Methodological issues in current practice may lead to bias in the development of biomarker combinations for predicting acute kidney injury

Allison Meisner[1], Kathleen F. Kerr[1], Heather Thiessen-Philbrook[2], Steven G. Coca[3] and Chirag R. Parikh[2,4]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA; [2]Section of Nephrology and Program of Applied Translational Research, Department of Medicine, New Haven, Connecticut, USA; [3]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA and [4]VA Medical Center, West Haven, Connecticut, USA

Individual biomarkers of renal injury are only modestly predictive of acute kidney injury (AKI). Using multiple biomarkers has the potential to improve predictive capacity. In this systematic review, statistical methods of articles developing biomarker combinations to predict AKI were assessed. We identified and described three potential sources of bias (resubstitution bias, model selection bias, and bias due to center differences) that may compromise the development of biomarker combinations. Fifteen studies reported developing kidney injury biomarker combinations for the prediction of AKI after cardiac surgery (8 articles), in the intensive care unit (4 articles), or other settings (3 articles). All studies were susceptible to at least one source of bias and did not account for or acknowledge the bias. Inadequate reporting often hindered our assessment of the articles. We then evaluated, when possible (7 articles), the performance of published biomarker combinations in the TRIBE-AKI cardiac surgery cohort. Predictive performance was markedly attenuated in six out of seven cases. Thus, deficiencies in analysis and reporting are avoidable, and care should be taken to provide accurate estimates of risk prediction model performance. Hence, rigorous design, analysis, and reporting of biomarker combination studies are essential to realizing the promise of biomarkers in clinical practice.

Kidney International advance online publication, 23 September 2015; doi:10.1038/ki.2015.283

KEYWORDS: biomarkers; combinations; diagnosis; NGAL; methods; prognosis

Correspondence: *Chirag R. Parikh, Program of Applied Translational Research, Department of Medicine, New Haven, Connecticut, USA and VA Medical Center, 60 Temple Street, Suite 6C, New Haven, Connecticut 06510, USA. E-mail: chirag.parikh@yale.edu*

Received 19 February 2015; revised 27 July 2015; accepted 31 July 2015

Acute kidney injury (AKI) is a frequent complication of hospitalized patients, particularly following cardiac surgery and critical illness.[1] AKI is associated with increased morbidity and mortality.[2,3] There is great interest in using biomarkers to predict risk of AKI, for several reasons. AKI is typically diagnosed based on changes in serum creatinine, a marker of renal function rather than injury,[4,5] which contributes to frequent delayed diagnosis or misdiagnosis.[5] It may be possible to use biomarkers to diagnose AKI earlier and/or more accurately than is possible with serum creatinine.[6] Biomarkers may also have an important role within the context of creatinine-defined AKI. When serum creatinine is used to diagnose AKI, the diagnosis is generally not made until several days after the injury, potentially too late to intervene.[7] It may be possible to use biomarkers to predict AKI prior to changes in serum creatinine, opening a therapeutic window. If biomarkers can be shown to accurately predict AKI, they could be used as inclusion criteria to enrich clinical trials or serve as intermediate outcomes.[7,8] Biomarkers that can accurately predict AKI and related complications could also potentially advance clinical care.[8,9]

Much work has been carried out to study associations between individual biomarkers and AKI.[8,10,11] Although many associations are strong and well-established, the predictive performance of these markers has been modest. AKI is a complex disease, and many possible modes of injury exist even in the relatively homogeneous setting of cardiac surgery.[1] Consequently, interest now centers on identifying combinations of injury markers that can predict AKI; such a strategy has been recommended in several reviews.[9,12–15]

The goals of this article are to provide an overview of current statistical practice in developing biomarker combinations for AKI and to discuss common issues surrounding the conduct of these analyses. In particular, we will consider the role of three potential sources of bias frequently encountered in the statistical evaluation of biomarker combinations: resubstitution bias, model selection bias, and bias due to center differences.

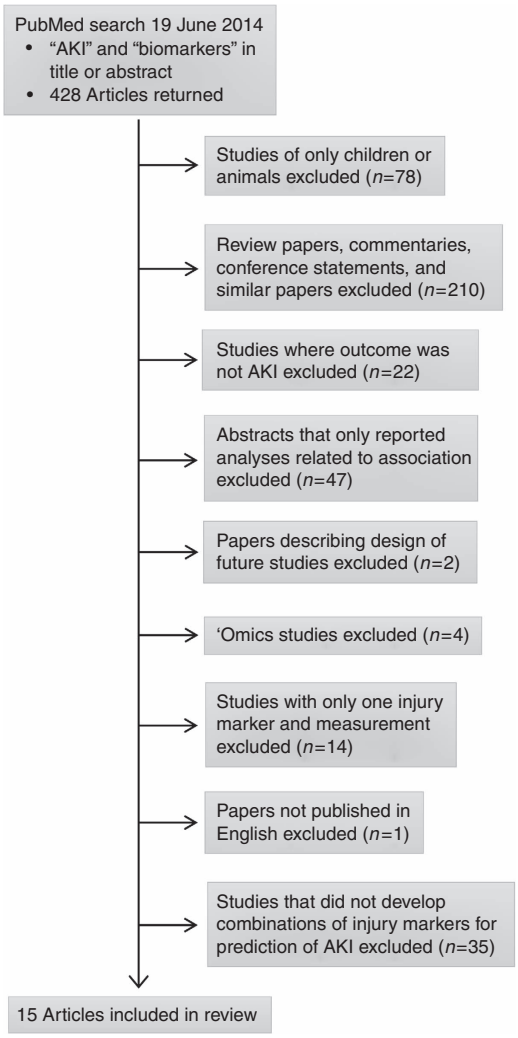

Figure 1 | Overview of study selection. AKI, acute kidney injury.

Resubstitution bias and model selection bias have previously been discussed at length.[16,17] Briefly, resubstitution bias arises when a data set is used to fit a predictive model, and then the model's performance is assessed by its apparent performance on the same data set—i.e., the data are 'resubstituted' into the model. Model selection bias results when several models are evaluated and the model with the best performance is chosen. Both resubstitution and model selection optimistically bias estimates of model performance unless methods are used to account for them. Note that resubstitution bias and model selection bias are widely known[18,19] but without standard terminology. These biases are commonly referred to jointly as 'optimistic bias,' but it is useful to distinguish the two sources of bias with separate labels.[17] Bias due to center differences can arise in studies involving multiple centers. In particular, differences by center can confound the estimate of model performance, biasing the results in either direction.[20] A challenge here is that not all differences among centers represent bias. For example, if one center tends to get sicker patients, and those patients tend to have both worse outcomes and correspondingly higher levels of an injury marker, this in itself does not present bias. However, suppose the center that tends to get sicker patients also uses different protocols for fluid administration that tends to either increase or decrease the measurement of a biomarker. Then the association of the biomarker with the outcome will either be over- or underestimated if data are simply pooled across centers.

Model selection bias and resubstitution bias are of particular concern in the development of biomarker combinations: when many marker measurements are available, both the size of the combination (number of marker measurements included in the combination) and the number of combinations considered may be quite large. Resubstitution bias is generally larger when the number of predictors in the model is large relative to the amount of data. Model selection bias is most worrisome when many models are considered.

The prevalence of these biases will be assessed through a literature review, and their potential impact will be explored by assessing the performance of published combinations in a large, independent study of AKI in cardiac surgery patients.

RESULTS

Literature search and study selection

Figure 1 summarizes our literature search. Briefly, 428 articles were screened, yielding 15 articles[10,21–34] after the exclusion criteria were applied.

Data extraction

Table 1 summarizes the 15 selected articles, with additional details provided in Supplementary Table 1 online. Eight of 15 articles (53.3%) were in the setting of cardiac surgery. All 15 articles relied on serum creatinine to define AKI. Table 2 presents the data related to potential sources of bias. None of the 15 papers explicitly stated the number of models considered; the numbers in Table 2 are likely to be a lower bound. It was often challenging to determine how the combination(s) presented was chosen and/or how the combination(s) was estimated.

Evaluation of biases

As indicated in Table 2, all papers were likely affected by at least one source of bias. Importantly, the reported performance of the combinations was generally good: in most cases, the area under the receiver operating characteristic curve (AUC) was above 0.8, and in a third of papers it exceeded 0.9.

In nearly all articles, the same data were used to fit and evaluate the models. In other words, most articles did not account for resubstitution bias. Furthermore, four papers had fewer than 10 events per marker in the final combination; in three papers, there were fewer than 15 events in total. In Parikh *et al.*[10] and Parikh *et al.*,[32] threefold cross-validation was used to address resubstitution bias. Cross-validation is a reasonable approach, although variants other than threefold