

REVIEW

Open Access

RiGoR: reporting guidelines to address common sources of bias in risk model development

Kathleen F Kerr^{1*}, Allison Meisner¹, Heather Thiessen-Philbrook², Steven G Coca³ and Chirag R Parikh⁴

Abstract

Reviewing the literature in many fields on proposed risk models reveals problems with the way many risk models are developed. Furthermore, papers reporting new risk models do not always provide sufficient information to allow readers to assess the merits of the model. In this review, we discuss sources of bias that can arise in risk model development. We focus on two biases that can be introduced during data analysis. These two sources of bias are sometimes conflated in the literature and we recommend the terms *resubstitution bias* and *model-selection bias* to delineate them. We also propose the RiGoR reporting standard to improve transparency and clarity of published papers proposing new risk models.

Keywords: Risk prediction, Reporting standards, Research design, Statistical bias

Introduction

There is currently broad interest in developing risk prediction models in medicine. However, recent reviews in a variety of fields have described a substantial number of flaws in the way risk models are developed and/or deficiencies in the way the work is reported [1-6]. An extensive review that spanned many fields of application found the vast majority of papers reporting risk models omitted important details such as: the extent and handling of missing data; key information on the study population; and the precise definition of the outcome or event of interest [6]. An evaluation of model calibration was typically absent. Additional issues include a tendency for models to be favorably evaluated when the model's developers are involved in validating the model [4].

Research reports related to risk prediction sometimes refer to “optimistic bias” or “optimism bias” [4,7,8]. Unfortunately, these terms are used to refer to a variety of problems in risk model development or assessment. It would be useful to have clear, distinctive, and descriptive names for different sources of bias that can affect scientific results. The first goal of this review is to propose terminology for referring to two sources of bias that are common in developing risk models. Both biases can arise

during data analysis, which makes them avoidable, at least in principle. The second goal of this paper is a proposal for a set of guidelines for reporting proposed new risk models. The guidelines should help readers evaluate the merits of new risk models and understand whether developers were attentive to avoiding common sources of bias.

Review

Common sources of bias in risk model development

Currently, two sources of bias that arise in developing risk prediction models from combinations of biomarkers and/or clinical variables are both called “optimistic bias.” We propose the terms “resubstitution bias” and “model-selection bias” as more precise and descriptive terms than “optimistic bias.” A predictive model will tend to perform better on the data that were used to fit or “train” the model than on new data. Resubstitution bias arises when the data that are used to fit a predictive model are used a second time to assess the performance of the model. Re-using the data in this way has been called resubstitution [9-15], so it is a modest extension to refer to the resulting bias as resubstitution bias. Since the ultimate goal of a risk prediction model is to estimate risks on new individuals, assessing model performance via resubstitution does not provide an unbiased or “honest” estimate of the model's predictive capacity.

Model-selection bias arises when many models are assessed, and the best performing model is reported.

* Correspondence: katiek@uw.edu

¹Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, USA

Full list of author information is available at the end of the article

This optimistic bias persists even if analysts have corrected for resubstitution bias in assessing the model. Occasionally, investigators have a single, pre-specified model that they fit with data. In this case, the resulting model is susceptible to resubstitution bias but not to model-selection bias. More typically, however, data analysts have a set of candidate predictors to choose from, which translates to a set of possible models. For example, if there are k candidate predictors and an analyst limits the set of possible models to linear logistic models, then the number of possible models is $2^k - 1$. For 20 candidate predictors this is over 1 million models, and we expect some of these to perform well by chance. Naïve assessment of the best-performing model is likely optimistic because this model is chosen because it performed best on the available data [16-19]. “Model-selection bias” refers to this particular source of optimism.

Although resubstitution bias and model-selection bias are well-known phenomena among methodologists and many data analysts [7,10,20], there is no standard terminology for referring to these sources of bias. We find the term “optimistic bias” inadequate for several reasons. “Optimistic bias” describes the direction of the bias and not the source of the bias, so it is insufficiently descriptive. By referring to multiple sources of bias with the same phrase researchers might claim to have addressed “optimistic bias” in developing a predictive model [21], when in fact they have only addressed one source of bias. Finally, in addition to resubstitution and model-selection there are additional phenomena referred to as “optimistic bias” [8], including the observation in psychology that people often underestimate personal risks [22].

We emphasize that resubstitution bias and model-selection bias are well-known among methodologists, and our modest contribution is our proposal for standard terms to refer to these issues. These terms have appeared occasionally (and rarely) in the literature [17,23,24] but are not in widespread use. Previously proposed terminology is “parameter uncertainty” and “model uncertainty” [20], where “model uncertainty” is said to lead to “selection bias.” However, this terminology is not standard and we find it less descriptive than the terms we propose. Furthermore, “selection bias” has an established meaning in epidemiology, where it refers to non-representative selection of study subjects.

Methodology for estimating the performance of a risk model that is not optimistically biased from resubstitution includes bootstrapping techniques, cross-validation, and using independent datasets for model development and validation [7,25]. Bootstrapping and cross-validation are computationally intensive, and employing them can surpass the abilities of some data analysts or software packages. Moreover, there are different varieties of bootstrapping and cross-validation and a lack of consensus

on the best procedure. A recent investigation [21] provides some much-needed guidance on the relative merits of different procedures for estimating the area under the ROC curve (AUC or “C statistic”) without resubstitution bias. Using independent datasets for model development and validation is computationally simpler, and provides stronger evidence in favor of a reported risk model if the validation dataset is from a separate study (“external validation”). More commonly, however, a validation dataset is created by data splitting – randomly partitioning the available data into a “training” dataset and a “test” or validation dataset. This strategy offers simplicity and flexibility in data analysis, but is criticized for its statistical inefficiency [7] because only part of the data inform development of the risk model. With data splitting there is an inherent tension between the amount of data allotted to the training dataset for developing the risk model, and the amount of data allotted to the test dataset for evaluating the risk model [26]. If the training dataset is too small a good risk model might not be found. On the other hand, if the test dataset is too small then estimates of model performance, while unbiased, are highly variable, making promising results less compelling. An advantage of having an independent validation dataset is that both resubstitution and model-selection bias are accounted for as long as the validation dataset is not used in any stage of model development, including variable selection.

Model-selection bias tends to be more difficult to address without an independent validation dataset. In principle, model-selection can be incorporated into a bootstrapping or cross-validation procedure, but this requires the use of an automated model-building process and further increases the computational complexity of using these methods.

Reporting standards

There have been several efforts to develop standards and guidelines for reporting various types of scientific studies. As summarized by McShane and colleagues [27], quality study reporting “cannot transform a poorly designed or analyzed study into a good one, but it can help to identify the poor studies.” Quality reporting is an important first step in improving the overall quality of risk model development work [4,6]. Reporting standards can additionally help guide researchers toward best practices. Table 1 presents our proposed RiGoR (Reporting Guidelines for Risk Models) Standards for reporting risk model development.

Previously published reporting standards that are related to risk model development are STARD [28], GRIPS [29], and REMARK [27]. The STARD initiative [28] assembled a comprehensive set of standards “to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in

Table 1 RiGoR: reporting guidelines for risk models

Section and topic	Item	Similar items		
		STARD	REMARK	GRIPS
Title/Abstract/Keywords	1	1	1	1
Introduction	2	2	20	2
Methods				
Participants	3	3	2	4,5,14
	4a	4		5
	4b	14	6	4
	5	5	6	4
Biomarker Data	6	8	4,5	
	7		4,5	7
	8	23		
Outcome variable	9		7	6
Statistical Methods	10	12		12
	11	9	11	8
	12a		8	9
	12b		10	9
	12c			10
	13		10	
	14a			10
	14b			10
	14c			10
	15			
	16	22		11
	17			
Results	18	15	13	15
	19			

Table 1 RiGoR: reporting guidelines for risk models (Continued)

	20	Report estimates of model performance with measures of uncertainty when possible (e.g., 95% confidence interval)	21	18,19
	21	Assess and report evidence of risk model calibration.		
Discussion	22	Discuss prospects of final risk model for satisfying the research goal	25	22, 23
	23	Discuss known and possible limitations to generalizability or applicability of risk model	19	21

the study (internal validity) and to evaluate its generalisability (external validity)" (<http://www.stard-statement.org/>). A primary result of the initiative is a 25 item checklist for articles reporting studies of diagnostic accuracy. The RiGoR guidelines are meant to emulate the contribution of STARD with a set of criteria tailored to the development of risk prediction instruments. The REMARK recommendations [27] were developed in the context of tumor markers with the potential to be used for prognosis. The focus of REMARK is markers for predicting time-to-event outcomes such as overall survival. In contrast, the focus of RiGoR is estimating patient risks of a binary outcome. The GRIPS statement [29] offers reporting standards focused on studies of risk prediction models that include genetic variants. The RiGoR guidelines are more general and more detailed.

In proposing the RiGoR standards, we both acknowledge and build upon the important previous efforts described above. For each RiGoR item, Table 1 notes similar STARD, REMARK, or GRIPS items. As Table 1 shows, most items are similar to criteria given in at least one of these previous reports. However, there are some notable exceptions. First, RiGoR includes a guideline that the calibration of a risk prediction model should be assessed and reported, as calibration is a necessary requirement for the validity of a model. While the importance of calibration is noted in many publications [6,30,31], it is not included in GRIPS. Second, our guidelines explicitly address resubstitution bias and model-selection bias, two common types of bias that can arise during risk model development.

There are items in the REMARK and GRIPS guidelines that are not included in RiGoR. In Appendix A we document our reasons for excluding these items.

Conclusions

In Epidemiology, common pitfalls in study design and data analysis commonly acquire standard names. Some examples include *immortal time bias* in survival analysis [32] and *lead time bias* in the evaluation of diagnostic screening tools [33]. *Publication bias* is a widely recognized issue in the scientific literature [34]. The most helpful terminology is descriptive; helps codify important concepts; and aids scientific communication. We believe the terms "resubstitution bias" and "model-selection bias" accomplish these goals.

In this article we have reviewed and discussed resubstitution bias and model-selection bias. We do not mean to suggest that they are the only two sources of bias that can affect risk model development. However, we believe resubstitution bias and model-selection bias deserve special attention because they are common. Furthermore, they are biases that arise during data analysis, which means, at least in principle, that they should be avoidable with use of proper methods of data analysis.

Other types of bias can enter into a study at earlier stages. For example, selection bias can inflate the performance of a proposed risk model if the cases in the dataset tend to be more severe than the population of cases, or controls tends to be healthier than the population of controls. Having an objective way to define the population of interest and to define the event of interest is an important aspect of a quality study. The RiGoR standards are designed to ensure that these and other important aspects of study design, conduct, and data analysis are documented.

Appendix A

We comment on items in the GRIPS and REMARK guidelines that do not appear in the RiGoR guidelines. Our purpose here is explain the rationale of the RiGoR guidelines and our comments should not be interpreted as criticisms of these important, previous efforts.

GRIPS

3. "Specify the study objectives and state the specific model(s) that is/are investigated. State if the study concerned the development of the model(s), a validation effort, or both."

The RiGoR guidelines apply to papers documenting the development of risk models.

13. "Describe all subgroups, interactions, and exploratory analyses that were examined."

20. "Present results of any subgroup, interaction, or exploratory analyses, whenever pertinent."

Item 13 is very broad, and item 20 is vague as to what constitutes "pertinent" results. In the RiGoR guidelines, we focus on sources of bias (e.g. model-selection bias and resubstitution bias) that are common in risk model development and provide specific guidelines to identify them.

17. “Report distributions of predicted risks and/or risk scores.”

Such distributions are one way to evaluate model performance. The RiGoR guidelines provide flexibility on metrics by which to evaluate risk models.

24. “State whether databases for the analyzed data, risk models, and/or protocols are or will become publicly available and if so, how they can be accessed.”

We did not include a similar item in the RiGoR guidelines because it is not crucial for assessing the study that was done.

25. “Give the source of funding and the role of the funders for the present study. State whether there are any conflicts of interest.”

We did not include an item like this in the RiGoR guidelines because it is not specific to risk model research.

REMARK

3. “Describe treatments received and how chosen (e.g. randomized or rule-based).”

This item is specific to prognosis in cancer.

9. “Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.”

This guideline is more appropriate for studies of association.

12. “Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.”

This item is not widely applicable to risk model development.

14. “Show the relation of the marker to standard prognostic variables.”

15. “Present univariate analyses showing the relation between the marker and outcome, with the estimated effect (e.g. hazard ratio and survival probability.) Preferably provide similar analyses for all other variables being analyzed. For the effect of a tumor marker on a time-to-event outcome, a Kaplan-Meier plot is recommended.”

16. “For key multivariable analyses, report estimated effects (e.g. hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.”

17. “Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variable are included, regardless of their statistical significance.”

Such information is tangential to the validity and utility of a risk model. Furthermore, in the context of extensive variable or model selection, estimated effects may be biased.

18. “If done, report results of further investigations, such as checking assumptions, sensitivity analyses, and internal validation.”

This guideline is somewhat broad and open-ended, so we chose not to include a similar item in the RiGoR guidelines.

Competing interests

The authors declared that they have no competing interests.

Authors' contributions

KFK drafted the manuscript. All authors contributed to the development of the reporting standards and reviewed and approved the final manuscript.

Acknowledgement

The research was supported by the NIH grant R01HL085757 (CRP) to fund the TRIBE-AKI Consortium to study novel biomarkers of acute kidney injury in cardiac surgery. CRP is also supported by NIH grant K24DK090203. SGC is supported by National Institutes of Health Grants K23DK080132 and R01DK096549. SGC and CRP are also members of the NIH-sponsored ASsess, Serial Evaluation, and Subsequent Sequelae in Acute Kidney Injury (ASSESS-AKI) Consortium (U01DK082185).

Author details

¹Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, USA. ²Kidney Clinical Research Unit Room ELL-101, Westminster Tower London Health Sciences Centre, 800 Commissioners Road East, London, ON N6C 6B5, Canada. ³Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1243, New York, NY 10029, USA. ⁴Yale University School of Medicine Program of Applied Translational Research, Temple Street, Suite 6C, New Haven, CT 06510, USA.

Received: 27 November 2014 Accepted: 24 December 2014

Published online: 24 January 2015

References

1. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst.* 2005;97(14):1043–55.
2. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Quality of reporting of cancer prognostic marker studies: association with reported prognostic effect. *J Natl Cancer Inst.* 2007;99(3):236–43.
3. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med.* 1993;118(3):201–10.
4. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ.* 2012;344:e3318.
5. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstant A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat.* 2012;133(1):1–10.
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. *J Clin Epidemiol.* 2013;66(3):268–77.
7. Harrell FEJ. *Regression Modeling Strategies.* New York: Springer; 2001.
8. Hammond T, Verbyla D. Optimistic bias in classification accuracy assessment. *Int J Remote Sens.* 1996;7(6):1261–6.
9. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst.* 2007;99(2):147–57.
10. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* 2005;21(15):3301–7.
11. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br J Cancer.* 2003;89(9):1599–604.
12. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? *Contemp Clin Trials.* 2013;36(2):636–41.
13. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics.* 2010;26(6):822–30.
14. Braga-Neto U, Hashimoto R, Dougherty ER, Nguyen DV, Carroll RJ. Is cross-validation better than resubstitution for ranking genes? *Bioinformatics.* 2004;20(2):253–8.

15. Way TW, Sahiner B, Hadjiiski LM, Chan HP. Effect of finite sample size on feature selection and classification: a simulation study. *Med Phys*. 2010;37(2):907–20.
16. Ding Y, Tang S, Liao SG, Jia J, Oesterreich S, Lin Y, et al. Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics*. 2014;30(22):3152–8.
17. Berrar D, Bradbury I, Dubitzky W. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*. 2006;22(10):1245–50.
18. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
19. Tibshirani RJ, Tibshirani R. A bias-correction for the minimum error rate in cross-validation. *Ann Appl Stat*. 2009;3(2):822–9.
20. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.
21. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. *Am J Epidemiol*. 2014;180(3):318–24.
22. Fontaine KR, Smith S. Optimistic bias in cancer risk perception: a cross-national study. *Psychol Rep*. 1995;77(1):143–6.
23. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J Clin Oncol*. 2004;22(14):2790–9.
24. Hathaway B, Landsittel DP, Gooding W, Whiteside TL, Grandis JR, Siegfried JM, et al. Multiplexed analysis of serum cytokines as biomarkers in squamous cell carcinoma of the head and neck patients. *Laryngoscope*. 2005;115(3):522–7.
25. Kerr KF, Meisner A, Thiessen-Philbrook H, Coca SG, Parikh CR. Developing risk prediction models for kidney injury and assessing incremental value for novel biomarkers. *Clin J Am Soc Nephrol*. 2014;9(8):1488–96.
26. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genomics*. 2011;4:31.
27. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Diagnostics SSoT-N-EWGoC: REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nat Clin Pract Oncol*. 2005;2(8):416–22.
28. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med*. 2003;138(1):40–4.
29. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ, Group G. Strengthening the reporting of Genetic Risk Prediction Studies: the GRIPS Statement. *PLoS Med*. 2011;8(3):e1000420.
30. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–38.
31. Pepe M, Janes H. Methods for Evaluating Prediction Performance of Biomarkers and Tests. In: Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, editors. *Risk Assessment and Evaluation of Predictions*. New York: Springer; 2013. p. 107–42.
32. Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf*. 2007;16(3):241–9.
33. Hutchison GB, Shapiro S. Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst*. 1968;41(3):665–81.
34. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263(10):1385–9.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

