Journal of Clinical Epidemiology ■ (2017) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Poor performance of clinical prediction models: the harm of commonly applied methods

Ewout W. Steyerberg[a,b,*], Hajime Uno[c], John P.A. Ioannidis[d,e,f,g], Ben van Calster[a,h], Collaborators

[a]*Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*
[b]*Department of Public Health, Erasmus MC, Rotterdam, The Netherlands*
[c]*Division of Population Sciences, Dana-Farber Cancer Institute, 02215 MA, Boston, USA*
[d]*Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA*
[e]*Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA*
[f]*Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA*
[g]*Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA*
[h]*Department of Development and Regeneration, KU Leuven, Leuven, Belgium*

Accepted 17 November 2017; Published online xxxx

Abstract

Objective: To evaluate limitations of common statistical modeling approaches in deriving clinical prediction models and explore alternative strategies.

Study Design and Setting: A previously published model predicted the likelihood of having a mutation in germline DNA mismatch repair genes at the time of diagnosis of colorectal cancer. This model was based on a cohort where 38 mutations were found among 870 participants, with validation in an independent cohort with 35 mutations. The modeling strategy included stepwise selection of predictors from a pool of over 37 candidate predictors and dichotomization of continuous predictors. We simulated this strategy in small subsets of a large contemporary cohort (2,051 mutations among 19,866 participants) and made comparisons to other modeling approaches. All models were evaluated according to bias and discriminative ability (concordance index, *c*) in independent data.

Results: We found over 50% bias for five of six originally selected predictors, unstable model specification, and poor performance at validation (median $c = 0.74$). A small validation sample hampered stable assessment of performance. Model prespecification based on external knowledge and using continuous predictors led to better performance ($c = 0.836$ and $c = 0.852$ with 38 and 2,051 events respectively).

Conclusion: Prediction models perform poorly if based on small numbers of events and developed with common but suboptimal statistical approaches. Alternative modeling strategies to best exploit available predictive information need wider implementation, with collaborative research to increase sample sizes.

Keywords: Validation; Prediction model; Regression analysis; Simulation; Sample size; Events per variable

1. Introduction

Prediction models are increasingly important in the current era of precision medicine [1]. Such models may inform patients on their individualized risk of developing disease, assist physicians in diagnostic workup, and provide a personalized prognosis by predicting outcomes of disease. The scientific research to develop and validate clinical prediction models has been criticized, with recent guidelines providing advice on transparent reporting and good practice [2].

Several systematic reviews have been performed with a focus on methodological biases in the development of prediction models [3–8]. Three problematic modeling aspects stood out in these reviews: (1) selection of predictors based on statistical significance (in 56–86% of models reviewed); (2) categorization of predictors (in 62–97% of models reviewed); and (3) inadequate sample size at model development (17–50% of models reviewed, Table S1). These

Conflict of interest: The authors have declared that no conflicts of interest exist.

E.W.S. is supported by U01 NS086294 from the NIH and by grant 602150 (CENTER-TBI) from the European Union's FP7 Program.

* Corresponding author. Tel.: +31715269700.

E-mail address: e.w.steyerberg@lumc.nl (E.W. Steyerberg).

https://doi.org/10.1016/j.jclinepi.2017.11.013
0895-4356/© 2017 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Simulations of the modeling strategy for a well-published prediction model showed severely biased effect estimates and poor predictive performance in independent data. The poor performance was caused by common but suboptimal statistical approaches: selection from a large set of candidate predictors based on statistical significance; dichotomization of continuous predictors; and development and validation in relatively small data sets.

What this adds to what was known?

- The impact of stepwise selection with small numbers of events is more detrimental than many may anticipate, while validation in small samples leads to unreliable assessment of model performance.

What is the implication and what should change now?

- The poor discrimination and poor calibration that is expected from models developed with rather standard statistical approaches in small data sets implies that we should have limited trust in many prediction models to support precision medicine.
- Modeling practices in small data sets need to improve immediately, including the prespecification of a limited set of (preferably continuous) predictors based on external knowledge, use of penalization techniques for regression models, and honest internal validation.
- Available prediction models require validation across different settings with hundreds of events, in addition to careful review of statistical methodology, prior to their dissemination and implementation in routine clinical practice.

approaches have been criticized in many theoretical and applied studies (Table S2). Nevertheless, they are still quite common. The developed models show spuriously promising results. Often, some external validation is performed, but this is based again on small sample size and this perpetuates the misinterpretation about the performance of the model [9—12]. This problem of small validation size is also common (46% in a recent review) [13]. Whenever external independent validation is subsequently performed with a large, rigorous study, this often shows disappointing performance [14,15]. This may be attributable to poor practice at model development rather than genuine differences between validation and development settings.

Indeed, these problematic approaches were used in the development and validation of a model that aimed to predict the likelihood of having a mutation in germline DNA mismatch repair genes at the time of diagnosis of colorectal cancer (CRC) (“MMRpredict”) [16]. This model was published in a prestigious journal (*the New England Journal of Medicine*). This may reflect that some problematic statistical procedures, such as stepwise selection of predictors from a wide set of candidate predictors, may be seen as good practice or unavoidable in developing prediction models. Furthermore, the model was developed with only 38 patients having the event of interest, and validation was done in an independent data set with only 35 events. Eventually, many years later, the MMRpredict model performed poorest compared with two competing prediction models in a validation study that included 5,755 CRC patients from 11 North American, European, and Australian cohorts [17]. This motivated the current methodological study in which we hypothesize that the rather standard modeling strategy that is exemplified by the case of MMRpredict causes poor interpretability, poor reproducibility, and poor performance of a prediction model. We aim to evaluate the impact of key modeling steps on the accuracy of estimated predictor effects and risk predictions and explore alternative modeling strategies.

2. Patients and methods

2.1. Clinical context

Hereditary nonpolyposis CRC (HNPCC, also called Lynch syndrome) is caused by inactivating mutations of DNA mismatch repair genes (including MSH2, MLH1, MSH6, and PMS2). Lynch syndrome accounts for approximately 3% of CRCs. If Lynch syndrome is diagnosed in patients with CRC (“probands”), they may benefit from more intensive post-treatment colonoscopic surveillance, more extensive surgery, and management of extracolonic cancer risks. Furthermore, family members of the proband who carry the same pathogenic gene mutation also benefit from cancer prevention strategies such as intensified surveillance to reduce the increased lifetime risk of developing CRC and other cancers [16]. Current clinical guidelines recommend the use of prediction models among patients with CRC to identify those at high risk of Lynch syndrome [18,19]. These prediction models quantify a proband’s risk of carrying a mismatch repair gene mutation and intend to support decision-making regarding genetic evaluation, including germline testing or molecular tumor testing. One such prediction model was based on logistic regression analysis of 870 patients diagnosed with CRC below the age of 55 years [16]. There were 38 mutations identified (4%). This MMRpredict model was validated in an independent cohort with 35 mutations among 155 patients.

We here perform an in-depth evaluation of the modeling strategy employed for the MMRpredict model. We analyze