Clinical prediction models cannot be trusted when common modeling approaches are followed

Ewout W. Steyerberg [1, 2], Hajime Uno [3], John P.A. Ioannidis [4,5,6,7], Ben van Calster [1,8]

and collaborators

Collaborators for this paper, identifiable in PubMed:

Chinedu Ukaegbu [3], Tara Dhingra [3], Sapna Syngal [3], Fay Kastrinos [9]

1 Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
2 Department of Public Health, Erasmus MC, Rotterdam, the Netherlands.
e.w.steyerberg@lumc.nl

3 Division of Population Sciences, Dana-Farber Cancer Institute, 02215 MA, Boston, USA
huno@jimmy.harvard.edu

4 Department of Medicine, Stanford University School of Medicine, Stanford, USA;
5 Department of Health Research and Policy, Stanford University School of Medicine, Stanford, USA;
6 Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, USA;
7 Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, USA. jioannid@stanford.edu

8 Department of Development and Regeneration, KU Leuven, Leuven, Belgium
ben.vancalster@kuleuven.be

9 Herbert Irving Comprehensive Cancer Center and Division of Digestive and Liver Diseases, Columbia University Medical Center, New York, NY, USA

Short title: Why most clinical prediction models cannot be trusted

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 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 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 discriminative ability (concordance index, *c*) in independent data.

Results: We found over 50% bias for 5 out of 6 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 pre-specification 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 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.

Key words: validation; prediction model; regression analysis; simulation; sample size; events per variable

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 work-up, 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] [4] [5] [6] [7] [8]. Three problematic modeling aspects stood out in these reviews: 1) selection of predictors based on statistical significance (in 56% to 86% of models reviewed); 2) categorization of predictors (in 62% to 97% of models reviewed); 3) inadequate sample size at model development (17% to 50% of models reviewed, Table S1). These 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] [10] [11] [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 germ-line DNA mismatch-repair genes at the time of diagnosis of colorectal cancer (“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 to two competing prediction models in a recent validation study that included 5,755 CRC patients from 11 North American, European, and Australian cohorts [17]. This motivated the current methodological study, where 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 aimed to evaluate the impact of key modeling steps on the accuracy of estimated predictor effects and risk predictions, and explore alternative modeling strategies.

Patients & Methods

Clinical context

Hereditary nonpolyposis colorectal cancer (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 colorectal cancers (CRC). 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 data from 19,866 patients with CRC who were tested for Lynch syndrome related mismatch repair genes (MLH1, MSH2, MSH6) at Myriad Genetics Laboratories [20] [21] [22]. Candidate predictors were defined following the Appendix of the original publication, where no specific rationale for the list was given [23]. Candidate predictors included age at diagnosis, sex, presence of other synchronous or metachronous CRC, endometrial cancer or other Lynch associated cancers (including gastric, kidney, and other cancers such as brain, melanoma, breast, ovarian, cervix, leukaemia, lymphoma or testis cancer). Family history included the number and youngest age at diagnosis of first degree relatives (FDR), and second-degree relatives (SDR) with CRC, endometrial, or other cancers. We could examine all candidate predictors for the MMRpredict model except three characteristics of the proband's CRC which were not available in our cohort: differentiation;

histology; and location of the tumor. A full model with all available candidate predictors required the estimation of 37 logistic regression coefficients (37 degrees of freedom, Table S3). Some of the age variables had missing values (<5%), which were imputed based on correlation with other variables in a single imputation procedure. Among the participants, 2,051 mutations were found in the MSH2, MLH1, or MSH6 genes. We performed simulation studies within this large cohort to assess the impact of choices in the modeling strategy underlying the MMRpredict model.

Modeling strategies

The original strategy for the development of MMRPredict included three elements that can affect model validity substantially [16].

1. Predictors for the model were selected in a stepwise manner based on statistical significance ($p<0.05$) from the set of candidate predictors, as specified in the Appendix of the original publication. A univariate screening of candidate predictors was followed by further selection from a multivariable logistic regression model. This practice is known to lead to chance findings [24] [25], exaggeration of true predictor effects [26], and optimistic expectations on model performance [27] [28] [29].
2. Continuous predictors were categorized (age at cancer below or above age 50 years). Such categorization of predictors causes a loss in information [30] [31].
3. Model development was based on data from 870 patients with 38 mutations (events) and validated in an independent cohort with 35 mutation carriers detected among 155 patients. The small cohorts and limited total events aggravate various problems at model development (events per variable, EPV, close to 1) and lead to large variability in statistical summary measures for performance [11] [27] [28] [29] [32] [33] [34].

We evaluated alternative modeling strategies including:

1. *Pre-specification of model structure*

   We summarize family history as a sum of the number of FDR (0, 1, 2+) and the number of SDR (0, 1, 2+), where SDR are weighted as half of the FDR for family history, reflecting the genetic distance between FDR and SDR. Hence, the family history can be summarized in a variable which ranges between 0 (FDR=0, SDR=0) and 3 (FDR >=2, SDR >=2). For family history, the degrees of freedom (df) decrease from 4 (for coding of FDR with 2 df and SDR with 2 df) to 1. We may also force the effect of the youngest age of CRC diagnosis in the proband, FDR, and SDR to be identical. The decrease in df is from 3 (for 3 age effects) to 1 df (for a summary effect). Such simplification can also be done for endometrial cancer, and other Lynch Syndrome - associated cancers. Modeling the family history and age effect for CRC, endometrial, and other Lynch Syndrome cancers could hence be achieved with 6 df rather than 21 df (for a model with 12 df for family history and 9 df for age effects) [20] [35].

2. *Avoidance of categorization*

   We may keep all continuous variables by default as linear terms in the prediction model [30] [31]. Non-linearity in effects may be evaluated in several ways, but was not considered here to prevent overfitting in relatively small development samples [9] [36] [37].

3. *Increase in the number of outcome events*

   A first alternative to a fixed split in a development and validation cohort is to base the final model on their combination, leading to 38+35=73 events for statistical modeling, with stratification by study [38]. We also simulate the situation that larger development and validation cohorts would be available. Following the principle of having at least 10 to 20 events per variable in the modeling process, we consider situations with 370 and 740 events at model development [39].

Simulation design

We draw 1,000 random samples for model development from our cohort with 19,866 patients, stratified by mutation status and hence fixing the event rate. The number of events ranged from 38 to 740. We validated the developed models in the remaining independent patients, not used for model development. We also examined small validation samples with 35 events, which were drawn at random from the independent patients. Different modeling strategies were followed, as outlined above. We evaluated bias in predictor effects on the logistic scale (i.e. with estimated regression coefficients, i.e. b=log(odds ratio)). Bias was defined as the difference between an estimated coefficient following a modeling strategy in a simulated sample and the coefficient in a model with the predictors in the full data set: ($b_{simulated}$ – b_{full})/ b_{full}. We also evaluated model stability (selection of predictors and variability between models), and predictive performance. Performance measures included measures for discrimination (separation provided by risk predictions, indicated by a concordance statistic, *c*) and calibration (reliability of risk predictions, indicated by the calibration slope, i.e. the regression coefficient of the linear predictor when used as the single predictor in a logistic regression model) [40] [41]. The c statistic is equivalent to the area under the ROC curve. It ranges between 0 and 1.0, and is over 0.5 if higher predictions are associated with higher risk of the event of interest [36]. The calibration slope is 1 at model development, and values below 1 reflect statistical overfitting: low predictions are too low and high predictions are too high [42].

We used R software for all analyses (version 3.3.2), after data preparation was done with SAS software (version 9.2). A single imputation procedure was performed with the `aregImpute` function. Logistic regression models were fit with the `lrm` function, `fastbw` for backward stepwise selection with $p<0.05$ from a multivariable model that included all candidate predictors with $p<0.05$ at univariable analysis, `unique.matrix` for counting the frequency of different models, and `val.prob` for model validation in independent data [36].

Results

Bias in predictor effects

Several characteristics of the proband were associated with the presence of mutations in the cohort of 19,866 patients with CRC (univariable analyses, Table 1). A history or presence of another CRC, endometrial cancer, or another Lynch-associated cancer each had odds ratios around 3. Among relatives, a CRC at young age was strongly predictive for Lynch Syndrome. In MMRpredict, the multivariable odds ratios were considerably larger for all predictors included in that model, e.g. an odds ratio of 9.5 for presence of another CRC in the proband, and an odds ratio of 46 for a FDR with CRC under 50 years [16]. The most remarkable finding was a multivariable odds ratio of 59 for a FDR with endometrial cancer, where we found a multivariable odds ratios of 2.8 in our cohort [95% confidence interval 2.5 – 3.2] (Table 1).

Our simulations illustrate that the large and clinically implausible estimates of predictor effects in MMRpredict might be partly attributed to stepwise selection as a modeling strategy (Figure 1). In samples of 870 probands with 38 mutation carriers we simulate the selection of predictors from the MMRpredict model. For male sex, we find that the effect was statistically non-significant in 79% of the simulated samples. In the 21% instances where the effect was statistically significant, we estimate an average odds ratio of 3. This estimate is substantially higher than the multivariable odds ratio of 1.7 that we found if the selected model was estimated in the full data set with 19,866 probands: a bias over 100% for this predictor at the log scale. Similarly, large bias was found for 4 other predictors (presence of more than one CRC, CRC < 50, FDR with CRC >50, endometrial cancer in FDR). Low bias (6%) was found for the predictor age of CRC diagnosis < 50 years in a first degree relative. This is explained by the 88% frequency of selection (predictor not statistically significant in only 12% of the simulated samples).

Model instability

Stepwise selection led not only to bias in predictor effects (Figure 1), but also to a wide variability in selected predictors (Figure 2). Typically, 3 or 4 predictors were selected per model (range: 1 to 11, Figure 2A). The most often included predictor was “FDR CRC<50” (88%, Figures 1 and 2B). Other predictors were less often selected (Figure 2B). Among the 5000 simulations with 38 events, 2174 different models were selected among 4601 with model convergence. The most frequently selected model (70 times, 1.5% of the simulations) contained two predictors: “FDR CRC<50” and “SDR CRC<50”. These were also the top two predictors over all selected models, where 1562 models were selected only once (34% of the simulations).

Model performance

The c statistic was 0.77 [0.76 – 0.78] for the refitted MMRpredict model in our large cohort (n=19,866) while it was 0.85 [0.77 – 0.93] in the original development cohort (n=870) and 0.82 [0.72 – 0.91] in the original validation cohort (n=155) [16]. The apparent performance was very optimistic for models developed with stepwise selection in simulated data sets with 870 patients and 38 events: median c=0.81 at development versus median c=0.74 at validation (Figure 3). The predictions were too extreme, with a calibration slope of 0.75 (ideal: 1.0, so 25% overfitting). Better performance was obtained if continuous predictors were used for the age of CRC in the proband and age of CRC in a FDR, rather than dichotomized versions. In the full data set, the discrimination for the refitted model increased from c=0.77 to c=0.82 [0.81 – 0.83] with continuous rather than dichotomized predictors.

Impact of number of events

With 38 events at model development, performance was estimated optimistically and with considerable uncertainty (Figure 3). Validation with 35 events led to large uncertainty in the performance estimates: the 95% range for the c statistic was 0.63 to 0.84 (median c=0.74), and 0.40 to 1.31 for the calibration slope (median c=0.75, Figure 3). The validation performance reported in the NEJM paper (c=0.82) is quite favorably placed within this expected range: 93% of the simulated models with 38 events would be expected to show a worse performance. If a large validation sample size were analyzed (over 2000 events), more stable performance estimates would be obtained, although the 95% range for the c statistic was still wide (e.g. 0.69 to 0.78).

Larger sample sizes for development led to substantially better performance. An analysis with 73 events, based on the hypothetical combination of development and validation cohorts (38+35 events) led to a median c statistic of 0.82 at development and 0.78 at validation (Figure 4). An even larger development set (370 events, for 10 events per candidate variable) would further improve model performance: c=0.830 at development and c=0.826 at validation(optimism in c statistic 0.004), and a calibration slope close to 1 (slope=0.96, Figure 4).

Pre-specification and continuous predictors

An previously proposed model included 9 predictors: male sex; synchronous or metachronous CRC; presence of endometrial, or other cancer; three summary variables for family history of CRC, endometrial, or other cancer; and two continuous summary variables for the age effects of CRC and endometrial cancer [21] [35]. If this model was estimated with 38 events, the median validated c was 0.836 (95% range 0.800-0.848, Figure 4). Predictions would be too extreme, as reflected in a calibration slope of 0.83 (95% range 0.59-1.14). With larger sample size, the validated performance increased rapidly to a median c statistic of 0.852 and perfect calibration (Figure 4).

Discussion

This study highlights problems with a number of key elements in prediction modeling strategies: selection of predictor variables based on statistical significance, dichotomization of predictors, and modeling in relatively small data sets. These elements are quite common in current scientific practice (Table S1), and lead to prediction models that cannot be trusted. The effects of predictors are exaggerated, while others are unduly discarded, and predictions are too extreme, invalidating reliable decision support. We hence call for immediate improvements in the practice of model development and validation.

Our study showed that small development and small validation samples lead to poor performance in terms of discrimination and calibration, and rather unstable estimates. The problems of small development samples also have been recognized in previous studies [7] [29] [32] [43] [44]. We add that the uncertainty of the validated performance estimates may be huge, since this uncertainty is determined by the combination of the variance in the development and validation sets. For MMRpredict, the original finding of a validated performance (c=0.82) close to the development performance (c=0.85) should not be interpreted as evidence for the validity of the prediction model [16]. We learn from Figure 3 that the validated performance has enormous uncertainty if only 35 events are present, with a c statistic ranging roughly between 0.6 and 0.9. Indeed, the 95% confidence interval was 0.72–0.91 for the reported validated c statistic of MMRPredict (c=0.82) when validated with 35 events [16]. Second, stepwise selection leads to biased regression coefficients with exaggerated prognostic effects for the predictors included in the prediction model (Winner’s curse, illustrated with Figure 1) [26]. Claims on the relevance of some characteristics and the irrelevance of other characteristics are misleading, unless the sample size are huge [27]. The selection of predictors was highly unstable, and any claims on independent effects of a specific set of predictors cannot be trusted [32] [45] [46] [47]. These issues are becoming better recognized by recent debates on the use and misuse of p-values in scientific research [48] [49] [50]. Third, dichotomization of continuous predictors for age at diagnosis led to a substantial loss of information, in line with theoretical expectations [30] [31] [51] [52] [53] [54]. The commonly used cut-off for an age at diagnosis below age 50 years as suspect for hereditary cancer should be reconsidered. It is unscientific to consider a patient with CRC at age 49 very different from a patient with CRC at age 51, but similar in risk as a patient with CRC at age 30 years.

Potential solutions

Various solutions to the development of more trustworthy prediction models have been proposed [9] [36] [39] [37] [42] [55]. Methodologists will agree that a sensible modeling strategy is especially needed if only a relatively small data set is available, commonly defined as a situation with less than 10 to 20 events per variable [39] [56]. Note that the number of candidate predictors needs to be considered here, with the corresponding effective degrees of freedom, rather than the degrees of freedom of predictors included in the final model [57]. The effective degrees of freedom increase by detailed model building, such as choosing optimal cut-offs, and examining various non-linear transformations for continuous predictors or statistical interactions. Pre-specification of a model may be attempted to save degrees of freedom, based on literature review and subject knowledge from clinical experts, with statistical testing for model specification limited as much as possible [39]. Some candidate predictors may be combined in summary variables, as illustrated for the case study with the effects of first and second degree family history, and the effect of age of cancer diagnosis [35]. Also, continuous predictors might best be considered as linear terms without testing for non-linearity, and potential statistical interaction terms ignored, if sample size is relatively small.

The benefits of reducing effective degrees of freedom need to be balanced against the loss of information by summarizing variables and other model simplifications [9] [36] [39] [58]. Far worse is the loss of information caused by stepwise selection where insufficient statistical power may easily lead to the exclusion of in fact relevant predictors [27] [32] [46] [56]. Application of a more lenient criterion for selection increases the statistical power for selection of relevant predictors, such as p<0.2 [26] or p<0.50 [29]. Similarly, models will be more informative with continuous rather than dichotomized predictors [30] [31] [51] [52] [53] [54] (Table S2). We note that the effects of continuous predictors can well be interpreted if appropriately scaled. For example, age may be coded per decade [9]. If a non-linear effect is modeled, a graphical display may be informative relatively easy to interpret [36] [59]. An example of such an attractive visual presentation was included in the prediction model derived from the GUSTO trial, where prognostic effects were plotted for 7 continuous predictors with spline transformations and effects summarized by comparing 75th to 25th percentiles [60].

Finally, large sample sizes, i.e. studies with many events, are needed to develop better models. Our case study illustrates that problems with model specification are less prominent if data sets with over 20 events per candidate predictor are available for analysis: relevant predictors will be identified, and performance stabilizes without signs of statistical overfitting. Validation sample sizes need to be large as well to give a reliable impression of performance. Our study confirms that less than 100 events at validation leads to rather unreliable estimates of performance, and that ideally at least 250 events should be present in a validation data set [9] [10] [11]. In the case study, the final MMRpredict model should have been based on the stratified combination of the development and validation data sets (38+35=73 events) [38]. This would have alleviated some of the unreliability and overfitting of the current MMRPredict model, but still be far too few events for reliable and accurate predictions [29]. A bootstrap validation might then have been performed repeating the full model specification strategy, producing a shrinkage factor that should be applied to prevent too extreme predictions in new patients [36] [38] [61] [62]. Ideally, cross-validation in multiple, large cohorts should be performed before a model is presented for clinical application, so as to get a better sense of what might be expected upon clinical application in different settings [38]. Rather than stepwise selection, a Lasso modeling or similar statistical penalization procedure should have been applied for a better balance between a small, clinically applicable model, while providing reliable predictions [29] [43] [44] [63] [64] [65]. Our recommendation is to use modern modeling approaches with penalization of estimated regression coefficients when model developers are confronted with sparse data with relatively few events. Moreover, honest internal validation approaches should be followed, that include all model specification steps. For example, if stepwise selection were used for the development of a model, e.g. with $p<0.20$ for selection of main effects of predictors [26] [29], a bootstrap cross-validation procedure should repeat this procedure in every bootstrap sample [61]. These model selection and estimation strategies require further study.

Increasing sample size, meta-analysis, and validation

A larger sample size for model development and validation may be obtained from prospective multicenter studies, or by combining individual patient data from different studies. Indeed, individual patient data meta-analysis (IPD MA) has become more common for prediction models, and opportunities will increase with the availability of "Big Data", including routinely collected data in electronic health records [66]. Our confirmations of minimum sample size requirements have implications for the design of multicenter studies and IPD MA of prediction models. Rather than expanding a single cohort, it may be more worthwhile to collect data from other cohorts once over 20 events per variable are available. For example, with 10 candidate predictors, it is more valuable to cross-validate a model in 5 cohorts with 200 events than analyzing a single cohort with 1,000 events [38].

The benefits of extending absolute numbers in IPD MA have to be balanced against possible sources of heterogeneity with respect to the clinical context, the definition of predictors, and the definition of the outcome event [67]. Heterogeneity will often be reflected in differences in baseline risk, even when accounting for different distributions of the predictors that are included in the model [66]. The advantages of IPD MA for prediction research are however numerous, such as the drive towards a consensus model rather than having a myriad of locally developed models with unclear qualities [68] [69]. Moreover, some differences between studies are needed to assess the generalizability of predictions rather than reproducibility, as examined in our simulations, where validation samples were drawn at random [67]. If model performance is consistently good in a variety of settings, this is strong evidence for the generalizability of a model [70] [13]. The possibility to perform cross-validations between studies is an important strength of IPD MA compared to development and validation of models in large single study settings [38] [71].

This cross-validation by cohort or other meaningful grouping, such by calender time [72], could not be performed in the current study, in contrast to earlier evaluations within the GUSTO trial [29] [56]. Substantial heterogeneity in baseline risk was observed among 11 cohorts included in another large external validation study of the MMRpredict model [17]. Here MMRPredict was compared to two competing models, PREMM$_{1,2,6}$ and MMRPro. The intention of these models is to support decision making on diagnostic work-up, including the ordering of tests for mutations in the mismatch repair genes in those classified as at relatively high risk. Such decision-support requires some degree of discrimination, while calibration is even more essential: poor calibration may lead to poorer decision making when guided by individualized predictions compared to a simple reference strategy such as testing all patients [73]. A model may have no clinical utility due to poor calibration [41]. Further study is needed on the extent that this problem can be prevented by applying shrinkage and penalization approaches in small data sets.

Conclusions

We conclude that prediction models have biased effect estimates and run a high risk of providing inaccurate predictions if developed with 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. Improvements may come from better statistical approaches, such as pre-specification of a limited set of predictors based on external knowledge, more refined statistical analysis, and from increased sample sizes, specifically in the context of collaborative IPD meta-analyses.

What's 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 is known

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

What is the implication, 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 pre-specification 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.

References

[1] Kattan MW, Hess KR, Amin MB, Lu Y, Moons KG, Gershenwald JE, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA: a cancer journal for clinicians. 2016;66:370-4.

[2] 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. Annals of internal medicine. 2015;162:W1-73.

[3] Mushkudiani NA, Hukkelhoven CW, Hernandez 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. Journal of clinical epidemiology. 2008;61:331-43.

[4] Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer investigation. 2009;27:235-43.

[5] Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC medicine. 2010;8:20.

[6] Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC medicine. 2011;9:103.

[7] 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 medicine. 2012;9:1-12.

[8] Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of clinical epidemiology. 2013;66:268-77.

[9] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.

[10] Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of clinical epidemiology. 2005;58:475-83.

[11] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine. 2016;35:214-26.

[12] Van Calster B, Steyerberg EW, Bourne T, Timmerman D, Collins GS. Flawed external validation study of the ADNEX model to diagnose ovarian cancer. Gynecologic oncology reports. 2016;18:49-50.

[13] 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 medical research methodology. 2014;14:40.

[14] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of clinical epidemiology. 2015;68:25-34.

[15] 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. Medical decision making : an international journal of the Society for Medical Decision Making. 1994;14:208-16.

[16] Barnetson RA, Tenesa A, Farrington SM, Nicholl ID, Cetnarskyj R, Porteous ME, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. The New England journal of medicine. 2006;354:2751-63.

[17] Kastrinos F, Ojha RP, Leenen C, Alvero C, Mercado RC, Balmana J, et al. Comparison of Prediction Models for Lynch Syndrome Among Individuals With Colorectal Cancer. Journal of the National Cancer Institute. 2016;108.

[18] Giardiello FM, Allen JI, Axilbund JE, Boland CR, Burke CA, Burt RW, et al. Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on colorectal cancer. Gastroenterology. 2014;147:502-26.

[19] Syngal S, Brand RE, Church JM, Giardiello FM, Hampel HL, Burt RW. ACG clinical guideline: Genetic testing and management of hereditary gastrointestinal cancer syndromes. The American journal of gastroenterology. 2015;110:223-62; quiz 63.

[20] Balmana 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.

[21] Kastrinos F, Steyerberg EW, Mercado R, Balmana J, Holter S, Gallinger S, et al. The PREMM(1,2,6) model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. Gastroenterology. 2011;140:73-81.

[22] Kastrinos F, Uno H, Ukaegbu C, Alvero C, McFarland A, Yurgelun MB, et al. Development and Validation of the PREMM5 Model for Comprehensive Risk Assessment of Lynch Syndrome. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2017:Jco2016696120.

[23] Barnetson RA. Appendix. http://www.nejm.org/doi/suppl/10.1056/NEJMoa053493/suppl_file/nejm_barnetson_2751sa1.pdf. 2006.

[24] Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2:e124.

[25] Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of clinical epidemiology. 1996;49:907-16.

[26] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology (Cambridge, Mass). 2008;19:640-8.

[27] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc, Ser A. 1995;158:419-66.

[28] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. Journal of clinical epidemiology. 1999;52:935-42.

[29] Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in medicine. 2000;19:1059-79.

[30] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25:127-41.

[31] Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in medicine. 2016;35:4124-35.

[32] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411-21.

[33] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology. 2014;14:137.

[34] Dekker FW, Ramspek CL, van Diepen M. Con: Most clinical risk scores are useless. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2017;32:752-5.

[35] Steyerberg EW, Balmana J, Stockwell DH, Syngal S. Data reduction for prediction: robust coding of age and family history for the risk of having a genetic mutation. Statistics in medicine. 2007;26:5545-56.

[36] Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer; 2015.

[37] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in medicine. 2007;26:5512-28.

[38] Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. Journal of clinical epidemiology. 2016;69:245-7.

[39] Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15:361-87.

[40] 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 (Cambridge, Mass). 2010;21:128-38.

[41] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of clinical epidemiology. 2016;74:167-76.

[42] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014;35:1925-31.

[43] Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Statistics in medicine. 2016;35:1159-77.

[44] Rahman MS, Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC medical research methodology. 2017;17:33.

[45] Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Statistics in medicine. 1989;8:771-83.

[46] Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. 1992;45:265-82.

[47] Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. Journal of clinical epidemiology. 2004;57:1138-46.

[48] Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70:129-33.

[49] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology. 2016;31:337-50.

[50] Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. European journal of epidemiology. 2017;32:21-9.

[51] Irwin JR, McClelland GH. Negative Consequences of Dichotomizing Continuous Predictor Variables. Journal of Marketing Research. 2003;40:366-71.

[52] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006;332:1080.

[53] Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. American Journal of Neuroradiology. 2011;32:437-40.

[54] Dawson NV, Weiss R. Dichotomizing Continuous Variables in Statistical Analysis. Medical Decision Making. 2012;32:225-6.

[55] Wynants L, Timmerman D, Verbakel JY, Testa A, Savelli L, Fischerova D, et al. Clinical Utility of Risk Models to Refer Patients with Adnexal Masses to Specialized Oncology Care: Multicenter External Validation Using Decision Curve Analysis. Clin Cancer Res. 2017;23:5082-90.

[56] 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. Medical decision making : an international journal of the Society for Medical Decision Making. 2001;21:45-56.

[57] Ye J. On measuring and correcting the effects of data mining and model selection. JASA. 1998;93:120-31.

[58] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

[59] Van Belle V, Van Calster B. Visualizing Risk Prediction Models. PloS one. 2015;10:e0132614.

[60] Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. Circulation. 1995;91:1659-68.
[61] 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. Journal of clinical epidemiology. 2003;56:441-7.
[62] Copas JB. Regression, prediction and shrinkage. J R Stat Soc, Ser B. 1983;45:311-54.
[63] Tibshirani R. Regression and shrinkage via the Lasso. J R Stat Soc, Ser B. 1996;58:267-88.
[64] Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. Int J Epidemiol. 2007;36:195 - 202.
[65] 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. Journal of clinical epidemiology. 2004;57:1262-70.
[66] Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ (Clinical research ed). 2016;353:i3140.
[67] 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. Journal of clinical epidemiology. 2015;68:279-89.
[68] 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. Statistics in medicine. 2013;32:3158-80.
[69] Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clinical research ed). 2016;353:i2416.
[70] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006;144:201-9.
[71] Ioannidis JP. How to make more published research true. PLoS medicine. 2014;11:e1001747.
[72] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of clinical epidemiology. 2003;56:1118-28.
[73] Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Medical decision making : an international journal of the Society for Medical Decision Making. 2015;35:162-9.

ES conceived and designed the study; analyzed and interpreted the data; wrote the paper

HU analyzed and interpreted the data; wrote the paper

JI interpreted the data; wrote the paper

BvC conceived and designed the study; interpreted the data; wrote the paper

Collaborators:

Chinedu Ukaegbu and Tara Dhingra prepared the data for statistical analysis;

Sapna Syngal and Fay Kastrinos provided clinical guidance for this project.

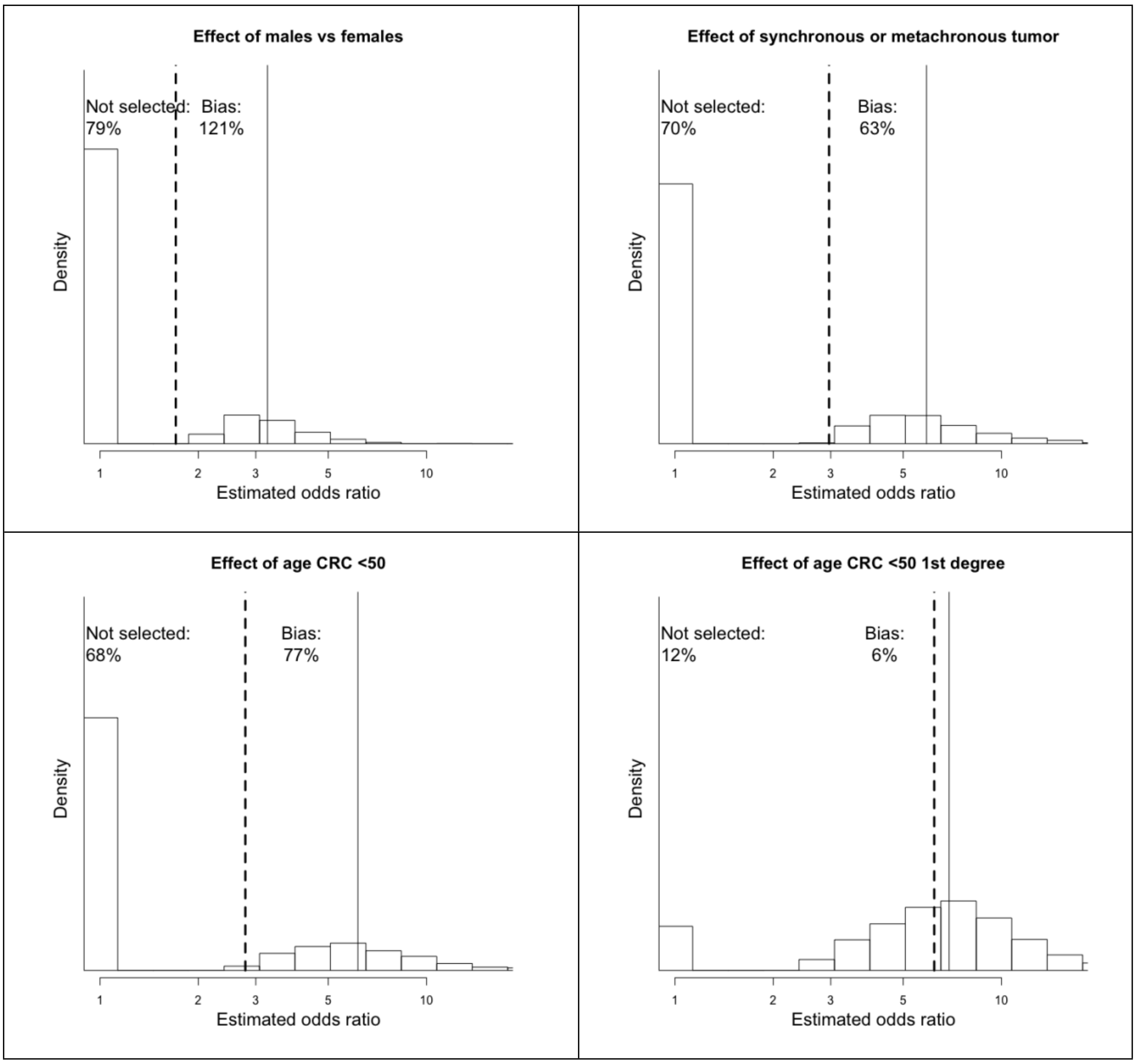
ES is supported by U01 NS086294 from the NIH and by grant 602150 (CENTER-TBI) from the European Union’s FP7 Programme.

Table 1 Associations of predictors of mutations in the MLH1, MSH2, or MSH6 genes among 19,866 probands with CRC. Univariable and multivariable odds ratios (OR) are shown with 95% confidence intervals after single imputation of missing values. The final column shows the odds ratios from univariate and multivariable analyses for the MMRPredict model.

Predictor	Missings	Non-carriers		Carriers		OR$_{\text{univariable}}$	OR$_{\text{multivariable}}$*	OR$_{\text{MMRPredict}}$
		n	%	n	%			uni / multivariable
		17,815	89.7	2051	10.3			
Proband								
Male	0	6491	86.9	977	13.1	1.59 [1.45 - 1.74]	1.54 [1.40 - 1.71]	2.24 / 2.57
Age CRC<50 yr	635 (3%)	11,141	88.0	1514	12.0	1.71 [1.54 - 1.89]	2.80 [2.49 - 3.14]	
Other cancers								
CRC	0	868	76.2	271	23.8	2.97 [2.57 - 3.44]	2.93 [2.48 - 3.45]	8.02 / 9.53
Endometrial cancer	0	1009	77.4	295	22.6	2.80 [2.44 - 3.21]		
Other Lynch cancer	0	832	75.4	272	24.6	3.12 [2.7 - 3.61]		
First degree relatives								
CRC	0	5616	80.6	1349	19.4	2.12 [2.02 - 2.22]	2.26 [1.97 - 2.59]	4.24 / 7.04
Age CRC<50 yr	382 (2%)	2042	67.9	964	32.1	6.87 [6.23 - 7.58]	10.0 [8.96 - 11.3]	36.0 / 46.26
Endometrial cancer	0	1058	76.6	323	23.4	2.56 [2.28 - 2.88]	2.83 [2.48 - 3.23]	- / 59.36
Age endometrial <50 yr	112 (1%)	582	74.2	202	25.8	3.22 [2.73 - 3.81]		
Other Lynch cancer	0	2738	85.3	471	14.7	1.48 [1.36 - 1.61]		
Age of other <50 yr	555 (3%)	962	78.9	257	21.1	2.54 [2.17 - 2.97]		
Second degree relatives								
CRC	0	5080	85.2	885	14.8	1.52 [1.45 - 1.59]		1.91 / -
Age CRC<50 yr	775 (4%)	1528	73.1	563	26.9	4.02 [3.60 - 4.49]		8.07 / -
Endometrial cancer	0	722	83.8	140	16.2	1.57 [1.35 - 1.82]		
Age endometrial <50 yr	119 (1%)	356	79.8	90	20.2	2.24 [1.77 - 2.83]		
Other Lynch cancer	0	2724	90.0	304	10.0	0.98 [0.89 - 1.08]		4.71**
Age of other <50 yr	710 (4%)	744	85.4	127	14.6	1.58 [1.28 - 1.96]		9.42***

* The multivariable logistic regression model was based on the selection in the MMRpredict model, with age at CRC in the proband coded as less than 50 years. The c statistic of this multivariable model, indicating discriminative ability, was 0.77 [95% confidence interval 0.76 – 0.78].
** Univariate odds ratio for gastric cancer at age over 50 years**, and less than 50 years ***, as provided in Supplementary Table 3A of the MMRPredict study.

Figure 1 Estimated odds ratios for 6 categorized predictors based on the MMRpredict model among 5,000 samples of 870 probands with 38 mutation carriers. The fraction of models without the predictor is indicated with an odds ratio of 1, e.g. 79% for males vs females. The average effect in models where the predictor was included is indicated with a solid vertical line, e.g. around 3 for males vs females. The average effect in the full data set of 19,866 probands is shown with a dotted vertical line, e.g. 1.7 for males vs females. The bias is 121% when calculated on the logistic scale (i.e. with log(odds ratio)).



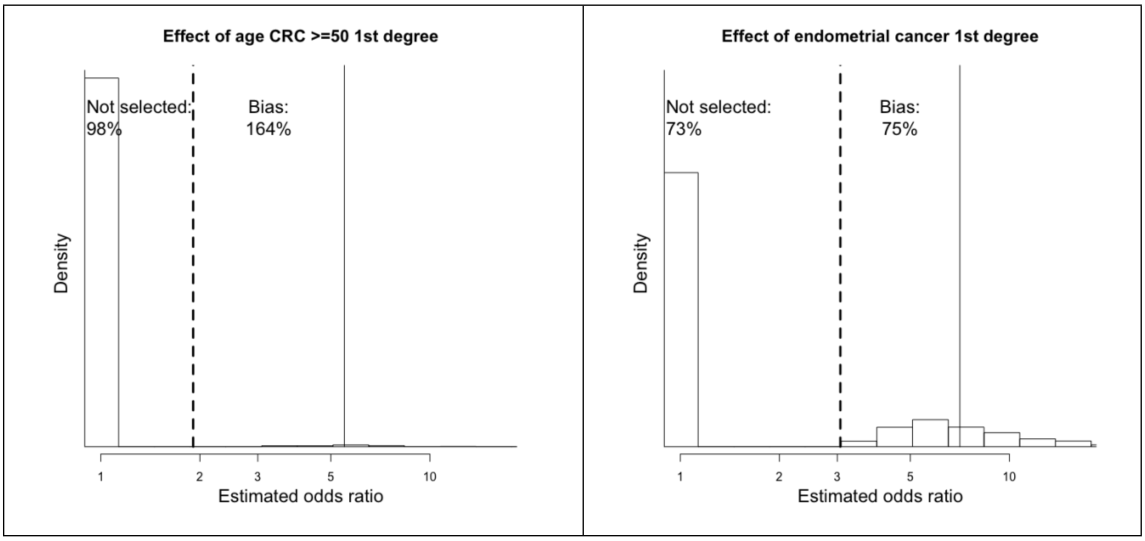


Figure 2 Number of predictors (panel A) and top 10 predictors (panel B) selected in models among 5,000 samples of 870 probands with 38 mutation carriers. FDR and SDR: First and second degree relatives; CRC: colorectal cancer; Endo: Endometrial cancer.

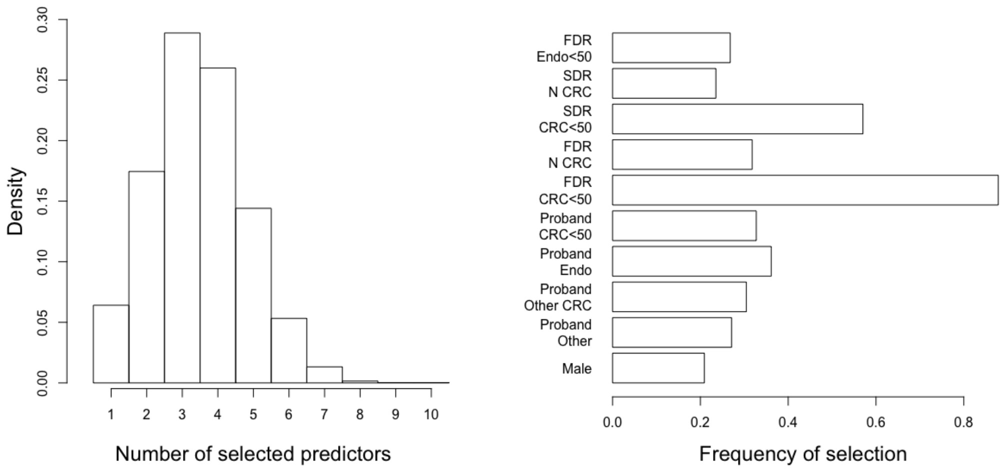


Figure 3 Estimated discriminative ability (C statistic) and calibration (slope) for models developed with stepwise selection in 5,000 samples of 870 probands with 38 mutation carriers ('events'). Samples were drawn for model development from a cohort with 19,866 probands with 2,051 events. Validation with 35 independent events (among 155 probands) led to far more variability in performance than validation with 2,013 independent events.

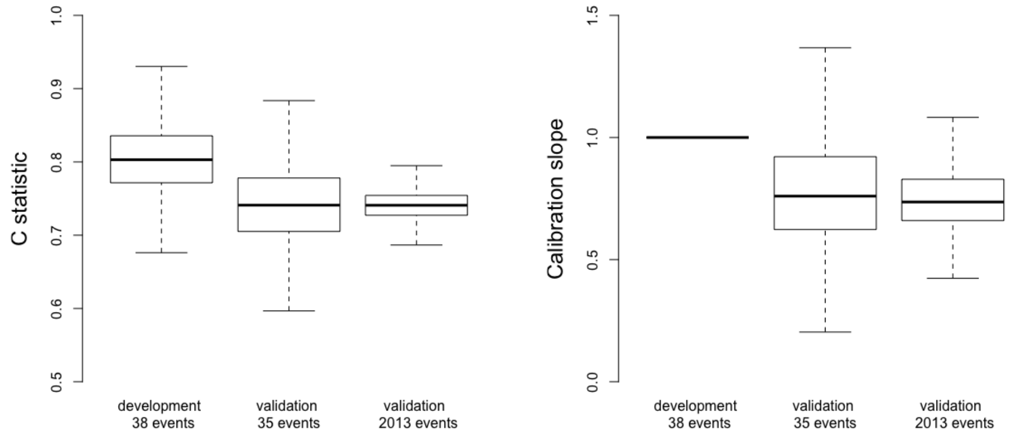
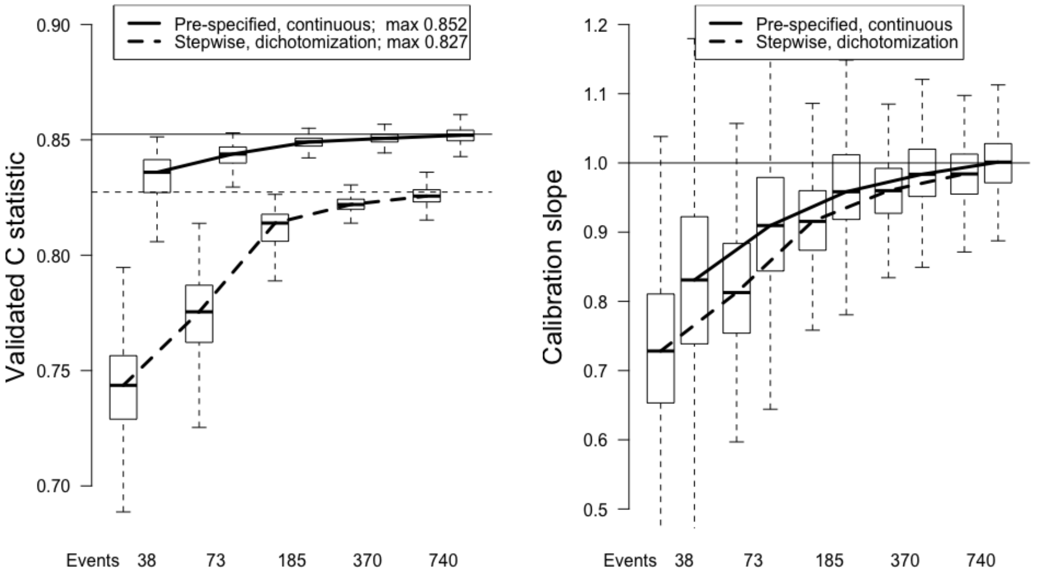


Figure 4 Impact of number of events in the development sample on estimates of model performance. Two modeling strategies were compared: stepwise selection with dichotomization versus pre-specification with continuous predictors. A total of 740 events corresponds to 20 events per variable (EPV=740/37) for the strategy with stepwise selection and dichotomization, and to 82 (EPV=740/9) for the strategy with pre-specified, continuous variables. A total of 185 events corresponds to EPV=5 and EPV=21 respectively.



Supplementary material

Table S1 Frequency of methodological issues in the development and validation of clinical prediction models in some recent systematic reviews (2008 – 2016)

First author	Year	Field	N models*	Significance testing for selection	Categorization	EPV<10
Mushkudiani [1]	2008	TBI	31	61%	79%**	NA
Altman [2]	2009	Breast cancer	53	57%	74%	NA
Mallett [3]	2010	Cancer	43	86%	97%	30%
Collins [4]	2011	Diabetes	39	56%	63%	21%
Bouwmeester [5]	2012	High IF papers	48	66%	80%	50%
Collins [6]	2013	Chronic kidney disease	14	57%	62%	17%

EPV: Events per variable
NA: not applicable, not clear from the review
* Total models in review; percentages refer to studies with item evaluated
** 22/28 models categorized age

Table S2 Overview of a selection of methodological studies considering statistical testing for model specification, categorization of continuous variables, and general modeling strategies.

First author	Year	Field	Key findings and conclusions
Statistical testing and stepwise selection			
Altman [7]	1989	primary biliary cirrhosis	Using 100 bootstrap samples using 17 candidate variables, the most frequently selected variables were those selected in the original analysis. Bootstrap confidence intervals were constructed for the estimated probability of surviving two years, which were markedly wider than those obtained from the original model.
Derksen [8]	1992	-	A Monte Carlo study was reported on the frequency with which authentic and noise variables are selected by automated subset algorithms. Results indicated that: (1) the degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model; (2) the number of candidate predictor variables affected the number of noise variables that gained entry to the model; (3) the size of the sample was of little practical importance in determining the number of authentic variables contained in the final model; and (4) the population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model.
Steyerberg [9]	1999	acute myocardial infarction	Bias by stepwise selection was studied with logistic regression in the GUSTO-I trial (40,830 patients). Random samples were drawn that included 3, 5, 10, 20, or 40 events per variable (EPV). Considerable overestimation of regression coefficients of selected covariables was found.
Austin [10]	2004	acute myocardial infarction	Using 1,000 bootstrap samples, backward elimination identified 940 unique models from 29 candidate variables for predicting mortality. Automated variable selection methods result in models that are unstable and not reproducible

Categorizing continuous variables			
MacCallum [11]	2002	-	The consequences of dichotomization for measurement and statistical analyses are illustrated and discussed. Dichotomization is rarely defensible and often will yield misleading results.
Irwin [12]	2003	Marketing	Marketing researchers frequently split (dichotomize) continuous predictor variables into two groups, as with a median split, before performing data analysis. The authors present the effect of dichotomizing continuous predictor variables with various nonnormal distributions and examine the effects of dichotomization on model specification and fit in multiple regression. The authors conclude that dichotomization has only negative consequences and should be avoided.
Altman [13]	2006	primary biliary cirrhosis	A prognostic model with bilirubin as a continuous explanatory variable explained 31% more of the variability in the data than when bilirubin distribution was split at the median.
Royston [14]	2006	primary biliary cirrhosis	Dichotomization may create rather than avoid problems, notably a considerable loss of power and residual confounding. In addition, the use of a data-derived 'optimal' cutpoint leads to serious bias. Dichotomization of continuous data is unnecessary for statistical analysis and in particular should not be applied to explanatory variables in regression models.
Naggara [15]	2011	unruptured intracranial aneurysms	Dichotomization leads to a considerable loss of power and incomplete correction for confounding factors. The use of data-derived “optimal” cut-points can lead to serious bias and should at least be tested on independent observations to assess their validity. Categorization of continuous data, especially dichotomization, is unnecessary. Continuous explanatory variables should be left alone in statistical models.
Dawson [16]	2012	Medical decision making	Many decisions are discrete: to admit a patient or not, to apply treatment or not. But models for understanding these decision problems must reflect our best science about the world, in which most causes and effects are continuous and not discrete. Dichotomization

			of continuous variables is strongly discouraged. If authors choose to present research findings in which dichotomization has been used, the authors must present evidence that the approach is superior to using the original continuous variable in this particular instance.
Collins [17]	2016		Categorising continuous predictors produces models with poor predictive performance and poor clinical usefulness. Categorising continuous predictors is unnecessary, biologically implausible and inefficient and should not be used in prognostic model development.
Modeling strategy			
Chatfield [18]	1995	-	Model uncertainty is caused by formulating, fitting, and checking a model on data in an iterative and interactive way. Model uncertainty leads to too narrow confidence and prediction intervals and bias in parameter estimates.
Steyerberg [19]	2000	acute myocardial infarction	Stepwise selection with a low alpha (for example, 0.05) led to a relatively poor model performance, when evaluated on independent data. Substantially better performance was obtained with full models with a limited number of important predictors, where regression coefficients were reduced with a shrinkage method. Incorporation of external information for selection and estimation improved the stability and quality of the prognostic models. Shrinkage methods in full models including prespecified predictors are recommended with incorporation of external information.
Babyak [20]	2004	-	Three common practices—automated variable selection, pretesting of candidate predictors, and dichotomization of continuous variables—are shown to pose a considerable risk for spurious findings in models. Alternative means of guarding against overfitting are discussed, including variable aggregation and the fixing of coefficients a priori. Techniques that account and correct for complexity, including shrinkage and penalization, are important in model development.

Table S3 Multivariable logistic regression model for all candidate predictors as considered for the MMRpredict model fitted in 19,866 probands with CRC.

Predictors	Coefficient	SE	p-value
Proband			
male gender	0.73	0.06	<0.0001
synchronous CRC	0.97	0.09	<0.0001
synchronous Other	1.23	0.13	<0.0001
Endometrial cancer	2.25	0.12	<0.0001
CRC agelt50	1.28	0.06	<0.0001
Endo agelt50	1.04	0.17	<0.0001
Other agelt50	0.01	0.18	0.94
Family history			
CRC			
CRC FDR ageht50	0.34	0.10	0.0004
CRC FDR agelt50	1.72	0.10	<0.0001
N FDR with CRC	0.35	0.05	<0.0001
CRC SDR ageht50	-0.20	0.10	0.042
CRC SDR agelt50	0.90	0.10	<0.0001
N SDR with CRC	0.24	0.05	<0.0001
Endometrial cancer			
Endo FDR ageht50	0.46	0.27	0.093
Endo FDR agelt50	0.59	0.29	0.040
N FDR with Endo	0.44	0.23	0.060
Endo SDR ageht50	0.21	0.35	0.54
Endo SDR agelt50	0.51	0.36	0.16
N SDR with Endo	0.12	0.28	0.66
Stomach cancer			
Stomach FDR ageht50	0.13	0.44	0.76
Stomach FDR agelt50	0.67	0.50	0.18
N SDR with Stomach	-0.13	0.38	0.73
Stomach SDR ageht50	0.61	0.47	0.19
Stomach SDR agelt50	1.35	0.53	0.011
N SDR with Stomach	-0.62	0.43	0.15
Urigenital cancer			
Urigenital FDR ageht50	2.22	0.81	0.006
Urigenital FDR agelt50	1.60	0.86	0.063
N FDR with Urigential	-1.88	0.78	0.016
Urigenital SDR ageht50	-0.52	0.58	0.38
Urigenital SDR agelt50	-1.00	0.75	0.18
N SDR with Urigenital	0.67	0.51	0.19
Other cancers			
Other FDR ageht50	-0.11	0.19	0.54
Other FDR agelt50	0.53	0.21	0.012
N FDR with Other	0.21	0.15	0.15
Other SDR ageht50	-0.06	0.20	0.78
Other SDR agelt50	0.22	0.26	0.40
N SDR with Other	0.06	0.16	0.69

FDR: First degree relative; SDR: Second degree relative; ageht50: age over 50; agelt50: age lower than 50.
The logistic regression model had 37 degrees of freedom. The c statistic was 0.833 [95% CI 0.823 – 0.843] in the full development set with n=19,866 and 2,051 events.

R code for key analyses

```
# draw random development samples
row.y1 <- sample(y1.rows, j)          # events, j==38
row.y0 <- sample(y0.rows, controls)   # non-events, controls ==870 – j

# Start univar screening in sel.x, varlist is list of candidate predictors
  for (p in (1:(length(varlist)))) {
      uni.fit <- lrm.fit(y=sel.y, x=sel.x[,p], tol=1e-2, maxit=20)
      p.cand[p] <- ifelse(uni.fit$fail,.99,uni.fit$stats[5])  }
 # End univar screen

# list of univar p < threshold; threshold == 0.05
list.cand.s <- ifelse(p.cand < p.threshold,T,F)

# make full data and selected data set
sel.data.full <- as.data.frame(cbind(fit.NEJM$y, xstart[,list.cand.s]))
sel.data      <- as.data.frame(cbind(sel.y, sel.x[,list.cand.s]))


sel.fit.full  <- lrm(V1~., data=sel.data.full, x=T, y=T, maxit=199)
sel.fit       <- lrm(V1~., data=sel.data, x=T, y=T, maxit=199)


# fastbw does the backward stepwise selection
selbw <- fastbw(sel.fit, type = "individual", rule = "p") # Stepwise, p<.05


# Fit stepwise selected models, from univariate selection
selbw.fit.full  <- lrm.fit(y=sel.fit.full$y,
 x=sel.fit.full$x[,selbw$factors.kept], maxit=199)

# this is the fit to be considered for validation performance, bw in small
 sample
selbw.fit       <- lrm.fit(y=sel.fit$y, x=sel.fit$x[,selbw$factors.kept],
 maxit=199)


# Validate in independent data, j3 indicated rows of small subsample
pval = as.matrix(sel.fit.full$x[-j3, selbw$factors.kept]) %*%
 selbw.fit$coefficients[-1]
val.prob(y=sel.fit.full$y[-j3], logit=pval, pl=F)
```

References Supplementary material

[1] Mushkudiani NA, Hukkelhoven CW, Hernandez 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. Journal of clinical epidemiology. 2008;61:331-43.
[2] Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer investigation. 2009;27:235-43.
[3] Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC medicine. 2010;8:20.
[4] Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC medicine. 2011;9:103.
[5] 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 medicine. 2012;9:1-12.
[6] Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of clinical epidemiology. 2013;66:268-77.
[7] Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Statistics in medicine. 1989;8:771-83.
[8] Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. 1992;45:265-82.
[9] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. Journal of clinical epidemiology. 1999;52:935-42.
[10] Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. Journal of clinical epidemiology. 2004;57:1138-46.
[11] MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7:19-40.
[12] Irwin JR, McClelland GH. Negative Consequences of Dichotomizing Continuous Predictor Variables. Journal of Marketing Research. 2003;40:366-71.
[13] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006;332:1080.
[14] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25:127-41.
[15] Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. American Journal of Neuroradiology. 2011;32:437-40.
[16] Dawson NV, Weiss R. Dichotomizing Continuous Variables in Statistical Analysis. Medical Decision Making. 2012;32:225-6.
[17] Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in medicine. 2016;35:4124-35.
[18] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc, Ser A. 1995;158:419-66.
[19] Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in medicine. 2000;19:1059-79.
[20] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411-21.

We declare that we have not conflicts of interest