Nephrol Dial Transplant (2017) 32: 752–755
doi: 10.1093/ndt/gfx073

Con: Most clinical risk scores are useless

Friedo W. Dekker, Chava L. Ramspek and Merel van Diepen

Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence and offprint requests to: Friedo W. Dekker; E-mail: F.W.Dekker@lumc.nl

ABSTRACT

While developing prediction models has become quite popular both in nephrology and in medicine in general, most models have not been implemented in clinical practice on a larger scale. This should be no surprise, as the majority of published models has been shown to be poorly reported and often developed using inappropriate methods. The main problems identified relate to either using too few candidate predictors (based on univariable $P < 0.05$) or too many (for the number of events), resulting in poorly performing prediction models. Guidelines on how to develop and test a prediction model all stress the importance of external validation to test discrimination and calibration in other populations, as prediction models usually perform less well in new subjects. However, external validity has not often been tested for prediction models in renal patients. Moreover, impact studies showing improved clinical outcomes when using a prediction model in routine clinical practice have been reported rarely. By and large, notwithstanding a few notable exceptions like the kidney failure risk equation prediction model, most models have not been validated externally or are at best inadequately reported, preventing them from be used in clinical practice. Therefore, we recommend researchers to spend more energy on validation and assessing the impact of existing models, instead of merely developing more models that will most likely never be used in clinical practice as well.

Keywords: clinical prediction models, nephrology, prediction research, prognosis, risk prediction

INTRODUCTION

'Prediction is very difficult, especially if it's about the future'. This famous quote attributed to quantum physicist and Nobel Prize winner Niels Bohr has lost nothing of its power since the last century. Patients are full of questions about their individual prognosis, they feel like everything can be found on the Internet, but they ask their doctor and blame him or her if the future turns out to be different. At the same time, day-to-day clinical reasoning and decision-making is highly based on each doctor's own expectation about the future for that patient, intuitively taking into account the patient's present health status as well as past experiences with similar cases.

As a specialized area in clinical research, developing clinical risk scores and prediction models has become increasingly popular over the last 10–20 years. A PubMed search on 'clinical prediction model' in nephrology resulted in 140 papers in 2016 while 10 years earlier only 14 were published. Also in general medicine, developing prediction models has become increasingly popular, though only a small fraction of these models is implemented in clinical practice [1].

Unfortunately, prediction research, particularly the development of prediction models or clinical risk scores, has proven to be quite prone to error [2]. In many prediction articles, the methods used are not up to standard, in spite of an extensive body of methodological literature [3–12]. In addition, the reporting of methods and results is often poor [13–15], making it hard or even impossible to judge methodological quality. Indeed, a systematic review of prediction studies in six high-impact general medical journals found that the majority of articles did not adhere to methodological recommendations and were lacking in reported details [16]. A recent systematic review on prediction models for cardiovascular disease risk published in the *British Medical Journal* showed similar results. The authors conclude there is an overabundance of these models, and even claim it is time to stop developing new prediction models in this area altogether, and to focus on making better use of available evidence [17]. These problems have been recognized by the scientific community, and attempts have been made to solve them through the development of more methodological articles and guidelines, such as the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guideline for reporting prediction studies [18, 19], which has been adopted by many leading medical journals. However, this has not halted the abundance of useless prediction models that are being published. We have identified a number of problem areas that contribute to the fact that few models are implemented, especially in nephrology, and discuss these below.

FLAWS AND PITFALLS IN MODEL DEVELOPMENT

Many methodological errors are made in the development of prediction models. Recently, a step-by-step plan for prediction modelling, simultaneously discussing common pitfalls, was published [20], confirming an ongoing need for such articles

and demonstrating the fact that methodological issues are far from solved. Other articles, such as the Elaboration and Explanation article of the TRIPOD guideline [19], also reviewed frequently made methodological mistakes, to educate and serve as a warning. Examples of such mistakes are: using too many candidate predictors for the number of events, dichotomizing of continuous variables, assuming linearity of predictors, applying a too strict selection criterion (e.g. $P < 0.05$) and selecting candidates based on univariable significance, to name just a few. In addition, reporting of methods—e.g. how were missing values handled—and results—e.g. the full prediction formula, including baseline hazard/intercept—are often lacking.

In the field of renal prediction specifically, things are no better. Errors abound and many nephrological prediction articles are of poor methodological quality. A systematic review of articles predicting chronic kidney disease (CKD) or end-stage renal disease found that most articles used inappropriate methods, potentially affecting predictive ability, and many were poorly reported, hindering judgment of usefulness of the presented models [21]. A methodological review on risk prediction for CKD warned against a number of common pitfalls particularly relevant in this field, such as accounting for competing events, measurement error in lab values and non-linear effects of glomerular filtration rate [22]. Recently, an article in a series on statistical methods for CKD cohorts focused on prediction modelling, specifying key methodological aspects illustrated by a CKD example, in an attempt to improve the quality of nephrological prediction research [1].

In addition, or perhaps as a consequence of their methodological shortcomings, many developed models show low predictive performance, as assessed by their calibration and discrimination. Even if discrimination (measured by the c-statistic) is not that bad, calibration can be poor or is not reported at all. Notwithstanding a few notable exceptions, including the Kidney Failure Risk Equation (KFRE), most models for predicting CKD or kidney failure suffer from poor on not reported calibration [23]. Furthermore, models for other important renal outcomes, such as cardiovascular events and mortality in CKD populations, have very limited predictive ability, as summarized in multiple reviews of renal prediction models [24, 25].

FLAWS AND PITFALLS IN MODEL VALIDATION

Validating a prediction model correctly and sufficiently has proven to be of utmost importance in order to judge whether the model has been adequately developed and to assess whether the model is applicable to different patient populations. Unfortunately, this seems to be a major pitfall within renal prediction research, in which validation is almost always incomplete and often poorly conducted [26]. Though the amount of prediction research has almost exponentially increased in the renal field, this mainly includes articles on the development of new prediction models or clinical risk scores and rarely concerns the validation of existing ones.

In prediction research one should always consider the important distinction between internal and external validation. The goal of internal validation is to uncover overfitting or other shortcomings in design or modelling methods. This is best done by bootstrapping the data in order to create multiple samples in which the model is tested within one cohort [27]. A more often used, but inferior, method is to simply split the sample into two (or more) groups and develop the model in one group and test it in the other. This method is unreliable, increases the risk of overfitting and wastes information as the model is only built on half of the available data.

Once the reproducibility of a model has been established through internal validation, the next step is to test the external validity of the model to see whether the model is transportable to other patient populations [28]. This is done by testing the performance of the model in a different independent patient cohort. This step is rarely performed; often researchers merely mention the need for further external validation in the discussion, without undertaking steps to do so. An exception herein is the validation of the KFRE by Tangri *et al.* [29]. An extensive validation study of this model was published in the *Journal of the American Medical Association*, assessing the external validity of two prediction models from the KFRE article in over 30 different cohorts participating in the CKD Prognosis Consortium [30]. Another example is the prediction of acute kidney failure after orthopaedic surgery developed by Bell *et al.*, where a concurrent external validation of their model is presented [31].

Nevertheless, the major lack of external validation studies in the renal research field hinders the implementation of prediction models in clinical practice, as one has no way of knowing in which populations the model is applicable since models almost always perform more poorly in external validation [32]. One could even go so far as to say that the development of a prediction model is useless if there are no plans for further external validation, as this model could never be safely implemented in practice [27]. Furthermore, to facilitate comparison between existing risk scores and models, it is essential that comprehensive validation studies are done in which multiple comparable models are validated in the same large independent cohort. Unfortunately, very few risk scores developed for similar outcomes and target populations have been directly validated and compared [33].

This lack of external validation is, however, not the only problem. Though it is being recognized that well-designed external validation studies are one of the main priorities in prediction research, what defines a 'well-designed' external validation study may not be so obvious and fundamental design issues have received only very limited attention [4, 34]. Little scientific effort, let alone empirical, has been put toward this subject, and a lot of questions remain unanswered, such as how to choose an appropriate validation cohort, how to determine if a model has been sufficiently externally validated and how to compare the external validation results of different risk scores. It is easy to influence the results of external validation through the chosen validation cohort. When the validation cohort is very similar to the derivation cohort, it is naturally more likely the risk score will perform well and the validation will be more a test of reproducibility than of transportability, discounting the original goal of external validation [35]. What is 'too similar' and 'different but related' is a significant issue, and one with profound consequences for external validation studies, however

there are no guidelines on what to adhere to concerning these problems, which makes it almost impossible for reviewers, clinicians and readers to navigate through and assess the quality of external validation studies.

THE LACK OF IMPACT STUDIES

Whilst proper development and validation are the first step towards clinical implementation of a risk score, these studies only gives us information on how well the model performs and tell us nothing about whether the implementation of the model would be beneficial to clinical practice. The development of these scores is based on an underlying assumption that accurately predicted estimated probabilities improve a clinician's decision-making or the patient's quality of life. The only way to uncover whether incorporation of a certain risk score in the decision-making process improves (or harms) patient care is by performing an impact study [36]. The impact of a validated risk score on health outcomes and cost-effectiveness should be studied separately from development and validation, preferably in the form of a randomized trial [6]. Impact studies might implement models or scores as an assistive role, by simply providing probabilities, leaving room for intuition and judgement from patient and doctor alike. Alternatively, impact studies can take a decisive approach in which each probability category suggests certain decisions [3].

Different studies have shown that accurate risk prediction may have an unpredictable effect on a clinician's decisions. A famous example is a study by Cameron and Naylor in which the impact of an educational intervention to increase the use of the Ottowa Ankle Rules was tested [37]. These rules are meant to guide clinicians in when it is necessary to X-ray an ankle, in order to reduce the amount of X-rays. Interestingly enough, the hospitals that received the educational intervention saw an increase in the amount of ankle X-rays, even though the clinicians reported to have found the educational intervention useful. In contrast, the control hospitals had a significant decrease in the amount of ankle X-rays. This example goes to show that clinicians and researchers alike cannot know whether using a risk score will improve patient care. Very few prediction rules have been formally analysed on impact to determine whether they improve outcomes. Within impact studies, measures of safety and efficiency are important outcomes to assess, but also the impact of the model when clinicians distrust or misuse the score [38]. A study by Kappen *et al.* illustrated through interviews with physicians that combining clinical experience with a probability generated by a score or model is often challenging and can be approached in many different ways [39]. One could imagine that patients have an even harder time interpreting these probabilities and using them in their favour.

As far as we are aware, no impact studies exist for prediction risk scores in CKD patients. Setting up such an impact study is an extremely costly and timely undertaking and can only be justified for extensively tested and validated models that might make a large-scale difference to clinical practice.

RECOMMENDATIONS

Notwithstanding a few notable examples like the KFRE prediction model developed by Tangri *et al.* [24], we see most prediction models in renal medicine are no better than an interesting first try. Most models have not been validated externally or are at best inadequately reported, preventing them from being used in clinical practice. Since researchers seem stuck on the first phases of prediction research and keep developing new models and risk scores, without reaching a further stage, it is understandable that most newly developed scores have been clinically useless thus far. In line with Damen *et al.*, we urge researchers to refrain from developing new models [17]. Redirection of that energy towards validation and assessing the impact of such models is essential if we want to clinically implement a model that has a positive influence on patient outcomes.

(See related articles by Tangri *et al.* Pro: Risk scores for chronic kidney disease progression are robust, powerful and ready for implementation. *Nephrol Dial Transplant* 2017; 32: 748–751; Zoccali. Moderator's view: Predictive models: a prelude to precision nephrology. *Nephrol Dial Transplant* 2017; 32: 756–758)

REFERENCES

1. Roy J, Shou H, Xie D *et al.* Statistical methods for cohort studies of CKD: prediction modeling. *Clin J Am Soc Nephrol*; 22 September 2016 [Epub ahead of print]
2. Van Diepen M, Ramspek CL, Jager KJ *et al.* Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrol Dial Transplant* 2017; 32: 1–5
3. Moons KG, Altman DG, Vergouwe Y *et al.* Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Br Med J* 2009; 338: b606
4. Altman DG, Vergouwe Y, Royston P *et al.* Prognosis and prognostic research: validating a prognostic model. *Br Med J* 2009; 338: b605
5. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361–387
6. Moons KG, Kengne AP, Grobbee DE *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; 98: 691–698
7. Moons KG, Kengne AP, Woodward M *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; 98: 683–690
8. Moons KG, Royston P, Vergouwe Y *et al.* Prognosis and prognostic research: what, why, and how? *Br Med J* 2009; 338: b375
9. Royston P, Moons KG, Altman DG *et al.* Prognosis and prognostic research: developing a prognostic model. *Br Med J* 2009; 338: b604
10. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, Springer-Verlag New York, 2009
11. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35: 1925–1931
12. Steyerberg EW, Vickers AJ, Cook NR *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21: 128–138
13. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *J Am Med Assoc* 1997; 277: 488–494
14. Meisner A, Kerr KF, Thiessen-Philbrook H *et al.* Methodological issues in current practice may lead to bias in the development of biomarker combinations for predicting acute kidney injury. *Kidney Int* 2016; 89: 429–438

15. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993; 118: 201–210
16. Bouwmeester W, Zuithoff NP, Mallett S *et al.* Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012; 9: 1–12
17. Damen JA, Hooft L, Schuit E *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Br Med J* 2016; 353: i2416
18. Collins GS, Reitsma JB, Altman DG *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br Med J* 2015; 350: g7594
19. Moons KG, Altman DG, Reitsma JB *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
20. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG* 2017; 124: 423–432
21. Collins GS, Omar O, Shanyinde M *et al.* A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013; 66: 268–277
22. Grams ME, Coresh J. Assessing risk in chronic kidney disease: a methodological review. *Nat Rev Nephrol* 2013; 9: 18–25
23. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med* 2012; 9: e1001344
24. Tangri N, Kitsios GD, Inker LA *et al.* Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med* 2013; 158: 596–603
25. Rigatto C, Sood MM, Tangri N. Risk prediction in chronic kidney disease: pitfalls and caveats. *Curr Opin Nephrol Hypertens* 2012; 21: 612–618
26. Collins GS, de Groot JA, Dutton S *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014; 14: 40
27. Bleeker SE, Moll HA, Steyerberg EW *et al.* External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003; 56: 826–832
28. Debray TP, Vergouwe Y, Koffijberg H *et al.* A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; 68: 279–289
29. Tangri N, Stevens LA, Griffith J *et al.* A predictive model for progression of chronic kidney disease to kidney failure. *J Am Med Assoc* 2011; 305: 1553–1559
30. Tangri N, Grams ME, Levey AS *et al.* Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *J Am Med Assoc* 2016; 315: 164–174
31. Bell S, Dekker FW, Vadiveloo T *et al.* Risk of postoperative acute kidney injury in patients undergoing orthopaedic surgery–development and validation of a risk score and effect of acute kidney injury on survival: observational cohort study. *Br Med J* 2015; 351: h5639
32. Siontis GC, Tzoulaki I, Castaldi PJ *et al.* External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015; 68: 25–34
33. Collins GS, Moons KG. Comparing risk prediction models. *Br Med J* 2012; 344: e3186
34. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; 19: 453–473
35. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; 130: 515–524
36. Wallace E, Smith SM, Perera-Salazar R *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
37. Cameron C, Naylor CD. No impact from active dissemination of the Ottawa Ankle Rules: further evidence of the need for local implementation of practice guidelines. *Can Med Assoc J* 1999; 160: 1165–1168
38. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; 144: 201–209
39. Kappen TH, van Loon K, Kappen MA *et al.* Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016; 70: 136–145

Received: 22.3.2017; Editorial decision: 22.3.2017

Nephrol Dial Transplant (2017) 32: 755–756
doi: 10.1093/ndt/gfx073a

Opponent's comments

Navdeep Tangri[1], Thomas Ferguson[2] and Paul Komenda[3]

[1]Department of Medicine and Community Health Sciences, University of Manitoba, Seven Oaks General Hospital, Winnipeg, Manitoba, Canada, [2]Department of Internal Medicine, University of Manitoba, Winnipeg, Manitoba, Canada and [3]Department of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada

Correspondence and offprint requests to: Navdeep Tangri; E-mail: ntangri@sogh.mb.ca

Dr Dekker and colleagues assert that prediction is difficult, and recommend that researchers focus efforts on validation and implementation studies rather than the development of new models. They argue that flaws in model development and questionable findings from impact studies have limited the clinical utility of most risk prediction models.

We would agree with Dr Dekker on the importance of external validation and studies of clinical impact. For predicting kidney failure, the field certainly needs to move on to evaluate the utility of the KFRE, rather than its discriminatory performance. However, we think significant deficits still remain for model development and early validation for predicting cardiovascular disease and early mortality on dialysis.

In our systematic review in 2013, we demonstrated a lack of accurate models for predicting cardiovascular events in patients with CKD, and recent reviews suggest a similar lack for predicting early mortality in dialysis [1]. In the absence of these models, clinicians may choose high-dose statins or other more