Chapter 10
Reporting Standards and Critical Appraisal of Prediction Models

Leonard Wee, Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, Mattea Welch, and Andre Dekker

10.1 Introduction

In the practice of modern medicine, it is often useful to be able to look into the future. Here are two illustrative situations that readers of this book chapter may already be familiar with:

(i) When meeting a patient in the consultation room, a physician may wish to foretell, *given the presence of a certain combination of risk factors, what is the likely long-term outcome (i.e. prognosis) of this particular disease?*
(ii) When faced with a choice of multiple feasible interventions to offer, a physician may wish to forecast, *given the particular characteristics of this patient and the specifics of their condition, what is the specific benefit that ought to be expected from each treatment option?*

L. Wee (✉) · A. Traverso · A. Dekker
School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands
e-mail: leonard.wee@maastro.nl

S. M. J. van Kuijk
Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Maastricht University Medical Center, Maastricht, The Netherlands

F. J. W. M. Dankers
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

M. Welch
Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_10

We take as given that quantitative clinical prediction models do already, and will continue to, play an important clinical role. In the first example, one attempts to offer a prognosis, which is dependent on the etiology and evolution of the disease, but has nothing to say about what an optimal treatment might be. In the second example, a model is used to project from the present time to a probable future outcome of treatment(s), and is useful for selecting an optimal treatment from a set of competing alternatives. For the purpose of reporting standards and critical appraisal, we shall not need to distinguish between predictions of prognosis (the former) and predictions of treatment outcome (the latter), since the subsequent discussions applies equally to both.

Transparent reporting is a necessary condition for taking prediction models from early development into widespread clinical use. The process involves progressive phases [1] from:

(i) **development**; where you intend to inform others about the creation of your model,
(ii) **validation**; where you demonstrate how your model performs in increasingly more generalizable conditions,
(iii) **updating/improving**; where you add new parameters and/or larger sample sizes to your model in an attempt to improve its accuracy and generalizability,
(iv) **assessment**; where you monitor the effect of the model on clinical workflows and assess health economic impacts within a controlled environment, and lastly,
(v) **implementation**; where you would deploy the model into widespread use and observe its long term effects in routine clinical practice.

Critical appraisal is the systematic and objective analysis of descriptions in a piece of published scientific research in order to determine: (i) the methodological soundness of the steps taken in the study to address its stated objectives, (ii) assumptions and decisions made during the conduct of the study that may have introduced bias into the results, and (iii) the relevance and applicability of this study to the research question in the mind of the reader. The central purpose of the appraisal is therefore to evaluate the likelihood that a model will be just as accurate and as precise in other studies (e.g. different patient cohorts, different investigators, different clinical settings) as it was proved within its own study. This requirement for model generalizability is known as **external validity**. This is a perspective distinct from **internal validity**, where a study is shown to be logically self-consistent and methodologically robust only within its own setting, using the guiding principles given in the previous chapters in Part 2.

Good quality of reporting about prediction models is essential at every step in translation to clinic, to adequately understand the potential risks of bias and potential generalizability of a model. Biased reporting could result in promising models not being brought rapidly into clinical practice, or worse, inappropriate models are used in clinical decision-making such that they cause harm to patients. Both ultimately lead to wasted resources in healthcare because physicians and patients are either deprived of a useful clinical tool or sub-optimal clinical decisions are made

due using a non-valid model. A more common problem that has now come to light is inadequate reporting [2], where there is insufficient documentation to reproduce the model and/or understand the limits of its validity.

10.1.1 Chapter Overview

The previous chapters in this book have primarily focused on internal validity of prediction models. Here, we shall switch our focus towards understanding external validity and consider the general process of critically appraising a published model. In the restricted scope of this chapter, we shall give attention to critical appraisal in development and validation studies. Issues pertaining to model update, impact assessment and clinical implementation are only briefly touched upon.

The content is organized as follows. We begin with a brief recapitulation of the methodological aspects of model development and model validation, emphasizing specific aspects that will be important for critical appraisal. We then introduce the ***TRIPOD*** (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) checklist [3, 4] for reporting and discuss the significance of its major elements in regards to reproducibility and validity. Our perspective next shifts towards critically appraising reports of predictive models that have been published in literature. There are common misunderstandings that TRIPOD can be either a checklist for designing a prediction modelling study or a checklist for critical appraisal, or both – it is in fact neither. We thus introduce the ***CHARMS*** (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) checklist [5], that was designed for critical appraisal and information extraction in evidence synthesis from multiple published studies.

Also given the restricted scope of this chapter, we will enter into a brief overview of systematic reviews of prediction modelling studies, however the specifics of quantitative meta-analysis of multiple models will be outside the current scope. References to methodological developments in this area and some guidelines on the topic will be provided.

10.2 Prediction Modelling Studies

Prediction modelling studies can be loosely categorized into development, validation, update, impact assessment and implementation studies. The quantity and robustness of clinically-derived evidence needed to support the model increases in roughly the same order. For reporting requirements and critical appraisal, we devote our attention on the first two – development and validation.

During model development, the primary focus is selecting from a measured set of characteristics (variously referred to as predictors, covariates, factors, features,

markers, etc.) and then combining them within a statistical framework in such a way as to yield dependable forecasts when new (hitherto unobserved) observations are given.

In contrast, model validation (with or without model update) refers to testing an already-developed model by exposing it to a diverse range of new inputs where the ground truth is already known, ideally as independently as possible using cohorts and clinical settings that are different from the one used to develop the initial model.

10.2.1 Development

We briefly recapitulate concepts that were discussed in previous chapters. In the main, our discussion is about *multivariate* predictive models, such that two or more predictors have a correlative mathematical relationship with some expected outcome (of a diagnosis, or a prognosis or from a treatment intervention).

Other methodological studies have already pointed out the importance of defining in a study protocol, as far in advance as possible, key aspects of the prediction modelling study such as its objectives, study design, patient population, clinically relevant outcomes, selected predictors, sample size considerations, and the intended statistical methods to be used [6–10]. As with any other kind of clinical study, internal review and iterative refinement of the protocol is highly desirable, since poor *ad hoc* decisions made during model development may often lead to biased results.

Principal among the potential biases in multivariate prediction modelling is the phenomenon of "overfitting" (also known as over-training) of a model such that an excessive number of predictors have been fitted to random fluctuations in the development cohort rather than to the true underlying signal. This caveat is of particular significance in an era of high throughput semi-automated measurements that extract very large numbers of potentially explanatory predictors (e.g., genomics, proteomics, metabolomics, radiomics, etc.) from a single source (e.g., blood, biopsy sample or radiological images). overfitting will become apparent when the predictive performance of the model in the development cohort is found to be generally over-optimistic when tested in fully independent cohorts; this often deals a fatal blow to the overall generalizability and widespread clinical utility to said model.

The risk of overfitting is exacerbated when multiple candidate predictors are combined with automatic predictor-selection algorithms that seek to optimize predictive performance within the development cohort [11]. This leads to rapid inflation of the false positive association risk, thus also leading to poor generalizability of models.

There are some sound strategies to mitigate risk of overfitting. Among these,*internal cross-validation* is widely practiced; the development cohort is divided into a (relatively larger) sub-cohort for fitting the model and a (relatively smaller) sub-cohort for testing the performance of the model. To avoid vagaries of sub-sampling, "*k*-folds" can be used where the development cohort is split even further into *k* equally-sized factions, then each of the *k* factions may be used one

after another as the internal validation cohort for a model developed on the remaining (*k*-1) factions. *Repeated cross-validations* may also be used simultaneously within *k*-fold cross-validation, such that investigators apply multiple random assignments of patients into the two initial sub-cohorts.

Dimensionality reduction is a powerful *a priori* method for reducing the risk of overfitting and increasing generalizability. If some predictors are known (by earlier experiments) to be highly irreproducible due to some unsolved instability in the measurement, or if the measured value differs greatly from one observer to another, it may turn out to be preferable to exclude these predictors from statistical analysis. This method sacrifices some potential explanatory power in favour of better reproducibility and wider generalizability of the finished model. Note however, that *a priori* dimensionality reduction should not utilize the intended primary outcome as the basis of eliminating predictors, otherwise there will be an attendant risk of contaminating the selected predictors with some implicit information correlated to the desired outcome.

A further possibility to reduce overfitting is toincrease the sample size, i.e. number of individual cases in the development cohort. An oft-quoted rule of thumb is "at least 10 events per predictor". That is, there should be an order of magnitude relationship between sample size and the number of pre-selected predictors. Note that adherence to the rule of thumb does not imply guaranteed protection against overfitting, merely that the risks of over-training one's model is somewhat reduced.

Increasing sample size or widening the patient enrolment is not always feasible. In retrospective modelling studies, it may be possible to return to the original repository of data and "mine" for additional cases. Likewise, in prospective studies, there may be sufficient resources to run case enrolment over a longer time interval or to expand recruitment. However, one generally encounters some sort of practical, logistic, regulatory or political barrier that limit the possibilities on increasing the sample size. With indiscriminate loosening of the inclusion criteria, there is an inherent danger of injecting excessive clinical heterogeneity into the development sample, for which there is no way to account for these variations using the existing predictors.

10.2.2 Validation

During model development, especially when using automated predictor selection algorithms, it is usually unavoidable that predictive performance of the model will be assessed on the same data that was used to construct the model. Interim assessments of performance in multivariate prediction models should at least test for calibration [4, 12] and discrimination. An appropriate discrimination metric would be the area under a receiver-operator curve in the case of binary outcomes, and the hazard ratio in the case of time-to-event predictions. However, this will not be sufficient to detect biases in the model; such interim evaluations will always be much too optimistic in regards to predictive performance.

The primary function of validation is to determine the limits of generalizability and transportability of the model. Therefore, after finalizing a model that is well-calibrated and properly fitted to the development cohort data, it is necessary to evaluate this model in other data that has hitherto never been "seen" before, i.e. an independent validation cohort. The observed characteristics for every instance in the validation cohort must be put into the model and its predictions shall be compared with the actual outcome. The validation cohort may differ from the development cohort in the following ways:

(i) **Time-shifted**; the validation cases may be collected by the same investigators as those that constructed the model, but the new cases were collected from a different time period;
(ii) **Institution-shifted**; the validation cases are assembled by a different team of investigators operating in a different hospital/institution, but usually retaining the same definitions of the input predictive factors.
(iii) **Setting-shifted**; the validation cases are collected in a different clinical practice setting on individuals with the same condition, but the definitions of the input predictive factors may be slightly different or slightly broader.
(iv) **Population-shifted**; thevalidation cases are from individuals presenting in an intentionally different medical context (e.g., different kind of index disease, or applying a model developed on adults to a paediatric population).

Each of these shifts progressively tests the validity of the model in increasingly generalized situations. A reason why model performance depends on time span, settings and populations can be traced to the *spectrum effect*; since most external validation cohorts involve relatively small samples, it would be unlikely that the distribution of predictor values would match in both cohorts. The results in validation thus appear "compressed" towards one or the other extreme of predicted outcomes.

As in model development, a validation study should also describe predictive performance in terms of calibration *on the instances in the validation cohort* and either discrimination (in the case of binary outcomes) or hazard ratios (in the case to time-to-event).

10.2.3 Updates

Following validation, a model might be shown to be transferable to a new situation, but this is generally not the case in the early history of model evolution. Updating a model (for example, adjusting the predictor coefficients) and/or retraining the model on new data can be validly performed to improve overall performance and increase generalizability. The caveat, however, is not to re-estimate the coefficients nor to re-calibrate the model using solely the validation data. In effect, this neglects the predictive potential contained in the development data.

Since validation cohorts typically contain fewer cases than development cohorts, doing so would risk rendering the updated model less generalizable and more susceptible to overfitting.

A model can be updated by shifting the baseline risk, rescaling the regression coefficients of the existing predictors, re-fitting the coefficients using added data or selecting a different set of predictors. Combinations of the above may also be applied. A suggested approach would be to first analyse the underlying statistical and clinical heterogeneities in the two data sets. Only if clinically meaningful, it would be advisable to combine individual records in both cohorts and re-develop a new model, either with or without fresh predictor selection. A new cohort would thus be required for independent validation.

10.2.4 Impact Assessment and Clinical Implementation

An assumption that needs to be challenged is that access to predictive models will lead to improved clinical care. The basis of the assumption is that predictive models could support medical decision-making and hence improve patient outcomes. This can only be properly tested in impact and implementation studies. Such studies could, among other possible endpoints, compare physicians' behavior, patient-centred outcomes and overall cost-effectiveness of care when using the predictive model versus without using such a model. This is only a reasonable prospect for models that have multiple validated and/or updated for better generalizability.

While the preferred study design may be individually randomized controlled clinical trials of long-term patient outcomes, there is indeed place for short-term process evaluation studies and cluster-randomized trials assessing health economic impact and behavioural changes amongst physicians. Randomization of individuals can sometimes be problematic due to contamination between groups; physicians having to alternate between using or not using the model may still retain some memory of the model outcomes from previous patients. If the study considers behavioral changes on the patients' side, as may be the case in model implementation studies in shared decision-making, one must be aware that patients are likely to exchange information about the model results with each other.

10.3 Reporting Your Own Work

It is assumed that the majority of readers will be interested in developing and independently validating models pertinent to their area of expertise. Quality reporting of any work in development and validation has the twofold objective of: (i) informing

others in your area of expertise about what models did (or did not) perform adequately under specifically constrained circumstances and, (ii) assists other investigators who may be attempting to reproduce and/or validate your prior work. Unbiased reporting of all work helps avoid wasteful duplication of efforts and accelerates the evolution of a model towards widespread utilization.

10.3.1 Purpose of Transparent Reporting Guidelines

The TRIPOD statement [3] (and its related explanation and elaboration document [4]) was developed as a consensusguideline for what a majority of investigators would consider essential for reporting of multivariate prediction modelling research. The statement contains 22 essential items, which are then summarized in a checklist that can be easily downloaded for use [13]. TRIPOD specifically focuses on studies involving development, validation or a mixture of both (with or without model updating). While most items are relevant to studies of both developmental and validation nature, a few items on the checklist are marked as only relevant to one or the other.

It is not productive here to examine each item in TRIPOD one by one. What we will focus on are the major themes that emerge from multiple items taken together, relating to methodological integrity and wider validity of your work.

10.3.2 Context

As in all other publication concerning clinical research, a clear explanation of context is required such that the reader fully understands what kind of patients, diseases, diagnoses or interventions and outcomes that the work will address. A summary of patient characteristics, eligibility, selection/inclusion method and any exclusion criteria is important to clarify the "case-mix" within which the model was developed/validated. A flow diagram detailing how many patients were lost and carried over to the next step of the process is essential, rather than a solitary number stating sample size. This can help to clarify if there had been any patient selection or systematic exclusion biases that might restrict the potential applicability of the model to other situations. Pertaining to potential case mixture mismatch during validation, it is also essential to discuss and compare (for example, with a suitable hypothesis test of group difference) the characteristics of the development and validation cohorts.

Study design is a further essential component of the context. It needs to be stated as clearly and as early as possible what is the ultimate clinical objective/outcome to be modelled (if building a model) and/or which specific model is being validated. If an update to an existing model is to be attempted, it should be stated whether the intention of the study was to attempt a model update, or whether

there had been a *post hoc* decision to introduce new data into the model. TRIPOD gives a classification system from Type 1 up to Type 4, akin of levels of evidence of external validity, based on whether all of the cohort data was used to construct the model, if there was in-cohort splitting or if an entirely separate data set was used to evaluate the model.

10.3.3 Sample Size, Predictors and Predictor Selection

Unlike conventional clinical trials with controls, there are no simple tools to calculate the required sample size for a multivariate prediction modelling study. In general, the number of predictors in the model has not been determined prior to conducting the statistical analysis for model building. In validation, the number of predictors in the existing model is known. In both cases, it will be necessary to justify whether the sample is sufficiently large in terms of the absolute number of target events. As a rough guide, it would prove difficult to defend or validate the performance of a predictive model if there are fewer than 10 target events in total in the subject cohort. An aforementioned "rule of thumb" – at least 10 events per predictor – will be a useful guide as to whether it is possible to develop/validate a model on a given cohort.

Therefore, it is essential to document the final number of target events available (after exclusion of unsuitable cases) and the number of predictors used. The source of the data should be clearly identified, be it retrospective data interrogation, prospective case enrolment or extraction from a disease registry. The source of patient data and the final sample size should be justified in regards to the objective of the study and intended application of the finished model.

In model development, there should be a very clear statement of the number of predictors before and after any kind of automated predictor selection algorithm has been applied. In regards to potential overfitting, the number of predictors available before predictor selection is a better surrogate for risk of overfitting, since a model optimization algorithm will generally expose this number of predictors to the target outcome. Whenever used, the predictor selection algorithm and model optimization process should be clearly documented in the methods section. At the end, the selected predictors should be unambiguously defined, including how and when the predictor was measured.

If performing model validation, it is also essential to document the manner in which the existing predictors have been measured. Major deviations from the prescribed predictor measurement method(s) must be clearly stated in the validation report. The method of calculating the predicted value must be reported. Furthermore, it is important to document whether or not the assessors of the actual outcome were blinded with respect to the calculated prediction. If assessors of outcome are aware of the individual prediction result, one should acknowledge that there is some risk of confirmation bias such that assessors may (without consciously intending to) bias their assessment towards (or against) the prediction.

10.3.4 Missing Data

Missing data (including unobserved predictors in a validation cohort) occupies a single item in the TRIPOD checklist, yet it may have a disproportionately strong impact on the outcome of a study. It is often the case that potentially useful predictors may contain some null values, either because information on some individuals was lost during data collection, was not measured or simply not disclosed in the source documents. In a validation cohort, it is possible that a required predictor has not been measured at all, or has been measured in an irreconcilable manner to the original work (for example, incompatible toxicity grading systems).

Previous chapters discussed in detail how data elements that are systematically missing can have a strong biasing effect on a model, therefore one must report how missing values (predictors) were managed, including any kind of data imputation method (if used). This applies equally to reports on model development and model validation.

10.3.5 Model Specification and Predictive Performance

The major portion of TRIPOD is concerned with reporting the performance of a prediction model or after update to an existing model. The model itself needs to be fully specified in terms of the type of statistical model used (e.g., Cox Proportional Hazards), the regression coefficients for all of the final predictors (also confidence intervals for each predictor) and an event rate at a fixed time point for each subgroup of individuals. If risk groups (stratification into different discrete categories based on result or time to event) are created, then it must also be clearly specified how the stratification was done.

Assuming the abovementioned details are easily located in your report, the readers will wish to know how well your model performed at its assigned task. Metrics will be required to demonstrate how well calibrated a model is, and how well it serves to discriminate between different outcomes. A calibration plot is the preferred format for the former, where predicted versus actual probabilities of outcomes are graphed against each other. There will be some choice in regards to a discrimination metric, where area under a receiver-operator curve is commonly reported for binary outcome classifications and hazard ratios derived from a Cox model is widely used for time to event models. The TRIPOD supplementary document also cites other options for quantifying the discriminating power of a model.

10.3.6 Model Presentation, Ease of Interpretation and Intended Impact

Lastly, the developer of a prediction model should clearly explain how and when it is intended to be used. Complex models with several predictors are often unwieldy to use without the aid of a computer. For instance, if amodel is meant to be used

on hospital ward rounds, then it needs to be presented in a form where it can be easily and unambiguously interpreted without the use of a computing device. Examples of suitable formats of model include nomogram charts and risk-score charts. If graphs or response curves are to be used as part of the model, discrete points on the curve should be made easily readable as a side table, since approximately interpolation from tabulated values is likely to be less error-prone that reading a graph by eye alone. In the present age of web-browser enabled personal phones, the option also exists to publish predictive models as interactive electronic interfaces; a number of such models are available for public access at the website: www.predictcancer.org.

In the discussion section of the report, in addition to addressing the limitations and likely limitations of applicability of the model, it is also important to explore the clinical significance of the model. For instance, which aspect of clinical practice or medical decision-making is likely to be affected by the use of this model? Specifically, a model should attempt to re-direct the course of medical care or change the way in which an individual's condition is being managed. Given this ambition, it is then possible to assess whether the predictive performance of the model and the intended context of use of the model will be fit for purpose. It is also important to consider how sensitive a model is to a particular measurement or observation – for example, would the predicted outcome change in a counterintuitive direction or disproportionate magnitude, relative to small uncertainties in measurement or rating of a given predictor? If the model is to intended to be used to support early diagnosis of a condition, then the reliable information needed to compute risk has to be available before the patient commences treatment or in-depth diagnostic investigation.

10.4 Critical Appraisal of Published Models

If we recall that the primary design principle of the TRIPOD checklist was to guide the reporting of prediction model development, validation and update studies, then it is clear that a complementary guidance document is required. The CHARMS checklist [5] was designed to provide guidance on how to search for multivariate modelling studies, how to select these on the basis of general validity and how to assess the applicability of a published model to a particular clinical problem.

There are two noteworthy distinctions between TRIPOD and CHARMS. First, TRIPOD does not prescribe how prediction modelling studies should be performed, merely how studies (regardless how well or poorly designed) ought to have common reported elements. Second, using TRIPOD as a checklist for critical appraisal is not helpful, since the presence or absence of a particular reported element does not necessarily connect with a risk of bias in the model. Critical appraisal emphasizesrisk of bias and broad applicability of a model, thus one must assess a reported model on the basis of what alternative methodological choices could have been made by the model developers, and whether their actual choices had led to a compromised model.

With a proliferation of predictive modelling papers, one could readily encounter multiple models all purporting to address the same target outcome. Some of these models may conflict with each other, and more than a few will suggest predictive power of quite divergent predictors for the desired outcome. Systematic search, assessment of bias and evidence synthesis from multiple published models is therefore an important, even necessary, effort to improving the state of clinical predictions as a scientific discipline.

10.4.1 Relevant Context of Prediction Modelling Studies

The CHARMS document consists of 2 parts. The first relates to framing a research question about prediction models, then defining a search strategy and to develop inclusion/exclusion criteria for what kind of studies to put into a review. Critical appraisal implies that the reviewer already has a research question or a clinical problem in mind, therefore it is essential to match the search and selection of modelling studies to fit the context of the research or clinical issue. This connects with the contextual elements of TRIPOD, such as whether the target condition, patient population, predictor measurements and primary outcomes of the published work actually match with the question in mind. This further includes considerations such as: (i) is the problem about making a diagnosis/prognosis or about selecting a particular intervention, (ii) at what time point in the clinical workflow is a prediction needed, and (iii) what kind of modelling study is required to answer the question – development, validation or update.

Following a concrete formulation of a research question about predictive models, it is then possible to design a literature search strategy [14–16], and establish inclusion/exclusion criteria for which papers to review.

10.4.2 Applicability and Risk of Bias

In addition to, but not mutually exclusive with, the abovementioned general assessments about the contextual relevance of a published study, CHARMS denotes certain elements as addressing the applicability of the study outside of its original setting and other elements as addressing the potential for biased findings about model performance. Naturally, some elements of critical appraisal address both.

Elements that address applicability of the model to other settings include:

(i) Did the modelling study select a representative source of individual data?
(ii) Were there differences in the treatments administered (if any) that does not match your question?
(iii) Will the predictors, its definitions and its methods of measurement match what you intend to do?

(iv) Does the desired outcome, its definition and its method of assessment match what you intend to do?
(v) Does the time point of the predicted event match what you intend to use the model for?
(vi) Is the performance of the model, in regards to calibration and discrimination, fit for purpose in regards to the clinical decisions that have to be made as a consequence of the prediction?

Some of the elements that address the risk of biased estimation of model behavior include:

(i) Was an appropriate study design used to collect information for model development? For example, a prospective longitudinal cohort design would be ideal for prognostic/treatment outcome prediction model development, but randomized clinical trials data, retrospective cohorts or registry extractions are often selected as pragmatic alternatives. The concern with randomized trials is that excessive selectivity of patients may not represent the wider population. Retrospective cohorts are highly susceptible to problems concerning handling of missing data. Registry extractions may yield large numbers of individual cases, but one needs to be mindful of the total number of target events together with significantly reduced detail of the observations/measurements.
(ii) Was the target outcome in the development and validation cohorts always defined the same way, objectively assessed in the same way and were the outcomes assessors blinded to the values of the candidate predictors? If the answer to one or more of these is no, then there is a risk that the model performance has been affected to some degree by interpretation bias, measurement bias and/or confirmation bias.
(iii) Was the number of candidate predictors and manipulation of the predictors during statistical analysis (e.g. premature dichotomization of continuous, categorical or ordinal values) reasonable for the number of target events seen? The former specifically relates to the risk of overfitting of the model on the development cohort (as we have discussed above) and the latter pertains to sensitivity of the model to arbitrary threshold cut-offs used for dichotomization.
(iv) Were missing values handled in an appropriate fashion? The risk of selection bias increases if a complete-cases analysis was used without testing whether the missing values were truly missing at random. If missing values had been imputed using surrogates of the target outcome, there is a risk of association bias since a correlation with the expected outcome has been introduced into the candidate predictors.
(v) Was predictor selection and regression coefficient fitting performed in a reasonable manner? There will be an elevated risk of predictor selection bias if single predictors with large (but spurious) correlation with the target outcome in univariate analysis are selected for inclusion into a multivariable model. A methodologically robust method is backwards stepwise multiple regression, such that predictors are recursively eliminated one by one to find the most

parsimonious model with the equivalent predictive performance as all the predictors. A modelling error occurs if the assumptions of the statistical model applied (e.g., constant hazard rate over time) is not actually met by the data.

(vi) Was the evaluation of model performance done in a sufficiently independent dataset? It is well known that evaluating a model in the same development cohort least to over-optimistic estimates of predictive performance. A validation cohort may be temporally or contextually shifted with respected to the development cohort, but failure to understand how the cohorts differ will lead to a biased assessment of the model. A related concern is whether the distribution of observed predictor values are equivalent in the development and validation cohorts.

10.4.3 Systematic Reviews and Meta-analyses

Reporting guidelines for systematic reviews of clinical trials, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines [17], are relatively mature and are being enforced by some journal editors. Similarly matured and widely applied guidelines for massed evidence synthesis on prediction modelling studies currently do not exist, but there is growing methodological research into the question [18].

Examples of systematic reviews of prediction models share a number of common themes as their clinical trials counterparts, chiefly: (i) a clear statement of the research question in terms of the population, context applicability and intended use of the models, (ii) a definitive search strategy for articles, with strict adherence to inclusion and exclusion criteria, (iii) assessment of the risk of bias in each included article and, (iv) an attempt at quantitative summary (i.e., meta-analysis) of performance metrics across all included articles. The potential sources of bias for prediction model development, validation and update stand quite distinctly apart from those in clinical trials, therefore the CHARMS checklist should still be used as the main conceptual component for formulating a systematic review of this kind. With rapid advances in "big data" and data sharing technologies, it becomes increasingly feasible that one may attempt to develop, validate and update predictive models using vast numbers of records gleaned either from electronic health records by accessing the individual cases in published models [12].

10.5 Conclusion

This chapter connects with the others by utilizing statistical concepts relating to model building and model testing that have been previously discussed, and acts as a bridge to further chapters that examine challenges and opportunities for bringing models into routine clinical use. This chapter may be used as a stand-alone source,

such that the reader understands the central matters in reporting on their own multivariable prediction models, and what key themes to look for when critically appraising published work on other models for validity and applicability to their own situation. Detailed checklists in the form of TRIPOD and CHARMS have been introduced, along with references to expansions and elaborations of such tools. Growing topics in methodological research such as clinical impact studies and evidence synthesis of multiple models (with and without a connection to "big data") have been briefly touched upon. Far from being a complete survey of reporting standards and critical appraisal, the driving motivation has been to equip the reader with insight of the most essential major themes, and to provide literature references where deeper detail on specific topics may be explored.

References

1. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. https://doi.org/10.1136/bmj.b606.
2. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 9(5):e1001221. https://doi.org/10.1371/journal.pmed.1001221.
3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55–63. https://doi.org/10.7326/M14-0697.
4. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–W73. https://doi.org/10.7326/M14-0698.
5. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 11(10):e1001744. https://doi.org/10.1371/journal.pmed.1001744.
6. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604. https://doi.org/10.1136/bmj.b604.
7. 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. https://doi.org/10.1136/heartjnl-2011-301246.
8. 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. https://doi.org/10.1136/heartjnl-2011-301247.
9. 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.
10. 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. https://doi.org/10.1186/1471-2288-14-40.
11. 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. https://doi.org/10.1093/jnci/86.11.829.

12. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140.
13. http://www.tripod-statement.org/TRIPOD/TRIPOD-Checklists
14. Haynes BR, McKibbon AK, Wilczynski NL, Walter SD, Were SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005;330:1179. https://doi.org/10.1136/bmj.38446.498542.8F.
15. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8:391–7.
16. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844. https://doi.org/10.1371/journal.pone.0032844.
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100. https://doi.org/10.1371/journal.pmed.1000100.
18. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.