

After developing a prediction model, it is strongly recommended to evaluate the performance of the model in other participant data than was used for the model development. External validation (Box C and Figure 1) (20, 26) requires that for each individual in the new participant data set, outcome predictions are made using the original model (that is, the published model or regression formula) and compared with the observed outcomes. External validation may use participant data collected by the same investigators, typically using the same predictor and outcome definitions and measurements, but sampled from a later period (temporal or narrow validation); by other investigators in another hospital or country (though disappointingly rare [27]), sometimes using different definitions and measurements (geographic or broad validation); in similar participants, but from an intentionally different setting (for example, a model developed in secondary care and assessed in similar participants, but selected from primary care); or even in other types of participants (for example, model developed in adults and assessed in children, or developed for predicting fatal events and assessed for predicting nonfatal events) (19, 20, 26, 28–30). In case of poor performance (for example, systematic miscalibration), when evaluated in an external validation data set, the model can be updated or adjusted (for example, recalibrating or adding a new predictor) on the basis of the validation data set (Box C) (2, 20, 21, 31).

Randomly splitting a single data set into model development and model validation data sets is frequently done to develop and validate a prediction model; this is often, yet erroneously, believed to be a form of external validation. However, this approach is a weak and inefficient form of internal validation, because not all available data are used to develop the model (23, 32). If the available development data set is sufficiently large, splitting by time and developing a model using data from one period and evaluating its performance using the data from the other period (temporal validation) is a stronger approach. With a single data set, temporal splitting and model validation can be considered intermediate between internal and external validation.

INCOMPLETE AND INACCURATE REPORTING

Prediction models are becoming increasingly abundant in the medical literature (9, 33, 34), and policymakers are increasingly recommending their use

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box C. Types of prediction model studies.

Prediction model development studies without validation* in other participant data aim to develop 1 (or more) prognostic or diagnostic prediction model(s) from the data set at hand: the development set. Such studies commonly aim to identify the important predictors for the outcome, assign the mutually adjusted weights per predictor in a multivariable analysis, develop a prediction model to be used for individualized predictions, and quantify the predictive performance (e.g., discrimination, calibration, classification) of that model in the development set. Sometimes, the development may focus on quantifying the incremental or added predictive value of a specific (e.g., newly discovered) predictor. In development studies, overfitting may occur, particularly in small development data sets. Hence, development studies ideally include some form of resampling techniques, such as bootstrapping, jack-knife, or cross-validation. These methods quantify any optimism in the predictive performance of the developed model and what performance might be expected in other participants from the underlying source population from which the development sample originated (see **Figure 1**). These resampling techniques are often referred to as "internal validation of the model," because no data other than the development set are used; everything is estimated "internally" with the data set at hand. Internal validation is thus always part of model development studies (see **Figure 1** and **Box F**).

Prediction model development studies with validation* in other participant data have the same aims as the previous type, but the development of the model is followed by quantifying the model's predictive performance in participant data other than the development data set (see **Figure 1**). This may be done in participant data collected by the same investigators, commonly using the same predictor and outcome definitions and measurements, but sampled from a later time period (so-called "temporal" or "narrow" validation); by other investigators in another hospital or country, sometimes using different definitions and measurements (geographic or broad validation); in similar participants, but from an intentionally chosen different setting (e.g., model developed in secondary care and tested in similar participants, but selected from primary care); or even in other types of participants (e.g., model developed in adults and tested in children, or developed for predicting fatal events and tested for predicting nonfatal events). Randomly splitting a single data set into a development and a validation data set is often erroneously referred to as a form of external validation* of the model. But this is an inefficient form of "internal" rather than "external" validation, because the 2 data sets only differ by chance (see **Figure 1).**

Model validation* studies without or with model updating aim to assess and compare the predictive performance of 1 (or more) existing prediction models by using participant data that were not used to develop the prediction model. When a model performs poorly, a validation study can be followed by updating or adjusting the existing model (e.g., recalibrating or extending the model by adding newly discovered predictors). In theory, a study may address only the updating of an existing model in a new data set, although this is unlikely and undesirable without first doing a validation of the original model in the new data set (see **Figure 1**).

* The term *validation*, although widely used, is misleading, because it indicates that model validation studies lead to a "yes" (good validation) or "no" (poor validation) answer on the model's performance. However, the aim of model validation is to evaluate (quantify) the model's predictive performance in either resampled participant data of the development data set (often referred to as *internal validation*) or in other, independent participant data that were not used for developing the model (often referred to as *external validation*).

in clinical practice guidelines (35–40). For some specific diseases, there is an overwhelming number of competing prediction models for the same outcome or target population. For example, there are over 100 prognostic models for predicting outcome after brain trauma (41), over 100 models for prostate cancer (42), over 60 models for breast cancer prognosis (43), 45 models for cardiovascular events after being diagnosed with diabetes (44), over 40 models for predicting prevalent and incident type 2 diabetes (45), and 20 models for predicting prolonged intensive care unit (ICU) stay after cardiac surgery (46).

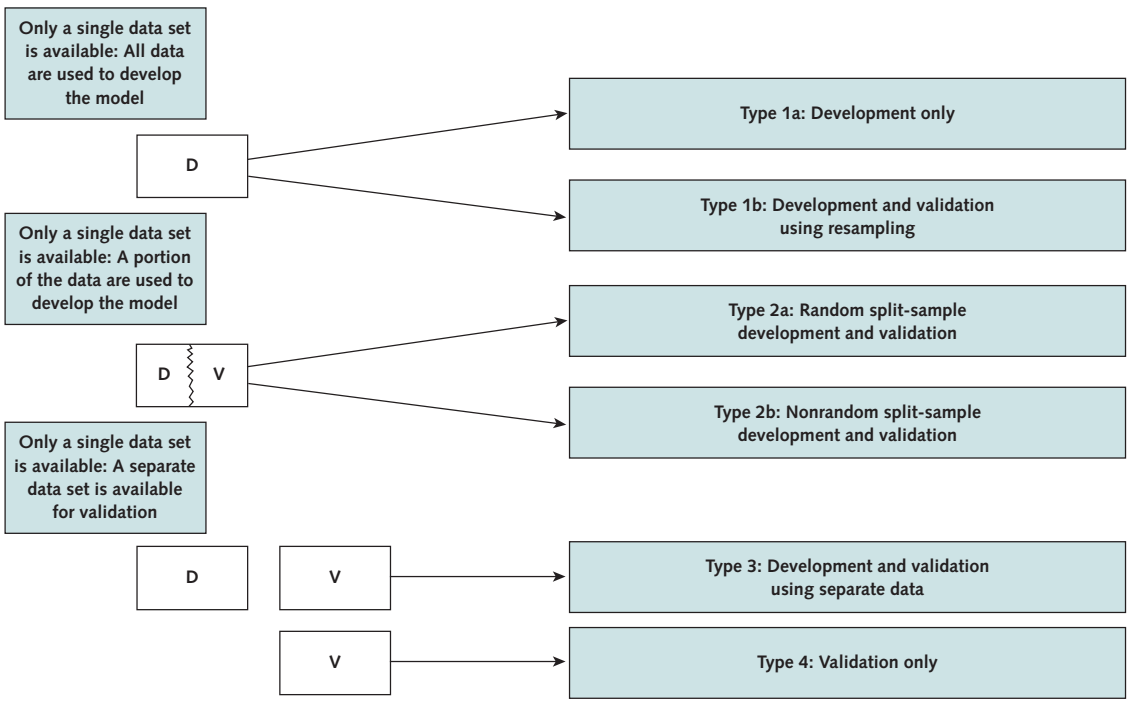
Given the abundance of published prediction models across almost all clinical domains, critical appraisal and synthesis of the available reports is a requirement to enable readers, care providers, and policymakers to judge which models are useful in which situations. Such an assessment, in turn, is possible only if key details of how prediction models were developed and validated are clearly reported (47, 48). Only then can generalizability and risk of bias of published prediction models be adequately assessed (49, 50), and subsequent researchers can replicate on the same data, if needed, the steps taken to obtain the same results (51, 52). Many reviews have illustrated, however, that the quality of published reports that describe the development or validation of prediction models across many different disease areas and different journals is poor (3, 34, 41, 43, 45, 46, 48, 53–95). For example, in a review of newly developed prediction models in the cancer literature (54, 55), reporting was disappointingly poor, with insufficient information provided about all aspects of model development. The same was found in a recent review of prediction models for prevalent or incident type 2 diabetes (45) and of prediction models published in 6 high-impact general medical journals (34).

Reporting guidelines for randomized trials (CONSORT [96]), observational studies (STROBE [97]), tumor marker studies (REMARK [98]), molecular epidemiology (STROBE-ME [99]), diagnostic accuracy (STARD [100]), and genetic risk prediction studies (GRIPS [101]) contain items that are relevant to all types of studies, including those developing or validating prediction models. The 2 guidelines most closely related to prediction models are REMARK and GRIPS. However, the focus of the REMARK checklist is primarily on prognostic factors and not prediction models, whereas the GRIPS statement is aimed at risk prediction using genetic risk factors and the specific methodological issues around handling large numbers of genetic variants.

To address a broader range of studies, we developed the TRIPOD guideline: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis. TRIPOD explicitly covers the development and validation of prediction models for both diagnosis and prognosis, for all medical domains and all types of predictors. TRIPOD also places considerable emphasis on model validation studies and the reporting requirements for such studies.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 1. Types of prediction model studies covered by the TRIPOD statement.



Analysis Type	Description
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as "internal validation", are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).
Type 2a	The data are randomly split into 2 groups: one to develop the prediction model, and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16).
Type 2b	The data are nonrandomly split (e.g., by location or time) into 2 groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a, because allows for nonrandom variation between the 2 data sets (6, 13, 17).
Type 3	Development of a prediction model using 1 data set and an evaluation of its performance on separate data (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data (13).
Types 3 and 4 are commonly referred to as "external validation studies." Arguably type 2b is as well, although it may be considered an intermediary between internal and external validation.	

D = development data; V = validation data.

THE TRIPOD STATEMENT

Prediction model studies can be subdivided into 5 broad categories (1, 8–10, 19, 20, 28, 33, 102–104): 1) prognostic or diagnostic predictor finding studies, 2) prediction model development studies without external validation, 3) prediction model development studies with external validation, 4) prediction model validation studies, and 5) model impact studies. TRIPOD addresses the reporting of prediction model studies aimed at developing or validating 1 or more prediction models (Box C). These development and validation studies can in turn be subdivided into various types (Figure 1). An increasing number of studies are evaluating the incremental value (103) of a specific predictor, to assess whether an existing prediction model may need to be updated or adjusted (22, 105, 106). TRIPOD also addresses such studies (Box C and Figure 1).

Prognostic or diagnostic predictor finding studies and model impact studies often have different aims, designs, and reporting issues compared with studies developing or validating prediction models. The for-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

mer commonly seek to identify predictors that independently contribute to the prediction of (that is, are associated with) a particular prognostic or diagnostic outcome. The aim is not to develop a final prediction model to be used for individualized predictions in other individuals. Prediction model impact studies aim to quantify the effect (impact) of using a prediction model on participant and physician decision making or directly on participant health outcomes, relative to not using the model (20, 102, 107). Model impact studies thus follow a comparative intervention design, rather than the single cohort design used in model development or validation studies, and are ideally a (cluster) randomized design. However, many items addressed in this reporting guideline do apply to these 2 other types of prediction research, although other reporting guidelines might serve them better. The REMARK guideline explicitly addresses the reporting of (single) prognostic factor studies (98, 108), and the CONSORT (96, 109) and STROBE (97) Statements are relevant guidelines for reporting of randomized or nonrandomized model impact studies, respectively.

Furthermore, TRIPOD primarily addresses prediction models for binary (for example, disease presence or absence) or time-to-event outcomes (for example, 10-year cardiovascular disease), because these are the most common types of outcomes to be predicted in medicine. However, outcomes may also be continuous measurements (for example, blood pressure, tumor size, percentage vessel stenosis, IQ scores, quality of life, or length of hospital stay), nominal outcomes (for example, the differential diagnosis rather than target disease present or absent; type of infection defined as viral, bacterial or no infection), or ordinal outcomes (for example, cancer stage, Glasgow Coma Scale [110], or Rankin scale [111]), for which prediction models may also be developed (2, 112). Most recommendations and reporting items in TRIPOD apply equally to the reporting of studies aimed at developing or validating prediction models for such outcomes.

Moreover, TRIPOD focuses on prediction models developed by regression modeling, because this is the approach by which most prediction models are developed, validated, or updated in medical research. However, most items equally apply to prediction tools developed, validated, or updated with other techniques, such as classification trees, neural networks, genetic programming, random forests, or vector machine learning techniques. The main difference in these other approaches over regression modeling is the method of data analysis to derive the prediction model. Problems of transparency in these nonregression modeling approaches are a particular concern, especially regarding reproducible research and implementation in practice.

DEVELOPMENT OF TRIPOD

We followed published guidance for developing reporting guidelines (113) and established a steering committee (Drs. Collins, Altman, Moons, and Reitsma) to organize and coordinate the development of TRIPOD. We conducted a systematic search of MEDLINE, EMBASE, PsycINFO, and Web of Science to identify any published articles making recommendations on reporting of multivariable prediction models or on methodological aspects of developing or validating a prediction model, reviews of published reports of multivariable prediction models that evaluated methodological conduct or reporting, and reviews of methodological conduct and reporting of multivariable models in general. From these studies, a list of 129 possible checklist items was generated. The steering group then merged related items to create a list of 76 candidate items.

Twenty-five experts with a specific interest in prediction models were invited by e-mail to participate in the Web-based survey and to rate the importance of the 76 candidate checklist items. Respondents (24 of 27) included methodologists, health care professionals, and journal editors. (In addition to the 25 meeting participants, the survey was also completed by 2 statistical editors from *Annals of Internal Medicine*.)

Twenty-four experts (22 of whom had participated in the survey) attended a 3-day meeting in Oxford, United Kingdom, in June 2011. This multidisciplinary group included statisticians, epidemiologists, methodologists, healthcare professionals, and journal editors (**Appendix**) (114). Several of the group also had experience in developing reporting guidelines for other types of clinical studies.

At the meeting, the results of the survey were presented, and each of the 76 candidate checklist items was discussed in turn. For each item, consensus was reached on whether to retain, merge with another item, or omit the item. Meeting participants were also asked to suggest additional items. After the meeting, the checklist was revised by the steering committee during numerous face-to-face meetings, and circulated to the participants to ensure it reflected the discussions. While making revisions, conscious efforts were made to harmonize our recommendations with other guidelines, and where possible we chose the same or similar wording for items.

THE TRIPOD STATEMENT: EXPLANATION AND ELABORATION

Aim and Outline of This Document

The TRIPOD Statement is a checklist of 22 items considered essential for good reporting of studies developing or validating multivariable prediction models (**Table 1**) (114). The items relate to the title and abstract (items 1 and 2), background and objectives (item 3), methods (items 4 through 12), results (items 13 through 17), discussion (items 18 through 20), and other information (items 21 and 22). The TRIPOD Statement covers studies that report solely development, both development and external validation, and solely external validation (with or without model updating) of a diagnostic or prognostic prediction model (**Box C**). Therefore, some items (denoted *D*) are relevant only for re-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 1. Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis*

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres	
	5b	D;V	Describe eligibility criteria for participants	
	5c	D;V	Give details of treatments received, if relevant	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted	
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors	
Sample size	8	D;V	Explain how the study size was arrived at	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	
	10c	V	For validation, describe how the predictions were calculated	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done	
Risk groups	11	D;V	Provide details on how risk groups were created, if done	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome)	
Model development	14a	D	Specify the number of participants and outcome events in each analysis	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point)	
	15b	D	Explain how to use the prediction model	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model	
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance)	

(*Continued on following page*)

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 1–Continued

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data)	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study	

* Items relevant only to the development of a prediction model are denoted by *D*, items relating solely to a validation of a prediction model are denoted by *V*, and items relating to both are denoted *D;V*. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

porting the development of a prediction model (items 10a, 10b, 14, and 15), and others (denoted *V*) apply only to reporting the (external) validation of a prediction model (items 10c, 10e, 12, 13c, 17, and 19a). All other items are relevant to all types of prediction model development and validation studies. Items relevant to all study types are denoted by *D;V*.

Discussion and explanation of all 22 items in the TRIPOD checklist (Table 1) are presented. We have split the discussion of a few complex and lengthy items into multiple parts to aid clarity.

The primary aim of this explanation and elaboration document is to outline a framework for improved reporting of prediction modeling studies. Many such studies are methodologically weak, however, so in this document, we also summarize the qualities of good (and the limitations of less good) studies, regardless of reporting.

Use of Examples

For each item, we present examples from published articles of both development and validation of prediction models, and often for both diagnosis and prognosis; they illustrate the type of information that is appropriate to report. Our use of a particular example does not imply that all aspects of the study were well conducted and reported, or that the methods being reported are necessarily the best methods to be used in prediction model research. Rather, the examples illustrate a particular aspect of an item that has been well reported in the context of the methods used by the study authors. Some of the quoted examples have been edited, with text omitted (denoted by . . .), text added (denoted by [ ]), citations removed, or abbreviations spelled out, and some tables have been simplified.

USE OF TRIPOD

Depending on the type of prediction model study (development, validation, or both), each checklist item (relevant to the study type) should be addressed somewhere in the report. If a particular checklist item cannot be addressed, acknowledgment that the information is unknown or irrelevant (if so) should be clearly reported. Although many of the items have a natural order and sequence in a report, some do not. We therefore do not stipulate a structured format, because this may also depend on journal formatting policies. Authors may find it convenient to report information for some of the requested items in a supplementary material section (for example, in online appendices).

To help the editorial process; peer reviewers; and, ultimately, readers, we recommend submitting the checklist as an additional file with the report, including indicating the pages where information for each item is reported. The TRIPOD reporting template for the checklist can be downloaded from www.tripod-statement.org.

Announcements and information relating to TRIPOD will be broadcast on the TRIPOD Twitter address (@TRIPODStatement). The Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network (www.equator-network.org) will help disseminate and promote the TRIPOD Statement.

THE TRIPOD CHECKLIST

Title and Abstract

Title

Item 1. Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. [D;V]

Examples

Development and validation of a clinical score to estimate the probability of coronary artery disease in men and women presenting with suspected coronary disease (115). [Diagnosis; Development; Validation]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Development and external validation of prognostic model for 2 year survival of non small cell lung cancer patients treated with chemoradiotherapy (116). [Prognosis; Development; Validation]

Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2 (117). [Prognosis; Validation]

Development of a prediction model for 10 year risk of hepatocellular carcinoma in middle-aged Japanese: the Japan Public Health Center based Prospective Study Cohort II (118). [Prognosis; Development]

Example With Additional Information

Development and validation of a logistic regression derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing (119). [Diagnosis; Development; Validation]

Examples of Well-Known Models

Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation (120). [Prognosis; Validation]

External validation of the SAPS II APACHE II and APACHE III prognostic models in South England: a multicentre study (121). [Prognosis; Validation]

Explanation

To enhance the retrieval by potential readers or systematic reviewers of multivariable prediction model studies, it is important that the title be as detailed as possible, without becoming too long. Authors may ideally address 4 main elements in their title:

Type of modeling study (development, validation or both).

Clinical context (diagnostic or prognostic).

Target population (individuals or patients for whom the model is intended).

Outcome that is predicted by the model.

Prediction model studies address model development (including internal validation; item 10b), external validation, or both (**Box C** and **Figure 1**). Authors should explicitly identify their study type by using these terms in the title. Likewise, if the study is updating an existing model, or focusing on the incremental value of a specific predictor, it is helpful to say so. Moreover, because many readers are interested in retrieving the available literature on a specific population or subpopulation of individuals or patients, or on a specific outcome in these persons, it is helpful also to include such identifying terms in the title.

Addressing these issues in a manuscript title is possible without creating long titles, as the above examples from published articles show. Studies including external validation, whether as the sole aim or in conjunction with developing a prediction model, should clearly indicate this in the title.

The terms *diagnostic* and *prognostic* are often not mentioned explicitly, but are implicitly covered by the description of the study population or outcomes. For example, the title that includes " . . . in men and women presenting with suspected coronary artery disease" clearly indicates that this is a study of a diagnostic model (115). Some prediction models are so well known by their given names only that titles of subsequent validation studies do not address the targeted population or predicted outcome. However, if the study is focusing on validating a well-known model in a different setting or predicting a different outcome, then this should be made clear in the title.

Sometimes the type of predictors (for example, predictors from patient history or physical examination), the timing of the prediction (for example, prediction of postoperative outcomes using preoperative patient characteristics), and the timing of the outcome (for example, 10-year risk for cardiovascular disease) can also be added to the title to further specify the nature of the study without unduly lengthening the title.

In a recent review of 78 external validation studies, only 27% (21 of 78) had the term "validation" or "validity" in the title of the article, and only 1 article explicitly stated in the title that the validation was carried out by independent investigators (122).

Abstract

Item 2. Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. [D;V]

Examples

OBJECTIVE: To develop and validate a prognostic model for early death in patients with traumatic bleeding.
DESIGN: Multivariable logistic regression of a large international cohort of trauma patients.
SETTING: 274 hospitals in 40 high, medium, and low income countries.
PARTICIPANTS: Prognostic model development: 20,127 trauma patients with, or at risk of, significant bleeding, within 8 hours of injury in the Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage (CRASH 2) trial. External validation: 14,220 selected trauma patients from the Trauma Audit and Research Network (TARN), which included mainly patients from the UK.
OUTCOMES: In hospital death within 4 weeks of injury.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

RESULTS: 3076 (15%) patients died in the CRASH 2 trial and 1765 (12%) in the TARN dataset. Glasgow coma score, age, and systolic blood pressure were the strongest predictors of mortality. Other predictors included in the final model were geographical region (low, middle, or high income country), heart rate, time since injury, and type of injury. Discrimination and calibration were satisfactory, with C statistics above 0.80 in both CRASH 2 and TARN. A simple chart was constructed to readily provide the probability of death at the point of care, and a web based calculator is available for a more detailed risk assessment (http://crash2.lshtm.ac.uk).

CONCLUSIONS: This prognostic model can be used to obtain valid predictions of mortality in patients with traumatic bleeding, assisting in triage and potentially shortening the time to diagnostic and lifesaving procedures (such as imaging, surgery, and tranexamic acid). Age is an important prognostic factor, and this is of particular relevance in high income countries with an aging trauma population (123). [Prognosis; Development]

OBJECTIVE: To validate and refine previously derived clinical decision rules that aid the efficient use of radiography in acute ankle injuries.

DESIGN: Survey prospectively administered in two stages: validation and refinement of the original rules (first stage) and validation of the refined rules (second stage).

SETTING: Emergency departments of two university hospitals.

PATIENTS: Convenience sample of adults with acute ankle injuries: 1032 of 1130 eligible patients in the first stage and 453 of 530 eligible patients in the second stage.

MAIN OUTCOME MEASURES: Attending emergency physicians assessed each patient for standardized clinical variables and classified the need for radiography according to the original (first stage) and the refined (second stage) decision rules. The decision rules were assessed for their ability to correctly identify the criterion standard of fractures on ankle and foot radiographic series. The original decision rules were refined by univariate and recursive partitioning analyses.

MAIN RESULTS: In the first stage, the original decision rules were found to have sensitivities of 1.0 (95% confidence interval [CI], 0.97 to 1.0) for detecting 121 malleolar zone fractures, and 0.98 (95% CI, 0.88 to 1.0) for detecting 49 midfoot zone fractures. For interpretation of the rules in 116 patients, kappa values were 0.56 for the ankle series rule and 0.69 for the foot series rule. Recursive partitioning of 20 predictor variables yielded refined decision rules for ankle and foot radiographic series. In the second stage, the refined rules proved to have sensitivities of 1.0 (95% CI, 0.93 to 1.0) for 50 malleolar zone fractures, and 1.0 (95% CI, 0.83 to 1.0) for 19 midfoot zone fractures. The potential reduction in radiography is estimated to be 34% for the ankle series and 30% for the foot series. The probability of fracture, if the corresponding decision rule were "negative," is estimated to be 0% (95% CI, 0% to 0.8%) in the ankle series, and 0% (95% CI, 0% to 0.4%) in the foot series.

CONCLUSION: Refinement and validation have shown the Ottawa ankle rules to be 100% sensitive for fractures, to be reliable, and to have the potential to allow physicians to safely reduce the number of radiographs ordered in patients with ankle injuries by one third. Field trials will assess the feasibility of implementing these rules into clinical practice (124). [Diagnosis; Validation; Updating]

Explanation

Abstracts provide key information that enables readers to assess the methodology and relevance of a study and give a summary of the findings. The abstract may be all that is readily available and helps readers to decide whether to read the full report. We recommend including at least the study objectives (ideally supported by a brief statement of background or rationale), setting, participants, sample size (and number of events), outcome, predictors, statistical analysis methods, results (for example, model performance and regression coefficients), and conclusions. A structured abstract is preferable, although requirements of specific journals vary.

The abstract should address the same attributes as the title (item 1), including whether the study concerns model development, model validation, or both; a diagnostic or prognostic model; the target population; and the outcome to be predicted. For model development studies, specifying all candidate predictors might not be feasible if too many were studied. In these instances, it may suffice to mention the total number considered and summarize in broad categories indicating when they were measured (for example, at history-taking and physical examination). The results section should ideally clarify the predictors included in the final model, along with measures of the model's predictive performance. This may not be necessary for complex models with many predictors, or studies validating a previously developed model in new data.

Informative abstracts and titles of prediction model studies enable researchers to locate relevant studies when conducting a literature search. A few search strategies for retrieving clinical prediction models have been published (125–127). They have recently been tested and slightly modified by independent researchers, who concluded that they miss few clinical predic-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

tion model studies (although they are less good at finding other types of prediction study) (128). Specific search filters for finding prediction model studies in the domain of primary care have also been developed (129).

Introduction

Background and Objectives

Item 3a. Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. [D;V]

Examples

> Confronted with acute infectious conjunctivitis most general practitioners feel unable to discriminate between a bacterial and a viral cause. In practice more than 80% of such patients receive antibiotics. Hence in cases of acute infectious conjunctivitis many unnecessary ocular antibiotics are prescribed. . . . To select those patients who might benefit most from antibiotic treatment the general practitioner needs an informative diagnostic tool to determine a bacterial cause. With such a tool antibiotic prescriptions may be reduced and better targeted. Most general practitioners make the distinction between a bacterial cause and another cause on the basis of signs and symptoms. Additional diagnostic investigations such as a culture of the conjunctiva are seldom done mostly because of the resulting diagnostic delay. Can general practitioners actually differentiate between bacterial and viral conjunctivitis on the basis of signs and symptoms alone? . . . A recently published systematic literature search summed up the signs and symptoms and found no evidence for these assertions. This paper presents what seems to be the first empirical study on the diagnostic informativeness of signs and symptoms in acute infectious conjunctivitis (130). [Diagnosis; Development]

> In the search for a practical prognostic system for patients with parotid carcinoma, we previously constructed a prognostic index based on a Cox proportional hazards analysis in a source population of 151 patients with parotid carcinoma from the Netherlands Cancer Institute. [The] **Table** . . . shows the pretreatment prognostic index PS1, which combines information available before surgery, and the post treatment prognostic index PS2, which incorporates information from the surgical specimen. For each patient, the index sums the properly weighted contributions of the important clinicopathologic characteristics into a number corresponding to an estimated possibility of tumor recurrence. These indices showed good discrimination in the source population and in an independent nationwide database of Dutch patients with parotid carcinoma. According to Justice et al, the next level of validation is to go on an international level. . . . For this purpose, an international database was constructed from patients who were treated in Leuven and Brussels (Belgium) and in Cologne (Germany), where the prognostic variables needed to calculate the indices were recorded, and predictions were compared with outcomes. In this way, we tried to achieve further clinical and statistical validation (131). [Prognosis; Validation]

> Any revisions and updates to a risk prediction model should be subject to continual evaluation (validation) to show that its usefulness for routine clinical practice has not deteriorated, or indeed to show that its performance has improved owing to refinements to the model. We describe the results from an independent evaluation assessing the performance of QRISK2 2011 on a large dataset of general practice records in the United Kingdom, comparing its performance with earlier versions of QRISK and the NICE adjusted version of the Framingham risk prediction model (117). [Prognosis; Validation]

Explanation

Multivariable prediction models can serve multiple purposes, so readers need a clear description of a model's rationale and potential use. Authors should describe the specific clinical context (such as decision) in which the model would be used. For example, a diagnostic prediction model may be used to help decide on the ordering of more invasive or burdensome tests in certain patients, and a prognostic model may inform patients with a certain condition about their future outcome or help judge subsequent treatment possibilities. This medical context and intended use of the model provides the rationale for their choice of patients (including setting) and to whom the results may be generalized, and what type of predictors would be available in this setting and therefore considered. The choice of outcome is a critical factor determining the clinical relevance of a model, and therefore the rationale for selecting a specific outcome should be given. Preferably, outcomes and duration of follow-up should be relevant to patients and clinical decision making.

Problems may arise if more expansive outcome definitions are used, thereby increasing the risk for labeling too many persons as high risk (132). A similar problem exists in the diagnostic setting if an abnormality on a new sensitive marker or high-resolution image becomes the new definition of disease, which may lead

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

to overdiagnosis and overtreatment (132, 133). A description of the medical context should also indicate any clinical decisions that may be informed by the predicted risk. Below are a few examples of the different uses of multivariable prediction models for both diagnostic and prognostic purposes.

Potential clinical uses of multivariable diagnostic models:

Decisions whether or not to order more invasive or costly diagnostic tests, or to refer patients to secondary care. Example: Ottawa rule for when to order radiography in patients with ankle injury (134, 135).

Decisions whether a certain target condition can be safely ruled out. Example: clinical decision rule and D-dimer to exclude deep venous thrombosis or pulmonary embolism (136).

Informing future parents about the likelihood that their unborn baby has trisomy 21. Example: triple tests during pregnancy (137).

Potential clinical uses of multivariable prognostic models:

Inform "healthy" individuals about their 10-year risk for cardiovascular disease. This information can be used to change unhealthy lifestyles. Examples: Framingham risk score (138), QRISK2 (139), and SCORE (140).

Inform patients diagnosed with a certain disease or patients undergoing a particular surgical procedure about their risk for having a poor outcome or complication, to decide on preemptive or therapeutic strategies. Example: indication for thrombolytic therapy based on 30-day mortality after an acute myocardial infarction (141).

When developing a model, researchers should mention, ideally on the basis of a literature review, whether related models (for example, for the same or similar intended use, participants, or outcomes) have already been developed (47). External validation studies generate valuable information about the performance of an existing, previously developed model in new patients. Authors should clearly state the existing model they were validating, citing the article, and state or restate the (potential) clinical use of this model. If other competing prediction models exist, authors should indicate why they only evaluated the selected model. Clearly, a comparative validation study (that is, evaluating multiple competing models [48]) on the same data set will generate additional information (47, 85). Any deliberate change in patient population, predictors, or outcomes in comparison with the study in which the model was developed should be highlighted (see also item 12), along with its rationale.

In a recent systematic review of external validation studies, 7 of 45 (16%) did not cite the original study developing the prediction model that was being evaluated (122).

Item 3b. Specify the objectives, including whether the study describes the development or validation of the model or both. [D;V]

Examples

The aim of this study was to develop and validate a clinical prediction rule in women presenting with breast symptoms, so that a more evidence based approach to referral—which would include urgent referral under the 2 week rule—could be implemented as part of clinical practice guidance (142). [Diagnosis; Development; Validation]

In this paper, we report on the estimation and external validation of a new UK based parametric prognostic model for predicting long term recurrence free survival for early breast cancer patients. The model's performance is compared with that of Nottingham Prognostic Index and Adjuvant Online, and a scoring algorithm and downloadable program to facilitate its use are presented (143). [Prognosis; Development; Validation]

Even though it is widely accepted that no prediction model should be applied in practice before being formally validated on its predictive accuracy in new patients no study has previously performed a formal quantitative (external) validation of these prediction models in an independent patient population. Therefore we first conducted a systematic review to identify all existing prediction models for prolonged ICU length of stay (PICULOS) after cardiac surgery. Subsequently we validated the performance of the identified models in a large independent cohort of cardiac surgery patients (46). [Prognosis; Validation]

Explanation

Study objectives are the specific aims or research questions that will be addressed in the study. By clearly specifying the objectives, often at the end of the introduction, the authors will provide the reader with the necessary background information to help critically appraise the study. For prediction model studies, the objectives should specify the purpose of prediction (diagnostic or prognostic), the outcome or type of outcome that is predicted, the setting and intended population the model will be used for, and the type of predictors that will be considered. Furthermore, authors should state whether the report concerns the development of a new model or the external validation of an existing model, or both.

Methods

Source of Data

Item 4a. Describe the study design or source of data (for example, randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. [D;V]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Examples

The population based sample used for this report included 2489 men and 2856 women 30 to 74 years old at the time of their Framingham Heart Study examination in 1971 to 1974. Participants attended either the 11th examination of the original Framingham cohort or the initial examination of the Framingham Offspring Study. Similar research protocols were used in each study, and persons with overt coronary heart disease at the baseline examination were excluded (144). [Prognosis; Development]

Data from the multicentre, worldwide, clinical trial (Action in Diabetes and Vascular disease: preterax and diamicron MR controlled evaluation) (ADVANCE) permit the derivation of new equations for cardiovascular risk prediction in people with diabetes. . . . ADVANCE was a factorial randomized controlled trial of blood pressure (perindopril indapamide versus placebo) and glucose control (gliclazide MR based intensive intervention versus standard care) on the incidence of microvascular and macrovascular events among 11,140 high risk individuals with type 2 diabetes . . . DIABHYCAR (The non insulin dependent diabetes, hypertension, microalbuminuria or proteinuria, cardiovascular events, and ramipril study) was a clinical trial of ramipril among individuals with type 2 diabetes conducted in 16 countries between 1995 and 2001. Of the 4912 randomized participants, 3711 . . . were suitable for use in validation. Definitions of cardiovascular disease in DIABHYCAR were similar to those in ADVANCE. . . . Predictors considered were age at diagnosis of diabetes, duration of diagnosed diabetes, sex, . . . and randomized treatments (blood pressure lowering and glucose control regimens) (145). [Prognosis; Development; Validation]

We did a multicentre prospective validation study in adults and an observational study in children who presented with acute elbow injury to five emergency departments in southwest England UK. As the diagnostic accuracy of the test had not been assessed in children we did not think that an interventional study was justified in this group (146). [Diagnosis; Validation]

We conducted such large scale international validation of the ADO index to determine how well it predicts mortality for individual subjects with chronic obstructive pulmonary disease from diverse settings, and updated the index as needed. Investigators from 10 chronic obstructive pulmonary disease and population based cohort studies in Europe and the Americas agreed to collaborate in the International chronic obstructive pulmonary disease Cohorts Collaboration Working Group (147). [Prognosis; Validation; Updating]

Explanation

A variety of data sources or study designs—here used as synonyms—can be used to develop or validate a prediction model. Detailed description of the design, how study participants were recruited and data collected, provides relevant details about the quality of the data, whether the proper statistical analyses have been performed, and the generalizability of the prediction model. The vulnerability for specific biases varies between designs.

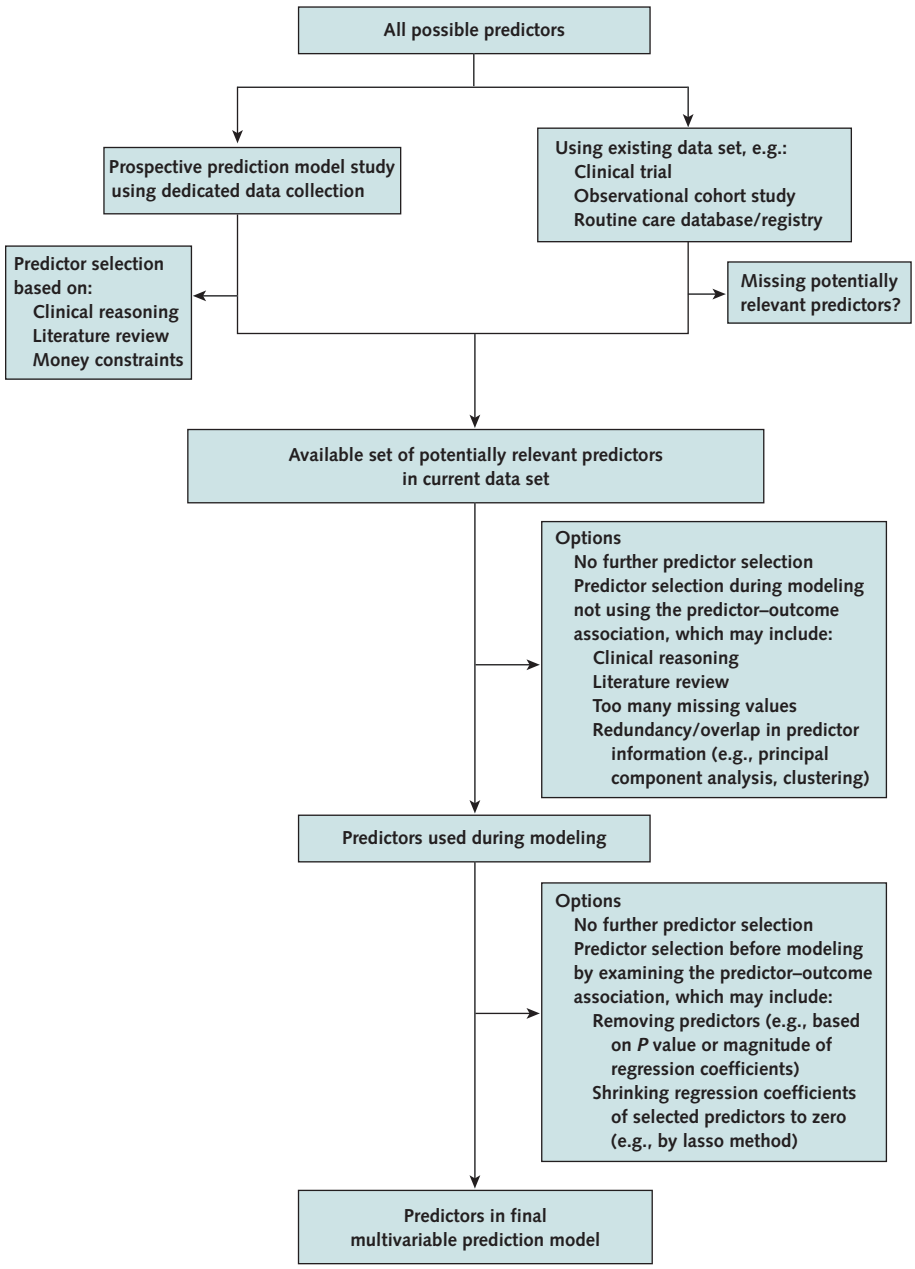
Diagnostic studies in principle study a cross-sectional relationship between the diagnostic predictors (patient characteristics and index test results) and the presence or absence of the outcome (the target condition of interest) (**Box A**). The natural design for such studies is a cross-sectional study, in which a group of patients with a certain characteristic is selected, usually defined as "suspected of having the target condition of interest" (148–151). There will often be an interval between measurement of the predictors and the outcome (reference standard). Ideally, this interval should be as short as possible and without starting any treatment within this period. Because of this short time period, and because one selects a group of patients with a similar characteristic (this is the definition of a cohort), there is debate about whether to label these studies "pure" cross-sectional studies or rather use the terms "diagnostic" (cross-sectional) cohort studies or "delayed" cross-sectional studies (152–154). Problems can arise if the interval between measurement of the predictors and of the outcome becomes too long, and certainly when intervening treatments are started; the disease status of some patients might change, thereby destroying the cross-sectional relationship of interest.

For efficiency, in some diagnostic modeling studies the reference standard is performed first, and the study uses all the cases (patients with the target condition) but a random sample of the noncases. Adjustment for sampling frequency is necessary to obtain unbiased absolute (diagnostic outcome) probabilities (152, 155–157). Such alternative sampling designs become attractive in situations where outcome (target condition) prevalence is low and the costs of measuring the predictors or index tests under study are high. A key issue here is whether the cases and noncases (controls) are representative of the cases and noncases that occur in the population of interest, that is, individuals suspected of having the target condition. A clear violation occurs in studies that include differential selection of typical, advanced cases and healthy controls (152, 155–157). Such participant selection may lead to overstatements about the clinical relevance of the study (158), and many measure of predictive performance will often be invalid (157).

The natural design of a prognostic study is a longitudinal cohort study, which can be prospective or retrospective (**Box A**) (1–3, 58, 103). Individuals enter the

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 2. Selection of predictors in a study of the development of a multivariable prediction model.



cohort on the basis of specific criteria, such as being diagnosed with a specific condition, undergoing a specific operation, or being pregnant. This is often referred to as *T = 0*, *baseline*, or *start point* (9). Subjects are then followed over time to determine whether they develop the outcome event of interest.

The preferred design is a prospective longitudinal cohort study. There is full control for ensuring that all relevant predictors and outcomes will be measured (Figure 2) and that the best method for measuring each predictor and outcome will be used, thereby minimizing the number of missing values and lost to follow-up.

In many studies, a model will be developed or validated using a data set that was originally designed and conducted for a different purpose. Although such a study may originally have been a prospective longitudinal cohort study, it may not have measured specific predictors or may have measured some predictors less well. Item 13b asks for detailed information on the number of missing values in potential predictors, and item 13a for patients lost to follow-up.

Randomized trials are a special subset of prospective longitudinal cohort studies, and can thus also be used for developing or validating prognostic models.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

However, authors should state how the intervention effect was accounted for (item 5c). There may be concerns about the generalizability of a model developed or validated by using data from a randomized trial, owing to (often) extensive exclusion criteria (1). One empirical analysis found that prognostic effects of emerging cardiovascular biomarkers (added beyond the traditional Framingham risk score) were stronger in data sets derived from observational studies than in data derived from randomized trials (159).

With international collaborations and data sharing becoming more commonplace, individual participant data from multiple studies are increasingly being used to develop and validate prediction models (89, 147, 160). Similarly, large existing data sets (for example, "big data" from national or international surveys or registries) are increasingly being used to develop and validate prediction models (139, 161, 162). For both of these data sources, data should be considered as clustered, because participants originate from different clusters (different cohorts, studies, hospitals, settings, regions, or countries), requiring a weighted approach when developing a prediction model. Recently, meta-analytical approaches have been proposed to account for such clustering of study participants (163–166). This involves accounting for different case mix reflected by different outcome prevalence (diagnosis) or incidence (prognosis) across cohorts, data sets, studies, hospitals, settings, regions, or countries, and thus accounting for different baseline probabilities or hazards (for example, by using random intercepts). But it also involves accounting for different case mix reflected by different predictor–outcome associations, by allowing for random predictor weights (regression coefficients) (163–167). Using individual participant data or other "big data" sources, enhances the possibility of developing and directly evaluating (externally validating) prediction models across hospitals, countries, or settings (**Figure 1**, study type 2b), again accounting for potential differences in intercept and predictor weights (164, 166). Extensions to commonly used measures of model performance to account for clustered data have also been recently proposed (167–171).

For reasons of efficiency or costs, sampling of patients rather than using the full cohort can be applied. Examples are case-cohort and nested case–control designs (172). Accurate reporting on the way in which patients (subjects) were sampled is required, because the sampling needs to be incorporated in the analysis to allow for proper estimation of the absolute probabilities of having or developing the outcome of interest (1, 103, 173–175). Selectively choosing or omitting participants may cast doubt on the representativeness of the sample to the population in which the model is to be applied and affect the generalizability of the prediction model.

The study design or source of data also provides relevant information about the setting and original purpose of collecting the data. The setting in combination with the eligibility criteria (item 5b) will help the reader to judge the generalizability of the model to the setting in which the reader may wish to use it.

Recent systematic reviews of prediction model studies have observed that studies often did not clearly indicate whether the sample was representative of the intended population, including whether all consecutive participants were included (34, 59, 84, 93, 176).

Item 4b. Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up. [D;V]

Examples

> This prospective temporal validation study included all patients who were consecutively treated from March 2007 to June 2007 in 19 phase I trials at the Drug Development Unit, Royal Marsden Hospital (RMH), Sutton, United Kingdom. . . . [A]ll patients were prospectively observed until May 31, 2008 (177). [Prognosis; Validation]

> All consecutive patients presenting with anterior chest pain (as a main or minor medical complaint) over a three to nine week period (median length, five weeks) from March to May 2001 were included. . . . Between October 2005 and July 2006, all attending patients with anterior chest pain (aged 35 years and over; $n = 1249$) were consecutively recruited to this study by 74 participating GPs in the state of Hesse, Germany. The recruitment period lasted 12 weeks for each practice (178). [Diagnosis; Development; Validation]

> The derivation cohort was 397 consecutive patients aged 18 years or over of both sexes who were admitted to any of four internal medicine wards at Donostia Hospital between 1 May and 30 June 2008 and we used no other exclusion criteria. The following year between 1 May and 30 June 2009 we recruited the validation cohort on the same basis: 302 consecutive patients aged 18 or over of both sexes who were admitted to any of the same four internal medicine wards at the hospital (179). [Prognosis; Development]

Explanation

Reporting the start and end dates of the period in which study participants were recruited places a study in a historical context. Readers can deduce relevant information, such as available tests and treatments in this period, and the state of the art of medical technologies to measure certain predictors. These dates, in combination with the total number of participants, may also indicate how selective enrolment was. Indeed, it is useful to report the number of enrolled participants per period (for example, per year); see item 13a.

As discussed in item 4a, time between predictor and outcome measurement is short in most diagnostic prediction studies. However, when a good reference

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 2. Example Table: Reporting Key Study Characteristics [Diagnosis; Development; Validation]

Characteristic	Swiss Population (*n* = 201)*	US Population (*n* = 258)*
Data collection period	December 1999 to February 2000	January to March 2002
Study design	Prospective cohort	Prospective cohort
Setting	University primary care clinic that serves an urban population of 150 000 in Lausanne, Switzerland	Emergency department or urgent care ambulatory patients in a large tertiary care university hospital in San Francisco, California
Inclusion criteria	Adult outpatients with influenza-like illness as determined by the primary care physician	Consecutive adults with symptoms of an acute respiratory tract infection (cough, sinus, pain, congestion/rhinorrhea, sore throat or fever) developing in past 3 weeks
Outcome	Presence of influenza A or B	Presence of influenza A or B
Reference standard	Culture	PCR
Prevalence of influenza	104 (52.8)	53 (20.5)
Men	101 (50)	103† (40)
Mean age (range), *y*	34.3 (17–86)	38.8 (18–90)
Fever	116 (58)	54 (21)
Cough	186 (93)	235 (91)
Sore throat	151 (75)	181 (70)
Myalgia	181 (90)	154 (60)
Rhinitis	163 (81)	185 (72)
Headache	169 (84)	190 (74)
Chills/sweating	166 (83)	158 (61)
Fatigue	184 (92)	197 (76)
Onset <48 hours	106 (33)	45 (17)

PCR = polymerase chain reaction.
From reference 181.
* Values are *n (%)* unless otherwise indicated.
† Out of 256 total patients.

standard does not exist, patients may be followed up to obtain a more reliable assessment of whether they had the target condition at the time the predictors were measured. In these situations, authors should report whether a maximum or minimum interval was allowed between measurement of the predictors and the final assessment to determine whether the target condition was present or absent.

In prognostic modeling studies, the duration of follow-up is critical when interpreting the performance of the model (items 6a and 13a). Follow-up from inclusion may be the same for all participants, in which case the study duration should be specified. Often, follow-up continues for all enrolled participants until the study is closed on a specific date (which should be reported). Efforts should then be made to establish participant status at that closing date; events that occur after that date are ignored.

Systematic reviews of prediction model studies have observed that key study dates are not always reported (43, 122, 176, 180). For example, in a systematic review of 61 studies developing or validating prognostic models for breast cancer, only 13 (12%) provided full information on the dates of start and end of recruitment and end of follow-up (43).

Participants

Item 5a. Specify key elements of the study setting (e.g., primary care, secondary care, general population), including number and location of centers. [D;V]

Examples

> We built on our previous risk prediction algorithm (QRISK1) to develop a revised algorithm . . . QRISK2. We conducted a prospective cohort study in a large UK primary care population using a similar method to our original analysis. We used version 19 of the QRESEARCH database (www.qresearch.org). This is a large validated primary care electronic database containing the health records of 11 million patients registered from 551 general practices (139). [Prognosis; Development; Validation]

See also Table 2.

Explanation

A detailed description of where and when study participants were recruited is particularly important so that others may judge the generalizability and usefulness of the models and to conduct further research (for example, validating or implementation of the model). "Where and when" refers not only to geographic location and calendar time, but also in which setting the participant data were collected (such as primary, secondary, tertiary, or emergency care, or general population), and adult or pediatric care. One cannot simply assume that prediction models can be directly transported from one type of setting or target population to another (19, 26, 28, 33).

Different settings have a different case mix, which commonly affects the generalizability and accuracy of prediction models; see item 4a (182–187). "Case mix" refers to the distribution of predictors, other relevant participant or setting characteristics, and the outcome prevalence (diagnosis) or incidence (prognosis), which may lead to different predictor–outcome associations potentially affecting the predictive accuracy of the model. It is well known, for example, that the predictive

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

performance of models developed in secondary care is usually lower when they are applied in a primary care setting (21, 183, 188). This is perhaps simply because primary or family care doctors selectively refer patients to secondary or tertiary care, such that the latter care populations show a narrower range of patient characteristics, a larger fraction of patients in later disease stages, and often higher outcome frequency (102, 189, 190).

Another setting difference is the transportability of prediction models from adult to pediatric care (102). For example, various prediction models have been developed to predict the risk for postoperative nausea and vomiting in adults scheduled for surgery under general anesthesia. When validated in children, the models' predictive ability was substantially decreased (191).

In general, models will be more generalizable when the case mix of the new population is within the case mix range of the development population (186). However, as we describe under item 10e (see also **Box C** and **Table 3**), one may adjust or update a previously developed prediction model that is applied in another setting to the local circumstances of the new setting to improve the model transportability.

We recommend presenting a table summarizing the key elements of the study characteristics for the development and any validation sample (192), to provide the reader insight into any differences in case mix and its potential consequences (item 5b). In addition, for validation studies, we suggest presenting a summary table of both the development and validation samples.

A systematic review of 48 studies developing or validating prognostic models for heart failure identified 10 studies (21%) failing to provide information on the number of centers (180).

Item 5b. Describe eligibility criteria for participants. [D;V]

Examples

> One hundred and ninety two patients with cutaneous lymphomas were evaluated at the Departments of Dermatology at the UMC Mannheim and the UMC Benjamin Franklin Berlin from 1987 to 2002. Eighty six patients were diagnosed as having cutaneous T cell lymphoma (CTCL) as defined by the European Organisation for Research and Treatment of Cancer classification of cutaneous lymphomas, including mycosis fungoides, Sezary Syndrome and rare variants. . . . Patients with the rare variants of CTCL, parapsoriasis, cutaneous pseudolymphomas and cutaneous B cell lymphomas were excluded from the study. . . . Staging classification was done by the TNM scheme of the mycosis fungoides Cooperative Group. A diagnosis of Sezary Syndrome was made in patients with erythroderma and >1000 Sezary cells mm) in the peripheral blood according to the criteria of the International Society for Cutaneous Lymphomas (ISCL) (193). [Prognosis; Development]

> Inclusion criteria were age 12 years and above, and injury sustained within 7 days or fewer. The authors selected 12 as the cutoff age because the emergency department receives, in the main, patients 12 years and above while younger patients were seen at a neighboring children's hospital about half a mile down the road from our hospital. In this, we differed from the original work by Stiell, who excluded patients less than 18 years of age. Exclusion criteria were: pregnancy, altered mental state at the time of consultation, patients who had been referred with an x ray study, revisits, multiply traumatized patients, and patients with isolated skin injuries such as burns, abrasions, lacerations, and puncture wounds (194). [Diagnosis; Validation]

Table 3. Overview of Different Approaches for Updating an Existing Prediction Model*

Method	Updating Method	Reason for Updating
0	No adjustment (the original prediction model)	–
1	Adjustment of the intercept (baseline risk/hazard)	Difference in the outcome frequency (prevalence or incidence) between development and validation sample
2	Method 1 + adjustment of all predictor regression coefficients by one overall adjustment factor (calibration slope)	The regression coefficients or combination thereof of the original model are overfitted or underfitted
3	Method 2 + extra adjustment of regression coefficients for predictors with different strength in the validation sample compared with the development sample	As in method 2, and the strength (regression coefficient) of one or more predictors may be different in the validation sample
4	Method 2 + selection of additional predictors (e.g., newly discovered markers)	As in method 2, and one or more potential predictors were not included in the original model, or a new predictor may need to be added
5	Reestimation of all regression coefficients, using the data of the validation sample only. If the development data set is also available, both data sets may be combined.	The strength of all predictors may be different in the validation sample, or the validation sample is much larger than the development sample
6	Method 5 + selection of additional predictors (e.g. newly discovered markers)	As in method 5, and one or more potential predictors were not included in the original model, or a new predictor may need to be added

* Information from references 31, 290, 372, and 373.

Explanation

Describing eligibility criteria is important to understand the potential applicability, and thus generalizabil-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

ity, of the prediction model. The selection process, describing who did or did not become a study participant, has implications regarding to whom the study results and predictions might be generalized.

For validation studies, it is useful to report whether the eligibility criteria for those studied were similar to or different from those used in the original derivation of the model. In the example above (194), a multivariable diagnostic prediction model to identify ankle fractures, originally derived in Canada, was validated in Asia. The authors described details of selection and contrasted them with those used in the original development study.

If some otherwise eligible participants were excluded because of missing data, this should be clearly reported. Simply omitting participants owing to missing data—either on predictors or outcome—and restricting the analysis only to those with completely observed predictor and outcome data may cause serious bias (195–201). Bias can arise because data are often not missing completely at random, but rather selectively (item 9).

Item 5c. Give details of treatments received, if relevant. [D;V]

Example

> Data from the multi-centre, worldwide, clinical trial (Action in Diabetes and Vascular disease: preterax and diamicron-MR controlled evaluation) (ADVANCE) permit the derivation of new equations for cardiovascular risk prediction in people with diabetes. . . . ADVANCE was a factorial randomized controlled trial of blood pressure (perindopril indapamide versus placebo) and glucose control (gliclazide MR based intensive intervention versus standard care) on the incidence of microvascular and macrovascular events among 11,140 high risk individuals with type 2 diabetes, recruited from 215 centres across 20 countries in Asia, Australasia, Europe and Canada. . . . Predictors considered were age at diagnosis of diabetes, duration of diagnosed diabetes, sex, systolic blood pressure, diastolic blood pressure, mean arterial blood pressure, pulse pressure, total cholesterol, high-density lipoprotein and non high-density lipoprotein and triglycerides, body mass index, waist circumference, Predictors waist to hip ratio, blood pressure lowering medication (i.e. treated hypertension), statin use, current smoking, retinopathy, atrial fibrillation (past or present), logarithmically transformed urinary albumin/creatinine ratio (ACR) and serum creatinine (S_{cr}), haemoglobin A1c (HbA1c), fasting blood glucose and randomized treatments (blood pressure lowering and glucose control regimens) (145). [Prognosis; Development; Validation]

Explanation

Cohorts for studying prognosis are defined by some presenting health state (202). In many prognostic studies, the participants have received preventive or curative interventions, either before or at the start of the follow-up period, which may influence their prognosis. An effective treatment will typically improve their prognosis, leading to a reduced probability of the outcome (203).

Developing a "pure baseline" prognostic model for predicting future outcomes of participants in a particular health state who have not been exposed to treatment is rarely possible. Frequently, participants will have received some treatment. Ideally, either all study participants receive the same treatment, such as a surgical procedure, or the treatment is chosen at random, as when prognostic models are based on randomized trial data; see item 4a (1, 204). Some prognostic models are explicitly developed and validated for patients receiving a particular intervention (205), but even here, there may be variation in co-interventions.

When randomized trial data are used, separate prognostic models can be fitted to those receiving the different interventions, especially in the presence of an effective intervention. Treatment could instead be included as a predictor in a model developed from all participants (item 7a); interactions between treatment and other predictors (item 10b) may be studied to allow for different predictions under different treatment strategies (1, 4). The focus here is not on the preventive or therapeutic effects of the intervention, but on their independent contribution to the outcome prediction. In many cases, however, the predictive effect of interventions is rather small compared with the important predictors, such as age, sex, and disease stage (1), so that treatment is excluded from the modeling or gets omitted during the predictor selection process.

In nonrandomized studies, not only is there variation in treatments received, but also a serious concern is that treatment choices for individuals may well have been influenced by the same predictors that are included in the statistical modeling (206). As with randomized trial data, treatment can still be considered as a predictor in the modeling, but the effect on the prediction model of treatment being influenced by other predictors cannot easily be judged. The preceding comments relate to treatments received before the start of the follow-up period. Treatments received at later times require very sophisticated models that are seldom applied in prediction model studies (207).

A rather different situation is when current treatment is used as a proxy for other underlying predictors. Examples include the use of antihypertensive or cholesterol-lowering medication as a proxy for hypertension or hypercholesterolemia, respectively, in cardiovascular risk models (17, 208). The consequences of this approach on the performance of prediction models is not yet fully understood.

In view of the above considerations, it is important for both developing and validating a prediction model to know which interventions the study participants

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

received that might have modified the probability for the outcome (203) (item 13b).

The issue of treatments is less relevant in most diagnostic prediction model studies, because these studies have a cross-sectional design in which the predictors and outcome are recorded at the same time (Box A). Sometimes, however, there is some interval between the predictor and outcome assessment (for example, when the outcome measure is based in part on follow-up information [209]). Then, any treatments received between the moment of prediction and outcome measurement represent relevant information and should be reported.

A recent review of 21 cardiovascular risk scores found that outcome-modifying interventions were not accounted for and that reporting of prior treatments was incomplete (203).

Outcome

Item 6a. Clearly define the outcome that is predicted by the prediction model, including how and when assessed. [D;V]

Examples

Outcomes of interest were any death, coronary heart disease related death, and coronary heart disease events. To identify these outcomes, cohort participants were followed over time using a variety of methods, including annual telephone interviews, triennial field center examinations, surveillance at ARIC community hospitals, review of death certificates, physician questionnaires, coroner/medical examiner reports, and informant interviews. Follow up began at enrollment (1987 to 1989) and continued through December 31, 2000. Fatal coronary heart disease included hospitalized and nonhospitalized deaths associated with coronary heart disease. A coronary heart disease event was defined as hospitalized definite or probable myocardial infarction, fatal coronary heart disease, cardiac procedure (coronary artery bypass graft, coronary angioplasty), or the presence of serial electrocardiographic changes across triennial cohort examinations. Event classification has been described in detail elsewhere [ref] (210). [Prognosis; Development]

Definite urinary tract infection was defined as $\geq 10^8$ colony forming units (cfu) per litre of a single type of organism in a voided sample $\geq 10^7$ cfu/L of a single organism in a catheter sample or any growth of a single organism in a suprapubic bladder tap sample. Probable urinary tract infection was defined as $\geq 10^7$ cfu/L of a single organism in a voided sample $\geq 10^6$ cfu/L of a single organism in a catheter sample $\geq 10^8$ cfu/L of two organisms in a voided sample or $\geq 10^7$ cfu/L of two organisms from a catheter sample (211). [Diagnosis; Development; Validation]

Patient charts and physician records were reviewed to determine clinical outcome. Patients generally were seen postoperatively at least every 3–4 months for the first year, semi annually for the second and third years, and annually thereafter. Follow up examinations included radiological imaging with computed tomography in all patients. In addition to physical examination with laboratory testing, intravenous pyelography, cystoscopy, urine cytology, urethral washings and bone scintigraphy were carried out if indicated. Local recurrence was defined as recurrence in the surgical bed, distant as recurrence at distant organs. Clinical outcomes were measured from the date of cystectomy to the date of first documented recurrence at computed tomography, the date of death, or the date of last follow up when the patient had not experienced disease recurrence (212). [Prognosis; Development]

Breast Cancer Ascertainment: Incident diagnoses of breast cancer were ascertained by self-report on biennial follow up questionnaires from 1997 to 2005. We learned of deaths from family members, the US Postal Service, and the National Death Index. We identified 1084 incident breast cancers, and 1007 (93%) were confirmed by medical record or by cancer registry data from 24 states in which 96% of participants resided at baseline (213). [Prognosis; Validation]

Explanation

For diagnostic models, the outcome is the presence or absence of a specific target condition at T0 (Box A). This diagnostic outcome is determined using a so-called reference standard, which should be the best available and well-accepted method for establishing the presence or absence of that condition (214). The rationale for the choice of reference standard should be stated. The reference standard can take many forms and may be a single test, a combination of tests, or some other method including a consensus based approach using an expert or outcome committee. Reference standard tests be laboratory, radiology, arthroscopy, angiography, or pathology assessments.

If relevant, blood or urine sampling methods, laboratory and imaging methods, technology, and definitions should be specified, including any cut-offs that were used to define the presence (or severity) of the target condition, as should rules to combine test results (composite reference standard) to establish the diagnostic outcome (215–217). Reasons for not using standard definitions and thresholds should be specified. If

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

multiple outcome examiners were used (such as when using a consensus-based outcome committee), the method for establishing the final diagnosis (for example, majority vote) should be described (215, 216).

In diagnostic modeling studies, the timing between assessment of the outcome relative to the assessment of the predictors should be specified, because of the potential for bias resulting from a change in the underlying condition since the predictor assessments (**Box A**). Furthermore, the order in which predictors and outcome were assessed should be explicitly reported (see also items 6b and 7b on the potential for bias in relation to nonblind assessments).

Ideally, the diagnostic outcome is verified for all participants using the same reference standard. This is not always possible. For example, it may be deemed unethical to subject patients to an invasive reference standard unless they have a positive result on 1 or more index tests. Two situations may then occur: partial verification, when outcome data are completely missing (item 9) for the subset of participants for whom there is no reference standard result, and differential verification, when patients who are not referred to the preferred reference standard are assessed using an alternative reference standard of differing, usually lower, accuracy (218, 219).

For instance, in cancer detection studies, pathology (reference standard) results are likely to be available only for those participants who have some positive index test. For the remaining participants, the alternative reference standard may be a period of follow-up that is long enough for cancers present at the time of index test measurement to become apparent (delayed verification), but not long enough for new incident cancers. Rules and procedures for partially or differentially verifying the outcome should be well described to allow assessment of the potential for so-called partial and differential verification bias (156, 218–220), and methods adjusting for these verification biases may be considered (219).

For prognostic models, common outcomes include death (from any cause or cause specific), nonfatal complications or events (for example, myocardial infarction, cancer recurrence, disease progression, or disease onset), and patient-centered outcomes (such as symptoms, functional status, and quality of life) (2). Combinations of outcomes are also used. For instance, events of interest for disease-free survival in cancer studies may include local recurrence, regional disease, distant metastases, and death (whichever occurs first) (221).

All outcomes should be defined unambiguously. If standard definitions are used (for example, International Classification of Diseases [ICD] codes), this should be stated and referenced, as well as any variations. Technical details provided in a study protocol or previous papers should be clearly referenced and ideally made available.

Prognostic studies follow participants over a period of time and document when the outcome occurs after the prognostic time origin (T0) (for example, the date of diagnosis or surgery; see **Box A**). Some studies assess all participants on their outcome status within a fixed period (for example, overall survival) and often at prespecified time points during follow-up (for example, 5- or 10-year cardiovascular disease risk); these time points should be clearly reported (222). Similarly, the frequency of outcome assessment during follow-up should also be clearly reported.

The data sources used to identify the occurrence of the outcomes, or loss to follow-up, should be specified. Examples include death registry data, hospital records, cancer registry data, clinical assessments, scans, or laboratory tests. For such outcomes as cause-specific mortality, the process by which cause of death was assigned should be clearly explained (for example, adjudication or end point committees). The composition of such committees and expertise of members should also be reported in brief (216).

A recent review of 47 studies reporting the development of prognostic models in cancer concluded that outcomes were poorly defined in 40% of studies (54). In 30%, it was unclear whether "death" referred to a cancer death or death from any cause. There was also inconsistency regarding which events were included in the definition of disease-free survival.

Item 6b. Report any actions to blind assessment of the outcome to be predicted. [D;V]

Examples

> All probable cases of serious bacterial infection were reviewed by a final diagnosis committee composed of two specialist paediatricians (with experience in paediatrics infectious disease and respiratory medicine) and in cases of pneumonia a radiologist. The presence or absence of bacterial infection [outcome] was decided blinded to clinical information [predictors under study] and based on consensus (211). [Diagnosis; Development; Validation]

> Liver biopsies were obtained with an 18 gauge or larger needle with a minimum of 5 portal tracts and were routinely stained with hematoxylin-eosin and trichrome stains. Biopsies were interpreted according to the scoring schema developed by the METAVIR group by 2 expert liver pathologists . . . who were blinded to patient clinical characteristics and serum measurements. Thirty biopsies were scored by both pathologists, and interobserver agreement was calculated by use of κ statistics (223). [Diagnosis; Development; Validation]

> The primary outcome [acute myocardial infarction coronary revascularization or death of cardiac or unknown cause within 30 days] was ascertained by investigators blinded to the predictor variables. If a diagnosis could not be assigned a cardiologist . . . reviewed all the clinical data and assigned an adjudicated outcome diagnosis. All positive and 10% of randomly selected negative outcomes were con-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

firmed by a second coinvestigator blinded to the standardized data collection forms. Disagreements were resolved by consensus (224). [Prognosis; Development]

Explanation

In prediction model studies, the outcome is ideally assessed while blinded to information about the predictors. The predictors may otherwise influence the outcome assessment, leading to a biased estimation of the association between predictors and outcome (148, 209, 225, 226). This risk is clearly less for objective outcomes, such as death from any cause or cesarean section. However, it is more relevant for outcome assessments requiring interpretation, such as cause-specific death.

Some outcomes are inherently difficult to assess or lack an established reference standard. Researchers may then explicitly want to use all available information for each patient (including information from predictors) to determine whether the outcome is indeed present or absent. In diagnostic research, this approach is known as *consensus diagnosis*, and adjudication or end-point committees are examples from prognostic and intervention research (item 6a) (149). If the explicit aim is to assess the incremental value of a particular predictor or comparing the performance of competing models (for example, when validating multiple models), the importance of blinded outcome assessment increases to prevent overestimation of a predictor's incremental value or to prevent biased preference for one model to another.

Researchers thus should carefully consider and clearly report which information was available to the assessors of the outcome and report any specific actions for blinding of the outcome assessment, if appropriate. However, systematic reviews have frequently reported a lack of information when trying to assess whether there was blind assessment of the outcome (34, 227).

Predictors

Item 7a. Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured. [D;V]

Examples

The following data were extracted for each patient: gender, aspartate aminotransferase in IU/L, alanine aminotransferase in IU/L, aspartate aminotransferase/alanine aminotransferase ratio, total bilirubin (mg/dl), albumin (g/dl), transferrin saturation (%), mean corpuscular volume (μm^3), platelet count ($\times 10^3/mm^3$), and prothrombin time(s). . . . All laboratory tests were performed within 90 days before liver biopsy. In the case of repeated test, the results closest to the time of the biopsy were used. No data obtained after the biopsy were used (228). [Diagnosis; Development]

Forty three potential candidate variables in addition to age and gender were considered for inclusion in the AMI [acute myocardial infarction] mortality prediction rules. . . . These candidate variables were taken from a list of risk factors used to develop previous report cards in the California Hospital Outcomes Project and Pennsylvania Health Care Cost Containment Council AMI "report card" projects. Each of these comorbidities was created using appropriate ICD 9 codes from the 15 secondary diagnosis fields in OMID. The Ontario discharge data are based on ICD 9 codes rather than ICD 9 CM codes used in the U.S., so the U.S. codes were truncated. Some risk factors used in these two projects do not have an ICD 9 coding analog (e.g., infarct subtype, race) and therefore were not included in our analysis. The frequency of each of these 43 comorbidities was calculated, and any comorbidity with a prevalence of <1% was excluded from further analysis. Comorbidities that the authors felt were not clinically plausible predictors of AMI mortality were also excluded (185). [Prognosis; Development; Validation]

Each screening round consisted of two visits to an outpatient department separated by approximately 3 weeks. Participants filled out a questionnaire on demographics, cardiovascular and renal disease history, smoking status, and the use of oral antidiabetic, antihypertensive, and lipid lowering drugs. Information on drug use was completed with data from community pharmacies, including information on class of antihypertensive medication. . . . On the first and second visits, blood pressure was measured in the right arm every minute for 10 and 8 minutes, respectively, by an automatic Dinamap XL Model 9300 series device (Johnson & Johnson Medical Inc., Tampa, FL). For systolic and diastolic BP, the mean of the last two recordings from each of the 2 visit days of a screening round was used. Anthropometrical measurements were performed, and fasting blood samples were taken. Concentrations of total cholesterol and plasma glucose were measured using standard methods. Serum creatinine was measured by dry chemistry (Eastman Kodak, Rochester, NY), with intra assay coefficient of variation of 0.9% and interassay coefficient of variation of 2.9%. eGFR [estimated glomerular filtration rate] was estimated using the Modification of Diet in Renal Disease (MDRD) study equation, taking into account gender, age, race, and serum creatinine. In addition, participants collected urine for two consecutive periods of 24 hours. Urinary albumin

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

concentration was determined by nephelometry (Dade Behring Diagnostic, Marburg, Germany), and UAE [urinary albumin excretion] was given as the mean of the two 24 hour urinary excretions. As a proxy for dietary sodium and protein intake, we used the 24 hour urinary excretion of sodium and urea, respectively (229). [Prognosis; Development]

Explanation

Predictors are typically obtained from participant demographic characteristics, medical history, physical examination, disease characteristics, test results, and previous treatment (1). Predictors should be fully defined, providing the units of measurement for any continuous predictors and all the categories for any categorical predictors (including whether categories have been combined). This is to ensure that readers and other investigators can potentially replicate the study and, more important, validate or implement the prediction model. If applicable, relevant sampling, laboratory and imaging methods should be specified, including any cut-offs that were used to define the presence (or severity) of a specific predictor, or rules to combine predictors (for example, mean blood pressure).

Authors should also explain how and when the predictors were measured. All predictors should be measured before or at the study time origin and known at the moment the model is intended to be used (1, 230, 231). Blood or tissue samples collected at or before the study time origin may be analyzed later; the important issue is when the samples were obtained and the predictors being used. Predictors measured after the time origin may be more appropriately examined as outcomes and not predictors unless time-dependent methods are used (232). However, statistical methods for handling predictors measured during follow-up (233, 234) are seldom used in prediction model studies. Predictor measurement methods (including assay and laboratory measurement methods) should be reported in a complete and transparent manner, with the level of detail that would enable it to be reproduced and evaluate the generalizability of the prediction model that includes such predictors.

In many studies developing prediction models, a large number of predictors are collected and available for statistical analysis (**Figure 2**). However, the larger the number of available predictors, the greater the chance of erroneously selecting weak and uninformative predictors in the final model, leading to so-called model overfitting and optimism (particularly in small data sets); see item 8. Moreover, smaller models are easier to apply in clinical practice than larger models. Reducing the number of available predictors before or during the analysis is therefore often necessary (2, 235); see item 10b. Reasons for omitting any available predictors from the set of predictors used in the modeling should be clearly reported (**Figure 2**).

Recent systematic reviews have highlighted frequent insufficient reporting to clearly identify all available predictors, the total number of predictors analyzed, and how and when they were selected (34, 43, 45, 53, 54, 73, 74, 80, 81, 87, 182). In a review of 29 prediction models in reproductive medicine, 34% of studies failed to provide an adequate description of the predictors (80).

Item 7b. Report any actions to blind assessment of predictors for the outcome and other predictors. [D;V]

Examples

A single investigator blinded to clinical data and echocardiographic measurements performed the quantitative magnetic resonance image analyses. [The aim was to specifically quantify the incremental diagnostic value of magnetic resonance beyond clinical data to include or exclude heart failure] (236). [Diagnosis; Development; Incremental value]

Blinded to [other] predictor variables and patient outcome [a combination of nonfatal and fatal cardiovascular disease and overall mortality within 30 days of chest pain onset], 2 board certified emergency physicians . . . classified all electrocardiograms [one of the specific predictors under study] with a structured standardized format . . . (224). [Prognosis; Development]

Investigators, blinded to both predictor variables and patient outcome, reviewed and classified all electrocardiograms in a structured format according to current standardized reporting guidelines. Two investigators blinded to the standardized data collection forms ascertained outcomes. The investigators were provided the results of all laboratory values, radiographic imaging, cardiac stress testing, and cardiac catheterization findings, as well as information obtained during the 30 day follow up phone call (237). [Diagnosis; Validation]

Explanation

Assessment of predictors may be influenced if assessors are not blinded to other observed information, either the outcome or other predictors (1, 225, 238–240). In a similar manner to blind outcome assessment (item 6b), the need for blind assessment of predictors is important, particularly those predictors that require subjective judgment and assessment, such as imaging, electrophysiology, and pathology results, and not so much for such predictors as age, sex, or quantitative laboratory values that are largely independent of observer interpretation.

Blinding to Outcome Information

Assessment of predictors should always be done without knowledge of the participant's outcome. Knowledge of the outcome will indirectly be included in or will contaminate the predictor assessment, thereby artificially increasing the association between

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

the predictor and outcome (1, 225, 239). Blinding predictor assessors for outcome information is inherent in follow-up studies for which outcomes are by design measured after the predictors, as is common for prognostic research. Potential bias due to predictor assessments contaminated by outcome knowledge, thus notably occurs in case–control studies, and in studies with a cross-sectional design in which predictors and outcomes are assessed close in time (225). Hence, this bias is more likely to occur in diagnostic modeling studies. It should thus be explicitly stated whether any outcome information was available when interpreting results of predictors (or index tests).

Blinding to Other Predictor Information

Assessors of predictors requiring interpretation may also be provided with other information (for example, prior information obtained during assessment of medical history or physical examination). Unlike blinding for the outcome information when assessing predictors, blinding for information of other predictors is not per se good or bad. The appropriateness depends on the research question and the potential clinical application of the specific predictors (209, 225, 226). Interpretation of subsequent predictors with knowledge of prior predictor information may be specifically designed, if in daily practice these subsequent predictors are always interpreted in view of this prior information. For example, predictors from additional imaging or electrophysiology measurements are commonly interpreted with knowledge of results from history-taking and physical examination.

Also, if the research purpose is to quantify the incremental value of a specific predictor to predictors that are in practice known anyhow, blinding the assessor of the former to the latter may be unwarranted. However, if a research purpose is to quantify whether a particular predictor or test may replace another predictor or test (for example, whether positron emission tomography–computed tomography may replace traditional scanning for the detection of cancer lesions in the lungs), mutually blinding the observers of both for each others' results is indicated to prevent contamination in both interpretations (225, 239). Nonblinding is likely to make both readings, and thus results, become more alike.

It should therefore be reported which predictor assessments, if any, were blinded to other predictor information, in relation to the study aim and where and how the predictors in the model will be used in practice.

Numerous systematic reviews have shown that blind assessment of predictors is either not carried out or not reported (3, 58, 67, 69, 95, 241). For example, only 47% of 137 studies describing the development prediction models in pediatrics explicitly reported blinding of the assessment of predictors.

Sample Size

Item 8. Explain how the study size was arrived at. [D;V]

Examples

We estimated the sample size according to the precision of the sensitivity of the derived decision rule. As with previous decision rule studies we prespecified 120 outcome events to derive a rule that is 100% sensitive with a lower 95% confidence limit of 97.0% and to have the greatest utility for practicing emergency physicians we aimed to include at least 120 outcome events occurring outside the emergency department (in hospital or after emergency department discharge). Review of quality data from the Ottawa hospital indicated that 10% of patients who presented to the emergency department with chest pain would meet outcome criteria within 30 days. We estimated that half of these events would occur after hospital admission or emergency department discharge. The a priori sample size was estimated to be 2400 patients (224). [Diagnosis; Development]

Our sample size calculation is based on our primary objective (i.e., to determine if preoperative coronary computed tomography angiograph has additional predictive value beyond clinical variables). Of our two objectives, this objective requires the largest number of patients to ensure the stability of the prediction model. . . . On the basis of the VISION Pilot Study and a previous non-invasive cardiac testing study that we undertook in a similar surgical population, we expect a 6% event rate for major perioperative cardiac events in this study. Table 2 presents the various sample sizes needed to test four variables in a multivariable analysis based upon various event rates and the required number of events per variable. As the table indicates, if our event rate is 6% we will need 1000 patients to achieve stable estimates. If our event rate is 4%, we may need up to 1500 patients. We are targeting a sample size of 1500 patients but this may change depending on our event rate at 1000 patients (242). [Prognosis; Development]

All available data on the database were used to maximise the power and generalisability of the results (243). [Diagnosis; Development]

We did not calculate formal sample size calculations because all the cohort studies are ongoing studies. Also there are no generally accepted approaches to estimate the sample size requirements for derivation and validation studies of risk prediction models. Some have suggested having at least 10 events per candidate variable for the derivation of a model and at least 100 events for validation studies. Since many studies to develop and validate prediction models are small a potential solution is to

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

have large scale collaborations as ours to derive stable estimates from regression models that are likely to generalize to other populations. Our sample and the number of events far exceeds all approaches for determining samples sizes and therefore is expected to provide estimates that are very robust (147). [Prognosis; Validation]

We calculated the study sample size needed to validate the clinical prediction rule according to a requirement of 100 patients with the outcome of interest (any intra-abdominal injury present), which is supported by statistical estimates described previously for external validation of clinical prediction rules. In accordance with our previous work, we estimated the enrolled sample would have a prevalence rate of intra-abdominal injury of 10%, and thus the total needed sample size was calculated at 1,000 patients (244). [Diagnosis; Validation]

Explanation

Although there is a consensus on the importance of having an adequate sample size for developing a prediction model, how to determine what counts as "adequate" is not clear. As for all medical research, a larger sample size yields more precise results. In the absence of bias, larger samples also yield more reliable findings. Crucially, in prediction studies (development and validation), the number of outcome events dictates the effective sample size; for a binary or time-to-event outcome, the effective sample size is the smaller of the 2 outcome frequencies. A large sample size, in terms of the number of individuals, may be inadequate if few individuals have the actual outcome.

Often, however, a data set may already be readily available with measurements on potential predictors and outcomes from an entire large cohort, and it would make sense to use the entire data set, regardless of whether it meets specific sample size calculations. If so, such circumstances should be clearly indicated rather than try to justify the sample size of the data set on the basis of arbitrary post hoc sample size calculations.

Development Study

As discussed under item 10b, a model's performance is likely to be overestimated when it is developed and assessed for its predictive accuracy on the same data set (23). That problem will be greatest with small sample sizes (25, 32, 112). Although the amount of optimism in the model can be estimated and adjusted for using internal validation and shrinkage techniques (discussed in item 10b), it is better to have a large sample in the first place. These concerns apply even when no predictor selection will be performed. They are far greater, however, when the predictors in the model will be selected from a large number of available predictors (**Figure 2**), especially when there are no strong predictors. With a small sample, there will be an increased risk for selecting spurious predictors (overfitting; item 10b) and an increased risk for failing to include important predictors (underfitting) (25, 26, 32, 112).

On the basis of some empirical investigations (245, 246) a rule of thumb for sample size was suggested that has been quite widely adopted. The rule is to have at least 10 outcome events per variable (EPV), or more precisely, per parameter estimated. Others, however, have suggested that the value of 10 is too strict (247) or indeed too lax (25, 32, 248, 249). In addition, it may be that the EPV is not the best basis for making these judgments (250). In principle, the sample size could be chosen instead to allow certain metrics of model performance to be estimated with a given precision. Measures that can be examined in this way include the *c*-index, R^2, Brier score, sensitivity and specificity, and many others (251–253).

In practice, researchers are often restricted to using an available data set. Such measures as EPV are often then just descriptive; however, the number of predictors analyzed may be reduced to control the EPV (**Box C**). Only for a planned prospective prediction model development study will the sample size be predetermined on statistical grounds, on the basis of the approaches mentioned above.

Authors should explain how the sample size was determined. If it is based on statistical considerations, these should be detailed. Frequently, sample size will be determined by practical considerations, such as time, availability of existing data, or cost. In these instances, it is helpful to discuss the adequacy of the sample size in relation to the number of predictors under study or the primary performance measures.

Validation Study

A validation study has a specific goal: quantifying the performance of an existing model in other data (**Box C** and **Figure 1**). Sample size requirements for validation studies are not well understood, and there is a dearth of empirical evidence to guide investigators. Sample size is therefore often determined by the available data, but in some cases it is possible to choose sample size on statistical grounds.

The limited empirical evidence to support investigators in guiding their sample size choice for validation studies suggests a minimum of 100 events and 100 nonevents (112, 254), whereas more than 250 events have been suggested as preferable (2). However, these suggestions have been based on limited simulation studies adopting a statistical hypothesis testing framework. Possible considerations include hypothesis testing (for example, test whether the calibration slope is <1, or a prespecified reduction in the *c*-index), or preferably a focus on the precision and accuracy of the performance measures in the new data.

Numerous systematic reviews have observed that prediction model studies, both development and validation studies, frequently provided no rationale for the sample size or any mention of overfitting (34, 54, 255).

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Missing Data

Item 9. Describe how missing data were handled (for example, complete-case analysis, single imputation, multiple imputation), with details of any imputation method. [D;V]

Examples

We assumed missing data occurred at random depending on the clinical variables and the results of computed tomography based coronary angiography and performed multiple imputations using chained equations. Missing values were predicted on the basis of all other predictors considered the results of computed tomography based coronary angiography as well as the outcome. We created 20 datasets with identical known information but with differences in imputed values reflecting the uncertainty associated with imputations. In total 667 (2%) clinical data items were imputed. In our study only a minority of patients underwent catheter based coronary angiography. An analysis restricted to patients who underwent catheter based coronary angiography could have been influenced by verification bias. Therefore we imputed data for catheter based coronary angiography by using the computed tomography based procedure as an auxiliary variable in addition to all other predictors. Results for the two procedures correlate well together especially for negative results of computed tomography based coronary angiography. This strong correlation was confirmed in the 1609 patients who underwent both procedures (Pearson r = 0.72). Since its data were used for imputation the computed tomography based procedure was not included as a predictor in the prediction models. Our approach was similar to using the results of computed tomography based coronary angiography as the outcome variable when the catheter based procedure was not performed (which was explored in a sensitivity analysis). However this approach is more sophisticated because it also takes into account other predictors and the uncertainty surrounding the imputed values. We imputed 3615 (64%) outcome values for catheter based coronary angiography. Multiple imputations were performed using Stata/SE 11 (StataCorp) (256). [Diagnosis; Development]

If an outcome was missing, the patient data were excluded from the analysis. Multiple imputation was used to address missingness in our nonoutcome data and was performed with SAS callable IVEware (Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI). Multiple imputation has been shown to be a valid and effective way of handling missing data and minimizes bias that may often result from excluding such patients. Additionally, multiple imputation remains valid even if the proportion of missing data is large. The variables included in the multiple imputation model were the 4 outcomes, age, sex, ICD-9 E codes, emergency department Glasgow coma score, out of hospital Glasgow coma score, Injury Severity Score, mechanism of trauma, and trauma team notification. Ten imputed data sets were created as part of the multiple imputation, and all areas under the receiver operating characteristic curve were combined across the 10 imputed data sets with a standard approach. Although there is no reported conventional approach to combining receiver operating characteristic curves from imputed data sets, we averaged the individual sensitivity and specificity data across the 10 imputed data sets and then plotted these points to generate the curves in our results (257). [Prognosis; Validation]

We split the data into development (training) and validation (test) data sets. The development data included all operations within the first 5 years; the validation data included the rest. To ensure reliability of data, we excluded patients who had missing information on key predictors: age, gender, operation sequence, and number and position of implanted heart valves. In addition, patients were excluded from the development data if they were missing information on >3 of the remaining predictors. Any predictor recorded for <50% of patients in the development data was not included in the modeling process, resulting in the exclusion of left ventricular end diastolic pressure, pulmonary artery wedge pressure, aortic valve gradient, and active endocarditis. Patients were excluded from the validation data if they had missing information on any of the predictors in the risk model. To investigate whether exclusions of patients as a result of missing data had introduced any bias, we compared the key preoperative characteristics of patients excluded from the study with those included. Any remaining missing predictor values in the development data were imputed by use of multiple imputation techniques. Five different imputed data sets were created (258). [Prognosis; Development; Validation]

Explanation

Almost all prediction model studies have some missing outcome or predictor data. Yet, few studies ex-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

plicitly discuss missing data, and even fewer attempt to address the issue statistically (34, 45, 53, 259). In the absence of a mention of missing data, it is reasonable to assume that participants with missing data have been omitted from any analyses, leading to a so-called complete-case analysis. Including only participants with complete data is not only inefficient (it may greatly reduce the sample size) but may also lead to biased results when the remaining individuals without missing data are not representative of the whole original study sample (that is, they are a selective subsample) (Box D). For studies developing or validating a prediction model, this selection bias will lead to different (biased) estimates of the predictor–outcome associations (in model development) and of the model's predictive performance (in model development and validation) compared with what would be obtained if the whole data set could be analyzed. Multiple imputation methods are now embedded in most commonly used statistical packages (Stata, R, SAS), allowing estimation (imputation) of any missing observation and subsequent analysis of the multiple imputed data sets. We refer the reader to existing guidance for combining estimates of interest (regression coefficients, predictive performance measures) for prediction model studies after multiple imputation (Box D).

Authors of prediction model studies are recommended to carefully report details about missing data (item 13b) and describe how these were handled (item 9). If individuals with any missing values are excluded from the analysis, then this should be clearly stated in the eligibility criteria (item 5b), with a rationale for the exclusion.

Key details to include when reporting how missing data were handled, on the basis of existing guidance (56, 200, 259), are presented in Table 4. For studies that both develop and validate a prediction model, authors should clearly indicate how missing data were handled for both data sets and describe any differences.

Systematic reviews evaluating the methodological conduct and reporting of studies developing and evaluating prediction models have consistently shown poor reporting of missing data and how they were handled (34, 43, 45, 53, 56, 59, 60, 62, 64, 66, 70, 71, 76, 78–84, 88, 93, 122, 176, 260, 261).

Statistical Analysis Methods

Item 10a. Describe how predictors were handled in the analyses. [D]

Examples

> For the continuous predictors age, glucose, and Hb [hemoglobin], a linear relationship with outcome was found to be a good approximation after assessment of nonlinearity using restricted cubic splines (262). [Prognosis]

> Fractional polynomials were used to explore presence of nonlinear relationships of the continuous predictors of age, BMI [body mass index], and year to outcome (258). [Prognosis]

Box D. Missing data.

Missing values, for either predictors or outcomes, occur in all types of medical research, including diagnostic and prognostic modeling studies and in both development and validation studies. Unless prompted to do otherwise, most statistical packages explicitly exclude individuals with any missing value on any of the data analyzed. The resulting so-called "available case" or "complete case" analysis is the most common approach to handle missing data. A small number of missing values in each of several study variables can, however, result in a large number of patients excluded from a multivariable analysis. Simply excluding records with missing data does not necessarily affect the validity of the results, if the deleted records are a completely random subset of the original study sample (195–200). However, if individuals with missing data are not representative of the original study sample a complete case analysis will be biased. The extent of the bias will depend on various factors including the number of individuals with missing data (10, 195–201, 492). Use of a separate category indicating missing data has been shown to bias results and is clinically nonsensical for prediction model studies and should be avoided (195, 196).

Data are described as "missing completely at random" (MCAR) if the probability that a specific observation is missing is not related to any observed study variables, predictors, or outcome. Data are "missing at random" (MAR) if missingness is related to other observed variables. Data are "missing not at random" (MNAR) if the probability of being missing depends on unobserved values, including possibly the missing value itself (493, 494). Although it is possible to verify the data to judge whether missing data are missing completely at random or associated with observed variables, it is generally impossible to prove that data are indeed MAR, let alone whether they are MNAR.

Instead of simply omitting all individuals with any missing value or using the missing indicator method, a more effective group of methods to deal with missingness that is related to observed variables, and thus assume a MAR mechanism, are so-called imputation techniques. Such imputation may include an overall mean or median imputation, a stratified or subgroup imputation, or using a multivariable model. The latter imputation approach can be done once (single imputation) or more than once (multiple imputation) (493–495).

Multiple imputation is advocated as the preferred imputation method, and also leads to more correct standard errors and *P* values; in single imputation, these are estimated too small (low), falsely increasing chance findings (103, 195–200, 492, 496). Multiple imputation involves creating multiple copies of the data set, with the missing values replaced by imputed values drawn from their predicted distribution by using the observed data (493, 497). Standard texts on multiple imputation typically suggest 5 or 10 imputations to be sufficient. However, more recently, it has been suggested that the number of imputations should be much larger and related to the fraction of missing information in the data (495). Finally, standard statistical analyses can be applied on each imputed data set which can then be combined (using the Rubin rule [494]) to produce an overall estimate of each regression coefficient or model performance measure (item 10d) (2, 498), while taking into account uncertainty in the imputed values (196–201, 492, 495, 499, 500).

> The nonlinear relationships between these predictor variables and lung cancer risk were estimated using restricted cubic splines. Splines for age, pack-years smoked, quit-time and smoking duration were prepared with knot

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 4. Key Information to Report About Missing Data

In the Methods section:

- A clear description of the method used to account for missing data on both predictors and outcome (e.g., complete case, single imputation, multiple imputation)
- Possible reasons for any missingness
- For imputation (single or multiple)–based analyses:
  - o Provide details of the software used (including any specific imputation routines–e.g., ICE, MICE, PROC MI, Amelia, aregImpute)
  - o List the variables that were included in the imputation procedure, including whether the outcome was included for imputing the predictors and vice versa
  - o Explain how continuous, binary, and categorical predictors were handled in the imputation model
  - o State whether any interactions were included in the imputation model
  - o Report the number of imputations if multiple imputation was used

In the Results section:

- The number of individuals with any missing value, 1 missing value, 2 missing values, etc.
- The number of missing values (per predictor and outcome)
- Comparison of the characteristics of individuals with any missing value and those with completely observed data. This provides some indication whether missingness on specific study variables (predictors or outcomes) was indeed missing completely at random or related to observed characteristics (Box D)

placement based on the percentile distributions of these variables in smokers only. Knots for age were at 55, 60, 64, and 72 years. Knots for pack-years were at 3.25, 23.25 and 63 pack-years. Knots for quit-time were at 0, 15, and 35 years. Knots for duration were at 8, 28, and 45 years (263). [Prognosis]

Explanation

Many predictors are measured as continuous variables. Researchers must decide how to handle these in the analysis. Although converting them into categorical predictors is ubiquitous in studies developing clinical prediction models, there are major concerns about the approach. Box E explains why continuous predictors should ideally be kept as continuous, and also examined to see whether they have a linear or nonlinear relationship with the outcome.

In the absence of a priori clinical consensus, authors who wish to categorize or even dichotomize continuous predictors are recommended to use a non–data-driven method. Choosing so-called optimal cut points on the basis of minimizing a *P* value should definitely be avoided (264, 265). Such data-driven approaches are highly flawed, leading to optimistic or spurious predictor–outcome associations contributing to model overfitting and thus optimistic model performance.

Categorical predictors may also be manipulated before the data are analyzed. In particular, categories may be grouped to eliminate sparse categories; for instance, rare histologic types may be combined into a single "other histology" category. Any revision of categories should be explained (see also item 10b).

Box E. Continuous predictors*.

Many predictors are recorded as continuous measurements, but are converted into categorical form for analysis by using 1 or more cut points (item 10a) (501). Common reasons are to simplify the analysis; to make it easier for clinicians to use the predictors or prediction model, because the predictor–outcome association is often unknown; and to facilitate graphical presentation (e.g., Kaplan–Meier curves). Although categorization of the estimated probabilities by the prediction models is required for decision making, it is important to recognise that categorization of continuous predictors that go into the model is unnecessary for statistical analysis. The perceived advantages of a simpler analysis come at a high cost, as explained below.

Categorization

Categorization allows researchers to avoid strong assumptions about the relationship between the predictor and outcome. However, this comes at the expense of throwing away information. The information loss is obviously greatest when the predictor is dichotomized (2 categories).

It is well known that the results (e.g., the model's predictive performance) can vary if different predictor cut points are used for splitting. If, however, the cut point is chosen on the basis of multiple analyses of the data, in particular taking the cut point that produced the smallest *P* value, then the *P* value for that predictor will be much too small and the performance of the prediction model will be overoptimistic (264).

Even with a prespecified cut point, dichotomization is statistically inefficient and is strongly discouraged (265, 502–505). Furthermore, if cut points are needed as an aid in classifying people into distinct risk groups, this should be done on the basis of the model's predicted probabilities or risks (30, 265).

Categorizing a continuous variable into 3 or more groups reduces the loss of information but is rarely done in clinical studies. Even so, cut points result in a model with step functions, which is inadequate to describe a smooth relationship (266).

Keeping Variables Continuous

A linear functional relationship is the most popular approach for keeping the continuous nature of a predictor. Often, that is an acceptable assumption, but it may be incorrect, leading to a misspecified model in which a relevant predictor may not be included or in which the assumed predictor–outcome relationship differs substantially from the unknown "true" relationship. A check for linearity can be done by investigating possible improvement of fit by allowing some form of nonlinearity. For a long time, quadratic or cubic polynomials were used to model nonlinear relationships, but the more general family of fractional polynomial (FP) functions provide a rich class of simple functions which often provide an improved fit (506). Determination of FP specification and model selection can be done simultaneously with a simple and understandable presentation of results (266, 297).

Spline functions, in particular restricted cubic splines, are another approach to investigate the functional relationship of continuous predictors (112). Restricted cubic splines are recommended over standard cubic spline functions, which are often poorly behaved in the tails of the predictor distribution, by restricting the tails to be linear (112, 507). They are extremely flexible, but no procedure for simultaneously selecting predictors and functional forms has yet found wide acceptance. Furthermore, even for a univariable spline model, reporting is usually restricted to the plot of the predictor–outcome relationship because presentation of the regression coefficients is often too complicated.

* The text of this box is substantially the same as Box 4 in reference 108.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Authors should clearly report how each predictor was handled in all the analyses. Specifically, the rationale (theoretical or clinical) for categorizing any continuous predictors should be reported, including specifying the cut points and how they were chosen. For any predictor that has been kept as continuous, authors should clarify whether they have been retained on the original scale or transformed (for example, log transformation). They should report whether each predictor was modeled as linear or nonlinear, with specification of the method if modeled as nonlinear (for example, by using fractional polynomials or restricted cubic splines). If a predictor was treated as linear, it is preferable to report whether the assumption of a linear relationship with the outcome was checked.

Extreme values may also be shifted to less extreme values to prevent undue leverage effects (2, 266). Authors should report whether they modified or omitted implausible observations, if done, such as omitting extreme outliers.

Although information about the way predictors were handled in the analysis is naturally part of Methods, it can be helpful also to show this information and the definitions of categories (item 7a) within the Results tables (items 13b and 13c).

Reviews of published modeling studies have consistently shown that categorization of continuous predictors is very common, with many dichotomizing all predictors (34, 41, 43, 45, 53, 54, 62, 63, 267, 268). A review of 11 studies for aneurysmal subarachnoid hemorrhage found that age was dichotomized for all models (81). A review of prediction models in cancer found that 12 of 45 (30%) did not provide the explicit coding of all the predictors in the final model (55). Other reviews have shown that how continuous predictors were handled in the analyses was often unclear (54, 64).

Item 10b. Specify type of model, all model-building procedures (including any predictor selection), and methods for internal validation. [D]

All the statistical methods used in the development of a prediction model should be reported. The general principle is that enough detail should be given such that a knowledgeable reader with access to the original data could verify the reported results (www.icmje.org). Moreover, the reader should be able to understand the reasons for the approaches taken.

Many possible analysis strategies can be followed when developing a prediction model. Choices are made at each step of the analysis (2, 112, 266, 269). Some decisions on modeling strategy need to be informed by the data, as well as by the medical context. For example, one may wish to develop a model with only a few major predictors to increase clinical applicability (items 3a, 19b, and 20), at the sacrifice of predictive performance.

A major problem in much prediction model research is that many different analyses may have been performed, but only the best prediction model is reported (that is, the one with best discrimination) (1). Such data-driven model selection can lead to selecting an overfitted model with optimistic model performance. This overfitting would become apparent if the model is evaluated in new data from the same underlying population (270). Hence, it is essential that authors provide a comprehensive view of the range of analyses that have been performed. If necessary, full specification of the statistical analyses can be given in supplementary material, including providing the computer code used to perform the analyses (item 21). Ideally, this code is accompanied with the individual participant data, permitting full reproducibility, although this may not be feasible unless open access to data is agreed on (271).

In the following sections, we consider specific aspects of model development analyses under several headings. Not all aspects will be relevant for some studies. More extensive discussions of statistical analysis methods for both binary and time-to-event outcomes can be found elsewhere (2, 12, 112, 266, 272–277).

1. Type of Model

Examples

> We used the Cox proportional hazards model in the derivation dataset to estimate the coefficients associated with each potential risk factor [predictor] for the first ever recorded diagnosis of cardiovascular disease for men and women separately (278). [Prognosis]

> All clinical and laboratory predictors were included in a multivariable logistic regression model (outcome: bacterial pneumonia) (279). [Diagnosis]

Explanation

Various types of model are used in medical prediction research (112). Most models are derived using multivariable regression. The logistic regression model is most often applied for binary endpoints, such as presence versus absence of disease in diagnostic models or short-term prognostic events (for example, 30-day mortality). The semi-parametric Cox proportional hazards regression model is most often applied for time-to-event outcomes in the case of longer-term prognostic outcomes (for example, 10-year cardiovascular disease risk), although fully parametric models can also be used for time-to-event data (280, 281).

Authors should clearly identify the regression model being used. If a logistic regression model is being used in place of a time-to-event approach for predicting longer-term prognostic outcomes, then a clear rationale should be reported. Developing (and validating) models predicting long-term outcomes using logistic regression inherently requires all participants to have been followed up for the entire time period.

Many variants of regression models are available for binary, multinomial, ordered, continuous, and other outcomes (2). Other types of prediction model include regression trees and machine learning techniques, such as neural networks and support vector machines (275). If such an alternative approach is used, it is recommended to provide a motivation for this choice.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

2. Predictor Selection Before Modeling

Examples

We chose risk factors based on prior meta-analyses and review; their ease of use in primary care settings; and whether a given risk factor was deemed modifiable or reversible by changing habits (i.e., smoking) or through therapeutic intervention; however, we were limited to factors that already had been used in the two baseline cohorts that constituted EPISEM (282). [Prognosis]

Candidate variables included all demographic, disease-related factors and patterns of care from each data source that have been shown to be a risk factor for mortality following an intensive care episode previously. Variables were initially selected following a review of the literature and consensus opinion by an expert group comprising an intensivist, general physician, intensive care trained nurse, epidemiologists, and a statistician. The identified set was reviewed and endorsed by 5 intensivists and a biostatistician who are familiar with the ANZICS APD (283). [Prognosis]

We selected 12 predictor variables for inclusion in our prediction rule from the larger set according to clinical relevance and the results of baseline descriptive statistics in our cohort of emergency department patients with symptomatic atrial fibrillation. Specifically, we reviewed the baseline characteristics of the patients who did and did not experience a 30-day adverse event and selected the 12 predictors for inclusion in the model from these 50 candidate predictors according to apparent differences in predictor representation between the 2 groups, clinical relevance, and sensibility. . . . [T]o limit colinearity and ensure a parsimonious model, Spearman's correlations were calculated between the clinically sensible associations within our 12 predictor variables. Specifically, Spearman's correlations were calculated between the following clinically sensible associations: (1) history of hypertension status and β-blocker and diuretic use, and (2) history of heart failure and β-blocker home use, diuretic home use, peripheral edema on physical examination, and dyspnea in the emergency department (284). [Prognosis]

Explanation

Often, more predictors are available than the investigator wishes to include in the final prediction model. Some form of predictor selection is therefore required, and a variety of approaches are available, each with strengths and weaknesses (**Figure 2**).

An obvious way to reduce a large set of potential predictors is to judge which ones to exclude a priori (item 7a). Here, external evidence may be sought, for example by critical consideration of relevant literature, ideally in the form of a formal systematic review. The knowledge of medical experts is also important to help reduce the number of candidate predictors.

Other possible considerations for predictor exclusion before the actual statistical modeling are that the predictor measurement was unreliable (58), or that relatively high financial costs or burden are associated with such measurement. In the latter case, a series of increasingly complex models can sometimes be developed with and without such predictors (262). Also, closely related predictors can sometimes be combined (for example, by using formal statistical clustering or principal component techniques) in a summary score, such as presence of atherosclerotic symptoms (285), or the association between predictors can be estimated (for example, by using correlation coefficients) in order to preselect 1 of the 2 predictors in the presence of collinearity.

3. Predictor Selection During Modeling

Example

We used multivariable logistic regression with backward stepwise selection with a *P* value greater than 0.05 for removal of variables, but we forced variables [predictors] that we considered to have great clinical relevance back into the model. We assessed additional risk factors [predictors] from clinical guidelines for possible additional effects (286). [Diagnosis]

Explanation

Even when some predictor preselection has been done as just described, there may still be more predictors remaining than one wishes to include in the prediction model (**Figure 2**). Further selection can be based on the predictive importance of each predictor, or simply fitting a model with retaining all remaining predictors (287).

One approach to predictor selection is to fit a model by choosing predictors on the basis of the strength of their unadjusted (univariable) association with the outcome that is to be predicted, or to preselect predictors before the multivariable modeling. The reasoning is that predictors with limited predictive value, based on nonsignificant univariable predictor–outcome association, can be dropped. Although quite common, that strategy is not recommended as a basis for selecting predictors, because important predictors may be rejected owing to nuances in the data set or confounding by other predictors (2, 112, 235). Thus a nonsignificant (unadjusted) statistical association with the outcome does not necessarily imply that a predictor is unimportant. However, if done, univariable predictor–outcome analyses should be reported, including the selection criteria (for example, significance level), and sample size (including the number of events) for each of the univariable analyses, because it is a form of predictor selection (items 13b and 14b and **Figure 2**).

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

A common procedure is to apply an automated variable selection method in the multivariable modeling. Several variants are available in most current software, including forward selection, backward elimination, and their combination. Backward elimination starts with a full model comprising all potential predictors; variables are sequentially removed from the model until a prespecified stopping rule (such as a *P* value or the Akaike information criterion [AIC]) is satisfied. Forward selection starts with an empty model, and predictors are sequentially added until a prespecified stopping rule is satisfied.

Backward elimination is generally preferred if automated predictor selection procedures are used because all correlations between predictors are considered in the modeling procedure (288). Use of automated predictor selection strategies during the multivariable modeling may yield overfitted and optimistic models, particularly when sample size is small (2, 23–25, 32, 112, 289, 290). The extent of overfitting due to the use of predictor selection strategies may be estimated, however, and accounted for in so-called internal validation procedures (**Box C** and **Figure 1**).

A critical issue in these automated predictor selection procedures is the criterion for predictors to be selected for inclusion in the model (2). Often, the predictor's significance level (α) is set to 0.05, as is common for hypothesis testing. However, simulation studies indicate that a higher value should be considered, particularly in small data sets (25). In such cases, use of the AIC for selection is an attractive option; it accounts for model fit while penalizing for the number of parameters being estimated and corresponds to using $\alpha = 0.157$ (2, 112, 291, 292).

Systematic reviews of multivariable prediction models have found that the strategy to build the prediction model was often unclear (34, 43, 54, 81, 182). For example, the approach in selecting predictors in the final models was unclear in 36% of 11 models for aneurysmal subarachnoid hemorrhage (81).

4. Interaction Terms

Example

> Clinically meaningful interactions were included in the model. Their significance was tested as a group to avoid inflating type I error. All interaction terms were removed as a group, and the model was refit if results were nonsignificant. Specifically, interactions between home use of β-blockers and diuretics and between edema on physical examination and a history of heart failure were tested (284). [Prognosis]

Explanation

Most prediction models include predictors as main effects, which assumes that effects of all predictors are additive. Note that additivity here is assumed on the scale of the modeling: on the log odds scale for logistic regression and on the log hazard scale for a Cox regression model. This additivity implies multiplicative effects on the original odds and hazard scales, respectively (273). The additivity assumption means that the predictive effect of each predictor is the same, regardless of the values of the other predictors. This assumption can formally be tested by assessing statistical interaction between the predictors (112). Few reported prediction models contain interactions, and it seems that few researchers examine them. This approach is generally reasonable because interaction terms rarely add to the predictive ability of the model.

If many interactions are examined and only the strongest included in the prediction model, this would contribute to model overfitting, leading to overly optimistic performance estimates (2). Authors should restrict examination of predictor interactions to a small number with prior rationale, rather than simply testing all possible interactions, particularly when the sample size is small. An alternative to interaction testing is to develop different models for different subgroups: for example, for men and women, or adults and children (278). Because of the drastic reduction in sample size and corresponding danger of model overfitting, this approach is rare and should only be considered when the sample size is large.

Survival models often also assume that predictor effects are constant over time (that is, that hazards are proportional). This is similar to assuming that there are no interactions of effects by time. Some consider testing of the proportional hazards assumption good statistical practice, whereas others warn of the risks for overfitting and optimism if models are adjusted based on statistically significant nonproportional effects, in a similar manner as described above with predictor selection strategies (2, 112).

Authors should report procedures for testing interactions and proportionality of hazards in survival models, if conducted.

5. Internal Validation

Example

> We assessed internal validity with a bootstrapping procedure for a realistic estimate of the performance of both prediction models in similar future patients. We repeated the entire modeling process including variable selection . . . in 200 samples drawn with replacement from the original sample. We determined the performances of the selected prediction model and the simple rule that were developed from each bootstrap sample in the original sample. Performance measures included the average area under the ROC curve, sensitivity and specificity for both outcome measures, and computed tomography reduction at 100% sensitivity for neurosurgical interventions within each bootstrap sample (286). [Diagnosis]

Explanation

The predictive performance of a model on the same data that was used to generate the results is referred to as the *apparent performance* of the model

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box F. Internal validation.

When developing a prediction model, several factors may lead to models yielding optimistic apparent performance. These factors include the inclusion of a large number of candidate predictors relative to the number of outcome events (small effective sample size), the use of predictor selection strategies (certainly in conjunction with small effective sample size), and categorization of continuous predictors (2, 12, 23–25, 32, 112, 290). It is therefore important that a more honest estimate of the model's performance from the development data set is obtained. This can be done using so-called "internal validation," preferably using resampling techniques, such as bootstrapping, or cross-validation methods.

Apparent Performance

The apparent performance of a prediction model refers to the performance estimated directly from the data set that was also used to develop the prediction model. The prediction model is tuned to the development data set, leading to optimistic (biased but stable) estimates of performance for small data sets; however, this optimism diminishes when the sample size become large (32).

Split-Sample Validation ("Data Splitting")

In the classical split-sample internal validation approach, the available development data set is divided into 2 data sets; one to develop the model and the other to validate the model (see **Figure 1** and **Box C**). Typically, the 2 data sets are created by randomly splitting the original data (e.g., 50:50 or 70:30). Despite this approach being ubiquitous in prediction model studies, it has several weaknesses: It is inefficient (because it does not use all available data for model development); the 2 data sets will be closely similar, because they vary only by chance (such that the model validation will probably show similar performance as in the development set); and different splits lead to different results notably in relatively small data sets (23, 25, 32, 295, 508). Furthermore, it is unclear how much data should be used to develop the model and how much should be set aside to evaluate the model (see item 8). Large sample sizes are required to make this approach reasonable, at which point the apparent performance will provide a reasonable estimate of model performance (2, 32). A better alternative, if the sample size is sufficiently large, is to split by time (temporal validation) or location (geographic validation) (19, 20, 26).

Cross-validation

Cross-validation is an extension of the split-sample technique to reduce the bias and variability of the performance estimates (32). For example, 10-fold cross-validation involves randomly splitting the data into 10 equally sized groups. The model is developed in 9 of the 10 groups, and its performance evaluated in the remaining group; this entire process is then repeated 10 times so that each of the 10 groups is used to test the model. The performance of the model is then taken as the average over the 10 iterations.

Bootstrap Validation

The bootstrap validation approach not only uses all of the data to develop the prediction model but also provides a mechanism to account for model overfitting or uncertainty in the entire model development process, thereby quantifying any optimism in the final prediction model. Also, it provides for estimating a so-called shrinkage factor that can be used to adjust the regression coefficients and apparent performance for optimism, such that in subsequent model validation studies and applications, better performance will be obtained. The bootstrap validation approach includes (2, 12):

1. Develop the prediction model using the entire original sample (size *n*) and determine the apparent performance.
2. Generate a bootstrap sample, by sampling *n* individuals with replacement from the original sample.
3. Develop a model using the bootstrap sample (applying all the same modeling and predictor selection methods, as in step 1):
   a. Determine the apparent performance (e.g., c-index) of this model on the bootstrap sample (bootstrap performance).
   b. Determine the performance of the bootstrap model in the original sample (test performance).
4. Calculate the optimism as the difference between the bootstrap performance and the test performance.
5. Repeat steps 2 through 4 at least 100 times.
6. Average the estimates of optimism in step 5, and subtract the value from the apparent performance obtained in step 1 to obtain an optimism-corrected estimate of performance.

There is evidence in high-dimensional settings (e.g., "omics" and genome-wide association studies) that cross-validation or bootstrapping is often inappropriately applied by failing to repeat all modeling steps in each cross-validation or bootstrap sample (299, 509, 510). This may result in an overoptimistic assessment of performance (299, 511). Other biases may also cumulatively contribute to inflated performance (512).

(12, 293, 294). Many prediction models are overfitted and their apparent performance optimistic, typically owing to the use of predictor selection strategies in small data sets (23–25, 32, 290, 295). A better initial assessment of the performance of a prediction model is gained by assessing its performance using resampling techniques, such as cross-validation or bootstrapping, all referred to as *internal validation* (Figure 1 and Box F) (12). We recommend that all model development studies include some form of internal validation, particularly if no additional external validation is performed.

Predictor selection based on predictive strength or *P* values in univariable and multivariable analyses often leads to considerable uncertainty in model structure (292, 296). The advantage of bootstrapping as an internal validation technique (instead of cross-validation) is that the effects of predictor selection strategies on the model building, and thus the extent of model overfitting and optimism, can be quantified by repeating the predictor selection process in each bootstrap sample (292, 296–298). Furthermore, bootstrapping provides an estimate of the so-called adjustment or correction

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

factor, by which the model (that is, regression coefficients) and its performance measures (item 16) can be shrunk and thus adjusted for overfitting (**Box F**). It is extremely important that all aspects of model fitting be incorporated into each random or bootstrap derivation sample, including selection of predictors, deciding on transformations, and tests of interaction with other variables or time. Omitting these steps is common in clinical research but can lead to biased assessments of fit, even in the validation sample (299, 300). Refitting the same predictors in each bootstrap sample (unless the model was built using all predictors, in a so-called full model approach) is not a valid approach. Authors should give details of any internal validation procedures.

Overfitting, optimism, and miscalibration can also be addressed and accounted for by applying shrinkage or penalization procedures (287, 290, 294, 301). The lasso method and variants of it are particularly popular when a model is developed with rare events or from a very large number of predictors and the sample size is small (24, 302, 303). However, its usefulness with a smaller number of predictors is less clear (291). If such a procedure was done, details should be given of the method used (for example, lasso, ridge regression, heuristic shrinkage).

Internal validation was reported for only 5 of 14 model development studies in a review of prediction model studies published in general medical journals (34), with similar findings found in other reviews (43, 53, 55, 64, 66, 71, 75, 76, 88, 93–95, 304, 305).

Item 10c. For validation, describe how the predictions were calculated. [V]

Examples

To evaluate the performance of each prostate cancer risk calculation, we obtained the predicted probability for any prostate cancer and for aggressive prostate cancer for each patient from the PRC [Prostate Cancer Prevention Trial risk calculator] (http://deb.uthscsa.edu/URORiskCalc/Pages/uroriskcalc.jsp) and from the SRC [Sunnybrook nomogram–based prostate cancer risk calculator] (www.prostaterisk.ca) to evaluate each prediction model performance (306). [Diagnosis]

To calculate the HSI [Hepatic Steatosis Index], we used the formula given by Lee et al [ref] to calculate the probability of having hepatic steatosis as follows:

$$\text{HS1} = \frac{e^{0.315 \times \text{BMI} + 2.421 \times \text{ALT-to-AST ratio} + 0.630 \times \text{DM} - 9.960}}{1 + e^{0.315 \times \text{BMI} + 2.421 \times \text{ALT-to-AST ratio} + 0.630 \times \text{DM} - 9.960}}$$

with presence of diabetes mellitus (DM) = 1; and absence of DM = 0. ALT and AST indicate alanine aminotransferase and aspartate aminotransferase, respectively (307). [Diagnosis]

Open source code to calculate the QCancer (Colorectal) scores are available from www.qcancer.org/colorectal/ released under the GNU Lesser General Public Licence, version 3 (308). [Prognosis]

Explanation

The preferred evaluation of the performance of an existing prediction model for a new set of individuals (**Box C** and **Figure 1**) relies on making predictions from the original model (as published), and comparing these predictions with the actual outcomes in the validation data set (that is, calibration and discrimination) (309); item 10d. It is therefore important that authors who evaluate the performance of an existing prediction model clearly state how they obtained the model predictions. This could include using the prediction model in full (all regression coefficients, including the intercept or baseline hazard for a particular time point), providing a link to a Web calculator, or including the computer code to implement the prediction model (item 14).

Some model development studies present multiple models or multiple presentations of the same model (for example, both a full regression model and a simplified score). If relevant, authors should clarify exactly which of these prediction models they evaluated.

Prediction models are often presented graphically as nomograms (item 15b) (310, 311). A nomogram permits rapid estimation for an individual participant without a calculator or computer, but obviously is inefficient for use in a validation study on a large numbers of participants. Authors should clearly explain whether the actual nomogram was used manually to obtain predictions or whether the underlying regression model was used.

Without access to the published prediction model, validation, recalibration, and updating are not possible. For example, the FRAX model for predicting the 10-year risk of osteoporotic or hip fracture (312), currently embedded in numerous clinical guidelines around the world (35, 37, 313), has not been published, making independent evaluation of the model impossible (314–316).

There are some misconceptions about how to validate an existing model. One is that validation involves repeating the whole modeling process on new data, including the predictor selection and estimation of the regression coefficients (and model performance), and then subsequently comparing these findings with those in the original model development study. Another misconception is to refit the final model in the previously developed and published model in the validation data. In both cases, the result would actually be another, new model and not the validation of an existing one (19, 20, 26, 28, 47, 309).

Authors will often validate their newly developed prediction model by using a separate data set (for example, recruited later in time or from another hospital). When performance in both the development and vali-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

dation data sets are deemed similar, it is not uncommon for the 2 data sets to be combined and a new prediction model developed on the combined data set (317). Although this is not bad practice per se, such a study is not a validation study, but rather a validation and model development or redevelopment, in that order. The new prediction model still requires further validation.

Item 10d. Specify all measures used to assess model performance and, if relevant, to compare multiple models. [D;V]

Numerous measures exist to assess and quantify the predictive performance of prediction models (26, 252, 253, 269) (**Boxes G** and **H**). Here, we here advocate the most widely used measures, which we encourage researchers to report. We divide these measures into the more traditional (statistical) measures; more recent measures that, to varying extents, incorporate clinical consequences of the predictions; and measures to explicitly estimate the incremental predictive value of a specific predictor beyond existing or established predictors or comparing different models.

1. Traditional Measures

Examples

> We assessed the predictive performance of the QRISK2- 2011 risk score on the THIN cohort by examining measures of calibration and discrimination. Calibration refers to how closely the predicted 10 year cardiovascular risk agrees with the observed 10 year cardiovascular risk. This was assessed for each 10th of predicted risk, ensuring 10 equally sized groups and each five year age band, by calculating the ratio of predicted to observed cardiovascular risk separately for men and for women. Calibration of the risk score predictions was assessed by plotting observed proportions versus predicted probabilities and by calculating the calibration slope.
>
> Discrimination is the ability of the risk score to differentiate between patients who do and do not experience an event during the study period. This measure is quantified by calculating the area under the receiver operating characteristic curve statistic; a value of 0.5 represents chance and 1 represents perfect discrimination. We also calculated the D statistic and R^2 statistic, which are measures of discrimination and explained variation, respectively, and are tailored towards censored survival data. Higher values for the D statistic indicate greater discrimination, where an increase of 0.1 over other risk scores is a good indicator of improved prognostic separation (117). [Prognosis; Validation]
>
> First, we compared the abilities of the clinical decision rule and the general practitioner judgement in discriminating patients with the disease from patients without the disease, using receiver operating characteristic (ROC) curve analysis. An area under the ROC curve (AUC) of 0.5 indicates no discrimination, whereas an AUC of 1.0 indicates perfect discrimination. Then, we constructed a calibration plot to separately examine the agreement between the predicted probabilities of the decision rule with the observed outcome acute coronary syndrome and we constructed a similar calibration plot for the predicted probabilities of the general practitioner. Perfect predictions should lie on the 45-degree line for agreement with the outcome in the calibration plot (318). [Diagnosis; Development]
>
> The accuracy of [the] internally validated and adjusted model was tested on the data of the validation set. The regression formula from the developed model was applied to all bakery workers of the validation set. The agreement between the predicted probabilities and the observed frequencies for sensitization (calibration) was evaluated graphically by plotting the predicted probabilities (x-axis) by the observed frequencies (y-axis) of the outcome. The association between predicted probabilities and observed frequencies can be described by a line with an intercept and a slope. An intercept of zero and a slope of one indicate perfect calibration. . . . The discrimination was assessed with the ROC area (319). [Diagnosis; Development]

Explanation

Two key aspects characterize the performance of a prediction model: calibration and discrimination. They should be reported in all prediction model papers (**Boxes G** and **H**).

Calibration reflects the agreement between predictions from the model and observed outcomes. Calibration is preferably reported graphically, with observed risks plotted on the *y*-axis against predicted risks on the *x*-axis, but may also be presented in a tabular format.

Discrimination refers to the ability of a prediction model to differentiate between those who do or do not experience the outcome event. The most general and widely reported measure of discrimination, for both logistic and survival models, is the concordance index (*c*-index), which equals the area under the receiver-operating characteristic curve for logistic prediction models. A number of different versions of the *c*-index exist (320); therefore, authors should clearly state which version is being calculated.

In addition to measures of discrimination and calibration, various other measures of *overall performance* can be reported; these include measures of explained variation (R^2) (321–329) and the Brier score (330–332). A large number of different approaches to estimate R^2 have been proposed; authors should clearly reference

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box G. Performance measures.

When we develop a risk prediction model, we should assess its performance. The most important considerations of a model's performance are discrimination and calibration (see item 10d and **Box H**). For model development studies, we are primarily interested in discrimination, because the model will be well calibrated (on average) by definition. In validation studies, assessment of both discrimination and calibration is fundamental (252, 513).

Calibration reflects the agreement between outcome predictions from the model and the observed outcomes. Informally, a model is said to be well calibrated if, for every group of, say, 100 individuals, each with a mean predicted risk of *x*%, close to *x* indeed have (diagnostic model) or develop (prognostic model) the outcome.

Calibration is preferably reported graphically with predicted outcome probabilities (on the *x*-axis) plotted against observed outcome frequencies (on the *y*-axis). This plot is commonly done by tenths of the predicted risk, and is preferably augmented by a smoothed (lowess) line over the entire predicted probability range, which is possible both for prediction models developed by logistic regression (112, 514) and by survival modeling (515) (see item 16). This plot displays the direction and magnitude of model miscalibration across the probability range, which can be combined with estimates of the calibration slope and intercept (515). Smoothed or by subgroups, a well-calibrated model shows predictions lying on or around the 45° line of the calibration plot; perfect calibration shows a slope of 1 and intercept of 0, although some caveats have recently been identified (516).

Calibration plots tend to show good calibration in the data set from which they were developed, and even perfect calibration when the smoothed method is used. They may be accompanied by a test for calibration intercept equals 0 and slope equals 1 (517, 518). Comparing predicted versus observed outcome probabilities may also be shown in tabular form (usually by tenths of predicted risk).

Finally, it is common to apply statistical tests for agreement between predicted and observed probabilities using the Hosmer–Lemeshow test or the counterpart tests for survival models including the Nam–D'Agostino test (519) or Grønnesby–Borgan test (520). Such tests have limited statistical power to evaluate poor calibration and are sensitive to the grouping and sample size (521–523): they are often nonsignificant for small *N* and nearly always significant for large *N*. Furthermore, they convey no indication of magnitude or direction of any miscalibration, hence the preference for calibration plots.

In addition, calibration (plots) may also be evaluated in relation to key predictors, such as age or sex subgroups (117, 524). Approaches for assessing calibration of multinomial prediction models have recently been proposed (525).

Discrimination refers to the ability of a prediction model to differentiate between those who do or do not experience the outcome event. A model has perfect discrimination if the predicted risks for all individuals who have (diagnostic) or develop (prognosis) the outcome are higher than those for all individuals who do not experience the outcome. Discrimination is commonly estimated by the so-called concordance index (*c*-index). The *c*-index reflects the probability that for any randomly selected pair of individuals, one with and one without the outcome, the model assigns a higher probability to the individual with the outcome (526). The *c*-index is identical to the area under the receiver-operating characteristic curve for models with binary endpoints, and can be generalized for time-to-event (survival) models accounting for censoring. For survival models, a number of different *c*-indices have been proposed (527); authors should state clearly which measure is used, including an appropriate reference. More recently, extensions to the *c*-index for models with more than 2 outcome categories (528), competing risks (529), and clustering have been proposed (170, 171).

Overall performance measures, such as explained variation (R^2) (321, 324–329) and the Brier score (330, 331), are sometimes reported in addition to the traditional measures of discrimination and calibration, although they are less intuitive. Moreover, a large number of different approaches to estimate R^2 have been proposed; it is therefore important that authors clearly reference the version they are calculating and reporting.

Classification measures, such as predictive values, sensitivity, and specificity, are performance measures after introducing 1 or more probability thresholds. Doing this, one can estimate accuracy or classification measures often reported in single diagnostic test or prognostic factor studies. However, such dichotomization and thus related classification measures lead to loss of information. Moreover, introducing such a threshold implies that it is relevant to clinical practice, which often is not the case.

Decision curve analysis (360, 363–366) offers insight into clinical consequences by determining the relationship between a chosen predicted probability threshold and the relative value of false-positive and false-negative results to obtain a value of net benefit of using the model at that threshold.

Net reclassification improvement (NRI) is commonly used to quantify whether adding a new predictor to an existing model is of benefit, but can also be used for comparing 2 nonnested models (339, 347, 348, 420, 530). The NRI is the net proportion of events reclassified correctly plus the net proportion of nonevents reclassified correctly. An upper bound on the NRI is the continuous NRI (i.e., no categories), which considers any change (increase or decrease) in predicted risk for each individual (347, 530).

Integrated discrimination improvement (IDI) is the difference in predicted probabilities in those who do and do not have (diagnosis) or develop (prognosis) the outcome (339). It estimates the magnitude of the probability improvements or worsening between 2 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.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Box H. Assessing performance of a Cox regression model.

For most types of regression-based prediction models, assessing their performance is straightforward. For logistic regression, for example, a common approach is to plot the observed probability of the outcome event against the predicted probability for multiple groups (often 10) defined by predicted risk (see **Box G** and items 10d and 15a). In such a "calibration plot," the model's discrimination is also indicated by the spread of the predicted probabilities across the risks groups (417); formal measures of discrimination can also be obtained (item 10d).

A comparable approach can be adopted for fully parametric models for time-to-event data, but these models are rarely used. Use of Cox regression predominates for such data, but assessing the calibration of a Cox model is not straightforward because a Cox model is not fully specified. The model allows estimation of relative differences in risk between patients with different characteristics, but because it does not estimate the baseline survival function, it does not estimate absolute risks (event probabilities) (309). The exception to this is when the focus of the Cox-based prediction model is on outcomes at a fixed time point (e.g., risk for cardiovascular death by 10 years). In this instance, only the baseline survival probability at the time point of interest is required, and discrimination and calibration can be assessed using the methods outlined in **Box G**.

Derivation of a Cox Model

A Cox model is specified by a set of predictors with their regression coefficients (log hazard ratios) (411, 531). The prognostic index (PI) is a weighted sum of the variables in the model, where the weights are the regression coefficients (see example in item 15b). The PI for an individual is then the log relative hazard compared with a hypothetical individual whose PI is zero (309).

When a new model is obtained using Cox regression, the PI can be used to examine the predicted survival for several risk groups. For example, patients could be split into 4 equal groups based on their PI values. Discrimination is seen visually from the spread of Kaplan–Meier curves for these risk groups, and numerical performance measures can be obtained (see item 10d). Calibration can be examined by superimposing survival curves derived directly from the Cox model (309, 373).

Validation of a Cox Model

In practice, the baseline survival function for a Cox model is never published. As a result, the external validation of a Cox model by different investigators is hindered because absolute risks cannot be estimated—specifically, calibration may not be assessed easily. Royston and Altman (309) suggested various analysis options in relation to the amount of information available from the derivation study.

Discrimination can be examined provided that the derivation model is specified at least as a set of predictors with their regression coefficients, as long as the precise coding of each predictor is specified. The value of the PI can then be calculated for each participant in the validation data set, and subsequently a regression analysis performed with the PI as a single covariate. With similar case mix, the discrimination in the validation data set is about the same as for the derivation data when the regression coefficient for the PI is approximately 1. If the slope in the validation data is < 1, discrimination is poorer; conversely, if it is > 1, discrimination is better.

If, in addition, Kaplan–Meier curves are shown for several risk groups in the derivation study, then a comparison between corresponding Kaplan–Meier plots for the derivation and validation data sets supports a rough assessment of model calibration. Good calibration may be inferred (by judgement and not a formal comparison) if the 2 sets of survival curves agree well. Such a calibration assessment is not a strict comparison between observed and predicted values, however, because the Cox model is not being used directly to predict survival probabilities. Without the baseline survival function, it is not possible to judge how good the calibration in an independent sample is (309).

the version they used. For models of survival (for example, developed using Cox regression), the D-statistic has recently been suggested as a measure of prognostic separation (333).

When comparing the performance of different models on the same data set, formal statistical tests, such as the DeLong test (334), can be used. However, this test is inappropriate if the compared models are nested (that is, if one model contains all predictors of the other, but has at least 1 additional predictor), and fitted in the same data set, for instance, if a model including clinical predictors plus a new molecular marker is compared with the clinical predictors alone (335).

Finally, it is widely recommended that any newly developed prediction model is compared to existing, published models, ideally in a quantitative manner (47, 48). In the absence of any direct comparison between 2 or more models on the same data set, it is difficult to decide, from all the available prediction models, which is potentially more useful. Numerous systematic reviews showed that very few studies developing or validating prediction models for the same outcome compare performance against other existing models (82).

Calibration is a key characteristic, and its assessment is widely recommended. However, many systematic reviews of multivariable prediction models have found that calibration is rarely reported (34, 41, 43, 55, 62, 63, 66, 73–82, 84, 86, 88, 90–92, 94, 122, 176, 180, 267, 336). For example, calibration was assessed in only 10 of 39 (26%) of studies of type 2 diabetes prediction models (45). Although discrimination is the performance measure most widely measured, this performance measure is also not always reported (74, 78, 81, 88, 122, 336, 337) (for example, in 44% of models for aneurysmal subarachnoid hemorrhage [81]). Very few studies compare model performance against other existing prediction models in the same data set (81, 82, 122).

2. Quantifying the Incremental Value of an Additional Predictor

Examples

We assessed the incremental prognostic value of biomarkers when added to the GRACE score by the likelihood ratio test. We used 3 complementary measures of discrimination improvement to assess the magnitude of the increase in model performance when individual biomarkers were added to GRACE: change in AUC (ΔAUC), integrated discrimination improvement (IDI), and continuous and categorical net reclassification improvement (NRI). To get a sense of clinical usefulness, we calculated the NRI (>0.02), which considers 2% as the minimum threshold for a meaningful change in predicted risk. Moreover, 2 categorical NRIs were applied with prespecified risk thresholds of 6% and 14%, chosen in accord with a previous study, or 5% and 12%, chosen in accord with the observed event rate in the present study. Categorical NRIs define upward and

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

downward reclassification only if predicted risks move from one category to another. Since the number of biomarkers added to GRACE remained small (maximum of 2), the degree of overoptimism was likely to be small. Still, we reran the ΔAUC and IDI analyses using bootstrap internal validation and confirmed our results (338). [Prognosis; Incremental Value]

Explanation

The advantage of multivariable analysis in contrast to single-marker or test research is that it generates direct evidence whether a test or marker has incremental value. However, quantifying the incremental value of adding a certain, often new, predictor to established predictors or even to an existing prediction model, by using the increase or improvement in the general, traditional performance measures (such as calibration, discrimination, or R^2), is difficult to interpret clinically (339, 340). Furthermore, there are concerns that such performance measures as the *c*-index are insensitive for assessing incremental value (341, 342), although its role as a descriptive measure still remains useful (343). Finally, statistical significance tests can be misleading, because statistically significant associations of new but weak predictors are easily found in a large sample.

New measures have therefore been proposed that are based on the concept of reclassification of individuals across predefined risk categories. Such reclassification tables show how individuals are reclassified (from low to high risk and vice versa) by a model with or without a particular predictor (344–346). The use of reclassification tables clearly relies on sensible thresholds to define the risk groups (item 11).

The net reclassification improvement (NRI) is currently a commonly used measure to quantify the extent of reclassification seen in such tables (339, 347, 348). The NRI can be used in model development when adding a certain predictor to established predictors or existing model (that is, the models are nested) and also in model validation when comparing nonnested models, provided that the compared models are sufficiently well calibrated for the data (349). Hence, before using the NRI, model calibration needs to be evaluated first to enable readers to judge the suitability of calculating the NRI.

The NRI has been shown to be highly sensitive to the selection of thresholds defining the risk categories (and is thereby open to manipulation), and there are several other caveats regarding its use, especially in models with suboptimal calibration (350–356). Hence, we recommend that if the NRI is calculated, it should always be accompanied by the underlying classification table stratified for participants with and without the outcome of interest (357); item 16. Concerns have also been raised that continuous NRI, which measures association rather than model improvement, is subject to overinterpretation and is sensitive to model miscalibration (346).

Alternative measures, such as the change in net benefit, the change in relative utility, and the weighted net reclassification improvement, have been suggested as preferable to the NRI. These 3 measures can be mathematically interconverted (349). Identifying suitable measures for quantifying the incremental value of adding a predictor to an existing prediction model remains an active research area, and finding clinically intuitive measures based on the model-based likelihood ratio test remains attractive (343, 358).

Systematic reviews found that studies on reclassification rarely provided a reference for the choice of risk thresholds (105). Furthermore, over one half of the studies failed to report calibration, and few provided information on the proportion of correct and incorrect reclassifications.

3. Utility Measures

Example

We used decision curve analysis (accounting for censored observations) to describe and compare the clinical effects of QRISK2-2011 and the NICE Framingham equation. A model is considered to have clinical value if it has the highest net benefit across the range of thresholds for which an individual would be designated at high risk. Briefly, the net benefit of a model is the difference between the proportion of true positives and the proportion of false positives weighted by the odds of the selected threshold for high risk designation. At any given threshold, the model with the higher net benefit is the preferred model (117). [Prognosis; Validation]

Explanation

Both discrimination and calibration are statistical properties characterizing the performance of a prediction model, but neither captures the clinical consequences of a particular level of discrimination or degree of miscalibration (359, 360). New approaches, such as decision curve analysis (361–363) and relative utility (364–366), offer insight to the clinical consequences or net benefits of using a prediction model at specific thresholds (349). They can also be used to compare the clinical usefulness of different models: for example, a basic and extended model fitted on the same data set, or even 2 different models (developed from 2 different data sets) validated on the same independent data set (367).

Item 10e. Describe any model updating (for example, recalibration) arising from the validation, if done. [V]

Examples

The coefficients of the [original diagnostic] expert model are likely subject to overfitting, as there were 25 diagnostic indicators originally under examination, but only 36 vignettes. To quantify the amount of overfitting, we determine [in our validation dataset] the shrinkage

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

factor by studying the calibration slope b when fitting the logistic regression model . . . :

logit (P (Y = 1)) = a + b * logit (p)

where [Y = 1 indicates pneumonia (outcome) presence in our validation set and] p is the vector of predicted probabilities. The slope b of the linear predictor defines the shrinkage factor. Well calibrated models have b ≈ 1. Thus, we recalibrate the coefficients of the genuine expert model by multiplying them with the shrinkage factor (shrinkage after estimation) (368). [Diagnosis; Model Updating; Logistic]

In this study, we adopted the [model updating] approach of "validation by calibration" proposed by Van Houwelingen. For each risk category, a Weibull proportional hazards model was fitted using the overall survival values predicted by the [original] UISS prediction model. These expected curves were plotted against the observed Kaplan-Meier curves, and possible differences were assessed by a "calibration model," which evaluated how much the original prognostic score was valid on the new data by testing 3 different parameters (α, β, and γ). If the joint null hypothesis on $\alpha = 0$, $\beta = -1$, and $\gamma = 1$ was rejected (i.e., if discrepancies were found between observed and expected curves), estimates of the calibration model were used to recalibrate predicted probabilities. Note that recalibration does not affect the model's discrimination accuracy. Specific details of this approach are reported in the articles by Van Houwelingen and Miceli et al (369). [Prognosis; Model Updating; Survival]

Results of the external validation prompted us to update the models. We adjusted the intercept and regression coefficients of the prediction models to the Irish setting. The most important difference with the Dutch setting is the lower Hb cutoff level for donation, which affects the outcome and the breakpoint in the piecewise linear function for the predictors previous Hb level. Two methods were applied for updating: recalibration of the model and model revision. Recalibration included adjustment of the intercept and adjustment of the individual regression coefficients with the same factor, that is, the calibration slope. For the revised models, individual regression coefficients were separately adjusted. This was done by adding the predictors to the recalibrated model in a step forward manner and to test with a likelihood ratio test ($p < 0.05$) if they had added value. If so, the regression coefficient for that predictor was adjusted further (370). [Diagnostic; Model Updating; Logistic]

Explanation

When validating (or applying) an existing prediction model in other individuals, the predictive performance is commonly poorer than the performance estimated in the individuals on whom the model was derived. The difference is likely to be greater if a more stringent form of validation is used (**Box C** and **Figure 1**): Reduced performance is more likely in a different geographic or setting validation, by different investigators, than in a temporal validation by the same investigators (2, 20, 21, 102, 290). When lower predictive accuracy is encountered, investigators may simply reject the existing model and refit the model on their validation set, or even develop a completely new model.

Although tempting, development of a new prediction model for the same outcomes or target populations is an unfortunate habit for various reasons (20, 31, 102, 290). First, developing a different model per time period, hospital, country, or setting makes prediction research localized. Second, health care providers will have great difficulty deciding which model to use in their practice. Third, validation studies often include fewer individuals than the corresponding development study, making the new model more subject to overfitting and perhaps even less generalizable than the original model. Finally, prior knowledge captured in the original (development) studies is not used optimally, which is counterintuitive to the notion that inferences and guidelines to enhance evidence-based medicine should be based on as much data as possible (371).

Before developing a new model from the validation data at hand, one may rather first try adjusting (that is, updating) the original prediction model to determine to what extent the loss in predictive accuracy may be overcome (85). An adjusted model combines the information captured in the original model with information from individuals in the validation set, and so is likely have improved transportability to other individuals.

There are several methods for updating prediction models (2, 20, 31, 102, 290, 372, 373). The methods vary in extensiveness, which is reflected by the number of parameters that are reestimated. Commonly, the development and validation data set differ in proportion of outcome events, yielding poor calibration of the original model in the new data. By adjusting the intercept or baseline hazard (if known) of the original model to the validation sample, calibration is often improved, requiring only 1 updated parameter and thus a small validation set (31, 290, 372, 373). More extensive updating methods vary from overall adjustment of all predictor weights by a single recalibration factor, adjustment of a particular predictor weight, or addition of a new predictor to reestimation of all individual regression coefficients. The last method may be indicated when the validation data set is much larger than the development data set.

Table 3 summarizes the different updating methods. Simple updating methods (1 and 2) only improve a model's calibration. To improve discrimination, meth-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

ods 3 to 6 are needed. Updated models, certainly when based on relatively small validation sets, still need to be validated before application in routine practice (20).

Finally, as noted in **Box C**, updating of an existing model in a new data set is not recommended without first quantifying the model's predictive performance in the new data (47). If the model has been updated, authors should state how and give the rationale for doing so.

Risk Groups

Item 11. Provide details on how risk groups were created, if done. [D;V]

Examples

> Once a final model was defined, patients were divided into risk groups in 2 ways: 3 groups according to low, medium, and high risk (placing cut points at the 25th and 75th percentiles of the model's risk score distribution); and 10 groups, using Cox's cut points. The latter minimize the loss of information for a given number of groups. Because the use of 3 risk groups is familiar in the clinical setting, the 3-group paradigm is used hereafter to characterize the model (374). [Prognosis; Development; Validation]

> One of the goals of this model was to develop an easily accessible method for the clinician to stratify risk of patients preparing to undergo head and neck cancer surgery. To this end, we defined 3 categories of transfusion risk: low (≤15%), intermediate (15%-24%) and high (≥25%). (375) [Prognosis; Validation]

> Patients were identified as high risk if their 10 year predicted cardiovascular disease risk was ≥20%, as per the guidelines set out by NICE (117). [Prognosis; Validation]

> Three risk groups were identified on the basis of PI [prognostic index] distribution tertiles. The low-risk subgroup (first tertile, PI ≤8.97) had event-free survival (EFS) rates at 5 and 10 years of 100 and 89% (95% CI, 60–97%), respectively. The intermediate-risk subgroup (second tertile, 8.97 < PI 10.06) had EFS rates at 5 and 10 years of 95% (95% CI, 85–98%) and 83% (95% CI, 64–93%), respectively. The high-risk group (third tertile, PI > 10.06) had EFS rates at 5 and 10 years of 85% (95% CI, 72–92%) and 44% (95% CI, 24–63%), respectively (376). [Prognosis; Development]

> Finally, a diagnostic rule was derived from the shrunken, rounded, multivariable coefficients to estimate the probability of heart failure presence, ranging from 0% to 100%. Score thresholds for ruling in and ruling out heart failure were introduced based on clinically acceptable probabilities of false-positive (20% and 30%) and false-negative (10% and 20%) diagnoses (377). [Diagnosis; Development; Validation]

Explanation

In many prediction model studies, risk groups are created using the probabilities from a multivariable prediction model. Often these are labeled, for example, as low-, intermediate-, and high-risk groups as part of the presentation of results or to aid clinical decision making (items 3a and 20).

There is no clear consensus on how to create risk groups, or indeed how many groups to use (43). If risk groups are constructed, authors should specify the boundaries used (that is, the range of predicted probabilities for each group) used and how they were chosen. If, however, the grouping is intended to aid decision making, authors should explain the rationale for the number of risk groups and choice of risk thresholds.

There are concerns that use of risk groups may not be in the best interest of patients (2, 112). Such groupings, although arbitrary, may become standardized despite lacking any rationale (for example, for the Nottingham Prognostic Index [378]). Also, the simplification of predictions means that the risks (probabilities) are assumed to be the same for all individuals within each category. Therefore, irrespective of the creation of any risk groups, reports should provide sufficient information (intercept and betas from a logistic regression model, nomograms, or Web-based calculators for detailed or more complex calculations) to enable calculation of subject-specific risks rather than only group-based risks (item 15a).

In a few cases, the risk groups may be formed based on external knowledge that suggests a different treatment or management plan based on specific risk thresholds (for example, whether a statin is indicated or not for preventing cardiovascular disease outcomes when the prognostic risk is above or below a certain threshold [117]). In most cases, however, there is no such explicit guidance based on estimated probabilities.

In a review of 47 prediction models in cancer, risk groups were created in 36 studies (76%), but the approach to create the groups was unclear or not reported in 17 studies (47%) (54). Other reviews have had similar findings (43).

Development Versus Validation

Item 12. For validation, identify any differences from the development study in setting, eligibility criteria, outcome, and predictors. [V]

Examples

> . . . the summed GRACE risk score corresponds to an estimated probability of all-cause

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

mortality from hospital discharge to 6 months. . . . [I]ts validity beyond 6 months has not been established. In this study, we examined whether this GRACE risk score calculated at hospital discharge would predict longer term (up to 4 years) mortality in a separate registry cohort . . . (379). [Prognosis; Different outcome]

The Wells rule was based on data obtained from referred patients suspected of having deep vein thrombosis who attended secondary care outpatient clinics. Although it is often argued that secondary care outpatients are similar to primary care patients, differences may exist because of the referral mechanism of primary care physicians. The true diagnostic or discriminative accuracy of the Wells rule has never been formally validated in primary care patients in whom DVT is suspected. A validation study is needed because the performance of any diagnostic or prognostic prediction rule tends to be lower than expected from data in the original study when it is applied to new patients, particularly when these patients are selected from other settings. We sought to quantify the diagnostic performance of the Wells rule in primary care patients and compare it with the results reported in the original studies by Wells and colleagues (188). [Diagnosis; Different setting]

When definitions of variables were not identical across the different studies (for example physical activity), we tried to use the best available variables to achieve reasonable consistency across databases. For example, in NHANES, we classified participants as "physically active" if they answered "more active" to the question, "Compare your activity with others of the same age." Otherwise, we classified participants as "not physically active." In ARIC, physical activity was assessed in a question with a response of "yes" or "no", whereas in CHS, we dichotomized the physical activity question into "no" or "low" versus "moderate" or "high" (380). [Prognosis; Different predictors]

As the NWAHS did not collect data on use of antihypertensive medications, we assumed no participants were taking antihypertensive medications. Similarly, as the BMES did not collect data on a history of high blood glucose level, we assumed that no participants had such a history (381). [Prognostic; Different Predictors]

Explanation

For studies that evaluate the performance of a prediction model on a separate data set, authors should clearly identify any differences, intended or not, that could potentially affect model transportability (26, 28).

Prediction models developed in one setting (such as primary care) or in a particular country are not necessarily equally useful in another setting (such as secondary care) or country (19–21, 26, 28, 33, 183, 382, 383). For example, the case mix (item 5a) tends to be different between primary and secondary care, with comparatively more signs and symptoms (and narrower ranges in predictor values) and more advanced disease status in secondary care (20, 21, 102).

Eligibility criteria may also differ unintentionally (for example, a wider or restricted age range), leading to some difference in case mix (186), or differ intentionally (for example, validating a prediction model in children that was developed in adult patients [191, 384]).

The outcome in a validation study may seem the same as in the development study, but the precise definition or method of measurement may differ. For instance, diabetes could be determined by using fasting glucose levels, oral glucose tolerance test, or self-reported diabetes (380, 385). Even when the outcome definition and measurement are the same, differences may nevertheless arise because conditions differ, for example owing to differences in the expertise of the observers (such as radiologists or pathologists), different laboratory procedures, or different imaging technologies.

As with setting and eligibility criteria, outcome differences may also be intentional. The objective of the study may be to assess whether a model can be used to predict a different outcome (379, 383, 386). Models developed to predict postoperative mortality after cardiac surgery have been evaluated to predict prolonged intensive care unit stay (46). Existing prediction models have also been evaluated for predicting the same outcome though at different time points (387): for example, the GRACE model predicting 6-month mortality in acute coronary syndrome patients (388), evaluated for predicting mortality at 4 years (379).

Finally, the definition and measurement of predictors may differ, again intentionally or not. When the definitions are the same, differences may arise because the conditions under which predictors are measured have changed. For example, a specific variable in blood may originally be measured using a laboratory method on venous blood but is validated by using a bedside, point-of-care assay on capillary blood (136, 389).

Authors of validation studies should also clearly report how the predictors have been coded. This includes providing the units of measurement for all continuous predictors and reporting how any categorical predictors have been coded (for example, for sex, with women coded as 0 and men coded as 1); see items 7a and 15a. Moreover, when using historical data to evaluate the performance of a prediction model, occasionally a predictor may not have been collected as the data were collected for a different purpose. Investigators may then use proxy predictors (46), impute the

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

predictor, or omit the predictor from the model (198). Omitting the predictor (equivalent to imputing a value of zero) should be avoided as the model predictions in the validation data become difficult to interpret (198)–for example, FRAX (312, 314).

It is therefore important that authors of validation studies clearly report whether there were any (intentional or not) modifications in setting, eligibility criteria, predictors, and outcome definition and measurement, or include a statement that the conditions, definitions, and measurements were identical to those of the development study. They should not merely list the eligibility criteria, outcome, and predictors, but clearly highlight any differences and how these were handled.

In a recent systematic review of external validation studies, it was unclear in 6 of 45 studies (13%) whether the definition of the outcome was same as the original outcome definition (122).

Results

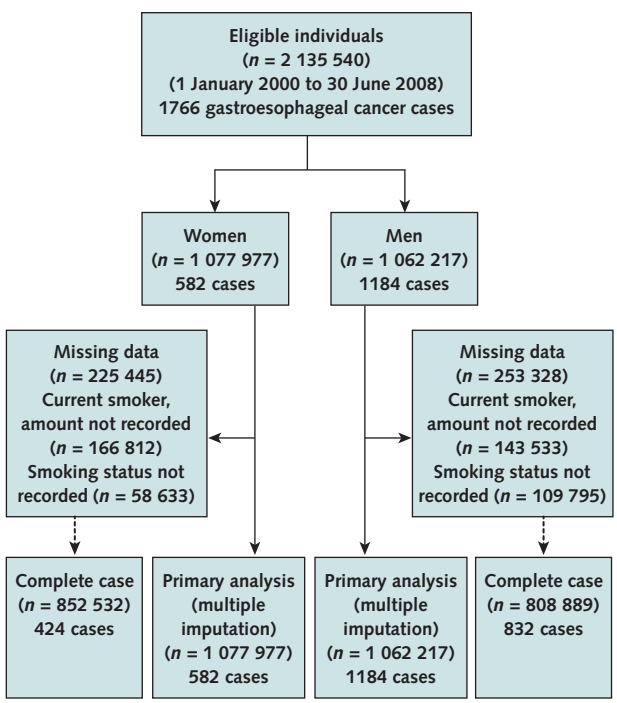
Participants

Item 13a. Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. [D;V]

Examples: Flow of Participants
See Figures 3 and 4.

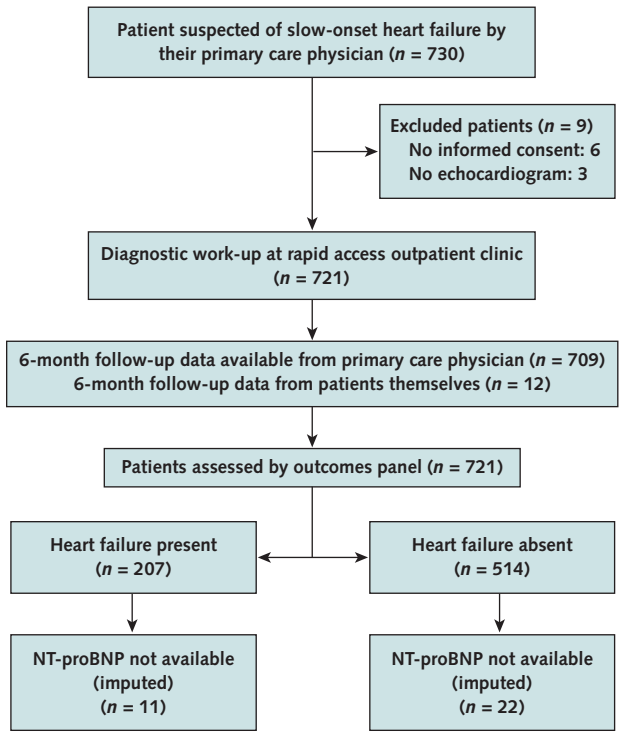
Examples: Follow-up Time

Figure 3. Example figure: participant flow diagram.



Reprinted from reference 390, with permission from Elsevier.

Figure 4. Example figure: participant flow diagram.



Reproduced from reference 377 with permission. NT-proBNP = *N*-terminal pro-brain natriuretic peptide.

> We calculated the 10 year estimated risk of cardiovascular for every patient in the THIN cohort using the QRISK2-2011 risk score . . . and 292 928 patients (14.1%) were followed up for 10 years or more (117). [Prognosis; Validation]
>
> At time of analysis, 204 patients (66%) had died. The median follow-up for the surviving patients was 12 (range 1-84) months (391). [Prognosis; Development]
>
> Median follow-up was computed according to the "reverse Kaplan Meier" method, which calculates potential follow-up in the same way as the Kaplan–Meier estimate of the survival function, but with the meaning of the status indicator reversed. Thus, death censors the true but unknown observation time of an individual, and censoring is an end-point (Schemper & Smith, 1996) (392). [Prognosis; Development]

Explanation

It is important for readers to understand the source of the study participants, including how they were selected from a larger initial group. Such information is vital to judge the context in which the prediction model can be validated or applied. Although the flow of participants in a study can be given in the text or a table, flow diagrams are a valuable way to clarify the derivation of the study sample in whom the model was developed or validated.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

The entrance point to the flow diagram is the source of participants, and successive steps in the diagram can relate eligibility criteria and data availability (108) (item 5b). It can be helpful also to include other information in the diagram, such as numbers of participants with missing observations and the numbers of outcome events.

For studies of prognosis, it is important to summarize the follow-up time of the included participants; this is often given as the median time. The method of calculating the median follow-up should be specified. One approach is the reverse Kaplan–Meier method, which uses data from all patients in the cohort (393). Here, the standard Kaplan–Meier method is used with the event indicator reversed, so that censoring becomes the outcome of interest (108). It may be helpful also to give the median follow-up of those patients who did not have the event (in other words, those with censored survival times). For models predicting the probability of an event by a particular time point, it is useful to report the number of individuals who have been observed until that time point.

For diagnostic studies with delayed disease verification as outcome (items 4a and 6b), reporting the median follow-up is important. If the study data were split into derivation and validation data sets, then it is helpful to provide all of the above information for each sample.

Recent systematic reviews of prediction model studies have observed that many do not report the number of outcome events (34, 45, 54, 85, 394). Other reviews have noted that studies often fail to report a summary of the follow-up (43).

Item 13b. Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. [D;V]

Examples

See Tables 5 and 6.

Explanation

A clear description of the distribution (prevalence, mean or median with standard deviation, or interquartile range) of the relevant characteristics of the study participants is important to judge the context, case mix, and setting of the study. Readers can judge whether the prediction model can potentially be validated in their data or even applied to their patients. It is not sufficient to report only the study inclusion criteria. Information should ideally be provided for all predictors, particularly those included in the final model, and also other important variables (such as demographic, clinical, or setting). Furthermore, the ranges of all continuous predictors, particularly those in the final model, should be clearly reported. In the absence of knowing the predictor ranges, it is unclear to whom the model may be applicable (item 15a).

The above information can most efficiently be shown in a table, which should also incorporate the number (percentage) of missing observations for each variable (see Table 4). If missing observations occur for just a few study variables, the information can be summarized in the text.

It is useful also to incorporate descriptive information about the outcome and, if a univariable analysis is

Table 5. Example Table: Participant Characteristics

Characteristic	Missing Values, *n (%)*	Value
Patients with confirmed PE	0	222 (23.0%)
General characteristics		
Mean age	0	60.6 y (SD, 19.4)
Mean weight	83 (8.6)	72.6 kg (SD, 16.1)
Men	0	403 (41.8%)
Risk factors		
Patients with family history of DVT or PE	6 (0.6)	102 (10.6%)
Patients with personal history of DVT or PE	2 (0.2)	166 (17.2%)
Patients with known congestive heart failure	0	95 (9.8%)
Patients with previous stroke	0	29 (3.0%)
Patients with COPD	0	99 (10.3%)
Patients who had surgery, fracture, or both within 1 mo	0	67 (6.9%)
Patients who were immobile within 1 mo	0	165 (17.1%)
Patients with active malignant condition	3 (0.3)	89 (9.2%)
Patients currently using oral contraceptive	1 (0.1)	69 (7.2%)
Pregnant or postpartum patients	0	10 (1.0%)
Symptoms		
Patients with syncope	2 (0.2)	68 (7.0%)
Patients with recent cough	0	197 (20.4%)
Patients with hemoptysis	0	43 (4.5%)
Patients with dyspnea	0	637 (66.0%)
Patients with chest pain	0	681 (70.6%)
Patients with unilateral lower-limb pain	0	138 (14.3%)
Clinical examination		
General signs		
Mean central temperature	37 (3.8)	36.9 °C (SD, 0.8)
Mean heart rate	4 (0.4)	86.3 beats/min (SD, 19.7)
Mean respiratory rate	59 (6.1)	20.2 cycles/min (SD, 7.0)
Mean systolic blood pressure	6 (0.6)	140 mm Hg (SD, 23)
Mean diastolic blood pressure	7 (0.7)	81 mm Hg (SD, 15)
Signs related to PE		
Patients with chronic venous insufficiency	3 (0.3)	199 (20.6%)
Patients with varicose veins	15 (1.6)	227 (23.5%)
Patients with unilateral edema and pain on deep venous palpation	0	51 (5.3%)
Patients with abnormal chest auscultation	2 (0.2)	158 (16.4%)
Patients with neck vein distention	2 (0.2)	108 (11.2%)

COPD = chronic obstructive pulmonary disease; DVT = deep venous thrombosis; PE = pulmonary embolism.
From reference 395.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 6. Example Table: Participant Characteristics

Characteristic	All Patients (*n* = 202)	TB* (*n* = 72)	No TB (*n* = 130)	*P* Value
Median age (IQR), *y*	32 (28–39)	32 (28–39)	33 (28–40)	0.59
Female sex, %	113 (56)	38 (53)	75 (58)	0.50
Newly diagnosed with HIV, %	53 (26)	14 (19)	39 (30)	0.10
Median CD4 count (IQR), *cells/μL*†	64 (23-191)	60 (70-148)	74 (26-213)	0.17
Taking cotrimoxazole prophylaxis, %‡	117 (58)	48 (67)	69 (53)	0.061
Taking antiretroviral therapy, %§	36 (18)	15 (21)	21 (16)	0.41
Took antibiotics before admission, %	134 (66)	51 (71)	83 (64)	0.31
2-month mortality, %¶	58 (32)	27 (42)	31 (26)	0.028

IQR = interquartile range; TB = tuberculosis.
From reference 396.
* Defined by any positive sputum or bronchoalveolar lavage mycobacterial culture on solid media.
† 4 responses missing.
‡ All but 1 patient had been taking cotrimoxazole for ≥1 month.
§ All patients reported taking antiretroviral therapy for ≥1 month.
¶ 8 patients with TB and 12 patients without TB were lost to follow-up.

done, to show summary statistics of the predictors and other relevant study variables across the different outcome categories (item 14b). Alternatively, one may show the outcome frequencies across predictor categories.

There is no evidence to suggest the reporting of participant characteristics or predictors is particularly poor. However, a few systematic reviews have identified studies that have failed to report such key information (43, 62, 71, 72, 122). In a recent review of 78 external validation studies evaluating the performance of 120 prediction models, the ranges of continuous predictors were reported in only 8% (10 of 120) of original studies developing the prediction model (122).

Item 13c. For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome). [V]

Examples
See Tables 7 and 8.

Explanation
A prediction model validation study is commonly done in participants similar to those in the original model development study (19, 20, 26, 28, 33). However, as discussed in item 12, the validation study population may differ intentionally from the development study. It is important to present demographic characteristics, predictors in the model, and outcome of the (validation) study participants, along with those reported in the original development study. Such information can most efficiently be presented in a table showing the distributions of these variables in the total samples, supplemented by specific participant groups (for example, by sex) if relevant. It is also helpful to report the number of missing observations for each of these variables in both data sets.

One might argue that for very well-known and long-existing models (such as the APACHE risk score or Framingham risk score), such comparison might not be needed. However, not all readers may be as familiar with these models, and comparison between the validation and original development data sets, or perhaps even previous validation studies, is still recommended.

Finally, if unintentional, authors should explain the reasons for any notable differences between the validation samples and the previous study samples (item 12), and later in the article consider the possible implications on the found results, such as the model's predictive performance in the validation set (items 16 and 18).

In a recent systematic review of 78 external validation studies (including development studies with an external validation), only 31 (40%) compared or discussed the characteristics of both the original development and external validation cohorts (122).

Model Development

Item 14a. Specify the number of participants and outcome events in each analysis. [D]

Examples
See Tables 9 and 10.

Explanation
As noted in item 8, the effective sample size in studies of prediction is the number of events and not the number of participants. The number of participants with the event relative to the number of predictors examined plays a central role when assessing the risk for overfitting in a particular study (items 8 and 10b).

In the presence of missing data, the number of participants and events will often vary across analyses, unless participants with any missing data are excluded or their data presented after any imputation (item 9). When deriving a new prediction model, authors frequently conduct analyses to examine the unadjusted association (often referred to as *univariable* or *bivariable association*) between a predictor and the outcome (item 14b). In these instances, if participants have any missing data and are excluded from the analyses (pairwise deletion), the number of participants will vary for the unadjusted association between each predictor and the outcome. Therefore, if univariable associations are reported, the number of participants without missing values for each predictor and the corresponding number of events in those participants should be presented.

Similarly, authors may derive or compare the performance of more than 1 multivariable model on the same data set. For example, one model might be based on routinely available predictors, and a second model might include additional predictors that are not routinely available (such as the result of a blood test). It is important to know the sample size and the number of events used to derive all models.

Readers need a clear understanding of which participants were included in each analysis. In particular, for studies developing a new prediction model, report-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 7. Example Table: Comparison of Participant Characteristics in Development and Validation Data [Development; Validation]

Characteristic	Derivation Cohort (*n* = 8820)	Internal Validation Cohort (*n* = 5882)	External Validation Cohort (*n* = 2938)
Demographic			
Median age (IQR), *y*	66 (56–74)	66 (57–75)	64 (55–72)
Male sex, *n (%)*	5430 (61.6)	3675 (62.5)	1927 (65.5)
Vascular risk factor, *n (%)*			
Hypertension	5601 (63.5)	3683 (62.6)	1987 (67.6)
Diabetes mellitus	1834 (20.8)	1287 (21.9)	720 (24.5)
Dyslipidemia	947 (10.7)	637 (10.8)	386 (13.1)
Atrial fibrillation	643 (7.3)	415 (7.1)	175 (6.0)
Coronary artery disease	1222 (13.9)	811 (13.8)	285 (9.7)
Peripheral artery disease	64 (0.7)	29 (0.5)	26 (0.9)
History of stroke/TIA	2795 (31.7)	1822 (31.0)	809 (27.5)
Smoking	3510 (39.8)	2326 (39.5)	1022 (34.8)
Heavy alcohol consumption	1346 (15.3)	921 (15.7)	372 (12.7)
Other coexistent condition, *n (%)*			
Congestive heart failure	169 (1.9)	121 (2.1)	24 (0.8)
Valvular heart disease	213 (2.4)	139 (2.4)	40 (1.4)
Chronic obstructive pulmonary disease	98 (1.1)	64 (1.1)	12 (0.4)
Hepatic cirrhosis	29 (0.3)	21 (0.4)	7 (0.2)
Peptic ulcer or previous GIB	283 (3.2)	195 (3.3)	76 (2.6)
Renal failure	7 (0.1)	4 (0.1)	3 (0.1)
Arthritis	266 (3.0)	176 (3.0)	45 (1.5)
Dementia	113 (1.3)	82 (1.4)	18 (0.6)
Cancer	150 (1.7)	109 (1.9)	54 (1.8)
Prestroke dependence (mRS ≥3), *n (%)*	809 (9.2)	535 (9.1)	0 (0.0)
Preadmission antiplatelet therapy, *n (%)*	1449 (16.4)	932 (15.8)	357 (12.2)
Preadmission anticoagulation therapy, *n (%)*	210 (2.4)	122 (2.1)	26 (0.9)
Median admission NIHSS score (IQR)	5 (2–9)	5 (2–9)	4 (2–8)
Median admission GCS score (IQR)	15 (14–15)	15 (14–15)	15 (15–15)
Median admission SBP (IQR), *mm Hg*	150 (134–163)	150 (135–162)	150 (135–167)
Median admission DBP (IQR), *mm Hg*	89 (80–95)	89 (80–95)	90 (80–98)
OCSP subtype, *n (%)*			
Partial anterior circulation infarction	4834 (54.8)	3327 (56.6)	1829 (62.3)
Total anterior circulation infarction	811 (9.2)	519 (8.8)	176 (6.0)
Lacunar infarction	1667 (18.9)	1074 (18.3)	246 (8.4)
Posterior circulation infarction	1508 (17.1)	962 (18.4)	687 (23.4)
Intravenous tPA within 3 h after onset, *n (%)*	108 (1.2)	73 (1.2)	137 (4.6)
Antithrombotic therapy on admission, *n (%)*	7371 (83.6)	4950 (84.2)	2550 (86.8)
Anticoagulation therapy on admission, *n (%)*	210 (2.4)	122 (2.1)	159 (5.4)
Median length of hospital stay (IQR), *d*	14 (10–20)	14 (10–20)	14 (11–18)
In-hospital GIB, *n (%)*	227 (2.6)	135 (2.3)	44 (1.5)

DBP = diastolic blood pressure; GCS = Glasgow Coma Score; GIB = gastrointestinal bleeding; IQR = interquartile range; mRS = modified Rankin Scale; NIHSS = National Institutes of Health Stroke Score; OCSP = Oxfordshire Community Stroke Project; SBP = systolic blood pressure; TIA = transient ischemic attack; tPA = tissue plasminogen activator.
From reference 397.

ing the number of events used to derive the model permits calculation of indicators of overfitting, such as EPV (items 8 and 10b). For development studies that have split the data into a development and validation data set, reporting the number of participants and outcome events for each data set is important.

Item 14b. If done, report the unadjusted association between each candidate predictor and outcome. [D]

Examples

See Table 11.

Explanation

Univariable analyses might be desirable to allow the reader confirmation of expected predictive relations based on previous studies, and to observe differences in a predictor's predictive accuracy from unadjusted (univariable) to adjusted (multivariable) analysis. This is for the same reasons as advocated for etiologic (causal) and nonrandomized intervention research, where both the so-called crude and adjusted associations are commonly reported (97, 401). These unadjusted results are a baseline against which to compare the adjusted results in the final multivariable prediction model.

For univariable analyses of binary endpoints (for example, 30-day mortality), authors should report risk ratios or odds ratios accompanied by confidence intervals. Similarly, for time-to-event outcomes, authors should report hazard ratios and associated confidence intervals. *P* values may also be presented, although they do not provide additional information beyond confidence intervals. Typically, such analyses are reported in tabular form, preferably in combination with the results (predictor–outcome associations) from the multivariable analysis.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 8. Example Table: Comparison of Participant Characteristics in Development and Validation Data [Validation]

Risk Predictor	QRESEARCH		THIN (External Validation)*		
	Development (n = 2 355 719)	Internal Validation (n = 1 238 971)	Women (n = 1 077 977)	Men (n = 1 062 217)	Overall (n = 2 140 194)
Median age (SD), *y*	50.1 (15.0)	50.1 (15.0)	49 (15.1)	47 (14.2)	48 (14.7)
Smoking status, *n (%)*					
Nonsmoker	1 194 692 (50.7)	624 788 (50.4)	477 785 (44.3)	369 315 (34.8)	847 100 (39.6)
Ex-smoker	427 246 (18.1)	229 516 (18.5)	123 037 (11.4)	155 961 (14.7)	278 998 (13.0)
Current smoker, amount not recorded	71 416 (3.0)	39 231 (3.2)	166 812 (15.5)	143 533 (13.5)	310 345 (14.5)
Light smoker (<10 cigarettes/d)	148 063 (6.3)	79 844 (6.4)	70 298 (6.5)	66 858 (6.3)	137 156 (6.4)
Moderate smoker (10–19 cigarettes/d)	179 931 (7.6)	95 754 (7.7)	106 203 (9.9)	102 868 (9.7)	209 071 (9.8)
Heavy smoker (≥ 20 cigarettes/d)	133 980 (5.7)	73 554 (5.9)	75 209 (7.0)	113 887 (10.7)	189 096 (8.8)
Not recorded	200 391 (8.5)	96 284 (7.8)	58 633 (5.4)	109 795 (10.3)	168 428 (7.9)
Current symptoms and symptoms in the preceding year, *n (%)*					
Current dysphagia	15 021 (0.6)	8165 (0.7)	10 391 (1.0)	8846 (0.8)	19 237 (0.9)
Current hematemesis	12 952 (0.5)	7119 (0.6)	4630 (0.4)	6162 (0.6)	10 792 (0.5)
Current abdominal pain	225 543 (9.6)	126 161 (10.2)	144 266 (13.4)	102 732 (9.7)	246 998 (11.5)
Current appetite loss	9978 (0.4)	6133 (0.5)	3317 (0.3)	2521 (0.2)	5838 (0.3)
Current weight loss	9998 (0.4)	5377 (0.4)	15 465 (1.4)	12 938 (1.2)	28 403 (1.3)
Hemoglobin <11 g/dL recorded in the last year	22 576 (1.0)	12 638 (1.0)	13 792 (1.3)	4563 (0.4)	18 355 (0.9)

THIN = The Health Improvement Network.
From reference 390.
* Compared with the original development cohort, the THIN cohort had more patients reporting abdominal pain and weight loss.

Authors should also provide number of participants included in each of the unadjusted analyses, if missing data are present (item 14a). For dichotomous or categorical predictors, authors should report how many participants experienced the outcome of interest per category.

However, we follow previous authors in advising against choosing predictors for inclusion into the multivariable model development on the basis of the unadjusted associations of each predictor with the outcome (2, 112, 235) (item 10b).

Model Specification

Item 15a. Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). [D]

Table 9. Example Table: Reporting the Sample Size and Number of Events for Multiple Models*

	Model A		Model B	
	Men	Women	Men	Women
Derivation cohort model estimates				
N	13 240	15 311	12 075	13 935
Number of events	466	215	425	189
	Beta	Beta	Beta	Beta
Age (1 y)	0.053	0.080	0.241	0.066
Smoker	0.466	0.776	2.453	0.784
Body mass index	–	–	–	–
Diabetes	–	–	0.528	0.778
SBP (10 mm Hg)	–	–	0.888	0.038
Total cholesterol (10 mg/dL)	–	–	0.061	0.077
HDL cholesterol (10 mg/dL)	–	–	−0.211	−0.272
Hypertension treatment by SBP >120 mm Hg	–	–	0.519	0.133
Age by smoking	–	–	−0.034	–
Age by SBP	–	–	−0.013	–
Cox 10-year event-free survival (%)	96.2	98.7	96.9	99.0
C-statistic	66.3	72.0	72.0	76.7
Model assessment in the validation cohort				
N	7955	9481	7955	9481
Number of events	263	147	263	147
C-statistic	66.0	69.6	71.0	73.8

HDL = high-density lipoprotein; SBP = systolic blood pressure.
From reference 398.
* β-Coefficients for the variables included in the simplified (model A) and complete (model B) models fitted in the derivation cohort for myocardial infarction or angina, and models' performance in the validation cohort by sex.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 10. Example Table: Reporting the Number of Events in Each Unadjusted Analysis

Characteristic	CDI Patients (n = 395), n (%)	Severe Course Due to CDI, n (%)*		Odds Ratio (95% CI)	P Value
		Yes	No		
Demographic					
Age					
<49 y	85 (22)	6 (13)	79 (23)	1 (reference)	0.01
50–84 y	275 (70)	31 (67)	237 (70)	1.72 (0.69–4.28)	
>85 y	35 (9)	9 (20)	23 (7)	5.15 (1.66–16.0)	
Male sex	220 (56)	24 (52)	191 (56)	0.85 (0.46–1.57)	0.59
Academic hospital	266 (67)	23 (50)	239 (71)	0.42 (0.22–0.28)	0.01
Department of diagnosis					
Other departments	293 (74)	35 (76)	251 (74)	1 (reference)	<0.01
Surgery	83 (21)	4 (9)	78 (23)	0.37 (0.13–1.07)	
Intensive care unit	19 (5)	7 (15)	10 (3)	5.02 (1.80–14.0)	
Medication and intervention history†					
Cytostatic agents	64 (16)	7 (15)	55 (16)	0.91 (0.39–2.15)	0.84
Immunosuppressive agents	172 (44)	21 (47)	146 (44)	1.13 (0.60–2.10)	0.71
Proton pump inhibitors	251 (64)	34 (76)	211 (63)	1.82 (0.89–3.71)	0.10
Recent abdominal surgery	110 (28)	4 (9)	105 (31)	0.21 (0.07–0.59)	<0.01
Recent admission	210 (55)	28 (61)	177 (54)	1.37 (0.71–2.49)	0.38
Antibiotic agents	335 (85)	34 (74)	293 (87)	0.44 (0.21–0.90)	0.03
Clinical					
Charlson Index					
0	59 (15)	7 (15)	52 (15)	1 (reference)	0.53
1–2	150 (38)	14 (30)	134 (40)	0.78 (0.30–2.03)	
3–4	120 (31)	15 (33)	101 (30)	1.10 (0.42–2.87)	
>5	64 (16)	10 (22)	50 (15)	1.49 (0.53–4.21)	
Diarrhea as reason for admission	104 (27)	23 (50)	78 (23)	3.31 (1.76–6.22)	<0.01
Healthcare-onset diarrhea	283 (72)	28 (61)	248 (74)	0.55 (0.29–1.04)	0.06
Fever	208 (60)	25 (66)	174 (59)	1.36 (0.67–2.76)	0.40
Hypotension	117 (30)	25 (63)	88 (30)	3.86 (1.94–7.68)	<0.01
Bloody diarrhea (macroscopic)	52 (15)	7 (16)	44 (15)	1.14 (0.48–2.71)	0.77
Laboratory					
Creatinine count before start of diarrhea					
<90	199 (58)	17 (43)	178 (61)	1 (reference)	0.05
>90	109 (32)	16 (40)	89 (30)	1.88 (0.91–3.90)	
Dialysis	33 (10)	7 (18)	25 (9)	2.93 (1.11–7.77)	

CDI = *Clostridium difficile* infection.
From reference 399.
* Outcome is missing for 10 patients (2.5%); therefore the maximum number of patients is 46 with a severe course and 339 without a severe course.
† Medication and intervention history was gathered from the 3 months before the start of diarrhea.

Examples

See Tables 12, 13, and 14.

Explanation

Prediction models should be presented in adequate detail to allow predictions for individuals, either for subsequent validation studies or in clinical practice (item 15b). For a binary outcome, this means presenting the regression coefficient or odds ratio for each predictor in the model and the intercept. It is good general practice also to report confidence intervals for each estimated coefficient (403), although these are not used in a validation study or in clinical practice. The same considerations apply to a parametric survival model for prognostic (long-term) time-to-event outcomes. If shrinkage methods have been used (item 10b), then both the original and shrunken regression coefficients should be reported.

The situation is rather different for the ubiquitous semi-parametric Cox regression model for a time-to-event outcome. Authors should present the regression coefficient or hazard ratio for each predictor in the model, along with its confidence interval. However, the Cox model has no intercept, and individual survival probabilities are estimated relative to an unspecified baseline survival function. Probabilities thus cannot be estimated from the regression coefficients alone.

To allow estimation of outcome probabilities for individuals at a specific time point, authors should report the cumulative baseline hazard (or baseline survival) for 1 or more clinically relevant time points (item 15b). In cardiovascular and oncologic research, 5-year or 10-year survival is often chosen, but other time points may also be relevant. Alternatively, authors developing prediction models using Cox regression should consider estimating and reporting the baseline hazard function using fractional polynomials or restricted cubic splines (297, 309, 373, 404).

It may not be easy to provide full details of a complex model (for example, the ICNARC model [405]). In other cases, models are updated regularly and continually made available on the Web rather than explicitly

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 11. Example Table: Unadjusted Association Between Each Predictor and Outcome*

Characteristic	Patients With an Outcome (n = 399)	Patients Without an Outcome (n = 15 881)	Univariate Odds Ratio (95% CI)	Multivariable Odds Ratio (95% CI)	P Value
Demographic					
Mean age (SD), *y*	81 (8)	75 (8)	1.8 (1.6–1.9)	1.6 (1.4–1.8)	<0.001
Male	41	38	1.2 (1.0–1.4)	1.3 (1.1–1.7)	0.008
Previous health care use					
Previous hospitalization due to pneumonia or influenza	16	1	22.4 (16.3–30.6)	8.1 (5.7–11.5)	<0.001
Mean outpatient visits (SD), *n*	26 (27)	11 (14)	2.4 (2.1–2.7)	1.5 (1.3–1.8)	<0.001
Comorbid condition					
Heart disease	50	24	3.2 (2.6–3.8)	1.2 (1.0–1.5)	0.10
Pulmonary disease	40	14	4.1 (3.3–5.0)	1.8 (1.4–2.3)	<0.001
Dementia or stroke	31	9	4.6 (3.7–5.8)	2.1 (1.6–2.7)	<0.001
Renal disease	13	4	4.0 (2.9–5.4)	1.5 (1.1–2.1)	0.02
Cancer	12	2	6.8 (4.9–9.4)	4.9 (3.4–7.0)	<0.001
Diabetes	19	12	1.8 (1.4–2.3)	–	–
Anemia	24	8	3.7 (2.9–4.7)	–	–
Nutritional deficiency	5	2	3.7 (2.4–5.9)	–	–
Vasculitis or rheumatologic disease	3	2	1.3 (0.7–1.3)	–	–
Immunodeficiency	2	1	2.0 (1.0–4.0)	–	–
Cirrhosis	1	0.3	3.1 (1.1–8.7)	–	–

From reference 400.
* Data are the percentage of patients, unless otherwise noted.

stated in journal articles (for example, QRISK2 [139]). Regardless of model complexity or frequency of model updating, we strongly recommend the specification of the full model in the peer-reviewed article or in a Web appendix. If the details of a model remain unpublished, it can never be validated, and indeed it is highly questionable whether such a model should be considered for clinical use (312, 314, 406).

In addition to reporting the explicit formula of the developed model, it is essential to know how all predictors were coded (see also item 7a). The measurement scale for all continuous predictors should be reported (for example, whether waist circumference is measured in centimeters or inches). If continuous predictors have been categorized (item 10a and Box E), then the cut points of all categories should be reported, including the lower and upper limit of the first and last category respectively, which are frequently not reported. For categorical predictors, authors should clearly indicate how these have been coded—for example, for sex, with women coded as 0 and men as 1.

Furthermore, the ranges of all continuous predictors should be clearly reported. In the absence of knowing the predictor ranges, it is unclear in whom the model may be applicable. For example, application of a prediction model that was developed by using partic-

Table 12. Example Table: Presenting the Full Prognostic (Survival) Model, Including the Baseline Survival, for a Specific Time Point*

	β Coefficient	SE	P Value
Age	0.15052	0.05767	0.009
Age^2	−0.00038	0.00041	0.35
Male sex	1.99406	0.39326	0.0001
Body mass index	0.01930	0.01111	0.08
Systolic blood pressure	0.00615	0.00225	0.006
Treatment for hypertension	0.42410	0.10104	0.0001
PR interval	0.00707	0.00170	0.0001
Significant cardiac murmur	3.79586	1.33532	0.005
Heart failure	9.42833	2.26981	0.0001
Male sex × age^2	−0.00028	0.00008	0.0004
Age × significant murmur	−0.04238	0.01904	0.03
Age × prevalent heart failure	−0.12307	0.03345	0.0002

From reference 402.
* $S_0(10)$ = 0.96337 (10-year baseline survival). β values are expressed per 1-unit increase for continuous variables and for the condition present in dichotomous variables.

Table 13. Example Table: Presenting the Full Diagnostic (Logistic) Model, Including the Intercept*

Intercept and Predictors	β†	Odds Ratio	95% CI
Intercept	−3.66		
Traditional baker	0.67	2.2	1.2–3.9
Nasoconjunctival symptoms in the past 12 mo	0.72	2.3	1.2–4.5
Asthma symptoms in the past 12 mo	0.63	2.0	0.9–4.4
Shortness of breath and wheeze	0.61	2.3	1.3–3.8
Work-related upper respiratory symptoms	0.47	1.7	0.9–3.1
Work-related lower respiratory symptoms	0.61	2.2	1.1–4.4
ROC area (95% CI)	0.75 (0.71–0.81)		

ROC = receiver-operating characteristic.
From reference 319.
* The predicted probability of wheat sensitization can be calculated using the following formula: *P(sensitization)* 1/(1 exp(–(–3.66 + traditional baker × 0.67 + nasoconjunctival symptoms in the past 12 mo × 0.72 + asthma symptoms in the past 12 mo × 0.63 + shortness of breath and wheeze × 0.61 + work-related upper respiratory symptoms × 0.47 + work-related lower respiratory symptoms × 0.61))). Predictor value is one when present and zero when absent.
† Regression coefficient multiplied with a shrinkage factor (obtained from the bootstrapping procedure) of 0.89.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 14. Example Table: Presenting Both the Original and Updated Prediction Model

Predictor	Original Model	Updated Model
Age (years)	−0.022	−0.017
Female sex	0.46	0.36
Current smoking	−0.63	−0.50
History of PONV or motion sickness	0.76	0.60
Lower abdominal or middle-ear surgery	0.61	–
Abdominal or middle-ear surgery*	–	0.48
Isoflurane and/or nitrous oxide anesthesia†	0.72	–
Inhalational anesthesia‡	–	0.35
Outpatient surgery§	–	−1.16
Intercept	0.15	0.12

PONV = postoperative nausea and vomiting.
From reference 187.
* In the updated model . . . this predictor replaced "lower abdominal or middle-ear surgery" from the original model. In the updated model . . . it included lower abdominal, upper abdominal, and laparoscopic surgery in addition to middle-ear surgery.
† As compared with intravenous anesthesia using propofol.
‡ As compared with intravenous anesthesia using propofol. In the updated model . . . this predictor replaced "isoflurane and/or nitrous oxide anesthesia" from the original model.
§ Predictor not included in the original model.

ipants aged 30 to 60 years to an individual aged 65 years is extrapolation (186).

Numerous systematic reviews have found that studies often report insufficient information to allow for validation or application of the model on other individuals (43, 62, 66, 88). For example, only 13 of 54 (24%) studies developing prognostic models for breast cancer (43) and 22 of 41 (54%) of models for predicting mortality in very premature infants (66) reported sufficient information for that purpose. Another review reported that none of the included studies presented the ranges of the continuous predictors in the models (53), and 2 recent systematic reviews found that even age ranges were frequently not reported (73, 74).

Item 15b. Explain how to use the prediction model. [D]

Examples

See Tables 15 to 17 and Figures 5 to 7.

Explanation

To allow individualized predictions, authors should explain how the developed model can be used to obtain predicted outcome probabilities or risks for an individual. Regression models yield a linear predictor, the weighted sum of the values of the predictors in the model (as measurement or codes), where the weights are the regression coefficients (item 15a). In the prognostic context, the linear predictor is often called a *prognostic index*. The regression coefficients from logistic regression are log odds ratios; for Cox models, they are log hazard ratios. Regression models also include an intercept (constant), except for the Cox model for time-to-event data.

The predicted probability of the outcome can be evaluated from any combination of predictor values. For a logistic regression model, the predicted probability of the outcome event is

$$\text{probability} = \exp(\beta_1 X_1 + \beta_x X_2 + \ \ldots \ + \beta_k X_k)/$$
$$(1 + \exp(\beta_1 X_1 + \beta_x X_2 + \ \ldots \ + \beta_k X_k))$$
$$= 1/(1 + \exp(-(\beta_1 X_1 + \beta_x X_2 + \ \ldots \ + \beta_k X_k)))$$

where β_j is the regression coefficient for predictor X_j and β_0 is the model intercept. It can help readers to provide this expression explicitly. Multiplying by 100

Table 15. Example Table: Presenting a Full Model, Including Baseline Survival for a Specific Time Point Combined With a Hypothetical Individual to Illustrate How the Model Yields an Individualized Prediction

[Simplified] Model B, the Reynolds Risk Score

10-year cardiovascular disease risk (%) = $[1 - 0.98634^{(\exp[B-22.325])}] \times 100\%$, where B = 0.0799 × age + 3.137 × natural logarithm (systolic blood pressure) + 0.180 × natural logarithm (high-sensitivity C-reactive protein) + 1.382 × natural logarithm (total cholesterol) − 1.72 × natural logarithm (high-density lipoprotein cholesterol) + 0.134 × hemoglobin A_{1c} (%) (if diabetic) + 0.818 (if current smoker) + 0.438 (if family history of premature myocardial infarction)

Clinical Example: Estimated 10-Year Risk for a 50-Year-Old Smoking Woman Without Diabetes, According to ATP III or to Clinical Simplified Model B (the Reynolds Risk Score)

Clinical Variables						Estimated 10-Year Risk, %	
Blood Pressure, *mm Hg*	Cholesterol, *mg/dL**			hsCRP, *mg/L*	Parental History†	ATP III Model	Simplified Model B
	Total	HDL	Non-HDL				
155/85	240	35	205	0.1	No	11.5	4.9
155/85	240	35	205	0.5	No	11.5	6.5
155/85	240	35	205	1.0	No	11.5	7.4
155/85	240	35	205	3.0	No	11.5	8.9
155/85	240	35	205	5.0	No	11.5	9.7
155/85	240	35	205	8.0	No	11.5	10.5
155/85	240	35	205	10.0	No	11.5	10.9
155/85	240	35	205	20.0	No	11.5	12.3

ATP = Adult Treatment Panel; HDL = high-density lipoprotein; hsCRP = high-sensitivity C-reactive protein.
From reference 208.
To convert cholesterol from mg/dL to mmol/L, multiply by 0.0259.
† Parental myocardial infarction event before age 60 years.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Table 16. Example Table: A Simple Scoring System From Which Individual Outcome Risks (Probabilities) Can Be Obtained*

To facilitate the calculation of an individual worker's risk, we developed a score chart. We multiplied the regression coefficients by 4 and rounded them to the nearest integer to form the scores for each of the predictors. The scores of predictors which are reviewed positively are added to calculate the "total score." This total score corresponds to risk for sick leave during follow-up.

			Total Score	Risk
Sick leave in the preceding 2 months				
None	0	. . .	≤1	10%–20%
0–1 week	2	. . .	2–3	20%–30%
>1 week	3	. . .	4–5	30%–40%
Intensity of shoulder pain (0–10)			6–7	40%–50%
0–3 points	0	. . .	8	50%–60%
4–6 points	2	. . .	9–10	60%–70%
7–10 points	3	. . .	11–12	70%–80%
Perceived cause: strain or overuse during regular activities	3	. . .	13–15	80%–90%
Reported psychological problems (anxiety, distress, depression)	6	. . .		
			+	
Total score		. . .		

From reference 407.
* The predicted probability of sick leave during 6 months was determined by P = 1/[1+ exp – (–1.72 + 0.53 × sick leave 0–1 week + 0.77 × sick leave >1 week + 0.50 × shoulder pain (4–6 points) + 0.65 × shoulder pain (7–10 points) + 0.68 × overuse due to usual activities + 1.38 × concomitant psychological problems)]. **Instruction:** If a predictor is scored positively, the given weight needs to be filled in. Subsequently the scores are added to calculate the 'Total score'. Using the table next to the score chart the risk (%) of sick leave for an individual patient can be determined based on his/her total score.

converts the probability into a percent risk (% risk = 100 × probability).

For prognostic models based on Cox regression, the predicted probability of the outcome occurring by a particular follow-up time *t* requires the prognostic index and also the estimated "baseline survival", $S_0(t)$ (112, 274, 411). The predicted survival probability is then calculated as $S_0(t)^{\exp(\beta_1 X_1+\beta_2 X_2+ \ldots +\beta_k X_k)}$, and the predicted probability of the outcome event as $1-S_0(t)^{\exp(\beta_1 X_1+\beta_2 X_2+ \ldots +\beta_k X_k)}$. These quantities can be multiplied by 100 to convert to percentages.

Studies developing new prediction models frequently aim to produce an easy-to-use simplified model or scoring system (45, 53), often referred to as a *bedside model*. By "simplified model or scoring system," we mean a simplified presentation format of the underlying regression model (rather than a reduced model in terms of fewer predictors, which is covered in item 10b). Many well-known prediction models have been transformed or simplified to ease their use in practice, such as the SCORE model for predicting the 10-year risk of fatal cardiovascular disease (140) or the Wells model for diagnosis of pulmonary embolism (412). Simplification may be done in several ways, for example by converting (such as rounding [413]) the regression coefficients for each predictor in the final model to easy-to-sum integers that are then related to outcome or survival probabilities, as shown in the examples above (414). An extreme form of rounding of regression coefficients is to give each predictor in the final model the same weight, and simply count the number of risk factors present. These easy-to-sum scores and corresponding outcome probabilities can be shown in tables or in graphs, as illustrated above.

Any simplification of a developed prediction model will, owing to the rounding, lead to some loss of predictive accuracy (1, 413). Hence, when authors convert an original model formula to a simplified scoring rule, it is useful to report the predictive accuracy measures (for example, the *c*-index) (items 10d and 16) before and after simplification. The reader may then judge to what extent the use of the simplified model leads a loss in predictive accuracy. If done, the simplified scoring must be based on the original scale of the regression coefficients (that is, log odds or log hazard scale) and not on any transformation of these coefficients, such as odds ratios or hazard ratios (415). In particular, for predictors with an associated odds or hazard ratio of 1 or less (that is, with a null or a protective/negative effect on the outcome), careful thought is required on how scores are assigned. In these instances, assigning a positive score will actually increase the overall score, indicating a higher likelihood of disease occurrence, whereas the associated contribution should be lower.

If a simplified scoring system is developed, authors should clearly detail the steps taken to create the simplified model and provide a clear description of how the score from the simplified model relates to outcome probabilities. The scoring system can clearly be presented in a table or chart, along with possible scores and their associated outcome probabilities. The values from a simplified model may be grouped to create risk or probability groups (item 11). In these instances, participants with values from the model in a particular range are all assigned to the same risk group, and thus all assigned the same risk. However, merely indicating that a participant is (for example) low, intermediate, or high risk, without quantifying the actual predicted risk associated with the each of the groups, is uninformative; score groups should be related to the corresponding (mean or range of) outcome probabilities, which could be the observed or predicted risks, or both.

Table 17. Example Table: Providing Full Detail to Calculate a Predicted Probability in an Individual

The resulting logit model after fitting to the training data can be expressed as

$$\log\left(\frac{P_{success}}{1-P_{success}}\right) = 2.66 + 1.48\text{IncompMisc} - 1.63\text{NilBleeding} - 0.07\text{Age}$$

where $P_{success}$ denotes the probability for a patient to have a successful expectant management. *IncompMisc* has value of 1 if the diagnosis at primary scan is incomplete miscarriage and 0 otherwise. *NilBleeding* is 1 if there is neither vaginal bleeding nor clots and 0 otherwise. Alternatively, the model can be represented in the following form for calculating the predictive probability for a patient to have a successful expectant management:

$$P_{success} = \frac{e^{2.66 + 1.48\text{IncompMisc} - 1.63\text{NilBleeding} - 0.07\text{Age}}}{1 + e^{2.66 + 1.48\text{IncompMisc} - 1.63\text{NilBleeding} - 0.07\text{Age}}}$$

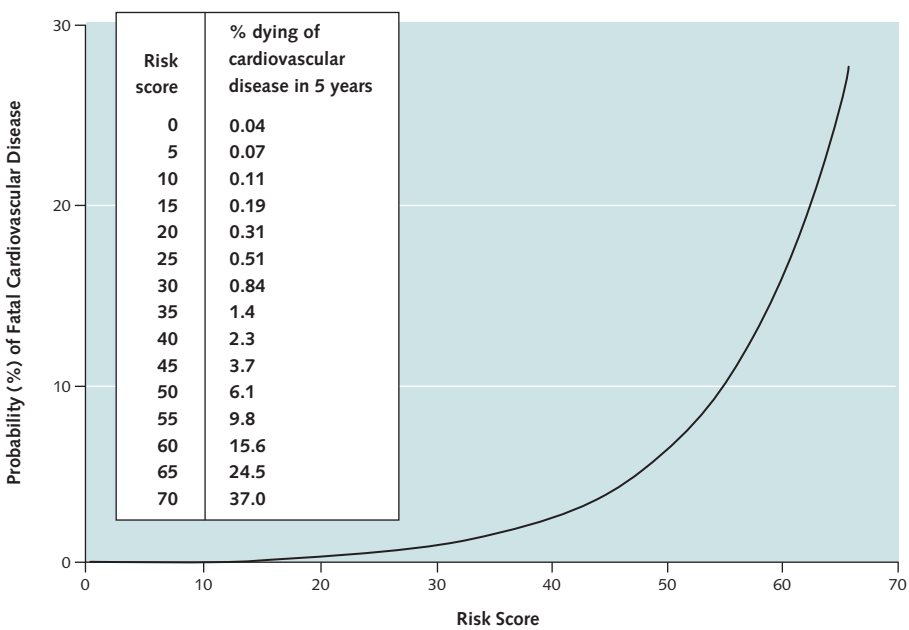
Information from reference 409.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 5. Example figure: a scoring system combined with a figure to obtain predicted probabilities for each score in an individual.

Women

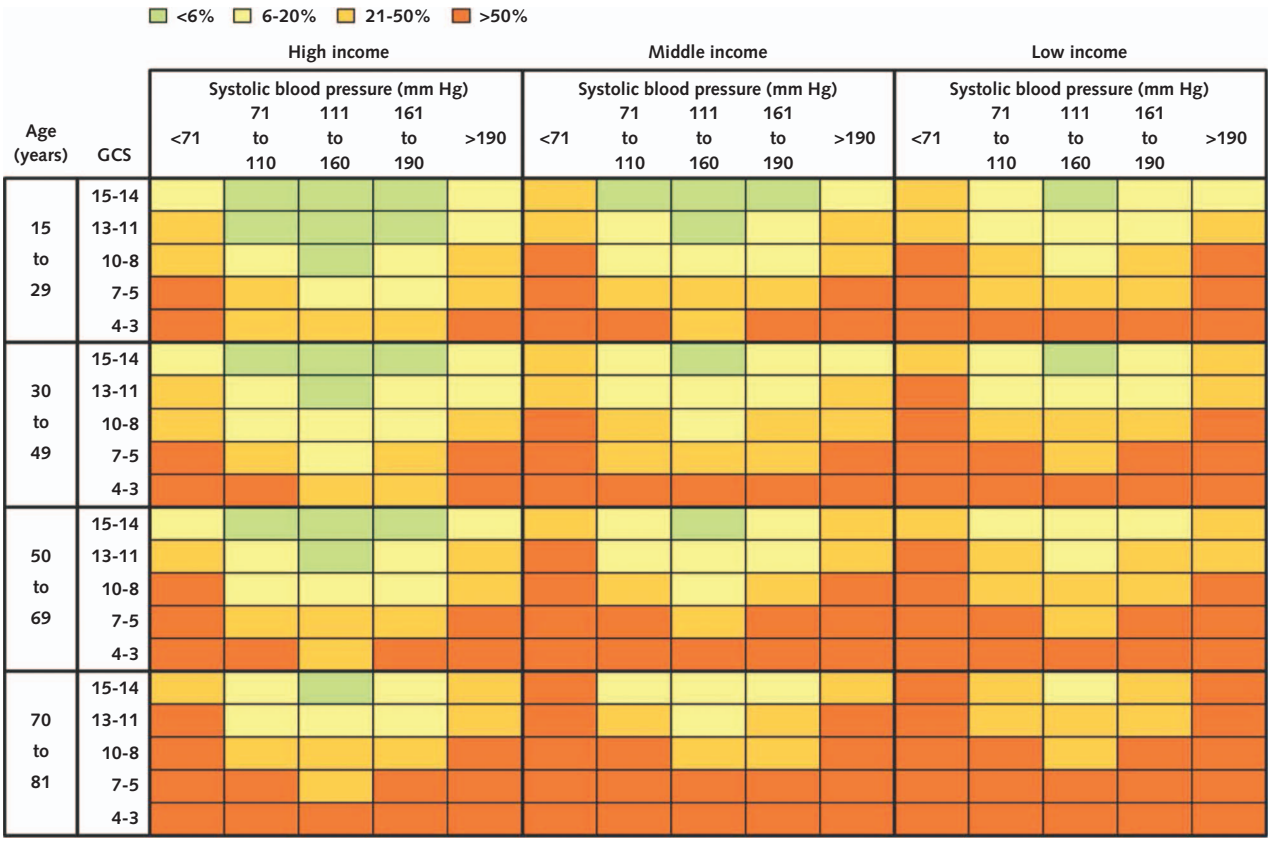
Risk factor	Addition to risk score											Risk score
Age (years)	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74				
Extra for cigarette smoker	0 +13	+5 +12	+9 +11	+14 +10	+18 +10	+23 +9	+27 +9	+32 +8				
Systolic blood pressure (mm Hg)	110–119 0	120–129 +1	130–139 +2	140–149 +3	150–159 +4	160–169 +5	170–179 +6	180–189 +8	190–199 +9	200–209 +10	>210 +11	
Total cholesterol concentration (mmol/L)	<5 0	5.0–5.9 0		6.0–6.9 +1		7.0–7.9 +1		8.0–8.9 +2		>9 +2		
Height (m)	<1.45 +6	1.45–1.54 +4		1.55–1.64 +3		1.65–1.74 +2		>1.75 0				
Creatinine concentration (μmol/L)	<50 0	50–59 +1	60–69 +1	70–79 +2	80–89 +2	90–99 +3	100–109 +3	>110 +4				
History of myocardial infarction	No 0 Yes +8											
History of stroke	No 0 Yes +8											
Left ventricular hypertrophy	No 0 Yes +3											
Diabetes	No 0 Yes +9											
Total risk score* =												



Reproduced from reference 408, with permission from BMJ Publishing Group.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 6. Example figure: a graphical scoring system to obtain a predicted probability in an individual.



For ease of use at the point of care, we developed a simple prognostic model. For this model, we included the strongest predictors with the same quadratic and cubic terms as used in the full model, adjusting for tranexamic acid. We presented the prognostic model as a chart that cross tabulates these predictors with each of them recoded in several categories. We made the categories by considering clinical and statistical criteria. In each cell of the chart, we estimated the risk for a person with values of each predictor at the mid-point of the predictor's range for that cell. We then coloured the cells of the chart in four groups according to ranges of the probability of death: <6%, 6-20%, 21-50%, and >50%. We decided these cut-offs by considering feedback from the potential users of the simple prognostic model and by looking at previous publications. GCS = Glasgow Coma Scale. Reproduced from reference 123, with permission from BMJ Publishing Group.

For survival models, Kaplan–Meier curves should be presented for each risk group, because they provide a natural description of variation in prognosis (for example, discrimination). Kaplan–Meier curves can usefully be supplemented by the total number of patients, the number of patients with the outcome, and a summary of the time to event (with confidence intervals) for each group.

Prediction models are sometimes presented as nomograms (310). This presentation format is not a simplification of a developed model, but rather a graphical presentation of the original mathematical regression formula (112). Such a format may be unfamiliar to many readers and potential users; therefore, it is important to provide a clear description on how to use the nomogram to obtain a prediction for an individual. A nomogram is not a substitute for fully reporting the regression equation (item 15a).

Finally, presenting clinical scenarios and giving a worked example of applying the prediction model to a hypothetical individual with a particular predictor profile may be instructive regardless of how the model is presented.

Model Performance

Item 16. Report performance measures (with confidence intervals) for the prediction model. [D;V]

Examples

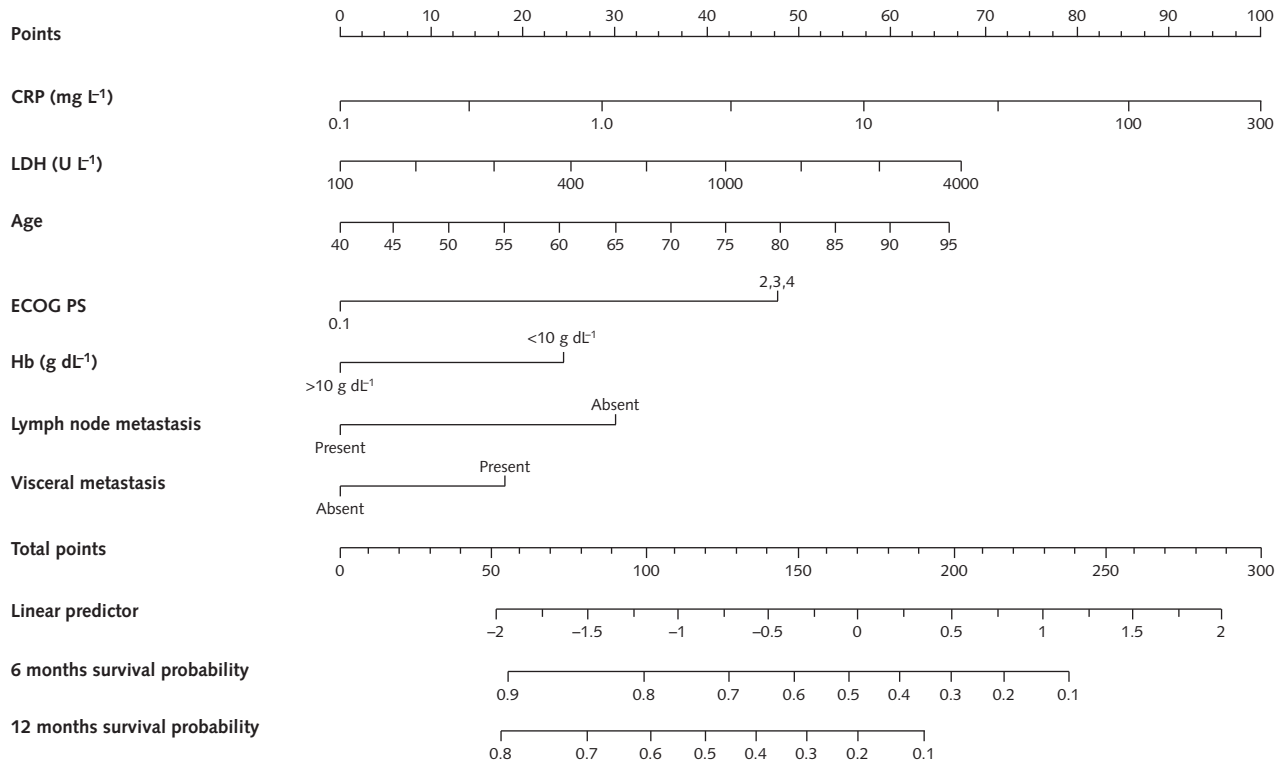
See **Figures 8** to **10** and **Table 18**.

Explanation

All performance measures described in the Methods section (item 10d) should be reported in the Results section, preferably with confidence intervals. If multiple models were developed or evaluated, then the performance measures for each model should be reported. For model development studies, results from internal validation should be reported, including any optimism-corrected performance measures (for example, reporting both the apparent and corrected *c*-index); item 10b and **Box F**. If a prediction model has been simplified (item 15b), the performance (for example, *c*-index) of the original model (for example, the full regression model), as well as any simplified model, should be reported.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 7. Example figure: a nomogram, and how to use it to obtain a predicted probability in an individual.



Nomogram for prediction of positive lymph nodes among patients who underwent a standard pelvic lymph node dissection. Instructions: Locate the patient's pretreatment prostate-specific antigen (PSA) on the initial PSA (IPSA) axis. Draw a line straight upward to the point's axis to determine how many points toward the probability of positive lymph nodes the patient receives for his PSA. Repeat the process for each variable. Sum the points achieved for each of the predictors. Locate the final sum on the Total Points axis. Draw a line straight down to find the patient's probability of having positive lymph nodes. ECOG = Eastern Cooperative Oncology Group; CRP = C-reactive protein; Hb = hemoglobin; LDH = lactate dehydrogenase; PS = performance status. Reprinted from reference 410, with permission from Elsevier.

In addition to reporting a single numerical quantification, graphical approaches to visualize discrimination between persons without and with the outcome are recommended, such as a histogram, density plot, or dot plot for each outcome group (417). For logistic regression models, a receiver-operating characteristic (ROC) curve may also be presented, but unless it depicts relevant predicted risks clearly labeled on the curve, it is largely uninformative and offers nothing over the *c*-statistic.

If more than 1 model was developed (for example, a basic and an extended model [418]) or evaluated on the same data set, one might compare their performance, which necessitates use of a statistical method that accounts for the fact that models were developed or validated on the same data (334, 335, 419).

Where the NRI has been calculated (acknowledging the caveats described in item 10d) to evaluate the incremental value of a specific predictor beyond a combination of existing predictors, authors should report separately both the event and nonevent components of the NRI separately (339), along with a single summary NRI (351, 357, 420, 421).

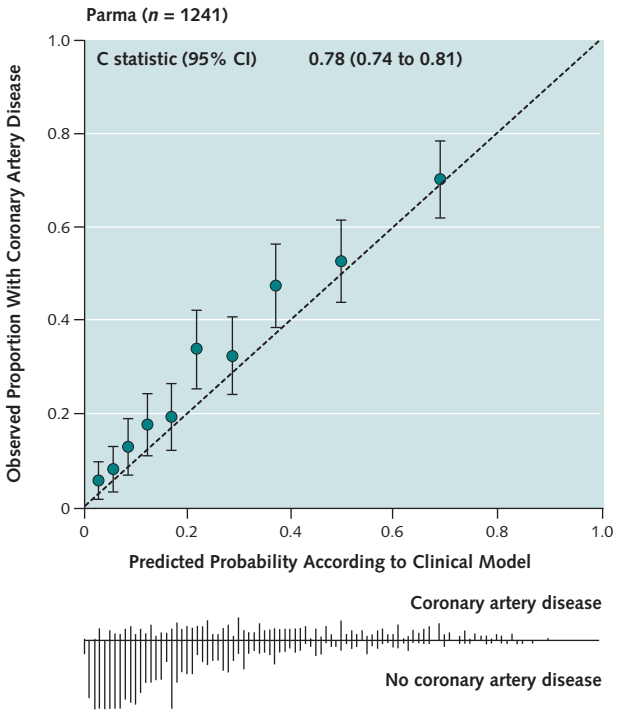
Decision analytic measures, such as net benefit or relative utility, are usually presented graphically rather than as a single numerical estimate (361–363, 422). The *x*-axis for such graphs is a measure of patient or clinician preference, such as the minimum probability of cancer at which a patient would opt for biopsy (138) or the number of patients a doctor would be willing to treat with a drug to prevent 1 cardiovascular event (117). The range of the *x*-axis should generally be chosen to represent reasonable variation in practice. For example, there is little justification for including 80% as a threshold for prostate cancer biopsy, because it assumes that some patients would not consider biopsy if given a 75% probability of cancer.

The *y*-axis, the net benefit, is the difference between the number of true-positive results and the number of false-positive results, weighted by a factor that gives the cost of a false-positive relative to a false-negative result. For example (**Figure 18**), if 2 models being compared at a particular threshold have a difference in net benefit of 0.005 (that is, model A [QRISK2-2011] minus model B [NICE Framingham]), then this is interpreted as the net increase in true-positive findings–that is, by using model A, 5 more true-positive outcomes are identified per 1000 individuals without increasing the number of false-positive findings.

Graphical representations of decision analytic measures should avoid devoting large portions of the graph to negative net benefit. Curves should be smoothed; if sample sizes are moderate, investigators can either apply statistical smoothing, or calculate net

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 8. Example figure: a calibration plot with c-statistic and distribution of the predicted probabilities for individuals with and without the outcome (coronary artery disease).



Reproduced from reference 256, with permission from BMJ Publishing Group.

benefit at more widely spaced intervals (for example, every 5%).

Model Updating

Item 17. If done, report the results from any model updating (i.e., model specification, model performance). [V]

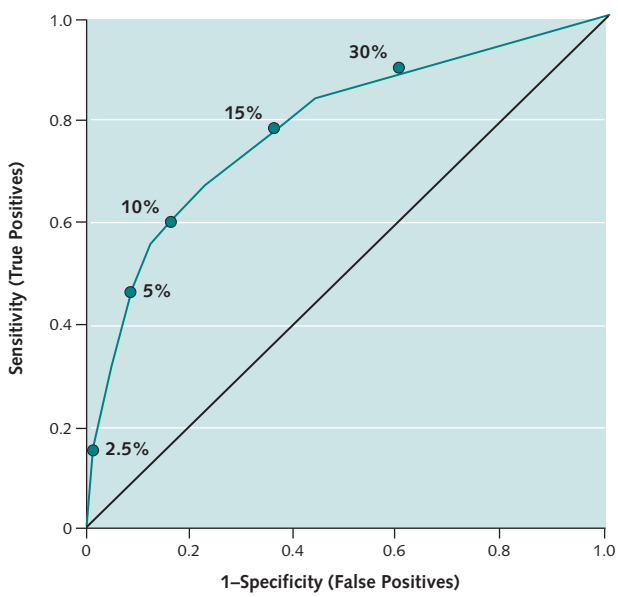
Examples

For the recalibrated models, all regression coefficients were multiplied by the slope of the calibration model (0.65 for men and 0.63 for women). The intercept was adjusted by multiplying the original value by the calibration slope and adding the accompanying intercept of the calibration model (−0.66 for men and −0.36 for women). To derive the revised models, regression coefficients of predictors that had added value in the recalibrated model were further adjusted. For men, regression coefficients were further adjusted for the predictors deferral at the previous visit, time since the previous visit, delta Hb, and seasonality. For women, regression coefficients were further adjusted for deferral at the previous visit and delta Hb . . . available as supporting information in the online version of this paper, for the exact formulas of the recalibrated and revised models to calculate the risk of Hb deferral) (370). [Diagnostic; Model Updating; Logistic]

The mis-calibration of Approach 1 indicated the need for re-calibration and we obtained a uniform shrinkage factor when we fitted logit(P(Y = 1)) = a + b*logit(p) in Approach 2. We obtained the estimates a = −1.20 and b = 0.11, indicating heavy shrinkage (368). [Diagnostic; Model Updating; Logistic]

Results of the performance of the original clinical prediction model compared with that of different models extended with genetic variables selected by the lasso method are presented in Table 3. Likelihood ratio tests were performed to test the goodness of fit between the two models. The AUC curve of the original clinical model was 0.856. Addition of TLR4 SNPs [single-nucleotide polymorphisms] to the clinical model resulted in a slightly decreased AUC. Addition of TLR9-1237 to the clinical model slightly increased the AUC curve to 0.861, though this was not significant (p = 0.570). NOD2 SNPs did not improve the clinical model (423). [Prognostic; Model Updating; Logistic]

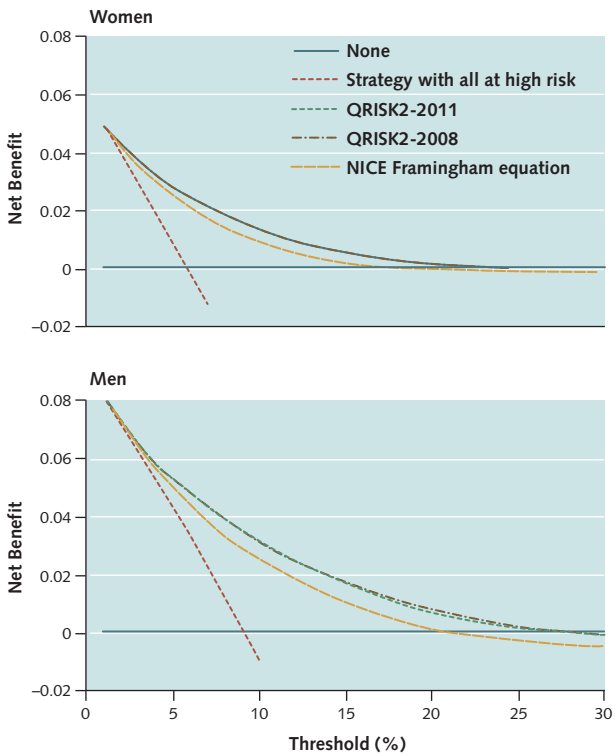
Figure 9. Example figure: a receiver-operating characteristic curve, with predicted risks labelled on the curve.



Receiver operating characteristic curve for risk of pneumonia . . . Sensitivity and specificity of several risk thresholds of the prediction model are plotted. Reproduced from reference 416, with permission from BMJ Publishing Group.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Figure 10. Example figure: a decision curve analysis.



The figure displays the net benefit curves for QRISK2-2011, QRISK2-2008, and the NICE Framingham equation for people aged between 35 and 74 years. At the traditional threshold of 20% used to designate an individual at high risk of developing cardiovascular disease, the net benefit of QRISK2-2011 for men is that the model identified five more cases per 1000 without increasing the number treated unnecessarily when compared with the NICE Framingham equation. For women the net benefit of using QRISK2-2011 at a 20% threshold identified two more cases per 1000 compared with not using any model (or the NICE Framingham equation). There seems to be no net benefit in using the 20% threshold for the NICE Framingham equation for identifying women who are at an increased risk of developing cardiovascular disease over the next 10 years. NICE = National Institute for Health and Care Excellence. Reproduced from reference 117, with permission from BMJ Publishing Group.

Explanation

The performance of an existing model with new data is often poorer than in the original development sample. The investigators may then decide to update or recalibrate the existing model in one of several ways (Table 3 and item 10e). If a prediction model has been updated on the basis of the validation study results, then authors should report all aspects of the model that have been updated. Depending on the method of updating (Table 3), this includes reporting the reestimated intercept, updated regression coefficients (for example, using the slope of the calibration plot of the original model in the validation set), or the estimated regression coefficients of the model, including any new predictors. Model updating is more complex in the context of Cox regression models (309, 373).

The updated model is in essence a new model. Updated models should thus also be reported in sufficient detail to enable readers to make predictions for individual patients (items 15a and 15b) in subsequent validation studies or in practice. The performance of the updated model should also be reported (item 16).

Discussion

Limitations

Item 18. Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data). [D;V]

Examples

The most important limitation of the model for predicting a prolonged ICU stay is its complexity. We believe this complexity reflects the large number of factors that determine a prolonged ICU stay. This complexity essentially mandates the use of automated data collection and calculation. Currently, the infrequent availability of advanced health information technology in most hospitals represents a major barrier to the model's widespread use. As more institutions incorporate electronic medical records into their process flow, models such as the one described here can be of great value.

Our results have several additional limitations. First, the model's usefulness is probably limited to the U.S. because of international differences that impact ICU stay. These differences in ICU stay are also likely to adversely impact the use of ICU day 5 as a threshold for concern about a prolonged stay. Second, while capturing physiologic information on day 1 is too soon to account for the impact of complica-

Table 18. Example of a Reclassification Table (With Net Reclassification Improvement and 95% CI) for a Basic and Extended Diagnostic Model Using a Single Probability Threshold*

	DVT Yes (n = 416)		
	Model 2 With D-Dimer		
Model 1 without D-Dimer	≤ 25	> 25	Total
≤ 25	92	123	215
>25	26	175	201
Total	118	298	416
	DVT No (n = 1670)		
	Model 2 With D-Dimer		
Model 1 without D-Dimer	≤ 25	> 25	Total
≤ 25	1223	116	1339
>25	227	104	331
Total	1450	220	1670

DVT = deep venous thrombosis.
From reference 367.
* The net reclassification improvement for addition of D-Dimer assay to the combination of history and physical examination results with the use of the numbers shown . . . was: (0.30–0.06) – (0.07–0.14) = 0.31 (95% CI, 0.24–0.36).

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

tions and response to therapy, day 5 may still be too early to account for their effects. Previous studies indicate that more than half of the complications of ICU care occur after ICU day 5. Third, despite its complexity, the model fails to account for additional factors known to influence ICU stay. These include nosocomial infection, do not resuscitate orders, ICU physician staffing, ICU acquired paralysis, and ICU sedation practices. Fourth, the model's greatest inaccuracy is the under-prediction of remaining ICU stays of 2 days or less. We speculate that these findings might be explained by discharge delays aimed at avoiding night or weekend transfers or the frequency of complications on ICU days 6 to 8 (424). [Prognosis; Development; Validation]

This paper has several limitations. First, it represents assessments of resident performance at 1 program in a single specialty. In addition, our program only looks at a small range of the entire population of US medical students. The reproducibility of our findings in other settings and programs is unknown. Second, we used subjective, global assessments in conjunction with summative evaluations to assess resident performance. Although our interrater reliability was high, there is no gold standard for clinical assessment, and the best method of assessing clinical performance remains controversial. Lastly, $r^2 = 0.22$ for our regression analysis shows that much of the variance in mean performance ratings is unexplained. This may be due to limited information in residency applications in such critical areas as leadership skills, teamwork, and professionalism (425). [Prognosis; Development]

Explanation

Even the best-conducted studies on prediction models are likely to have many limitations to consider. Yet, many articles published in even the most influential journals do not report limitations (426). Moreover, it is common in studies of molecular classifiers to witness overinterpretation of results without proper consideration of the caveats stemming from the study design and findings (158).

When asked after its publication, many coauthors of a paper feel that the printed discussion does not fully express their views, and they often add that the published discussion is short in listing limitations and caveats (427). Nevertheless, it has been repeatedly argued that the explicit acknowledgment of limitations is one of the key aspects of a scientific work, and a most valuable part of the discussion of a scientific paper (428, 429). Acknowledgment of limitations strengthens, rather than weakens, the research.

Limitations need to be placed into perspective, and an effort should be made to characterize what might be the impact of each separate problem on the results of the study. In some cases, the impact may be too uncertain and practically impossible to fathom, whereas in other cases the direction of the bias introduced can be safely predicted and the impact assessed with substantial certainty.

Limitations may pertain to any aspect of the study design and conduct or analysis of the study. This may include (430), but is not limited to, the types of the study populations, selection of participants (representativeness), selection of predictors, robustness of definitions and procedures used in data collection both for predictors and for outcomes, sample size (especially in comparison with the complexity and number of predictors under study and outcome events), length of follow-up and ascertainment methods, multiplicity of analyses, missing data, overfitting, internal validation processes, and differences between the development and validation cohorts (if applicable). It should be discussed whether limitations affected the model development, model validation, or both, and what the overall impact on the credibility, applicability, and generalizability of the multivariable model would be expected to be.

For example, omission of relevant well-known predictors should be explicitly acknowledged, and these known predictors should be catalogued. It may be useful to clarify, if possible, whether predictors that have been omitted need to be considered in parallel in future studies or practical applications, or the included predictors already capture sufficiently the information from the omitted predictors. In another example, if overfitting is suspected, this should also be acknowledged, and some statement should be made about how serious the problem is expected to be; how much the performance of the model is overestimated; and whether this overestimation should affect the decision to apply the model in practice versus waiting for some further validation, model updating (including recalibration; items 10e and 17), or a continuous improvement strategy (for example, QRISK2 [117, 431–433]) that would alleviate these concerns.

Papers that develop a model in a single population without any validation in a different population should mention the lack of external validation as a major limitation by default, besides any other limitations that may exist.

Interpretation

Item 19a. For validation, discuss the results with reference to performance in the development data, and any other validation data. [V]

Examples

The ABCD2 score was a combined effort by teams led by Johnston and Rothwell, who merged two separate datasets to derive high-risk clinical findings for subsequent stroke. Rothwell's dataset was small, was derived from patients who had been referred by primary care physicians and used predictor variables

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

scored by a neurologist one to three days later. Johnston's dataset was derived from a retrospective study involving patients in California who had a transient ischemic attack.

Subsequent studies evaluating the ABCD2 score have been either retrospective studies or studies using information from databases. Ong and colleagues found a sensitivity of 96.6% for stroke within seven days when using a score of more than two to determine high risk, yet 83.6% of patients were classified as high-risk. Fothergill and coworkers retrospectively analyzed a registry of 284 patients and found that a cutoff score of less than 4 missed 4 out of 36 strokes within 7 days. Asimos and colleagues retrospectively calculated the ABCD2 score from an existing database, but they were unable to calculate the score for 37% of patients, including 154 of the 373 patients who had subsequent strokes within 7 days. Sheehan and colleagues found that the ABCD2 score discriminated well between patients who had a transient ischemic attack or minor stroke versus patients with transient neurologic symptoms resulting from other conditions, but they did not assess the score's predictive accuracy for subsequent stroke. Tsivgoulis and coworkers supported using an ABCD2 score of more than 2 as the cutoff for high risk based on the results of a small prospective study of patients who had a transient ischemic attack and were admitted to hospital. The systematic review by Giles and Rothwell found a pooled AUC of 0.72 (95% CI 0.63–0.82) for all studies meeting their search criteria, and an AUC of 0.69 (95% CI 0.64–0.74) after excluding the original derivation studies. The AUC in our study is at the low end of the confidence band of these results, approaching 0.5 (434). [Prognosis]

Explanation

When the study presents the validation of an existing model, it should be clearly discussed whether the currently validated model is identical to the one previously developed or whether there were any differences, and if so, why (item 12). The performance of the model in the validation study should be discussed and placed in context to the model performance in the original development study and with any other existing validation studies of that model. One should highlight the main results, as well as any biases that may have affected the comparison.

When the validation study shows a different (usually poorer) performance, reasons should be discussed to enhance interpretation. For example, reasons may include differences in case mix, predictor and outcome definition or measurement, or follow-up time (if applicable). When more than 1 model is validated in a single data set—a so-called comparative validation—the main results should be highlighted, including any biases that may have affected the comparison (47, 48).

Item 19b. Give an overall interpretation of the results considering objectives, limitations, results from similar studies, and other relevant evidence. [D;V]

Examples

Our models rely on demographic data and laboratory markers of CKD [chronic kidney disease] severity to predict the risk of future kidney failure. Similar to previous investigators from Kaiser Permanente and the RENAAL study group, we find that a lower estimated GFR [glomerular filtration rate], higher albuminuria, younger age, and male sex predict faster progression to kidney failure. In addition, a lower serum albumin, calcium, and bicarbonate, and a higher serum phosphate also predict a higher risk of kidney failure and add to the predictive ability of estimated GFR and albuminuria. These markers may enable a better estimate of measured GFR or they may reflect disorders of tubular function or underlying processes of inflammation or malnutrition.

Although these laboratory markers have also previously been associated with progression of CKD, our work integrates them all into a single risk equation (risk calculator and **Table 5**, and smartphone app, available at www.qxmd.com/Kidney-Failure-Risk-Equation). In addition, we demonstrate no improvement in model performance with the addition of variables obtained from the history (diabetes and hypertension status) and the physical examination (systolic blood pressure, diastolic blood pressure, and body weight). Although these other variables are clearly important for diagnosis and management of CKD, the lack of improvement in model performance may reflect the high prevalence of these conditions in CKD and imprecision with respect to disease severity after having already accounted for estimated GFR and albuminuria (261). [Prognosis; Development; Validation]

Explanation

Interpretation of the study results places the findings in context of other evidence. That other evidence could include previous similar studies on the same multivariable model, previous studies on different models with the same or similar outcome, and other types of evidence that may be considered relevant. Often, there are many other prediction models that aim to serve the same or similar purposes. For example, in a single year, data were published on 240 assessments of 118 different predictive tools for mortality alone (65).

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

When there are many available prediction models for the same target populations or outcomes, a systematic juxtaposition of the model at hand against previously developed models would be useful in identifying the strengths and weaknesses of the new model. Such a comparison would ideally depend on a systematic review of previous studies, if such a review is available (47, 48, 435). Otherwise, the authors need to consider performing at least an informal review of the previous evidence and discuss the main studies that may be competing against the current study, in terms of informing the evidence base and action plans for further validation studies or adoption into practice. Differences in model building, predictors considered, applicable populations and settings, and performance and strength of the validation process may be particularly useful to comment on. Additional considerations may also have a bearing on the interpretation of the results—these include limitations of the study (as discussed in item 18); whether the initial objectives of the study were met, and if not, why; and aspects of feasibility of using the proposed model in diverse settings and how its introduction may be expected to fit into or alter other medical practices.

In some cases, other relevant evidence may be interesting to consider. For example, there may exist useful data on the biological plausibility of predictors included in the model, or other data that may offer insights about why some predictors are particularly important in their model. An empirical study suggests that authors tend to be very nonsystematic in evoking biological plausibility evidence to support the inclusion of specific predictors in their models (436). Efforts should be made to give balanced views and discuss both supportive and contrarian evidence, whenever such exists.

Implications

Item 20. Discuss the potential clinical use of the model and implications for future research. [D;V]

Examples

> The likelihood of influenza depends on the baseline probability of influenza in the community, the results of the clinical examination, and, optionally, the results of point of care tests for influenza. We determined the probability of influenza during each season based on data from the Centers for Disease Control and Prevention. A recent systematic review found that point of care tests are approximately 72% sensitive and 96% accurate for seasonal influenza. Using these data for seasonal probability and test accuracy, the likelihood ratios for flu score 1, a no-test/test threshold of 10% and test/treat threshold of 50%, we have summarized a suggested approach to the evaluation of patients with suspected influenza in Table 5. Physicians wishing to limit use of anti-influenza drugs should consider rapid testing even in patients who are at high risk during peak flu season. Empiric therapy might be considered for patients at high risk of complications (181). [Diagnosis; Development; Validation; Implications for Clinical Use]

> To further appreciate these results, a few issues need to be addressed. First, although outpatients were included in the trial from which the data originated, for these analyses we deliberately restricted the study population to inpatients, because the PONV [postoperative nausea and vomiting] incidence in outpatients was substantially less frequent (34%) and because different types of surgery were performed (e.g. no abdominal surgery). Accordingly, our results should primarily be generalized to inpatients. It should be noted that, currently, no rules are available that were derived on both inpatients and outpatients. This is still a subject for future research, particularly given the increase of ambulatory surgery (437). [Prognosis; Incremental Value; Implications for Clinical Use]

> Our study had several limitations that should be acknowledged. We combined data from 2 different populations with somewhat different inclusion criteria, although the resulting dataset has the advantage of greater generalizability because it includes patients from 2 countries during 2 different flu seasons and has an overall pretest probability typical of that for influenza season. Also, data collection was limited to adults, so it is not clear whether these findings would apply to younger patients. Although simple, the point scoring may be too complex to remember and would be aided by programming as an application for smart phones and/or the Internet (181). [Diagnosis; Development; Validation; Limitations; Implications for Research]

Explanation

In the Discussion section, an author has both the opportunity and the responsibility to discuss the implications of the work at several levels. Prediction models can have different potential uses. In item 3a (background and rationale for the model), researchers are encouraged to describe these for their models. The Discussion section is then the natural place to discuss the potential clinical application in the light of the study findings. Clearly, for newly developed models, it may be more difficult to formally discuss how the model could be used in practice, because validation studies may be the logical next step. Indeed, authors should be discouraged from recommending application of a model on the basis of an initial development study only.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

Similarly, clinical guidelines should not recommend the use of nonvalidated prediction models. Rather, clinical recommendations should be based on the availability and synthesis of the evidence on the accuracy of a model in other participant data, and thus of the transportability of a model.

We stress that external model-validation studies, even prospectively designed studies, do not indicate the extent to which the use of such models affect medical decision making or improve health outcomes. The effect on decision making, clinician behavior, and patient outcomes can only be evaluated in comparative (preferably randomized [438–440]) studies rather than single-cohort, model-validation studies (20, 28, 33). Unfortunately, external model-validation studies are rare, let alone model-impact studies (441, 442).

Among the topics that may be discussed with regard to "To whom do results apply?" include setting (primary care, hospital), geographic location, age, sex, and clinical features of the medical problem for which prediction is being assessed. A second topic concerns how the rule might actually be applied: for example, whether the goal of a validated diagnostic model is to confirm or to exclude disease, which cut-offs in the prediction rule might be used for each goal, and what the possible consequences are (for example, further workup, or false-positive or false-negative findings).

Beyond the immediate possible implications, specific suggestions for further research could be based on the limitations of the current study, regarding such issues as the need to validate a newly developed model in other data sets, power of the rule to accomplish its goals (and potential usefulness of other predictors), selection of thresholds to guide clinical management, and problems in practical applications.

Other Information

Supplementary Information

Item 21. Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets. [D;V]

Examples

> The design and methods of the RISK-PCI trial have been previously published [ref]. Briefly, the RISK-PCI is an observational, longitudinal, cohort, single, center trial specifically designed to generate and validate an accurate risk model to predict major adverse cardiac events after contemporary pPCI [primary percutaneous coronary intervention] in patients pretreated with 600 mg clopidogrel. Patients were recruited between February 2006 and December 2009. Informed consent was obtained from each patient. The study protocol conforms to the ethical guidelines of the Declaration of Helsinki. It was approved by a local research ethics committee and registered in the Current Controlled Trials Register–ISRCTN83474650–(www.controlled-trials.com/ISRCTN83474650) (443). [Prognosis; Development]

> User-friendly calculators for the Reynolds Risk Scores for men and women can be freely accessed at www.reynoldsriskscore.org (444). [Prognosis; Incremental Value]

> Open source codes to calculate the QFracture-Scores are available from www.qfracture.org released under the GNU lesser general public licence-version 3. (315). [Prognosis; Validation]

Explanation

All research on humans should be protocol-driven (445, 446). A protocol for a prediction model study should start with a clear research aim, followed by the study design, description of predictors and outcome, and a statistical analysis plan. Studies to develop or validate a prediction model will benefit from careful construction of a detailed protocol before analysis is conducted; these protocols are occasionally published (447–464). If published, readers can compare what was planned with what was actually done. If not, we encourage authors to send their protocol with their submission to a journal and possibly include it with the published article as an online appendix to help reviewers evaluate the published report.

To use a prediction model, either in daily practice or for further research, requires that the model be published in sufficient detail (items 15a, 15b, and 16) to allow probability predictions to be made for an individual and for researchers to validate and update the prediction model. In addition, authors are encouraged to provide details on how to access any Web calculators and standalone applications (for example, for electronic devices, such as tablets) that have been developed (for example, www.outcomes-umassmed.org/GRACE/). In rare instances where the full prediction is too complex to report in sufficient detail in the published report (or supplementary material) or if the model is to be continually updated (for example, QRISK2 [139]), details on where full access to the underlying computer source code to calculate predictions should be made available.

There is increasing appreciation that where possible, data sets and computer code should be made publicly available for reproducing analyses (27, 465–467), but also to allow individual participant data to be combined and meta-analyzed (468–473). Guidance for preparing and sharing raw clinical data with other scientists has been developed that will assist authors (271). The exemplar study by Marchionni and colleagues (474) provides a prototypic template for reproducible development of a prognostic model demonstrating that transparency of the full process can be achieved. If possible, authors should provide details for access to the source code used for the data analyses.

There is currently no mandatory requirement for registration of observational studies; there has been recent support for the idea (475–478), but also opposition (479–481). Many clinical trials registries, including ClinicalTrials.gov, explicitly state that observational studies can be registered (482). Although there are clear difficulties associated with detailed preplanning

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

of some types of observational studies, those concerns should not apply to (prospective) prediction model studies collecting new participant data for the purpose of developing or validating a prediction model (476).

Funding

Item 22. Give the source of funding and the role of the funders for the present study. [D;V]

Examples

The Reynolds Risk Score Project was supported by investigator-initiated research grants from the Donald W. Reynolds Foundation (Las Vegas, Nev) with additional support from the Doris Duke Charitable Foundation (New York, NY), and the Leducq Foundation (Paris, France). The Women's Health Study cohort is supported by grants from the National Heart, Lung, and Blood Institute and the National Cancer Institute (Bethesda, Md) (208). [Prognosis; Development]

The Clinical and Translational Service Center at Weill Cornell Medical College provided partial support for data analyses. The funding source had no role in the design of our analyses, its interpretation, or the decision to submit the manuscript for publication (380). [Diagnosis; Development; Validation]

Explanation

Studies of prediction, even prospective studies, tend to receive little or no funding, which has been suggested to contribute to the large number of poor-quality studies: many are conducted without any peer review during the planning phase, when funding is usually sought (472).

Authors should disclose all sources of funding received for conducting the study and state what role of the funder had in the design, conduct, analysis, and reporting of the study. If the funder had no involvement, the authors should state so. Similarly, if the study received no external funding, the authors should clearly say so. For models that are incorporated in guidelines, it is important to show the potential financial and other conflicts of interest of all guideline development members, not just those involved in the prediction model development (316, 483, 484).

CONCLUDING REMARKS

Studies addressing prediction models are abundant, with the number of publications describing the development, validation, updating, or extension of prediction models showing no sign of abating. The TRIPOD Statement aims to provide helpful guidance for the reporting of studies developing or validating (without or with updating) 1 or more prediction models, either for diagnostic or prognostic purposes. Only with full and transparent reporting can the strengths and weaknesses of a study be revealed, thereby facilitating its interpretation and making it usable (485–487). Complete reporting also underpins future prediction model studies, notably allowing researchers to validate and compare existing prediction models. It can also contribute to and enhance the uptake and implementation of validated prediction models for use in daily practice. The TRIPOD Statement will be useful in guiding peer reviewers and journal editors in the evaluation of articles on prediction model studies. TRIPOD may also aid in the design, conduct, and analysis of prediction model studies.

TRIPOD was developed by a multidisciplinary group of 24 experts, including several who were also part of the CONSORT (96), STROBE (97, 99), PRISMA (488), REMARK (98), GRIPS (101), STREGA (489), STARD (100), ARRIVE (490), and CARE (491) reporting guidelines. Using this collective experience of developing consensus-based guidelines with expert subject knowledge, we adhered to published guidance on developing reporting guidelines (113). For each item in the checklist, we have provided extensive discussion, providing the rationale and including illustrative examples of good reporting. Where possible, we have referred to relevant empirical evidence from reviews of publications. Furthermore, we have included several boxes to provide additional discussion on key issues in developing and validating prediction models.

Some may argue that TRIPOD will increase the workload for the authors, reviewers, and journals, but we believe following TRIPOD will probably reduce review time, reduce requests for revisions, and help to ensure a fair review process (108). The items included in the checklist reflect numerous discussions to reach consensus on the minimal set of information to report to enable an informed assessment of study quality, risks of bias and clinical relevance, and enable the results to be used (532).

Reporting guidelines have also mistakenly been suggested to stifle research creativity. Like other reporting guidelines, TRIPOD does not dictate how analyses should be carried out, but rather aims to ensure the relevant information is reported.

Finally, the TRIPOD Statement should be viewed as an evolving document that will require continual assessment, and if necessary refinement, as methodology for prediction model studies continues to evolve. The TRIPOD Web site (www.tripod-statement.org) will provide a forum for discussion, suggestions for improving the checklist and this explanation and elaboration document, and resources relevant to prediction model studies. We also envisage encouraging translations of the checklist and making them available on the Web site. Announcements and information relating to TRIPOD will be broadcast on the TRIPOD Twitter address (@TRIPODStatement). TRIPOD will also be linked to and promoted by the EQUATOR Network library for health research reporting (www.equator-network.org).

From Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands; Cen-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

tre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford, United Kingdom; Stanford Prevention Research Center, School of Medicine, School of Humanities and Sciences, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California; Screening and Test Evaluation Program (STEP), School of Public Health, Sydney Medical School, University of Sydney, Sydney, Australia; Erasmus MC–University Medical Center Rotterdam, Rotterdam, the Netherlands; Memorial Sloan Kettering Cancer Center, New York, New York; and University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M14-0698.

Grant Support: There was no explicit funding for the development of this checklist and guidance document. The consensus meeting in June 2011 was partially funded by a National Institute for Health Research Senior Investigator Award held by Dr. Altman, Cancer Research UK (grant C5529), and the Netherlands Organization for Scientific Research (ZONMW 918.10.615 and 91208004). Drs. Collins and Altman are funded in part by the Medical Research Council (grant G1100513). Dr. Altman is a member of the Medical Research Council Prognosis Research Strategy (PROGRESS) Partnership (G0902393/99558).

Requests for Single Reprints: Karel G.M. Moons, PhD, Julius Centre for Health Sciences and Primary Care, UMC Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands; e-mail, K.G.M.Moons@umcutrecht.nl.

Current Author Addresses: Drs. Moons and Reitsma: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands.

Drs. Altman and Collins: Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford OX3 7LD, United Kingdom.

Dr. Ioannidis: Stanford Prevention Research Center, School of Medicine, Stanford University, 291 Campus Drive, Room LK3C02, Li Ka Shing Building, 3rd Floor, Stanford, CA 94305-5101.

Dr. Macaskill: Screening and Test Evaluation Program (STEP), School of Public Health, Edward Ford Building (A27), Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia.

Dr. Steyerberg: Department of Public Health, Erasmus MC–University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, the Netherlands.

Dr. Vickers: Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, Box 44, New York, NY 10065.

Dr. Ransohoff: Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, 4103 Bioinformatics, CB 7080, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7080.

Author Contributions: Conception and design: K.G.M. Moons, D.G. Altman, J.B. Reitsma, P. Macaskill, G.S. Collins.

Analysis and interpretation of the data: K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins.

Drafting of the article: K.G.M. Moons, D.G. Altman, J.B. Reitsma, G.S. Collins.

Critical revision of the article for important intellectual content: K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins.

Final approval of the article: K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins.

Provision of study materials or patients: K.G.M. Moons, D.G. Altman, J.B. Reitsma, G.S. Collins.

Statistical expertise: K.G.M. Moons, D.G. Altman, J.B. Reitsma, P. Macskill, E.W. Steyerberg, A.J. Vickers, G.S. Collins.

Obtaining of funding: K.G.M. Moons, D.G. Altman, G.S. Collins.

Administrative, technical, or logistic support: K.G.M. Moons, G.S. Collins.

Collection and assembly of data: K.G.M. Moons, D.G. Altman, G.S. Collins.

References

1. **Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG.** Prognosis and prognostic research: what, why, and how? BMJ. 2009; 338:b375. [PMID: 19237405]
2. **Steyerberg EW.** Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
3. **Wasson JH, Sox HC, Neff RK, Goldman L.** Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985; 313:793-9. [PMID: 3897864]
4. **Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al.** Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ. 2011; 343:d5888. [PMID: 21968126]
5. **Hayward RA, Kent DM, Vijan S, Hofer TP.** Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol. 2006;6:18. [PMID: 16613605]
6. **Kattan MW, Vickers AJ.** Incorporating predictions of individual patient risk in clinical trials. Urol Oncol. 2004;22:348-52. [PMID: 15283895]
7. **Kent DM, Hayward RA.** Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007;298:1209-12. [PMID: 17848656]
8. **Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10: e1001380. [PMID: 23393429]
9. **Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10: e1001381. [PMID: 23393430]
10. **Royston P, Moons KG, Altman DG, Vergouwe Y.** Prognosis and prognostic research: developing a prognostic model. BMJ. 2009; 338:b604. [PMID: 19336487]
11. **Collins GS, Altman DG.** Identifying patients with undetected renal tract cancer in primary care: an independent and external validation of QCancer® (Renal) prediction model. Cancer Epidemiol. 2013; 37:115-20. [PMID: 23280341]
12. **Harrell FE Jr, Lee KL, Mark DB.** Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy,

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

and measuring and reducing errors. Stat Med. 1996;15:361-87. [PMID: 8668867]
13. **Canet J, Gallart L, Gomar C, Paluzie G, Vallès J, Castillo J, et al; ARISCAT Group.** Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010; 113:1338-50. [PMID: 21045639]
14. **Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al.** EuroSCORE II. Eur J Cardiothorac Surg. 2012;41:734-44. [PMID: 22378855]
15. **Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, et al.** An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007;30:510-5. [PMID: 17327313]
16. **Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P.** Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ. 2009;338:b880. [PMID: 19297312]
17. **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;117:743-53. [PMID: 18212285]
18. **North RA, McCowan LM, Dekker GA, Poston L, Chan EH, Stewart AW, et al.** Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. BMJ. 2011;342:d1875. [PMID: 21474517]
19. **Altman DG, Vergouwe Y, Royston P, Moons KG.** Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338: b605. [PMID: 19477892]
20. **Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al.** Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691-8. [PMID: 22397946]
21. **Toll DB, Janssen KJ, Vergouwe Y, Moons KG.** Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61:1085-94. [PMID: 19208371]
22. **Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B.** Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2012; 42:216-28. [PMID: 21726217]
23. **Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG.** Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003; 56:441-7. [PMID: 12812818]
24. **Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD.** Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000; 19:1059-79. [PMID: 10790680]
25. **Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD.** Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45-56. [PMID: 11206946]
26. **Altman DG, Royston P.** What do we mean by validating a prognostic model? Stat Med. 2000;19:453-73. [PMID: 10694730]
27. **Ioannidis JPA, Khoury MJ.** Improving validation practices in "omics" research. Science. 2011;334:1230-2. [PMID: 22144616]
28. **Justice AC, Covinsky KE, Berlin JA.** Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-24. [PMID: 10075620]
29. **McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS.** Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA. 2000;284:79-84. [PMID: 10872017]
30. **Taylor JM, Ankesrt DP, Andridge RR.** Validation of biomarker-based risk prediction models. Clin Cancer Res. 2008;14:5977-83. [PMID: 18829476]
31. **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. [PMID: 18083464]
32. **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;54:774-81. [PMID: 11470385]
33. **Reilly BM, Evans AT.** Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201-9. [PMID: 16461965]
34. **Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al.** Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9:1-12. [PMID: 22629234]
35. **Rabar S, Lau R, O'Flynn N, Li L, Barry P; Guideline Development Group.** Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698. [PMID: 22875946]
36. National Institute for Health and Care Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. Clinical guideline CG67. London: National Institute for Health and Care Excellence; 2008. Accessed at http://guidance.nice.org.uk/CG67 on 30 October 2011.
37. National Osteoporosis Foundation. Clinician's guide to prevention and treatment of osteoporosis. Washington DC: National Osteoporsis Foundation; 2010. Accessed at http://nof.org/hcp/clinicians-guide on 17 January 2013.
38. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. Circulation. 2002;106:3143-421. [PMID: 12485966]
39. **Goldstein LB, Adams R, Alberts MJ, Appel LJ, Brass LM, Bushnell CD, et al; American Heart Association; American Stroke Association Stroke Council.** Primary prevention of ischemic stroke: a guideline from the American Heart Association/American Stroke Association Stroke Council: cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group; Cardiovascular Nursing Council; Clinical Cardiology Council; Nutrition, Physical Activity, and Metabolism Council; and the Quality of Care and Outcomes Research Interdisciplinary Working Group. Circulation. 2006;113: e873-923. [PMID: 16785347]
40. **Lackland DT, Elkind MS, D'Agostino R Sr, Dhamoon MS, Goff DC Jr, Higashida RT, et al; American Heart Association Stroke Council; Council on Epidemiology and Prevention; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; Council on Peripheral Vascular Disease; Council on Quality of Care and Outcomes Research.** Inclusion of stroke in cardiovascular risk prediction instruments: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2012;43:1998-2027. [PMID: 22627990]
41. **Perel P, Edwards P, Wentz R, Roberts I.** Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. 2006;6:38. [PMID: 17105661]
42. **Shariat SF, Karakiewicz PI, Margulis V, Kattan MW.** Inventory of prostate cancer predictive tools. Curr Opin Urol. 2008;18:279-96. [PMID: 18382238]
43. **Altman DG.** Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 2009;27:235-43. [PMID: 19291527]
44. **van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al.** Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart. 2012;98:360-9. [PMID: 22184101]
45. **Collins GS, Mallett S, Omar O, Yu LM.** Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103. [PMID: 21902820]
46. **Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG.** Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. Circulation. 2010;122:682-9. [PMID: 20679549]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

47. **Collins GS, Moons KG.** Comparing risk prediction models. BMJ. 2012;344:e3186. [PMID: 22628131]
48. **Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP.** Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318. [PMID: 22628003]
49. **Seel RT, Steyerberg EW, Malec JF, Sherer M, Macciocchi SN.** Developing and evaluating prediction models in rehabilitation populations. Arch Phys Med Rehabil. 2012;93(8 Suppl):S138-53. [PMID: 22840880]
50. **Green SM, Schriger DL, Yealy DM.** Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. Ann Emerg Med. 2014;64:286-91. [PMID: 24530108]
51. **Laine C, Goodman SN, Griswold ME, Sox HC.** Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146:450-3. [PMID: 17339612]
52. **Groves T, Godlee F.** Open science and reproducible research. BMJ. 2012;344:e4383. [PMID: 22736475]
53. **Collins GS, Omar O, Shanyinde M, Yu LM.** A systematic review finds prediction models for chronic kidney were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013;66:268-77. [PMID: 23116690]
54. **Mallett S, Royston P, Dutton S, Waters R, Altman DG.** Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20. [PMID: 20353578]
55. **Mallett S, Royston P, Waters R, Dutton S, Altman DG.** Reporting performance of prognostic models in cancer: a review. BMC Med. 2010;8:21. [PMID: 20353579]
56. **Burton A, Altman DG.** Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91:4-8. [PMID: 15188004]
57. **Concato J, Feinstein AR, Holford TR.** The risk of determining risk with multivariable models. Ann Intern Med. 1993;118:201-10. [PMID: 8417638]
58. **Laupacis A, Sekar N, Stiell IG.** Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94. [PMID: 9020274]
59. **Steurer J, Haller C, Häuselmann H, Brunner F, Bachmann LM.** Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. PLoS One. 2011;6:e19994. [PMID: 21625596]
60. **van Dijk WD, Bemt L, Haak-Rongen S, Bischoff E, Weel C, Veen JC, et al.** Multidimensional prognostic indices for use in COPD patient care. A systematic review. Respir Res. 2011;12:151. [PMID: 22082049]
61. **Hayden JA, Côté P, Bombardier C.** Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med. 2006;144:427-37. [PMID: 16549855]
62. **Meads C, Ahmed I, Riley RD.** A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132:365-77. [PMID: 22037780]
63. **Mushkudiani NA, Hukkelhoven CW, Hernández AV, Murray GD, Choi SC, Maas AI, et al.** A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol. 2008;61:331-43. [PMID: 18313557]
64. **Rehn M, Perel P, Blackhall K, Lossius HM.** Prognostic models for the early care of trauma patients: a systematic review. Scand J Trauma Resusc Emerg Med. 2011;19:17. [PMID: 21418599]
65. **Siontis GC, Tzoulaki I, Ioannidis JP.** Predicting death: an empirical evaluation of predictive tools for mortality. Arch Intern Med. 2011;171:1721-6. [PMID: 21788535]
66. **Medlock S, Ravelli ACJ, Tamminga P, Mol BW, Abu-Hanna A.** Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6:e23441. [PMID: 21931598]
67. **Maguire JL, Kulik DM, Laupacis A, Kuppermann N, Uleryk EM, Parkin PC.** Clinical prediction rules for children: a systematic review. Pediatrics. 2011;128:e666-77. [PMID: 21859912]
68. **Kulik DM, Uleryk EM, Maguire JL.** Does this child have appendicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. J Clin Epidemiol. 2013;66:95-104. [PMID: 23177898]
69. **Kulik DM, Uleryk EM, Maguire JL.** Does this child have bacterial meningitis? A systematic review of clinical prediction rules for children with suspected bacterial meningitis. J Emerg Med. 2013;45:508-19. [PMID: 23910166]
70. **Jacob M, Lewsey JD, Sharpin C, Gimson A, Rela M, van der Meulen JH.** Systematic review and validation of prognostic models in liver transplantation. Liver Transpl. 2005;11:814-25. [PMID: 15973726]
71. **Hussain A, Choukairi F, Dunn K.** Predicting survival in thermal injury: a systematic review of methodology of composite prediction models. Burns. 2013;39:835-50. [PMID: 23384617]
72. **Haskins R, Rivett DA, Osmotherly PG.** Clinical prediction rules in the physiotherapy management of low back pain: a systematic review. Man Ther. 2012;17:9-21. [PMID: 21641849]
73. **Echouffo-Tcheugui JB, Kengne AP.** Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS Med. 2012;9:e1001344. [PMID: 23185136]
74. **Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP.** Risk models to predict hypertension: a systematic review. PLoS One. 2013;8:e67370. [PMID: 23861760]
75. **Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A.** Risk prediction models of breast cancer: a systematic review of model performances. Breast Cancer Res Treat. 2012;133:1-10. [PMID: 22076477]
76. **van Oort L, van den Berg T, Koes BW, de Vet RH, Anema HJ, Heymans MW, et al.** Preliminary state of development of prediction models for primary care physical therapy: a systematic review. J Clin Epidemiol. 2012;65:1257-66. [PMID: 22959592]
77. **Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, et al.** Risk prediction models for patients with chronic kidney disease: a systematic review. Ann Intern Med. 2013;158:596-603. [PMID: 23588748]
78. **van Hanegem N, Breijer MC, Opmeer BC, Mol BW, Timmermans A.** Prediction models in women with postmenopausal bleeding: a systematic review. Womens Health (Lond Engl). 2012;8:251-62. [PMID: 22554173]
79. **Minne L, Ludikhuize J, de Jonge E, de Rooij S, Abu-Hanna A.** Prognostic models for predicting mortality in elderly ICU patients: a systematic review. Intensive Care Med. 2011;37:1258-68. [PMID: 21647716]
80. **Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, et al.** Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update. 2009;15:537-52. [PMID: 19435779]
81. **Jaja BN, Cusimano MD, Etminan N, Hanggi D, Hasan D, Ilodigwe D, et al.** Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. Neurocrit Care. 2013;18:143-53. [PMID: 23138544]
82. **Wlodzimirow KA, Eslami S, Chamuleau RA, Nieuwoudt M, Abu-Hanna A.** Prediction of poor outcome in patients with acute liver failure-systematic review of prediction models. PLoS One. 2012;7:e50952. [PMID: 23272081]
83. **Phillips B, Wade R, Stewart LA, Sutton AJ.** Systematic review and meta-analysis of the discriminatory performance of risk prediction rules in febrile neutropaenic episodes in children and young people. Eur J Cancer. 2010;46:2950-64. [PMID: 20621468]
84. **Rubin KH, Friis-Holmberg T, Hermann AP, Abrahamsen B, Brixen K.** Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. J Bone Miner Res. 2013;28:1701-17. [PMID: 23592255]
85. **Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al.** Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ. 2012;345:e5900. [PMID: 22990994]
86. **Braband M, Folkestad L, Clausen NG, Knudsen T, Hallas J.** Risk scoring systems for adults admitted to the emergency department: a systematic review. Scand J Trauma Resusc Emerg Med. 2010;18:8. [PMID: 20146829]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

87. **Maguire JL, Boutis K, Uleryk EM, Laupacis A, Parkin PC.** Should a head-injured child receive a head CT scan? A systematic review of clinical prediction rules. Pediatrics. 2009;124:e145-54. [PMID: 19564261]
88. **Vuong K, McGeechan K, Armstrong BK, Cust AE.** Risk prediction models for incident primary cutaneous melanoma: a systematic review. JAMA Dermatol. 2014;150:434-44. [PMID: 24522401]
89. **Ahmed I, Debray TP, Moons KG, Riley RD.** Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3. [PMID: 24397587]
90. **Huen SC, Parikh CR.** Predicting acute kidney injury after cardiac surgery: a systematic review. Ann Thorac Surg. 2012;93:337-41. [PMID: 22186469]
91. **Calle P, Cerro L, Valencia J, Jaimes F.** Usefulness of severity scores in patients with suspected infection in the emergency department: a systematic review. J Emerg Med. 2012;42:379-91. [PMID: 22142675]
92. **Usher-Smith JA, Emery J, Kassianos AP, Walter FM.** Risk prediction models for melanoma: a systematic review. Cancer Epidemiol Biomarkers Prev. 2014;23:1450-63. [PMID: 24895414]
93. **Warnell I, Chincholkar M, Eccles M.** Predicting perioperative mortality after oesophagectomy: a systematic review of performance and methods of multivariate models. Br J Anaesth. 2014 Sep 17. [Epub ahead of print]. [PMID: 25231768]
94. **Silverberg N, Gardner AJ, Brubacher J, Panenka W, Li JJ, Iverson GL.** Systematic review of multivariable prognostic models for mild traumatic brain injury. J Neurotrauma. 2014 Sep 15. [Epub ahead of print]. [PMID: 25222514]
95. **Delebarre M, Macher E, Mazingue F, Martinot A, Dubos F.** Which decision rules meet methodological standards in children with febrile neutropenia? Results of a systematic review and analysis. Pediatr Blood Cancer. 2014;61:1786-91. [PMID: 24975886]
96. **Schulz KF, Altman DG, Moher D; CONSORT Group.** CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. [PMID: 20332509]
97. **von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative.** Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335:806-8. [PMID: 17947786]
98. **McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics.** Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst. 2005;97:1180-4. [PMID: 16106022]
99. **Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M, et al.** STrengthening the Reporting of OBservational studies in Epidemiology - Molecular Epidemiology (STROBE-ME): an extension of the STROBE statement. Eur J Clin Invest. 2012;42:1-16. [PMID: 22023344]
100. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; Standards for Reporting of Diagnostic Accuracy.** Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Radiology. 2003;226:24-8. [PMID: 12511664]
101. **Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ; GRIPS Group.** Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Clin Invest. 2011;41:1004-9. [PMID: 21434891]
102. **Moons KG, Altman DG, Vergouwe Y, Royston P.** Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. [PMID: 19502216]
103. **Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al.** Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683-90. [PMID: 22397945]
104. **Labarère J, Bertrand R, Fine MJ.** How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med. 2014;40:513-27. [PMID: 24570265]
105. **Tzoulaki I, Liberopoulos G, Ioannidis JP.** Use of reclassification for assessment of improved prediction: an empirical evaluation. Int J Epidemiol. 2011;40:1094-105. [PMID: 21325392]
106. **Peters SA, Bakker M, den Ruijter HM, Bots ML.** Added value of CAC in risk stratification for cardiovascular events: a systematic review. Eur J Clin Invest. 2012;42:110-6. [PMID: 21644944]
107. **Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al; International Diagnostic and Prognosis Prediction (IDAPP) Group.** Framework for the impact analysis and implementation of clinical prediction rules (CPRs). BMC Med Inform Decis Mak. 2011;11:62. [PMID: 21999201]
108. **Altman DG, McShane LM, Sauerbrei W, Taube SE.** Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med. 2012;10:51. [PMID: 22642691]
109. **Campbell MK, Elbourne DR, Altman DG; CONSORT Group.** CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328:702-8. [PMID: 15031246]
110. **Teasdale G, Jennett B.** Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;2:81-4. [PMID: 4136544]
111. **Farrell B, Godwin J, Richards S, Warlow C.** The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: final results. J Neurol Neurosurg Psychiatry. 1991;54:1044-54. [PMID: 1783914]
112. **Harrell FE.** Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001.
113. **Moher D, Schulz KF, Simera I, Altman DG.** Guidance for developers of health research reporting guidelines. PLoS Med. 2010;16:e1000217. [PMID: 20169112]
114. **Collins GS, Reitsma JB, Altman DG, Moons KG.** Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis: the TRIPOD statement. Ann Intern Med. 2014;162:55-63.
115. **Morise AP, Haddad WJ, Beckner D.** Development and validation of a clinical score to estimate the probability of coronary artery disease in men and women presenting with suspected coronary disease. Am J Med. 1997;102:350-6. [PMID: 9217616]
116. **Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, et al.** Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2009;74:355-62. [PMID: 19095367]
117. **Collins GS, Altman DG.** Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344:e4181. [PMID: 22723603]
118. **Michikawa T, Inoue M, Sawada N, Iwasaki M, Tanaka Y, Shimazu T, et al; Japan Public Health Center-based Prospective Study Group.** Development of a prediction model for 10-year risk of hepatocellular carcinoma in middle-aged Japanese: the Japan Public Health Center-based Prospective Study Cohort II. Prev Med. 2012;55:137-43. [PMID: 22676909]
119. **Morise AP, Detrano R, Bobbio M, Diamond GA.** Development and validation of a logistic regression-derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing. J Am Coll Cardiol. 1992;20:1187-96. [PMID: 1401621]
120. **D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P; CHD Risk Prediction Group.** Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA. 2001;286:180-7. [PMID: 11448281]
121. **Beck DH, Smith GB, Pappachan JV, Millar B.** External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med. 2003;29:249-56. [PMID: 12536271]
122. **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. [PMID: 24645774]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

123. **Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al.** Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012; 345:e5166. [PMID: 22896030]
124. **Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M, et al.** Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. JAMA. 1993; 269:1127-32. [PMID: 8433468]
125. **Holland JL, Wilczynski NL, Haynes RB; Hedges Team.** Optimal search strategies for identifying sound clinical prediction studies in EMBASE. BMC Med Inform Decis Mak. 2005;5:11. [PMID: 15862125]
126. **Ingui BJ, Rogers MA.** Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8:391-7. [PMID: 11418546]
127. **Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R; Hedges Team.** Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc. 2003:728-32. [PMID: 14728269]
128. **Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG.** Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844. [PMID: 22393453]
129. **Keogh C, Wallace E, O'Brien KK, Murphy PJ, Teljeur C, McGrath B, et al.** Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. J Clin Epidemiol. 2011;64:848-60. [PMID: 21411285]
130. **Rietveld RP, ter Riet G, Bindels PJ, Sloos JH, van Weert HC.** Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. BMJ. 2004; 329:206-10. [PMID: 15201195]
131. **Poorten VV, Hart A, Vauterin T, Jeunen G, Schoenaers J, Hamoir M, et al.** Prognostic index for patients with parotid carcinoma: international external validation in a Belgian-German database. Cancer. 2009;115:540-50. [PMID: 9137571]
132. **Moynihan R, Glassock R, Doust J.** Chronic kidney disease controversy: how expanding definitions are unnecessarily labelling many people as diseased. BMJ. 2013;347:f4298. [PMID: 23900313]
133. **Moynihan R, Henry D, Moons KG.** Using evidence to combat overdiagnosis and overtreatment: evaluating treatments, tests, and disease definitions in the time of too much. PLoS Med. 2014;11: e1001655. [PMID: 24983872]
134. **Dowling S, Spooner CH, Liang Y, Dryden DM, Friesen C, Klassen TP, et al.** Accuracy of Ottawa Ankle Rules to exclude fractures of the ankle and midfoot in children: a meta-analysis. Acad Emerg Med. 2009;16:277-87. [PMID: 19187397]
135. **Bachmann LM, Kolb E, Koller MT, Steurer J, ter Riet G.** Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ. 2003;326:417. [PMID: 12595378]
136. **Büller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R, et al; AMUSE (Amsterdam Maastricht Utrecht Study on thromboEmbolism) Investigators.** Safely ruling out deep venous thrombosis in primary care. Ann Intern Med. 2009;150:229-35. [PMID: 19221374]
137. **Sparks AB, Struble CA, Wang ET, Song K, Oliphant A.** Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. Am J Obstet Gynecol. 2012;206:319.e1-9. [PMID: 22464072]
138. **Ankerst DP, Boeck A, Freedland SJ, Thompson IM, Cronin AM, Roobol MJ, et al.** Evaluating the PCPT risk calculator in ten international biopsy cohorts: results from the Prostate Biopsy Collaborative Group. World J Urol. 2012;30:181-7. [PMID: 22210512]
139. **Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al.** Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008; 336:1475-82. [PMID: 18573856]
140. **Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al; SCORE Project Group.** Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24:987-1003.
141. **Califf RM, Woodlief LH, Harrell FE Jr, Lee KL, White HD, Guerci A, et al.** Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J. 1997;133:630-9. [PMID: 9200390]
142. **McCowan C, Donnan PT, Dewar J, Thompson A, Fahey T.** Identifying suspected breast cancer: development and validation of a clinical prediction rule. Br J Gen Pract. 2011;61:e205-14. [PMID: 21619744]
143. **Campbell HE, Gray AM, Harris AL, Briggs AH, Taylor MA.** Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the UK. Br J Cancer. 2010;103:776-86. [PMID: 20823886]
144. **Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB.** Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837-47. [PMID: 9603539]
145. **Kengne AP, Patel A, Marre M, Travert F, Lievre M, Zoungas S, et al; ADVANCE Collaborative Group.** Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. Eur J Cardiovasc Prev Rehabil. 2011;18:393-8. [PMID: 21450612]
146. **Appelboam A, Reuben AD, Benger JR, Beech F, Dutson J, Haig S, et al.** Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. BMJ. 2008;337:a2428. [PMID: 19066257]
147. **Puhan MA, Hansel NN, Sobradillo P, Enright P, Lange P, Hickson D, et al; International COPD Cohorts Collaboration Working Group.** Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. BMJ Open. 2012;2:6. [PMID: 23242246]
148. **Knottnerus JA.** *The Evidence Base of Clinical Diagnosis*. London: BMJ Books; 2002.
149. **Knottnerus JA, Muris JW.** Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol. 2003;56: 1118-28. [PMID: 14615003]
150. **Grobbee DE, Hoes AW.** Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. London: Jones and Bartlett Publishers; 2009.
151. **Sackett DL, Tugwell P, Guyatt GH.** *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2d ed. Boston: Little, Brown; 1991.
152. **Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG.** Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:48. [PMID: 18644127]
153. **Knottnerus JA, Dinant GJ.** Medicine based evidence, a prerequisite for evidence based medicine. BMJ. 1997;315:1109-10. [PMID: 9374881]
154. **Knottnerus JA, van Weel C, Muris JW.** Evaluation of diagnostic procedures. BMJ. 2002;324:477-80. [PMID: 11859054]
155. **Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM.** Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51:1335-41. [PMID: 15961549]
156. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]
157. **van Zaane B, Vergouwe Y, Donders AR, Moons KG.** Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. BMC Med Res Methodol. 2012;12:166. [PMID: 23114025]
158. **Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JP, Hernández-Aguado I.** Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem. 2009;55:786-94. [PMID: 19233907]
159. **Tzoulaki I, Siontis KC, Ioannidis JP.** Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ. 2011;343: d6829. [PMID: 22065657]
160. **Greving JP, Wermer MJ, Brown RD Jr, Morita A, Juvela S, Yonekura M, et al.** Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol. 2014;13:59-66. [PMID: 24290159]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

161. **Collins GS, Altman DG.** Predicting the adverse risk of statin treatment: an independent and external validation of Qstatin risk scores in the UK. Heart. 2012;98:1091-7. [PMID: 22689714]
162. **Glickman SW, Shofer FS, Wu MC, Scholer MJ, Ndubuizu A, Peterson ED, et al.** Development and validation of a prioritization rule for obtaining an immediate 12-lead electrocardiogram in the emergency department to identify ST-elevation myocardial infarction. Am Heart J. 2012;163:372-82. [PMID: 22424007]
163. **Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KG.** Incorporating published univariable associations in diagnostic and prognostic modeling. BMC Med Res Methodol. 2012;12:121. [PMID: 22883206]
164. **Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW.** Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med. 2012;31:2697-712. [PMID: 22733546]
165. **Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD.** Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? PLoS One. 2013;8:e60650. [PMID: 23585842]
166. **Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD.** A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158-80. [PMID: 23307585]
167. **Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y.** Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol. 2013;13:19. [PMID: 23414436]
168. **Bouwmeester W, Moons KG, Happen TH, van Klei WA, Twisk JW, Eijkemans MJ, et al.** Internal validation of risk models in clustered data: a comparison of bootstrap schemes. Am J Epidemiol. 2013;177:1209-17. [PMID: 23660796]
169. **Rosner B, Qiu W, Lee ML.** Assessing discrimination of risk prediction rules in a clustered data setting. Lifetime Data Anal. 2013;19:242-56. [PMID: 23263872]
170. **van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y.** Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol. 2014;14:5. [PMID: 24423445]
171. **van Klaveren D, Steyerberg EW, Vergouwe Y.** Interpretation of concordance measures for clustered data. Stat Med. 2014;33:714-6. [PMID: 24425541]
172. **Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L.** Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol. 2013;13:113. [PMID: 24034146]
173. **Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E.** Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol. 2012;175:715-24. [PMID: 22396388]
174. **Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K.** Case-cohort design in practice—experiences from the MORGAM Project. Epidemiol Perspect Innov. 2007;4:15. [PMID: 18053196]
175. **Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al.** Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol. 2014;2:19-29. [PMID: 24622666]
176. **Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al.** Risk prediction models for mortality in ambulatory heart failure patients: a systematic review. Circ Heart Fail. 2013;6:881-9. [PMID: 23888045]
177. **Arkenau HT, Barriuso J, Olmos D, Ang JE, de Bono J, Judson I, et al.** Prospective validation of a prognostic score to improve patient selection for oncology phase I trials. J Clin Oncol. 2009;27:2692-6. [PMID: 19332724]
178. **Ronga A, Vaucher P, Haasenritter J, Donner-Banzhoff N, Bösner S, Verdon F, et al.** Development and validation of a clinical prediction rule for chest wall syndrome in primary care. BMC Fam Pract. 2012;13:74. [PMID: 22866824]
179. **Martinez JA, Belastegui A, Basabe I, Goicoechea X, Aguirre C, Lizeaga N, et al.** Derivation and validation of a clinical prediction rule for delirium in patients admitted to a medical ward: an observational study. BMJ Open. 2012;2:e001599. [PMID: 22983876]
180. **Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al.** Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail. 2014;2:440-6. [PMID: 25194291]
181. **Ebell MH, Afonson AM, Gonzales R, Stein J, Genton B, Senn N.** Development and validation of a clinical decision rule for the diagnosis of influenza. J Am Board Fam Med. 2012;25:55-62. [PMID: 22218625]
182. **Counsell C, Dennis M.** Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12:159-70. [PMID: 11641579]
183. **Knottnerus JA.** Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002;55:1201-6. [PMID: 12547450]
184. **Moreno R, Apolone G.** Impact of different customization strategies in the performance of a general severity score. Crit Care Med. 1997;25:2001-8. [PMID: 9403750]
185. **Tu JV, Austin PC, Walld R, Roos L, Agras J, McDonald KM.** Development and validation of the Ontario acute myocardial infarction mortality prediction rules. J Am Coll Cardiol. 2001;37:992-7. [PMID: 11263626]
186. **Vergouwe Y, Moons KG, Steyerberg EW.** External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971-80. [PMID: 20807737]
187. **Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG.** Adaptation of clinical prediction models for application in local settings. Med Decis Making. 2012;32:E1-10. [PMID: 22427369]
188. **Oudega R, Hoes AW, Moons KG.** The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med. 2005;143:100-7. [PMID: 16027451]
189. **Knottnerus JA, Leffers P.** The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol. 1992;45:1143-54. [PMID: 1474411]
190. **Knottnerus JA.** The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making. 1987;7:139-48. [PMID: 3613914]
191. **Eberhart LH, Morin AM, Guber D, Kretz FJ, Schäuffelen A, Treiber H, et al.** Applicability of risk scores for postoperative nausea and vomiting in adults to paediatric patients. Br J Anaesth. 2004;93:386-92. [PMID: 15247114]
192. **Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG.** A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2014 Aug 29. [Epub ahead of print]. [PMID: 25179855]
193. **Klemke CD, Mansmann U, Poenitz N, Dippel E, Goerdt S.** Prognostic factors and prediction of prognosis by the CTCL Severity Index in mycosis fungoides and Sézary syndrome. Br J Dermatol. 2005;153:118-24. [PMID: 16029336]
194. **Tay SY, Thoo FL, Sitoh YY, Seow E, Wong HP.** The Ottawa Ankle Rules in Asia: validating a clinical decision rule for requesting X-rays in twisting ankle and foot injuries. J Emerg Med. 1999;17:945-7. [PMID: 10595876]
195. **Donders AR, van der Heijden GJ, Stijnen T, Moons KG.** Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-91. [PMID: 16980149]
196. **Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG.** Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184:1265-9. [PMID: 22371511]
197. **Janssen KJ, Donders AR, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al.** Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol. 2010;63:721-7. [PMID: 20338724]
198. **Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Chen Q, Grobbee DE, et al.** Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55:994-1001. [PMID: 19282357]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

199. **Moons KG, Donders RA, Stijnen T, Harrell FE Jr.** Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092-101. [PMID: 16980150]
200. **Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al.** Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. [PMID: 19564179]
201. **Vergouwe Y, Royston P, Moons KG, Altman DG.** Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63:205-14. [PMID: 19596181]
202. **Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al; PROGRESS Group.** Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:35595. [PMID: 23386360]
203. **Liew SM, Doust J, Glasziou P.** Cardiovascular risk scores do not account for the effect of treatment: a review. Heart. 2011;97:689-97. [PMID: 21474616]
204. **Simon R, Altman D, G.** Statistical aspects of prognostic factor studies in oncology. Br J Cancer. 1994;69:979-85. [PMID: 8198989]
205. **Landefeld CS, Goldman L.** Major bleeding in outpatients treated with warfarin: incidence and prediction by factors known at the start of outpatient therapy. Am J Med. 1989;87:144-52. [PMID: 2787958]
206. **Schuit E, Groenwold RH, Harrell FE Jr, de Kort WL, Kwee A, Mol BW, et al.** Unexpected predictor-outcome associations in clinical prediction research: causes and solutions. CMAJ. 2013;185:E499-505. [PMID: 23339155]
207. **Wong J, Taljaard M, Forster AJ, Escobar GJ, van Walraven C.** Addition of time-dependent covariates to a survival model significantly improved predictions for daily risk of hospital death. J Eval Clin Pract. 2013;19:351-7. [PMID: 22409151]
208. **Ridker PM, Buring JE, Rifai N, Cook NR.** Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA. 2007;297:611-9. [PMID: 17299196]
209. **Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM.** A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62:797-806. [PMID: 19447581]
210. **Massing MW, Simpson RJ Jr, Rautaharju PM, Schreiner PJ, Crow R, Heiss G.** Usefulness of ventricular premature complexes to predict coronary heart disease events and mortality (from the Atherosclerosis Risk In Communities cohort). Am J Cardiol. 2006;98:1609-12. [PMID: 17145219]
211. **Craig JC, Williams GJ, Jones M, Codarini M, Macaskill P, Hayen A, et al.** The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses. BMJ. 2010;340:c1594. [PMID: 20406860]
212. **Todenhofer T, Renninger M, Schwentner C, Stenzl A, Gakis G.** A new prognostic model for cancer-specific survival after radical cystectomy including pretreatment thrombocytosis and standard pathological risk factors. BJU Int. 2012;110(11 Pt B):E533-40. [PMID: 22578156]
213. **Boggs DA, Rosenberg L, Pencina MJ, Adams-Campbell LL, Palmer JR.** Validation of a breast cancer risk prediction model developed for Black women. J Natl Cancer Inst. 2013;105:361-7. [PMID: 23411594]
214. **Knottnerus JA, Buntinx F.** The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. Hoboken, NJ: Wiley-Blackwell; 2009.
215. **Naaktgeboren CA, de Groot JA, van Smeden M, Moons KG, Reitsma JB.** Evaluating diagnostic accuracy in the face of multiple reference standards. Ann Intern Med. 2013;159:195-202. [PMID: 23922065]
216. **Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al.** Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10:e1001531. [PMID: 24143138]
217. **Naaktgeboren CA, Bertens LC, van Smeden M, Groot JA, Moons KG, Reitsma JB.** Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605. [PMID: 24162938]
218. **de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al.** Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770. [PMID: 21810869]
219. **de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, et al.** Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. Am J Epidemiol. 2012;175:847-53. [PMID: 22422923]
220. **Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM.** Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76. [PMID: 16477057]
221. **Rouzier R, Pusztai L, Delaloge S, Gonzalez-Angulo AM, Andre F, Hess KR, et al.** Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer. J Clin Oncol. 2005;23:8331-9. [PMID: 16293864]
222. **Elliott J, Beringer T, Kee F, Marsh D, Willis C, Stevenson M.** Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery. J Clin Epidemiol. 2003;56:788-95. [PMID: 12954472]
223. **Adams LA, Bulsara M, Rossi E, DeBoer B, Speers D, George J, et al.** Hepascore: an accurate validated predictor of liver fibrosis in chronic hepatitis C infection. Clin Chem. 2005;51:1867-73. [PMID: 16055434]
224. **Hess EP, Brison RJ, Perry JJ, Calder LA, Thiruganasambandamoorthy V, Agarwal D, et al.** Development of a clinical prediction rule for 30-day cardiac events in emergency department patients with chest pain and possible acute coronary syndrome. Ann Emerg Med. 2012;59:115-25. [PMID: 21885156]
225. **Moons KG, Grobbee DE.** When should we remain blind and when should our eyes remain open in diagnostic studies? J Clin Epidemiol. 2002;55:633-6. [PMID: 12160909]
226. **Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM.** Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;iii, ix-51. [PMID: 18021577]
227. **Kaijser J, Sayasneh A, Van Hoorde K, Ghaem-Maghami S, Bourne T, Timmerman D, et al.** Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update. 2014;20:449-52. [PMID: 24327552]
228. **Kaul V, Friedenberg FK, Braitman LE, Anis U, Zaeri N, Fazili J, et al.** Development and validation of a model to diagnose cirrhosis in patients with hepatitis C. Am J Gastroenterol. 2002;97:2623-8. [PMID: 12385450]
229. **Halbesma N, Jansen DF, Heymans MW, Stolk RP, de Jong PE, Gansevoort RT ; PREVEND Study Group.** Development and validation of a general population renal risk score. Clin J Am Soc Nephrol. 2011;6:1731-8. [PMID: 21734089]
230. **Beyersmann J, Wolkewitz M, Schumacher M.** The impact of time-dependent bias in proportional hazards modelling. Stat Med. 2008;27:6439-54. [PMID: 18837068]
231. **van Walraven C, Davis D, Forster AJ, Wells GA.** Time-dependent bias was common in survival analyses published in leading clinical journals. J Clin Epidemiol. 2004;57:672-82. [PMID: 15358395]
232. **Rochon J.** Issues in adjusting for covariates arising postrandomization in clinical trials. Drug Inf J. 1999;33:1219-28.
233. **D'Agostino RB.** Beyond baseline data: the use of time-varying covariates. J Hypertens. 2008;26:639-40. [PMID: 18327070]
234. **Scheike TH.** Time-varying effects in survival analysis. In: Rao CR, ed. *Advances in Survival Analysis*. Amsterdam: Elsevier; 2004:61-8.
235. **Sun GW, Shook TL, Kay GL.** Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49:907-16. [PMID: 8699212]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

236. **Rutten FH, Vonken EJ, Cramer MJ, Moons KG, Velthuis BB, Prakken NH, et al.** Cardiovascular magnetic resonance imaging to identify left-sided chronic heart failure in stable patients with chronic obstructive pulmonary disease. Am Heart J. 2008;156:506-12. [PMID: 18760133]
237. **Hess EP, Perry JJ, Calder LA, Thiruganasambandamoorthy V, Body R, Jaffe A, et al.** Prospective validation of a modified thrombolysis in myocardial infarction risk score in emergency department patients with chest pain and possible acute coronary syndrome. Acad Emerg Med. 2010;17:368-75. [PMID: 20370775]
238. **Begg CB.** Bias in the assessment of diagnostic tests. Stat Med. 1987;6:411-23. [PMID: 3114858]
239. **Elmore JG, Wells CK, Howard DH, Feinstein AR.** The impact of clinical history on mammographic interpretations. JAMA. 1997;277:49-52. [PMID: 8980210]
240. **Loy CT, Irwig L.** Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA. 2004;292:1602-9. [PMID: 15467063]
241. **Loewen P, Dahir K.** Risk of bleeding with oral anticoagulants: an updated systematic review and performance analysis of clinical prediction rules. Ann Hematol. 2011;90:1191-200. [PMID: 21670974]
242. **Sheth T, Butler C, Chow B, Chan MT, Mitha A, Nagele P, et al; CTA VISION Investigators.** The coronary CT angiography vision protocol: a prospective observational imaging cohort study in patients undergoing non-cardiac surgery. BMJ Open. 2012;2:e001474. [PMID: 22855630]
243. **Hippisley-Cox J, Coupland C.** Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2012;62:e38-e45. [PMID: 22520674]
244. **Holmes JF, Mao A, Awasthi S, McGahan JP, Wisner DH, Kuppermann N.** Validation of a prediction rule for the identification of children with intra-abdominal injuries after blunt torso trauma. Ann Emerg Med. 2009;54:528-33. [PMID: 19250706]
245. **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;48:1503-12. [PMID: 8543964]
246. **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;49:1373-9. [PMID: 8970487]
247. **Vittinghoff E, McCulloch CE.** Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165:710-8. [PMID: 17182981]
248. **Feinstein AR.** *Multivariable Analysis*. New Haven, CT: Yale University Press; 1996.
249. **Schumacher M, Holländer N, Schwarzer G, Binder H, Sauerbrei W.** Prognostic factor studies. In: Crowley J, Hoering A, eds. *Handbook of Statistics in Clinical Oncology*. 3rd ed. London: Chapman and Hall/CRC; 2012:415-70.
250. **Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV.** Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64:993-1000. [PMID: 21411281]
251. **Jinks RC.** Sample size for multivariable prognostic models. PhD thesis. University College London; 2012.
252. **Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al.** Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. [PMID: 20010215]
253. **Steyerberg EW, Calster BV, Pencina MJ.** Performance measures for prediction models and markers: evaluation of predictions and classifications. Rev Esp Cardiol (Engl Ed). 2011;64:788-94. [PMID: 24775954]
254. **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;58:475-83. [PMID: 15845334]
255. **Audigé L, Bhandari M, Kellam J.** How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand. 2004;75:184-94. [PMID: 15180234]
256. **Genders TS, Steyerberg EW, Hunink MG, Nieman K, Galema TW, Mollet NR, et al.** Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ. 2012;344:e3485. [PMID: 22692650]
257. **Thompson DO, Hurtado TR, Liao MM, Byyny RL, Gravitz C, Haukoos JS.** Validation of the Simplified Motor Score in the out-of-hospital setting for the prediction of outcomes after traumatic brain injury. Ann Emerg Med. 2011;58:417-25. [PMID: 21803448]
258. **Ambler G, Omar RZ, Royston P, Kinsman R, Keogh BE, Taylor KM.** Generic, simple risk stratification model for heart valve surgery. Circulation. 2005;112:224-31. [PMID: 15998680]
259. **Mackinnon A.** The use and reporting of multiple imputation in medical research–a review. J Intern Med. 2010;268:586-93. [PMID: 20831627]
260. **Hussain A, Dunn KW.** Predicting length of stay in thermal burns: a systematic review of prognostic factors. Burns. 2013;39:1331-40. [PMID: 23768707]
261. **Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al.** A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011;305:1553-9. [PMID: 21482743]
262. **Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al.** Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5:e165. [PMID: 18684008]
263. **Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al.** Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. J Natl Cancer Inst. 2011;103:1058-68. [PMID: 21606442]
264. **Altman DG, Lausen B, Sauerbrei W, Schumacher M.** Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829-35. [PMID: 8182763]
265. **Royston P, Altman DG, Sauerbrei W.** Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127-41. [PMID: 16217841]
266. **Royston P, Sauerbrei W.** Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester: John Wiley; 2008.
267. **Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW.** Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke. 2011;42:1482-8. [PMID: 21474812]
268. **Lubetzky-Vilnai A, Ciol M, McCoy SW.** Statistical analysis of clinical prediction rules for rehabilitation interventions: current state of the literature. Arch Phys Med Rehabil. 2014;95:188-96. [PMID: 24036159]
269. **Steyerberg EW, Vergouwe Y.** Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925-31. [PMID: 24898551]
270. **Ioannidis JP.** Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8. [PMID: 18633328]
271. **Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG.** Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials. 2010;11:9. [PMID: 20113465]
272. **Hosmer DW, Lemeshow S.** *Applied Logistic Regression*. New York: Wiley; 2000.
273. **Vittinghoff E.** Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer; 2005.
274. **Hosmer DW, Lemeshow S, May S.** *Applied Survival Analysis: Regression Modelling of Time-To-Event Data*. Hoboken, NJ: Wiley-Interscience; 2008.
275. **Hastie T, Tibshirani R, Friedman JH.** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2001.

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

276. **Kuhn M, Johnson K.** *Applied Predictive Modelling*. New York: Springer; 2013.
277. **Andersen PK, Skovgaard LT.** *Regression With Linear Predictors*. New York: Springer; 2010.
278. **Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P.** Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335:136. [PMID: 17615182]
279. **Moreno L, Krishnan JA, Duran P, Ferrero F.** Development and validation of a clinical prediction rule to distinguish bacterial from viral pneumonia in children. Pediatr Pulmonol. 2006;41:331-7. [PMID: 16493666]
280. **Anderson KM, Odell PM, Wilson PW, Kannel WB.** Cardiovascular disease risk profiles. Am Heart J. 1991;121:293-8. [PMID: 1985385]
281. **Royston P, Parmar MK.** Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21:2175-97. [PMID: 12210632]
282. **Hans D, Durosier C, Kanis JA, Johansson H, Schott-Pethelaz AM, Krieg MA.** Assessment of the 10-year probability of osteoporotic hip fracture combining clinical risk factors and heel bone ultrasound: the EPISEM prospective cohort of 12,958 elderly women. J Bone Miner Res. 2008;23:1045-51. [PMID: 18302507]
283. **Bohensky MA, Jolley D, Pilcher DV, Sundararajan V, Evans S, Brand CA.** Prognostic models based on administrative data alone inadequately predict the survival outcomes for critically ill patients at 180 days post-hospital discharge. J Crit Care. 2012;27:422.e11-21. [PMID: 22591572]
284. **Barrett TW, Martin AR, Storrow AB, Jenkins CA, Harrell FE, Russ S, et al.** A clinical prediction model to estimate risk for 30-day adverse events in emergency department patients with symptomatic atrial fibrillation. Ann Emerg Med. 2011;57:1-12. [PMID: 20728962]
285. **Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in 't Veld AJ, Schalekamp MA, Habbema JD.** A clinical prediction rule for renal artery stenosis. Ann Intern Med. 1998;129:705-11. [PMID: 9841602]
286. **Smits M, Dippel DW, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, et al.** Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. Ann Intern Med. 2007;146:397-405. [PMID: 17371884]
287. **Moons KG, Donders AR, Steyerberg EW, Harrell FE.** Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol. 2004;57:1262-70. [PMID: 15617952]
288. **Mantel N.** Why stepdown procedures in variable selection? Technometrics. 1970;12:621-5.
289. **Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al.** External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56:826-32. [PMID: 14505766]
290. **Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD.** Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23:2567-86. [PMID: 15287085]
291. **van Houwelingen HC, Sauerbrei W.** Cross-validation, shrinkage and variable selection in linear regression revisited. Open J Statist. 2013;3:79-102.
292. **Sauerbrei W, Boulesteix AL, Binder H.** Stability investigations of multivariable regression models derived from low- and high-dimensional data. J Biopharm Stat. 2011;21:1206-31. [PMID: 22023687]
293. **Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA.** Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3:143-52. [PMID: 6463451]
294. **van Houwelingen JC, Le Cessie S.** Predictive value of statistical models. Stat Med. 1990;9:1303-25. [PMID: 2277880]
295. **Molinaro AM, Simon R, Pfeiffer RM.** Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21:3301-7. [PMID: 15905277]
296. **Chatfield C.** Model uncertainty, data mining and statistical inference. J R Stat Soc A. 1995;158:419-66.
297. **Sauerbrei W, Royston P, Binder H.** Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26:5512-28. [PMID: 18058845]
298. **Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC.** Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Meth. 2007;7:33. [PMID: 17629912]
299. **Castaldi PJ, Dahabreh IJ, Ioannidis JP.** An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011;12:189-202. [PMID: 21300697]
300. **Varma S, Simon R.** Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. [PMID: 16504092]
301. **Vach K, Sauerbrei W, Schumacher M.** Variable selection and shrinkage: comparison of some approaches. Stat Neerl. 2001;55:53-75.
302. **Lin IF, Chang WP, Liao YN.** Shrinkage methods enhanced the accuracy of parameter estimation using Cox models with small number of events. J Clin Epidemiol. 2013;66:743-51. [PMID: 23566374]
303. **Ambler G, Seaman S, Omar RZ.** An evaluation of penalised survival methods for developing prognostic models with rare events. Stat Med. 2012;31:1150-61. [PMID: 21997569]
304. **Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK.** Prognostic indices for older adults: a systematic review. JAMA. 2012;307:182-92. [PMID: 22235089]
305. **Spelt L, Andersson B, Nilsson J, Andersson R.** Prognostic models for outcome following liver resection for colorectal cancer metastases: a systematic review. Eur J Surg Oncol. 2012;38:16-24. [PMID: 22079259]
306. **Nam RK, Kattan MW, Chin JL, Trachtenberg J, Singal R, Rendon R, et al.** Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. J Clin Oncol. 2011;29:2959-64. [PMID: 21690464]
307. **Meffert PJ, Baumeister SE, Lerch MM, Mayerle J, Kratzer W, Völzke H.** Development, external validation, and comparative assessment of a new diagnostic score for hepatic steatosis. Am J Gastroenterol. 2014;109:1404-14. [PMID: 24957156]
308. **Collins GS, Altman DG.** Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). Br J Cancer. 2012;107:260-5. [PMID: 22699822]
309. **Royston P, Altman DG.** External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33. [PMID: 23496923]
310. **Iasonos A, Schrag D, Raj GV, Panageas KS.** How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26:1364-70. [PMID: 18323559]
311. **Zivanovic O, Jacks LM, Iasonos A, Leitao MM Jr, Soslow RA, Veras E, et al.** A nomogram to predict postresection 5-year overall survival for patients with uterine leiomyosarcoma. Cancer. 2012;118:660-9. [PMID: 21751199]
312. **Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, et al.** The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int. 2007;18:1033-46. [PMID: 17323110]
313. **Papaioannou A, Morin S, Cheung AM, Atkinson S, Brown JP, Feldman S, et al; Scientific Advisory Council of Osteoporosis Canada.** 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ. 2010;182:1864-73. [PMID: 20940232]
314. **Collins GS, Michaëlsson K.** Fracture risk assessment: state of the art, methodologically unsound, or poorly reported? Curr Osteoporos Rep. 2012;10:199-207. [PMID: 22688862]
315. **Collins GS, Mallett S, Altman DG.** Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. BMJ. 2011;342:d3651. [PMID: 21697214]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

316. **Järvinen TL, Jokihaara J, Guy P, Alonso-Coello P, Collins GS, Michaëlsson K, et al.** Conflicts at the heart of the FRAX tool. CMAJ. 2014;186:165-7. [PMID: 24366895]
317. **Balmaña J, Stockwell DH, Steyerberg EW, Stoffel EM, Deffenbaugh AM, Reid JE, et al.** Prediction of MLH1 and MSH2 mutations in Lynch syndrome. JAMA. 2006;296:1469-78. [PMID: 17003395]
318. **Bruins Slot MH, Rutten FH, van der Heijden GJ, Geersing GJ, Glatz JF, Hoes AW.** Diagnosing acute coronary syndrome in primary care: comparison of the physicians' risk estimation and a clinical decision rule. Fam Pract. 2011;28:323-8. [PMID: 21239470]
319. **Suarthana E, Vergouwe Y, Moons KG, de Monchy J, Grobbee D, Heederik D, et al.** A diagnostic model for the detection of sensitization to wheat allergens was developed and validated in bakery workers. J Clin Epidemiol. 2010;63:1011-9. [PMID: 20189762]
320. **Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ.** On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105-17. [PMID: 21484848]
321. **Akazawa K.** Measures of explained variation for a regression model used in survival analysis. J Med Syst. 1997;21:229-38. [PMID: 9442437]
322. **Choodari-Oskooei B, Royston P, Parmar MK.** A simulation study of predictive ability measures in a survival model I: explained variation measures. Stat Med. 2012;31:2627-43. [PMID: 21520455]
323. **Heller G.** A measure of explained risk in the proportional hazards model. Biostatistics. 2012;13:315-25. [PMID: 22190711]
324. **Korn EL, Simon R.** Measures of explained variation for survival data. Stat Med. 1990;9:487-503. [PMID: 2349402]
325. **Mittlböck M, Schemper M.** Explained variation for logistic regression. Stat Med. 1996;15:1987-97. [PMID: 8896134]
326. **Royston P.** Explained variation for survival models. Stata Journal. 2006;6:83-96
327. **Schemper M.** Predictive accuracy and explained variation. Stat Med. 2003;22:2299-308. [PMID: 12854094]
328. **Schemper M, Henderson R.** Predictive accuracy and explained variation in Cox regression. Biometrics. 2000;56:249-55. [PMID: 10783803]
329. **Schemper M, Stare J.** Explained variation in survival analysis. Stat Med. 1996;15:1999-2012. [PMID: 8896135]
330. **Gerds T, Schumacher M.** Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J. 2006;6:1029-40. [PMID: 17240660]
331. **Rufibach K.** Use of Brier score to assess binary predictions. J Clin Epidemiol. 2010;63:938-9. [PMID: 20189763]
332. **Gerds TA, Cai T, Schumacher M.** The performance of risk prediction models. Biom J. 2008;50:457-79. [PMID: 18663757]
333. **Royston P, Sauerbrei W.** A new measure of prognostic separation in survival data. Stat Med. 2004;23:723-48. [PMID: 14981672]
334. **DeLong ER, DeLong DM, Clarke-Pearson DL.** Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45. [PMID: 3203132]
335. **Demler OV, Pencina MJ, D'Agostino RB Sr.** Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31:2577-87. [PMID: 22415937]
336. **Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP.** Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. Anesthesiology. 2013;119:959-81. [PMID: 24195875]
337. **Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM.** Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care. 2014;52:751-65. [PMID: 25023919]
338. **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;59:1497-505. [PMID: 23818444]
339. **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;27:157-72. [PMID: 17569110]
340. **Cook NR.** Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928-35. [PMID: 17309939]
341. **Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al; American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council.** Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119:2408-16. [PMID: 19364974]
342. **Cook NR.** Assessing the incremental role of novel and emerging risk factors. Curr Cardiovasc Risk Rep. 2010;4:112-9. [PMID: 20640227]
343. **Vickers AJ, Cronin AM, Begg CB.** One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13. [PMID: 21276237]
344. **Cook NR, Ridker PM.** Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009;150:795-802. [PMID: 19487714]
345. **Cook NR, Paynter NP.** Performance of reclassification statistics in comparing risk prediction models. Biom J. 2011;53:237-58. [PMID: 21294152]
346. **Cook NR.** Clinically relevant measures of fit? A note of caution. Am J Epidemiol. 2012;176:488-91. [PMID: 22875759]
347. **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;176:473-81. [PMID: 22875755]
348. **Pencina MJ, D'Agostino RB, Vasan RS.** Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med. 2010;48:1703-11. [PMID: 20716010]
349. **Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW.** Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. Med Decis Making. 2013;33:490-501. [PMID: 23313931]
350. **Hilden J, Gerds TA.** A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33:3405-14. [PMID: 23553436]
351. **Pepe MS.** Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173:1327-35. [PMID: 21555714]
352. **Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJ, Uitterlinden AG, Witteman JC, et al.** Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol. 2010;172:353-61. [PMID: 20562194]
353. **Mühlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB.** Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol. 2013;28:25-33. [PMID: 23179629]
354. **Pepe M, Fang J, Feng Z, Gerds T, Hilden J.** *The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement with Miscalibrated or Overfit Models*. UW Biostatistics Working Paper Series. Working Paper 392. Madison, WI: University of Wisconsin; 2013.
355. **Vickers AJ, Pepe M.** Does the net reclassification improvement help us evaluate models and markers? Ann Intern Med. 2014;160:136-7. [PMID: 24592500]
356. **Hilden J.** Commentary: On NRI, IDI, and "good-looking" statistics with nothing underneath. Epidemiology. 2014;25:265-7. [PMID: 24487208]
357. **Leening MJ, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW.** Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med. 2014;160:122-31. [PMID: 24592497]
358. **Al-Radi OO, Harrell FE Jr, Caldarone CA, McCrindle BW, Jacobs JP, Williams MG, et al.** Case complexity scores in congenital heart surgery: a comparative study of the Aristotle Basic Complexity

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

score and the Risk Adjustment in Congenital Heart Surgery (RACHS-1) system. J Thorac Cardiovasc Surg. 2007;133:865-75. [PMID: 17382616]
359. **Localio AR, Goodman S.** Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012;157:294-5. [PMID: 22910942]
360. **Van Calster B, Vickers AJ.** Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2014 Aug 25. [Epub ahead of print]. [PMID: 25155798]
361. **Vickers AJ.** Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat. 2008;62: 314-20. [PMID: 19132141]
362. **Vickers AJ, Cronin AM, Kattan MW, Gonen M, Scardino PT, Milowsky MI, et al; International Bladder Cancer Nomogram Consortium.** Clinical benefits of a multivariate prediction model for bladder cancer: a decision analytic approach. Cancer. 2009;115:5460-9. [PMID: 19823979]
363. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. [PMID: 17099194]
364. **Baker SG.** Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42. [PMID: 19843888]
365. **Baker SG, Cook NR, Vickers A, Kramer BS.** Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009; 172:729-48. [PMID: 20069131]
366. **Baker SG, Kramer BS.** Evaluating a new marker for risk prediction: decision analysis to the rescue. Discov Med. 2012;14:181-8. [PMID: 23021372]
367. **Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM.** Quantifying the added value of a diagnostic test or marker. Clin Chem. 2012;58:1408-17. [PMID: 22952348]
368. **Held U, Bové DS, Steurer J, Held L.** Validating and updating a risk model for pneumonia–a case study. BMC Med Res Methodol. 2012;12:99. [PMID: 22817850]
369. **Cindolo L, Chiodini P, Gallo C, Ficarra V, Schips L, Tostain J, et al.** Validation by calibration of the UCLA integrated staging system prognostic model for nonmetastatic renal cell carcinoma after nephrectomy. Cancer. 2008;113:65-71. [PMID: 18473356]
370. **Baart AM, Atsma F, McSweeney EN, Moons KG, Vergouwe Y, de Kort WL.** External validation and updating of a Dutch prediction model for low hemoglobin deferral in Irish whole blood donors. Transfusion. 2014;54(3 Pt 2):762-9. [PMID: 23607909]
371. **Chalmers I, Glasziou P.** Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374:86-9. [PMID: 19525005]
372. **Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG.** A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth. 2009;56:194-201. [PMID: 19247740]
373. **van Houwelingen HC.** Validation. calibration, revision and combination of prognostic survival models. Stat Med. 2000;19:3401-15. [PMID: 11122504]
374. **Manola J, Royston P, Elson P, McCormack JB, Mazumdar M, Négrier S, et al; International Kidney Cancer Working Group.** Prognostic model for survival in patients with metastatic renal cell carcinoma: results from the International Kidney Cancer Working Group. Clin Cancer Res. 2011;17:5443-50. [PMID: 21828239]
375. **Krupp NL, Weinstein G, Chalian A, Berlin JA, Wolf P, Weber RS.** Validation of a transfusion prediction model in head and neck cancer surgery. Arch Otolaryngol Head Neck Surg. 2003;129:1297-302. [PMID: 14676155]
376. **Morra E, Cesana C, Klersy C, Barbarano L, Varettoni M, Cavanna L, et al.** Clinical characteristics and factors predicting evolution of asymptomatic IgM monoclonal gammopathies and IgM-related disorders. Leukemia. 2004;18:1512-7. [PMID: 15322559]
377. **Kelder JC, Cramer MJ, van Wijngaarden J, van Tooren R, Mosterd A, Moons KG, et al.** The diagnostic value of physical examination and additional testing in primary care patients with suspected heart failure. Circulation. 2011;124:2865-73. [PMID: 22104551]
378. **Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al.** A prognostic index in primary breast cancer. Br J Cancer. 1982;45:361-6. [PMID: 7073932]
379. **Tang EW, Wong CK, Herbison P.** Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome. Am Heart J. 2007;153:29-35. [PMID: 17174633]
380. **Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al.** Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med. 2009;151:775-83. [PMID: 19949143]
381. **Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al.** AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010;192:197-202. [PMID: 20170456]
382. **Starmans R, Muris JW, Fijten GH, Schouten HJ, Pop P, Knottnerus JA.** The diagnostic value of scoring models for organic and non-organic gastrointestinal disease, including the irritable-bowel syndrome. Med Decis Making. 1994;14:208-16. [PMID: 7934707]
383. **Tzoulaki I, Seretis A, Ntzani EE, Ioannidis JP.** Mapping the expanded often inappropriate use of the Framingham Risk Score in the medical literature. J Clin Epidemiol. 2014;67:571-7. [PMID: 24513280]
384. **Harrison DA, Rowan KM.** Outcome prediction in critical care: the ICNARC model. Curr Opin Crit Care. 2008;14:506-12. [PMID: 18787441]
385. **Kanaya AM, Wassel Fyr CL, de Rekeneire N, Schwartz AV, Goodpaster BH, Newman AB, et al.** Predicting the development of diabetes in older adults: the derivation and validation of a prediction rule. Diabetes Care. 2005;28:404-8. [PMID: 15677800]
386. **Stephens JW, Ambler G, Vallance P, Betterridge DJ, Humphries SE, Hurel SJ.** Cardiovascular risk and diabetes. Are the methods of risk prediction satisfactory? Eur J Cardiovasc Prev Rehabil. 2004;11: 521-8. [PMID: 15580065]
387. **Cogswell R, Kobashigawa E, McGlothlin D, Shaw R, De Marco T.** Validation of the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) pulmonary hypertension prediction model in a unique population and utility in the prediction of long-term survival. J Heart Lung Transplant. 2012;31:1165-70. [PMID: 23062726]
388. **Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, et al; GRACE Investigators.** A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. JAMA. 2004;291:2727-33. [PMID: 15187054]
389. **Geersing GJ, Erkens PM, Lucassen WA, Büller HR, Cate HT, Hoes AW, et al.** Safe exclusion of pulmonary embolism using the Wells rule and qualitative d-dimer testing in primary care: prospective cohort study. BMJ. 2012;345:e6564. [PMID: 23036917]
390. **Collins GS, Altman DG.** Identifying patients with undetected gastro-oesophageal cancer in primary care: external validation of QCancer® (Gastro-Oesophageal). Eur J Cancer. 2013;49:1040-8. [PMID: 23159533]
391. **de Vin T, Engels B, Gevaert T, Storme G, De Ridder M.** Stereotactic radiotherapy for oligometastatic cancer: a prognostic model for survival. Ann Oncol. 2014;25:467-71. [PMID: 24355488]
392. **Bernasconi P, Klersy C, Boni M, Cavigliano PM, Calatroni S, Giardini I, et al.** World Health Organization classification in combination with cytogenetic markers improves the prognostic stratification of patients with de novo primary myelodysplastic syndromes. Br J Haematol. 2007;137:193-205. [PMID: 17408458]
393. **Schemper M, Smith TL.** A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17:343-6. [PMID: 8889347]
394. **Echouffo-Tcheugui JB, Woodward M, Kengne AP.** Predicting a post-thrombolysis intracerebral hemorrhage: a systematic review. J Thromb Haemost. 2013;11:862-71. [PMID: 23469771]
395. **Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H, et al.** Prediction of pulmonary embolism in the emer-

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

gency department: the revised Geneva score. Ann Intern Med. 2006; 144:165-71. [PMID: 16461960]
396. **Davis JL, Worodria W, Kisembo H, Metcalfe JZ, Cattamanchi A, Kawooya M, et al.** Clinical and radiographic factors do not accurately diagnose smear-negative tuberculosis in HIV-infected inpatients in Uganda: a cross-sectional study. PLoS One. 2010;5:e9859. [PMID: 20361038]
397. **Ji R, Shen H, Pan Y, Wang P, Liu G, Wang Y, et al; China National Stroke Registry (CNSR) Investigators.** Risk score to predict gastrointestinal bleeding after acute ischemic stroke. BMC Gastroenterol. 2014;14:130. [PMID: 25059927]
398. **Marrugat J, Subirana I, Ramos R, Vila J, Marin-Ibanez A, Guembe MJ, et al; FRESCO Investigators.** Derivation and validation of a set of 10-year cardiovascular risk predictive functions in Spain: the FRESCO Study. Prev Med. 2014;61:66-74. [PMID: 24412897]
399. **Hensgens MP, Dekkers OM, Goorhuis A, LeCessie S, Kuijper EJ.** Predicting a complicated course of *Clostridium difficile* infection at the bedside. Clin Microbiol Infect. 2014;20:O301-8. [PMID: 24188103]
400. **Hak E, Wei F, Nordin J, Mullooly J, Poblete S, Nichol KL.** Development and validation of a clinical prediction rule for hospitalization due to pneumonia or influenza or death during influenza epidemics among community-dwelling elderly persons. J Infect Dis. 2004;189: 450-8. [PMID: 14745702]
401. **Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al; STROBE Initiative.** Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007;18:805-35. [PMID: 18049195]
402. **Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, et al.** Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373:739-45. [PMID: 19249635]
403. **Lang TA, Altman DG.** Basic statistical reporting for articles published in clinical medical journals: the SAMPL guidelines. In: Smart P, Maisonneuve H, Polderman A, eds. *Science Editors' Handbook*. European Association of Science Editors; 2013.
404. **Binder H, Sauerbrei W, Royston P.** Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med. 2013;32:2262-77. [PMID: 23034770]
405. **Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K.** A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. Crit Care Med. 2007;35: 1091-8. [PMID: 17334248]
406. **Brady AR, Harrison D, Black S, Jones S, Rowan K, Pearson G, et al.** Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the United Kingdom. Pediatrics. 2006;117:e733-42. [PMID: 16510615]
407. **Kuijpers T, van der Windt DA, van der Heijden GJ, Twisk JW, Vergouwe Y, Bouter LM.** A prediction rule for shoulder pain related sick leave: a prospective cohort study. BMC Musculoskelet Disord. 2006;7:97. [PMID: 17150087]
408. **Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP.** A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. BMJ. 2001;323:75-81. [PMID: 11451781]
409. **Casikar I, Lu C, Reid S, Condous G.** Prediction of successful expectant management of first trimester miscarriage: development and validation of a new mathematical model. Aust N Z J Obstet Gynaecol. 2013;53:58-63. [PMID: 23405997]
410. **Godoy G, Chong KT, Cronin A, Vickers A, Laudone V, Touijer K, et al.** Extent of pelvic lymph node dissection and the impact of standard template dissection on nomogram prediction of lymph node involvement. Eur Urol. 2011;60:195-201. [PMID: 21257258]
411. **Bradburn MJ, Clark TG, Love SB, Altman DG.** Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. Br J Cancer. 2003;89:431-6. [PMID: 12888808]
412. **Wells P, Anderson D, Rodger M, Ginsberg J, Kearon C, Gent M, et al.** Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED d-dimer. Thromb Haemost. 2000;83:416-20. [PMID: 10744147]
413. **Cole TJ.** Scaling and rounding regression-coefficients to integers. Appl Stat. 1993;42:261-8.
414. **Sullivan LM, Massaro JM, D'Agostino RB Sr.** Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004;23:1631-60. [PMID: 15122742]
415. **Moons KG, Harrell FE, Steyerberg EW.** Should scoring rules be based on odds ratios or regression coefficients? J Clin Epidemiol. 2002;55:1054-5. [PMID: 12464384]
416. **Nijman RG, Vergouwe Y, Thompson M, van Veen M, van Meurs AH, van der Lei J, et al.** Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. BMJ. 2013;346:f1706. [PMID: 23550046]
417. **Royston P, Altman DG.** Visualizing and assessing discrimination in the logistic regression model. Stat Med. 2010;29:2508-20. [PMID: 20641144]
418. **Taş U, Steyerberg EW, Bierma-Zeinstra SM, Hofman A, Koes BW, Verhagen AP.** Age, gender and disability predict future disability in older people: the Rotterdam Study. BMC Geriatrics. 2011;11: 22. [PMID: 21569279]
419. **Hanley JA, McNeil BJ.** A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839-43. [PMID: 6878708]
420. **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. [PMID: 21204120]
421. **Pepe MS, Janes H.** Reporting standards are needed for evaluations of risk reclassification. Int J Epidemiol. 2011;40:1106-8. [PMID: 21571811]
422. **Vickers AJ, Cronin AM.** Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Semin Oncol. 2010;37:31-8. [PMID: 20172362]
423. **Sanders MS, de Jonge RC, Terwee CB, Heymans MW, Koomen I, Ouburg S, et al.** Addition of host genetic variants in a prediction rule for post meningitis hearing loss in childhood: a model updating study. BMC Infect Dis. 2013;13:340. [PMID: 23879305]
424. **Kramer AA, Zimmerman JE.** A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Med Inform Decis Mak. 2010;10:27. [PMID: 20465830]
425. **Neely D, Feinglass J, Wallace WH.** Developing a predictive model to assess applicants to an internal medicine residency. J Grad Med Educ. 2010;2:129-32. [PMID: 21975899]
426. **Ioannidis JP.** Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol. 2007;60:324-9. [PMID: 17346604]
427. **Horton R.** The hidden research paper. JAMA. 2002;287:2775-8. [PMID: 12038909]
428. **Docherty M, Smith R.** The case for structuring the discussion of scientific papers. BMJ. 1999;318:1224-5. [PMID: 10231230]
429. **Ioannidis JP.** Research needs grants, funding and money–missing something? Eur J Clin Invest. 2012;42:349-51. [PMID: 22050119]
430. **Janssens AC, Ioannidis JP, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al.** Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. Eur J Clin Invest. 2011;41:1010-35. [PMID: 21434890]
431. **Collins GS.** Cardiovascular disease risk prediction in the UK. Primary Care Cardiovascular Journal. 2013;6:125-8.
432. **Collins GS, Altman DG.** An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. BMJ. 2009;339:b2584. [PMID: 19584409]
433. **Collins GS, Altman DG.** An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010;340:c2442. [PMID: 20466793]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

434. **Perry JJ, Sharma M, Sivilotti ML, Sutherland J, Symington C, Worster A, et al.** Prospective validation of the ABCD2 score for patients in the emergency department with transient ischemic attack. CMAJ. 2011;183:1137-45. [PMID: 21646462]
435. **Clarke M, Chalmers I.** Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? JAMA. 1998;280:280-2. [PMID: 9676682]
436. **Ioannidis JP, Polyzos NP, Trikalinos TA.** Selective discussion and transparency in microarray research findings for cancer outcomes. Eur J Cancer. 2007;43:1999-2010. [PMID: 17629475]
437. **Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ.** Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? Anesth Analg. 2005;100:1525-32. [PMID: 15845719]
438. **Kappen TH, Moons KG, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, van Klei WA.** Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. Anesthesiology. 2014;120:343-54. [PMID: 24105403]
439. **Poldervaart JM, Reitsma JB, Koffijberg H, Backus BE, Six AJ, Doevendands PA, et al.** The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. BMC Cardiovasc Disord. 2013;13:77. [PMID: 24070098]
440. **Hutchings HA, Evans BA, Fitzsimmons D, Harrison J, Heaven M, Huxley P, et al.** Predictive risk stratification model: a progressive cluster-randomised trial in chronic conditions management (PRISMATIC) research protocol. Trials. 2013;14:301. [PMID: 24330749]
441. **Ioannidis JP.** More than a billion people taking statins? Potential implications of the new cardiovascular guidelines. JAMA. 2014;311:463-4. [PMID: 24296612]
442. **Ioannidis JP, Tzoulaki I.** What makes a good predictor? The evidence applied to coronary artery calcium score. JAMA. 2010;303:1646-7. [PMID: 20424257]
443. **Mrdovic I, Savic L, Krljanac G, Asanin M, Perunicic J, Lasica R, et al.** Predicting 30-day major adverse cardiovascular events after primary percutaneous coronary intervention. The RISK-PCI score. Int J Cardiol. 2013;162:220-7. [PMID: 21663982]
444. **Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR.** C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation. 2008;118:2243-51. [PMID: 18997194]
445. World Medical Association. Declaration of Geneva. Accessed at www.wma.net/en/30publications/10policies/g1/ on 24 June 2008.
446. Council for International Organizations of Medical Sciences. International ethical guidelines for biomedical research involving human subjects. Bull Med Ethics. 2002;182:17-23. [PMID: 14983848]
447. **Arnold DH, Gebretsadik T, Abramo TJ, Sheller JR, Resha DJ, Hartert TV.** The Acute Asthma Severity Assessment Protocol (AASAP) study: objectives and methods of a study to develop an acute asthma clinical prediction rule. Emerg Med J. 2012;29:444-50. [PMID: 21586757]
448. **Azagra R, Roca G, Encabo G, Prieto D, Aguye A, Zwart M, et al.** Prediction of absolute risk of fragility fracture at 10 years in a Spanish population: validation of the WHO FRAX tool in Spain. BMC Musculoskelet Disord. 2011;12:30. [PMID: 21272372]
449. **Collins SP, Lindsell CJ, Jenkins CA, Harrell FE, Fermann GJ, Miller KF, et al.** Risk stratification in acute heart failure: rationale and design of the STRATIFY and DECIDE studies. Am Heart J. 2012;164:825-34. [PMID: 23194482]
450. **Hafkamp-de Groen E, Lingsma HF, Caudri D, Wijga A, Jaddoe VW, Steyerberg EW, et al.** Predicting asthma in preschool children with asthma symptoms: study rationale and design. BMC Pulm Med. 2012;12:65. [PMID: 23067313]
451. **Hess EP, Wells GA, Jaffe A, Stiell IG.** A study to derive a clinical decision rule for triage of emergency department patients with chest pain: design and methodology. BMC Emerg Med. 2008;8:3. [PMID: 18254973]
452. **Horisberger T, Harbarth S, Nadal D, Baenziger O, Fischer JE.** G-CSF and IL-8 for early diagnosis of sepsis in neonates and critically ill children–safety and cost effectiveness of a new laboratory prediction model: study protocol of a randomized controlled trial [ISRCTN91123847]. Crit Care. 2004;8:R443-50. [PMID: 15566590]
453. **Liman TG, Zietemann V, Wiedmann S, Jungehuelsing GJ, Endres M, Wollenweber FA, et al.** Prediction of vascular risk after stroke–protocol and pilot data of the Prospective Cohort with Incident Stroke (PROSCIS). Int J Stroke. 2013;8:484-90. [PMID: 22928669]
454. **Mann DM, Kannry JL, Edonyabo D, Li AC, Arciniega J, Stulman J, et al.** Rationale, design, and implementation protocol of an electronic health record integrated clinical prediction rule (iCPR) randomized trial in primary care. Implement Sci. 2011;6:109. [PMID: 21929769]
455. **Meijs MF, Bots ML, Vonken EJ, Cramer MJ, Melman PG, Velthuis BK, et al.** Rationale and design of the SMART Heart study: a prediction model for left ventricular hypertrophy in hypertension. Neth Heart J. 2007;15:295-8. [PMID: 18030317]
456. **Mrdovic I, Savic L, Perunicic J, Asanin M, Lasica R, Marinkovic J, et al.** Development and validation of a risk scoring model to predict net adverse cardiovascular outcomes after primary percutaneous coronary intervention in patients pretreated with 600 mg clopidogrel: rationale and design of the RISK-PCI study. J Interv Cardiol. 2009;22:320-8. [PMID: 19515084]
457. **Nee RJ, Vicenzino B, Jull GA, Cleland JA, Coppieters MW.** A novel protocol to develop a prediction model that identifies patients with nerve-related neck and arm pain who benefit from the early introduction of neural tissue management. Contemp Clin Trials. 2011;32:760-70. [PMID: 21718803]
458. **Pita-Fernández S, Pértega-Diaz S, Valdés-Cañedo F, Seijo-Bestilleiro R, Seoane-Pillado T, Fernández-Rivera C, et al.** Incidence of cardiovascular events after kidney transplantation and cardiovascular risk scores: study protocol. BMC Cardiovasc Disord. 2011;11:2. [PMID: 21639867]
459. **Sanfelix-Genoves J, Peiro S, Sanfelix-Gimeno G, Giner V, Gil V, Pascual M, et al.** Development and validation of a population-based prediction scale for osteoporotic fracture in the region of Valencia, Spain: the ESOSVAL-R study. BMC Public Health. 2010;10:153. [PMID: 20334639]
460. **Siebeling L, ter Riet G, van der Wal WM, Geskus RB, Zoller M, Muggensturm P, et al.** ICE COLD ERIC–International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts–study protocol for an international COPD cohort study. BMC Pulm Med. 2009;9:15. [PMID: 19419546]
461. **Canadian CT Head and C-Spine (CCC) Study Group.** Canadian C-Spine Rule study for alert and stable trauma patients: I. Background and rationale. CJEM. 2002;4:84-90. [PMID: 17612425]
462. **Canadian CT Head and C-Spine (CCC) Study Group.** Canadian C-Spine Rule study for alert and stable trauma patients: II. Study objectives and methodology. CMAJ. 2002;4:185-93. [PMID: 17609004]
463. **van Wonderen KE, van der Mark LB, Mohrs J, Geskus RB, van der Wal WM, van Aalderen WM, et al.** Prediction and treatment of asthma in preschool children at risk: study design and baseline data of a prospective cohort study in general practice (ARCADE). BMC Pulm Med. 2009;9:13. [PMID: 19368704]
464. **Waldron CA, Gallacher J, van der Weijden T, Newcombe R, Elwyn G.** The effect of different cardiovascular risk presentation formats on intentions, understanding and emotional affect: a randomised controlled trial using a web-based risk formatter (protocol). BMC Med Inform Decis Mak. 2010;10:41. [PMID: 20673347]
465. **Laine C, Guallar E, Mulrow C, Taichman DB, Cornell JE, Cotton D, et al.** Closing in on the truth about recombinant human bone morphogenetic protein-2: evidence synthesis, data sharing, peer review, and reproducible research. Ann Intern Med. 2013;158:916-8. [PMID: 23778911]
466. **Peng RD.** Reproducible research and *Biostatistics*. Biostatistics. 2009;10:405-8. [PMID: 19535325]
467. **Keiding N.** Reproducible research and the substantive context. Biostatistics. 2010;11:376-8. [PMID: 20498225]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

468. **Vickers AJ.** Whose data set is it anyway? Sharing raw data from randomized trials. Trials. 2006;7:15. [PMID: 16704733]
469. **Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, et al.** Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer. 2003;88:1191-8. [PMID: 12698183]
470. **Riley RD, Sauerbrei W, Altman DG.** Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. Br J Cancer. 2009;100:1219-29. [PMID: 19367280]
471. **Riley RD, Simmonds MC, Look MP.** Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60:431-9. [PMID: 17419953]
472. **Hemingway H, Riley RD, Altman DG.** Ten steps towards improving prognosis research. BMJ. 2009;339:b4184. [PMID: 20042483]
473. **Groves T.** BMJ policy on data sharing. BMJ. 2010;340:c564. [PMID: 20110310]
474. **Marchionni L, Afsari B, Geman D, Leek JT.** A simple and reproducible breast cancer prognostic test. BMC Genomics. 2013;14:336. [PMID: 23682826]
475. **Loder E, Groves T, Macauley D.** Registration of observational studies. BMJ. 2010;340:c950. [PMID: 20167643]
476. **Chavers S, Fife D, Wacholtz M, Stang P, Berlin J.** Registration of Observational Studies: perspectives from an industry-based epidemiology group. Pharmacoepidemiol Drug Saf. 2011;20:1009-13. [PMID: 21953845]
477. Should protocols for observational studies be registered? Lancet. 2010;375:348. [PMID: 20113809]
478. **Altman DG.** The time has come to register diagnostic and prognostic research. Clin Chem. 2014;60:580-2. [PMID: 24520099]
479. The registration of observational studies–when metaphors go bad. Epidemiology. 2010;21:607-9. [PMID: 20657291]
480. **Sørensen HT, Rothman KJ.** The prognosis of research. BMJ. 2010;340:c703. [PMID: 20164129]
481. **Vandenbroucke JP.** Registering observational research: second thoughts. Lancet. 2010;375:982-3. [PMID: 20304239]
482. **Williams RJ, Tse T, Harlan WR, Zarin DA.** Registration of observational studies: Is it time? CMAJ. 2010;182:1638-42. [PMID: 20643833]
483. **Lenzer J.** Majority of panelists on controversial new cholesterol guideline have current or recent ties to drug manufacturers. BMJ. 2013;347:f6989. [PMID: 24264770]
484. **Lenzer J, Hoffman JR, Furberg CD, Ioannidis JP ; Guideline Panel Review Working Group.** Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. BMJ. 2013;347:f5535. [PMID: 24046286]
485. **Simera I.** Get the content right: following reporting guidelines will make your research paper more complete, transparent and usable. J Pak Med Assoc. 2013;63:283-5. [PMID: 23894916]
486. **Simera I, Kirtley S, Altman DG.** Reporting clinical research: guidance to encourage accurate and transparent research reporting. Maturitas. 2012;72:84-7. [PMID: 22440533]
487. **Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG.** Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. 2010;8:24. [PMID: 20420659]
488. **Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group.** Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151:264-9. [PMID: 19622511]
489. **Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al; STrengthening the REporting of Genetic Association Studies.** STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med. 2009; 6:e22. [PMID: 19192942]
490. **Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG; NC3Rs Reporting Guidelines Working Group.** Animal research: reporting in vivo experiments: the ARRIVE guidelines. J Gene Med. 2010;12:561-3. [PMID: 20607692]
491. **Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D; CARE Group.** The CARE guidelines: consensus-based clinical case reporting guideline development. J Med Case Rep. 2013;7:223. [PMID: 24228906]
492. **Marshall A, Altman DG, Royston P, Holder RL.** Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7. [PMID: 20085642]
493. **Little RJ, Rubin DB.** *Statistical Analysis With Missing Data*. Hoboken, NJ: Wiley; 2002.
494. **Rubin DB.** *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons; 1987.
495. **White IR, Royston P, Wood AM.** Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30:377-99. [PMID: 21225900]
496. **Harel O, Pellowski J, Kalichman S.** Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. AIDS Behav. 2012;16:1382-93. [PMID: 22223301]
497. **Schafer JL.** Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3-15. [PMID: 10347857]
498. **Marshall A, Altman DG, Holder RL, Royston P.** Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57. [PMID: 19638200]
499. **van Buuren S, Boshuizen HC, Knook DL.** Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681-94. [PMID: 10204197]
500. **Wood AM, White IR, Royston P.** How should variable selection be performed with multiply imputed data? Stat Med. 2008;27:3227-46. [PMID: 18203127]
501. **Turner EL, Dobson JE, Pocock SJ.** Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. Epidemiol Perspect Innov. 2010;7:9.
502. **van Walraven C, Hart RG.** Leave 'em alone–why continuous variables should be analyzed as such. Neuroepidemiology. 2008;30: 138-9. [PMID: 18421216]
503. **Vickers AJ, Lilja H.** Cutpoints in clinical chemistry: time for fundamental reassessment. Clin Chem. 2009;55:15-7. [PMID: 19028819]
504. **Bennette C, Vickers A.** Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21. [PMID: 22375553]
505. **Dawson NV, Weiss R.** Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Making. 2012;32: 225-6. [PMID: 22457338]
506. **Royston P, Altman DG.** Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43:429-67.
507. **Harrell FE Jr, Lee KL, Pollock BG.** Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80:1198-202. [PMID: 3047407]
508. **Schumacher M, Binder H, Gerds T.** Assessment of survival prediction models based on microarray data. Bioinformatics. 2007;23: 1768-74. [PMID: 17485430]
509. **Subramanian J, Simon R.** Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102:464-74. [PMID: 20233996]
510. **Dupuy A, Simon RM.** Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99:147-57. [PMID: 17227998]
511. **Boulesteix AL.** Validation in bioinformatics and molecular medicine. Brief Bioinform. 2011;12:187-8. [PMID: 21546447]
512. **Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL.** Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26:1990-8. [PMID: 20581402]
513. **Vickers AJ, Cronin AM.** Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology. 2010;76:1298-301. [PMID: 21030068]

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019

514. **Austin PC, Steyerberg EW.** Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33:517-35. [PMID: 24002997]
515. **Crowson CS, Atkinson EJ, Therneau TM.** Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2014 Apr 7. [Epub ahead of print]. [PMID: 23907781]
516. **Vach W.** Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66:1296-301. [PMID: 24021610]
517. **Miller ME, Hui SL, Tierney WM.** Validation techniques for logistic-regression models. Stat Med. 1991;10:1213-26. [PMID: 1925153]
518. **Cox DR.** Two further applications of a model for binary regression. Biometrika. 1958;45:562-5.
519. **D'Agostino RB, Nam BH.** Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. *Handbook of Statistics, Survival Methods*. Amsterdam: Elsevier; 2004:1-25.
520. **Grønnesby JK, Borgan O.** A method for checking regression models in survival analysis based on the risk score. Lifetime Data Anal. 1996;2:315-28. [PMID: 9384628]
521. **Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G.** One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat. 2000;5:251-3. [PMID: 11055275]
522. **Kramer AA, Zimmerman JE.** Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med. 2007;35:2052-6. [PMID: 17568333]
523. **Marcin JP, Romano PS.** Size matters to a model's fit. Crit Care Med. 2007;35:2212-3. [PMID: 17713369]
524. **Bannister CA, Poole CD, Jenkins-Jones S, Morgan CL, Elwyn G, Spasic I, et al.** External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations. Diab Care. 2014;37:537-45. [PMID: 24089541]
525. **Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B.** Assessing calibration of multinomial risk prediction models. Stat Med. 2014;33:2585-96. [PMID: 24549725]
526. **Cook NR.** Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54:17-23. [PMID: 18024533]
527. **Pencina MJ, D'Agostino RB Sr, Song L.** Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543-53. [PMID: 22344892]
528. **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;31:2610-26. [PMID: 22733650]
529. **Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA.** Concordance for prognostic models with competing risks. Biostatistics. 2014;15:526-39. [PMID: 24493091]
530. **Pencina MJ, D'Agostino RB Sr, Demler OV.** Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med. 2012;31:101-13. [PMID: 22147389]
531. **Bradburn MJ, Clark TG, Love SB, Altman DG.** Survival analysis part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. Br J Cancer. 2003;89:605-11. [PMID: 12915864]
532. **Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al.** Critical appraisal and data extraction for the systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744. [PMID: 25314315]

Appendix: Members of the TRIPOD Group

Gary Collins (University of Oxford, Oxford, United Kingdom); Douglas Altman (University of Oxford, Oxford, United Kingdom); Karel Moons (University Medical Center Utrecht, Utrecht, the Netherlands); Johannes Reitsma (University Medical Center Utrecht, Utrecht, the Netherlands); Virginia Barbour (*PLoS Medicine*, United Kingdom and Australia); Nancy Cook (Division of Preventive Medicine, Brigham & Women's Hospital, Boston, Massachusetts); Joris de Groot (University Medical Center Utrecht, Utrecht, the Netherlands); Trish Groves (*BMJ*, London, United Kingdom); Frank Harrell, Jr. (Vanderbilt University, Nashville, Tennessee); Harry Hemingway (University College London, London, United Kingdom); John Ioannidis (Stanford University, Stanford, California); Michael W. Kattan (Cleveland Clinic, Cleveland, Ohio); André Knottnerus (Maastricht University, Maastricht, the Netherlands, and *Journal of Clinical Epidemiology*); Petra Macaskill (University of Sydney, Sydney, Australia); Susan Mallett (University of Oxford, Oxford, United Kingdom); Cynthia Mulrow (*Annals of Internal Medicine*, American College of Physicians, Philadelphia, Pennsylvania); David Ransohoff (University of North Carolina at Chapel Hill, Chapel Hill, North Carolina); Richard Riley (University of Birmingham, Birmingham, United Kingdom); Peter Rothwell (University of Oxford, Oxford, United Kingdom); Patrick Royston (Medical Research Council Clinical Trials Unit at University College London, London, United Kingdom); Willi Sauerbrei (University of Freiburg, Freiburg, Germany); Ewout Steyerberg (University Medical Center Rotterdam, Rotterdam, the Netherlands); Ian Stiell (University of Ottawa, Ottawa, Ontario, Canada); Andrew Vickers (Memorial Sloan Kettering Cancer Center, New York).

Downloaded from https://annals.org by Phoenix Childrens Hospital user on 06/06/2019