HHS Public Access

Author manuscript

Kidney Int. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Kidney Int. 2016 February ; 89(2): 429–438. doi:10.1038/ki.2015.283.

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

Allison Meisner, MA[1], **Kathleen F. Kerr, PhD**[1], **Heather Thiessen-Philbrook, MMath**[2], **Steven G. Coca, DO, MS**[3], and **Chirag R. Parikh, MD, PhD**[4]

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

[2]Kidney Clinical Research Unit Room ELL-101, Westminster Tower, London Health Sciences Centre, 800 Commissioners Road East, London, ON, Canada, N6C 6B5

[3]Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1243, New York, NY 10029

[4]Program of Applied Translational Research, Yale University School of Medicine and VA Medical Center, 60 Temple Street, Suite 6C, New Haven, CT 06510

Abstract

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 acute kidney injury 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.

Correspondence: Dr. Chirag R. Parikh, Section of Nephrology, Yale University School of Medicine, 60 Temple Street, Suite 6C, New Haven, CT 06510. ; Email: chirag.parikh@yale.edu. Telephone: 203-737-2676. Fax: 203-764-8373

Disclosures: None

Author Contributions

AM, SGC, HTP, KFK and CRP conceived of the review and analysis. AM conducted the literature review and performed the analysis. HTP assisted in extracting information from the selected articles. KFK and CRP provided statistical expertise and support for the review and analysis. All authors contributed to interpretation of the results. AM drafted the manuscript, with substantial revising by KFK, HTP, SGC and CRP. All authors approved the final version of the manuscript. The corresponding author had full access to the data and final responsibility for the decision to submit.

Keywords

acute kidney injury; cardiovascular disease

Introduction

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 play 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 done 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.

Resubstitution bias and model selection bias have previously been discussed at length ([16, 17]). Briefly, resubstitution bias arises when a dataset is used to fit a predictive model, and then the model's performance is assessed by its apparent performance on the same dataset; that is, 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 under-estimated 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. 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 (ROC) 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]), three-fold cross-validation was used to address resubstitution bias. Cross-validation is a reasonable approach, although variants other than 3-fold cross-validation have been shown to perform better ([35]). However,

the purpose of this article is not to critique every methodological choice of authors, but rather to examine whether “big picture” issues and common sources of biases were addressed.

A third of papers considered 10 models or more to arrive at the final combination(s), increasing the likelihood of model selection bias in the estimate of the performance of that combination. As noted above, the number of models reported in Table 2 is likely to be a conservative estimate. Thus, the performance results provided by some of the other papers may also be affected by model selection bias.

Three articles involved multi-center cohorts; none addressed the possibility of bias due to center differences, either in the analysis or in the discussion of limitations. Certainly, if center differences do not exist (perhaps due to careful design and/or conduct of the study) then there will be no bias; however, it is important in a multi-center study to consider whether results might be affected by center differences.

Importantly, while most articles acknowledged that the reported study had low power/sample size, no study explicitly acknowledged resubstitution bias or model selection bias as a possible limitation.

Replication in TRIBE-AKI Data

We were able to assess the performance of biomarker combinations from seven papers ([25], [27], [28], [30], [31], [33], [34]) in the Translational Research Investigating Biomarker Endpoints in AKI (TRIBE-AKI) study data ([10]). TRIBE-AKI involves 1219 adults undergoing cardiac surgery at six centers. Table 3 gives the seven articles and the performances of the published marker combinations when applied to TRIBE-AKI. The performance in TRIBE-AKI was typically more modest than the published estimate of performance. This difference was not due to a low number of events in TRIBE-AKI: for each article’s definition of AKI, there were at least 56 “AKI cases” in TRIBE-AKI (Supplementary Table 1). We note the important limitation that for 4 of these 7 articles, the study included general intensive care unit (ICU) patients, which represent a more heterogeneous population than the population of patients undergoing cardiac surgery who comprise the TRIBE-AKI study ([25], [27], [28], [33]). We report the assays used to measure the biomarkers in TRIBE-AKI and in each study in Supplementary Table 2.

We present details on the designs of the included studies, and the extent of replication in TRIBE-AKI, in Supplementary Table 1. Our goal was to match the exclusion criteria in each study when applying published marker combinations to the TRIBE-AKI data. However we were sometimes limited by a lack of detail regarding the exclusion criteria in the published articles. Furthermore, some exclusion criteria were built into the study design of TRIBE-AKI, precluding a perfect match. In some cases, we were further limited by the data collected by TRIBE-AKI. We note also that incomplete reporting often made it difficult to determine whether urine markers had been normalized to urine creatinine or whether markers were transformed. Three articles reported the estimated combination ([27], [30], [34]), while we had to re-estimate the combination in the remaining four articles.

We note that in Luo et al. ([31]), cases were matched on age, sex, and admission time. Furthermore, in Siew et al. ([33]), cases and controls were frequency-matched on categories of eGFR. Such restricted sampling may have played a role in the difference between the reported AUC and the AUC in TRIBE-AKI for these papers. Neither paper addressed this aspect of study design in evaluating the performance of the combination(s), despite the complications that matching introduces when the goal is prediction ([20, 36]).

Discussion

We have provided an overview of the current state of biomarker combinations research in the setting of AKI risk prediction, highlighting the potential role of three common sources of bias in 15 published studies. Each of the 15 papers was susceptible to at least one source of bias, and 8 were potentially affected by two sources of bias. Three of the 15 articles involved a multi-center study, and none of these three discussed the possibility of bias due to center differences. In several cases, inadequate reporting made assessment of the articles challenging and could be remedied by following recently proposed guidelines (Reporting Guidelines for Risk Models, RiGoR ([17]) and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, TRIPOD ([37])).

In two-thirds of papers, the reported AUC was quite good (above 0.8). When possible, we applied published models to TRIBE-AKI data, yielding an estimate of model performance which we could compare to the published model performance. The performance in TRIBE-AKI was more modest than the published results in 6 out of 7 cases. This in itself is not evidence of bias; differences between the TRIBE-AKI study and the published study could also explain some differences. Publication bias ([38]), a concept related to model selection bias, may also play a role. However, most studies did not account for resubstitution bias, and studies that considered many candidate models did not report addressing model selection bias. Therefore, we posit that these biases likely explain at least part of the reported estimates of good model performances. In addition, bias due to center differences may have affected the published estimates for the three studies involving multiple centers. Notably, the single study in which model performance was higher in TRIBE-AKI was the study with the lowest reported AUC by a wide margin.

We chose to focus on biases commonly encountered in the evaluation of combinations of biomarkers for risk prediction. However, we provide more extensive guidance on the design and analysis of biomarker combination studies in Table 4 ([16, 17, 37, 39–68]) and an expanded discussion of sources of bias that can affect these studies in Supplementary Table 3 ([16, 17, 37, 39–50, 52, 54–58, 60, 61, 63, 66]).

We have not addressed issues with commonly used definitions in AKI, including the use of creatinine-defined AKI and the dichotomization of continuous changes in creatinine. Important work on these topics has been done ([69–71]) and future research should focus on clinically meaningful outcomes that utilize all available information. However, since the most widely-used definitions of AKI are based on dichotomizations of changes in creatinine, our survey of current practice does not address these issues.

The Food and Drug Administration (FDA) has recently approved the use of urine [tissue inhibitor of metalloproteinases-2 (TIMP-2)]*[insulin-like growth factor-binding protein 7 (IGFBP7)] to estimate risk of developing AKI ([72]). This biomarker panel was developed in stages ([73]). In the first stage, investigators screened 340 individual biomarkers and biomarker combinations formed by multiplying concentrations of 2, 3, or 4 markers. By pre-specifying how biomarkers were to be combined, this approach had the advantage of avoiding resubstitution bias, with the possible disadvantage of a potentially large loss in predictive capacity by not allowing the data to inform the combination. In the second stage, the selected biomarker panel was evaluated in independent data, providing an estimate of model performance unaffected by model selection bias. Both the development and validation studies involved multiple centers, so the potential for bias due to center differences remains. The panel, urine [TIMP-2]*[IGFBP7], has been subsequently evaluated in cardiac surgery patients ([74]), and future investigations may further expand its application. The papers reporting the development and evaluation of this panel were excluded from the present review because the approach of selecting biomarkers and creating “supermarkers” by multiplying biomarker values is distinct from the methods typically used to develop biomarker combinations.

It is essential to thoroughly evaluate a risk model prior to adoption in clinical practice, and an important component of model evaluation is an accurate estimate of model performance. External validation – using independent data to assess model performance – is desirable but often not feasible in early stages of model development. Internal validation, i.e. using the data at hand to assess the performance of a model, is a more practical alternative. In particular, internal validation can be used to avoid or correct for resubstitution bias ([60]). Model selection bias can be more challenging to account for, although some methodology has been developed ([75–79]). Both sources of bias are potentially large and, at the very least, should be acknowledged. Finally, in the case of multicenter studies, it is possible to account for differences by center and obtain an unbiased estimate of model performance ([80]).

External validation results are frequently disappointing ([81]); the gap between apparent and externally validated performance may be due in part to optimistic bias resulting from resubstitution or model selection. Without careful design and rigorous statistical analysis, studies of biomarker combinations will continue to be published with (often optimistically) biased estimates of model performance, leading to disappointment after considerable time and resources have been invested in external validation.

Methods

Literature Search and Study Selection

We searched PubMed for all articles published before June 19, 2014 with both “AKI” and “biomarkers” in the title or abstract. Abstracts were reviewed, followed by consideration of the full text of potentially relevant articles. Articles were excluded if they satisfied any of the following conditions: ([1]) study of animals or children only; ([2]) review paper, commentary or conference statement; ([3]) outcome of interest was not AKI; ([4]) abstracts only reported results related to association, not prediction; ([5]) description of a future study; ([6]) ‘omics

study; ([7]) only one injury marker measured a single time; ([8]) not published in English; ([9]) did not consider combinations of injury markers for prediction of AKI.

Because the focus of this review is combinations of biomarkers, we considered only studies with at least two injury biomarker measurements. Murray et al. discussed AKI biomarkers at length, including distinguishing between functional markers and known injury markers ([7]). We used the list of injury markers provided in Murray et al. as the basis for exercising condition [7] above.

Data Extraction

For each of the articles satisfying the inclusion criteria, we collected the following information: markers studied, method used to combine biomarkers, setting giving rise to AKI cases, definition of AKI used, number of models fit, number of AKI cases, number of markers/measurements considered in each combination ("size of the combination"), whether the study included multiple centers, whether resubstitution bias was addressed, reported model performance, and estimated biomarker combination score(s) (if reported). We allowed for combinations that included functional markers in addition to injury markers. We define "many models" as 10 or more, and "few events" as less than 10 events per biomarker in the combination. These thresholds are somewhat arbitrary, so we report the actual numbers to aid interpretation. When more than one combination and/or outcome was reported, we considered the "primary" combination and/or outcome based on the presentation of the results in the abstract, if such a determination was possible.

In the context of predictive models, discrimination refers to the ability of a model to distinguish individuals with and without the outcome of interest. The most common measure of discrimination is the area under the receiver operating characteristic (ROC) curve, also known as AUC. In particular, AUC was reported in all the articles in our review. We therefore focus on AUC as the primary measure of model performance, while also acknowledging that there are important aspects of model performance not captured by AUC ([46]).

Evaluation of Biases

For each article in our review, we evaluated the evidence of whether the report was likely to be affected by the three types of bias discussed: resubstitution bias, model selection bias, and bias due to confounding by center. Our evaluation was based on the number of models fit, number of AKI cases, size of the combination(s), whether the study included multiple centers, and whether resubstitution bias was considered. The number of models fit gives an indication of whether model selection bias is likely to be present, while a small number of cases and/or large combinations can exacerbate resubstitution bias. Bias due to center differences is only a concern when multiple centers are involved.

Replication in TRIBE-AKI Data

When possible, we applied the published models to data from the Translational Research Investigating Biomarker Endpoints in AKI (TRIBE-AKI) study ([10]). Briefly, TRIBE-AKI involves adults undergoing cardiac surgery at six academic medical centers. Urine and

plasma were collected preoperatively and daily for up to six post-operative days. On the first post-operative day, urine was collected every six hours. Patients with evidence of AKI prior to surgery, pre-operative serum creatinine above 4.5 mg/dL or end-stage renal disease (ESRD) were excluded, leaving 1219 subjects. Biomarkers measured in TRIBE-AKI include serum neutrophil gelatinase-associated lipocalin (pNGAL), urine interleukin 18 (uIL18), urine neutrophil gelatinase-associated lipocalin (uNGAL), urine albumin, urine kidney injury molecule-1 (uKIM1), urine liver-type fatty acid binding protein (uLFABP), urine cystatin C and urine creatinine. Participants at each TRIBE-AKI study site provided informed consent and the protocols were approved by the respective institutional review boards.

Applying published models to the TRIBE-AKI cohort provides an unbiased assessment of model performance that can be compared to the published performance. We could only apply published models to TRIBE-AKI if the markers in the model were measured in TRIBE-AKI at similar time point(s) relative to the episode of AKI. When it was not possible to precisely match the timing of urine collection, we chose the next closest option(s). If the estimated combination was provided in the article, we applied that combination to the TRIBE-AKI data; if this was not given, we re-estimated the combination in our data based on how it was estimated in the original article. If the article did not specify how the combination was estimated, we used logistic regression. We replicated the exclusion criteria, outcome definition, timing of biomarker measurement, and form of the biomarkers (i.e., transformation or normalization for urine creatinine) as much as possible. For papers providing the estimated combination, we calculated the apparent and center-adjusted AUCs ([80]) of the combination in TRIBE-AKI data. When the combination had to be re-estimated in TRIBE-AKI, the AUC estimate was center-adjusted ([80]) and corrected for resubstitution bias (“optimism-corrected”) using a bootstrapping procedure with 1000 replications ([60]). We estimated 95% confidence intervals by bootstrapping.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

None

Grants: The research was supported by the NIH grant RO1HL085757 (CRP) to fund the Translational Research Investigating Biomarker Endpoints in AKI (TRIBE-AKI) Consortium to study novel biomarkers of acute kidney injury in cardiac surgery. AM is supported by the NIH grant RO1HL085757. 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). The opinions, results, and conclusions reported in this article are those of the authors and are independent of the funding sources. The results presented in this article have not been published previously in whole or part.

References

1. Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol. 2006; 1:19–32. [PubMed: 17699187]

2. Coca SG, Yusuf B, Shlipak MG, et al. Long-term risk of mortality and other adverse outcomes after acute kidney injury: a systematic review and meta-analysis. Am J Kidney Dis. 2009; 53:961–73. [PubMed: 19346042]

3. Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. Kidney Int. 2012; 81:442–8. [PubMed: 22113526]

4. Coca SG, Yalavarthy R, Concato J, Parikh CR. Biomarkers for the diagnosis and risk stratification of acute kidney injury: a systematic review. Kidney Int. 2008; 73:1008–16. [PubMed: 18094679]

5. Siew ED, Ware LB, Ikizler TA. Biological markers of acute kidney injury. J Am Soc Nephrol. 2011; 22:810–20. [PubMed: 21493774]

6. Pickering JW, Endre ZH. Linking Injury to Outcome in Acute Kidney Injury: A Matter of Sensitivity. PLoS One. 2013; 8:e62691. [PubMed: 23626850]

7. Murray PT, Mehta RL, Shaw A, et al. Potential use of biomarkers in acute kidney injury: report and summary of recommendations from the 10th Acute Dialysis Quality Initiative consensus conference. Kidney Int. 2014; 85:513–21. [PubMed: 24107851]

8. Koyner JL, Parikh CR. Clinical utility of biomarkers of AKI in cardiac surgery and critical illness. Clin J Am Soc Nephrol. 2013; 8:1034–42. [PubMed: 23471130]

9. Devarajan P. Emerging biomarkers of acute kidney injury. Contrib Nephrol. 2007; 156:203–12. [PubMed: 17464129]

10. Parikh CR, Coca SG, Thiessen-Philbrook H, et al. Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. J Am Soc Nephrol. 2011; 22:1748–57. [PubMed: 21836143]

11. Devarajan P, Murray P. Biomarkers in acute kidney injury: are we ready for prime time? Nephron Clin Pract. 2014; 127:176–9. [PubMed: 25343845]

12. Halawa A. The early diagnosis of acute renal graft dysfunction: a challenge we face. The role of novel biomarkers. Ann Transplant. 2011; 16:90–8. [PubMed: 21436782]

13. Lisowska-Myjak B. Serum and urinary biomarkers of acute kidney injury. Blood Purif. 2010; 29:357–65. [PubMed: 20389065]

14. Sprenkle P, Russo P. Molecular markers for ischemia, do we have something better then creatinine and glomerular filtration rate? Arch Esp Urol. 2013; 66:99–114. [PubMed: 23406805]

15. Martensson J, Martling CR, Bell M. Novel biomarkers of acute kidney injury and failure: clinical applicability. Br J Anaesth. 2012; 109:843–50. [PubMed: 23048068]

16. Kerr KF, Meisner A, Thiessen-Philbrook H, et al. Developing risk prediction models for kidney injury and assessing incremental value for novel biomarkers. Clin J Am Soc Nephrol. 2014; 9:1488–96. [PubMed: 24855282]

17. Kerr KF, Meisner A, Thiessen-Philbrook H, et al. RiGoR: reporting guidelines to address common sources of bias in risk model development. Biomark Res. 2015; 3(1):2. [PubMed: 25642328]

18. Steyerberg, EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer; New York: 2009.

19. Harrell, FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; New York: 2001.

20. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. Biometrics. 2008; 64:1–9. [PubMed: 17501939]

21. Vaidya VS, Waikar SS, Ferguson MA, et al. Urinary biomarkers for sensitive and specific detection of acute kidney injury in humans. Clin Transl Sci. 2008; 1:200–8. [PubMed: 19212447]

22. Han WK, Wagener G, Zhu Y, et al. Urinary biomarkers in the early detection of acute kidney injury after cardiac surgery. Clin J Am Soc Nephrol. 2009; 4:873–82. [PubMed: 19406962]

23. Liangos O, Tighiouart H, Perianayagam MC, et al. Comparative analysis of urinary biomarkers for early detection of acute kidney injury following cardiopulmonary bypass. Biomarkers. 2009; 14:423–31. [PubMed: 19572801]

24. Che M, Xie B, Xue S, et al. Clinical usefulness of novel biomarkers for the detection of acute kidney injury following elective cardiac surgery. Nephron Clin Pract. 2010; 115:c66–72. [PubMed: 20173352]

25. de Geus HR, Bakker J, Lesaffre EM, le Noble JL. Neutrophil gelatinase-associated lipocalin at ICU admission predicts for acute kidney injury in adult patients. Am J Respir Crit Care Med. 2011; 183:907–14. [PubMed: 20935115]
26. Katagiri D, Doi K, Honda K, et al. Combination of two urinary biomarkers predicts acute kidney injury after adult cardiac surgery. Ann Thorac Surg. 2012; 93:577–83. [PubMed: 22269724]
27. Kokkoris S, Parisi M, Ioannidou S, et al. Combination of renal biomarkers predicts acute kidney injury in critically ill adults. Ren Fail. 2012; 34:1100–8. [PubMed: 22889061]
28. Cho E, Yang HN, Jo SK, et al. The role of urinary liver-type fatty acid-binding protein in critically ill patients. J Korean Med Sci. 2013; 28:100–5. [PubMed: 23341719]
29. Kambhampati G, Ejaz NI, Asmar A, et al. Fluid balance and conventional and novel biomarkers of acute kidney injury in cardiovascular surgery. J Cardiovasc Surg. 2013; 54:639–46. [PubMed: 24002394]
30. Liu S, Che M, Xue S, et al. Urinary L-FABP and its combination with urinary NGAL in early diagnosis of acute kidney injury after cardiac surgery in adult patients. Biomarkers. 2013; 18:95–101. [PubMed: 23167703]
31. Luo Q, Zhou F, Dong H, et al. Implication of combined urinary biomarkers in early diagnosis of acute kidney injury following percutaneous coronary intervention. Clin Nephrol. 2013; 79:85–92. [PubMed: 23110770]
32. Parikh CR, Thiessen Heather P, Garg AX, et al. Performance of Kidney Injury Molecule-1 and Liver Fatty Acid-Binding Protein and Combined Biomarkers of AKI after Cardiac Surgery. Clin J Am Soc Nephrol. 2013; 8:1079–1088. [PubMed: 23599408]
33. Siew ED, Ware LB, Bian A, et al. Distinct injury markers for the early detection and prognosis of incident acute kidney injury in critically ill adults with preserved kidney function. Kidney Int. 2013; 84:786–94. [PubMed: 23698227]
34. Zeng XF, Li JM, Tan Y, et al. Performance of urinary NGAL and L-FABP in predicting acute kidney injury and subsequent renal recovery: a cohort study based on major surgeries. Clin Chem Lab Med. 2014; 52:671–8. [PubMed: 24293449]
35. Smith GC, Seaman SR, Wood AM, et al. Correcting for optimistic prediction in small data sets. Am J Epidemiol. 2014; 180(3):318–24. [PubMed: 24966219]
36. Pepe MS, Fan J, Seymour CW, et al. Biases introduced by choosing controls to match risk factors of cases in biomarker research. Clin Chem. 2012; 58:1242–51. [PubMed: 22730452]
37. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015; 162:W1–W73. [PubMed: 25560730]
38. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990; 263(10):1385–9. [PubMed: 2406472]
39. Mallett S, Timmer A, Sauerbrei W, Altman DG. Reporting of prognostic studies of tumour markers: a review of published articles in relation to REMARK guidelines. Br J Cancer. 2010; 102(1):173–80. [PubMed: 19997101]
40. 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. [PubMed: 23116690]
41. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med. 2006; 144:427–37. [PubMed: 16549855]
42. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer. 1994; 69(6):979–985. [PubMed: 8198989]
43. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014; 11(10):e1001744. [PubMed: 25314315]
44. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS Med. 2012; 9(11):e1001344. [PubMed: 23185136]
45. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst. 2009; 101(21):1446–52. [PubMed: 19815849]

46. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014; 35(29):1925–31. [PubMed: 24898551]
47. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ. 2003; 326(7379):41–4. [PubMed: 12511463]
48. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012; 98(9):683–90. [PubMed: 22397945]
49. Altman DG. Systematic reviews of evaluations of prognostic variables. BMJ. 2001; 323(7306): 224–8. [PubMed: 11473921]
50. McShane LM, Altman DG, Sauerbrei W, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer. 2005; 93(4):387–91. [PubMed: 16106245]
51. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013; 10(2):e1001381. [PubMed: 23393430]
52. Pepe MS, Feng Z, Janes H, et al. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst. 2008; 100(20):1432–8. [PubMed: 18840817]
53. Parikh CR, Butrymowicz I, Yu A, et al. Urine stability studies for novel biomarkers of acute kidney injury. Am J Kidney Dis. 2014; 63(4):567–72. [PubMed: 24200462]
54. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, et al. Risk prediction models of breast cancer: a systematic review of model performances. Breast Cancer Res Treat. 2012; 133(1):1–10. [PubMed: 22076477]
55. McGinn TG, Guyatt GH, Wyer PC, et al. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA. 2000; 284(1):79–84. [PubMed: 10872017]
56. Tangri N, Kitsios GD, Inker LA, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. Ann Intern Med. 2013; 158(8):596–603. [PubMed: 23588748]
57. McShane LM, Altman DG, Sauerbrei W. Identification of clinically useful cancer prognostic factors: what are we missing? J Natl Cancer Inst. 2005; 97(14):1023–5. [PubMed: 16030294]
58. Jelizarow M, Guillemot V, Tenenhaus A, et al. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010; 26(16):1990–8. [PubMed: 20581402]
59. Knottnerus JA. Challenges in dia-prognostic research. J Epidemiol Community Health. 2002; 56(5):340–1. [PubMed: 11964428]
60. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15(4): 361–87. [PubMed: 8668867]
61. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014; 14:40. [PubMed: 24645774]
62. Petretta M, Cuocolo A. Prediction models for risk classification in cardiovascular disease. Eur J Nucl Med Mol Imaging. 2012; 39(12):1959–69. [PubMed: 23053326]
63. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg. 2015; 102(3):148–58. [PubMed: 25627261]
64. Simon R. Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. J Natl Cancer Inst. 2006; 98(17):1169–71. [PubMed: 16954463]
65. Tripepi G, Heinze G, Jager KJ, et al. Risk prediction models. Nephrol Dial Transplant. 2013; 28(8):1975–80. [PubMed: 23658248]
66. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JPA. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012; 344:e3318. [PubMed: 22628003]
67. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999; 130(6):515–24. [PubMed: 10075620]

68. Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Cancer Inst. 2005; 97(12):866–7. [PubMed: 15956642]
69. Pickering JW, Endre ZH. The clinical utility of plasma neutrophil gelatinase-associated lipocalin in acute kidney injury. Blood Purif. 2013; 35(4):295–302. [PubMed: 23712081]
70. Pickering JW, Endre ZH. The definition and detection of acute kidney injury. J Renal Inj Prev. 2014; 3(1):21–5. [PubMed: 25340159]
71. Pickering JW, Frampton CM, Endre ZH. Evaluation of trial outcomes in acute kidney injury by creatinine modeling. Clin J Am Soc Nephrol. 2009; 4(11):1705–15. [PubMed: 19729431]
72. FDA allows marketing of the first test to assess risk of developing acute kidney injury. FDA. Sep 5. 2014 Available from: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm412910.htm.
73. Kashani K, Al-Khafaji A, Ardiles T, et al. Discovery and validation of cell cycle arrest biomarkers in human acute kidney injury. Crit Care. 2013; 17:R25. [PubMed: 23388612]
74. Meersch M, Schmidt C, Van Aken H, et al. Urinary TIMP-2 and IGFBP7 as early biomarkers of acute kidney injury and renal recovery following cardiac surgery. PLoS One. 2014; 9(3):e93460. [PubMed: 24675717]
75. Berrar D, Bradbury I, Dubitzky W. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics. 2006; 22:1245–50. [PubMed: 16500931]
76. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006; 7:91. [PubMed: 16504092]
77. Tibshirani RJ, Tibshirani R. A bias-correction for the minimum error rate in cross-validation. Ann Appl Stat. 2009; 3(2):822–9.
78. Bernau C, Augustin T, Boulesteix AL. Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. Biometrics. 2013; 69:693–702. [PubMed: 23845182]
79. Boulesteix A-L, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009; 9:85. [PubMed: 20025773]
80. Janes H, Longton G, Pepe M. Accommodating covariates in ROC analysis. Stata J. 2009; 9:17–39. [PubMed: 20046933]
81. Ioannidis JP. Biomarker failures. Clin Chem. 2013; 59:202–4. [PubMed: 22997282]

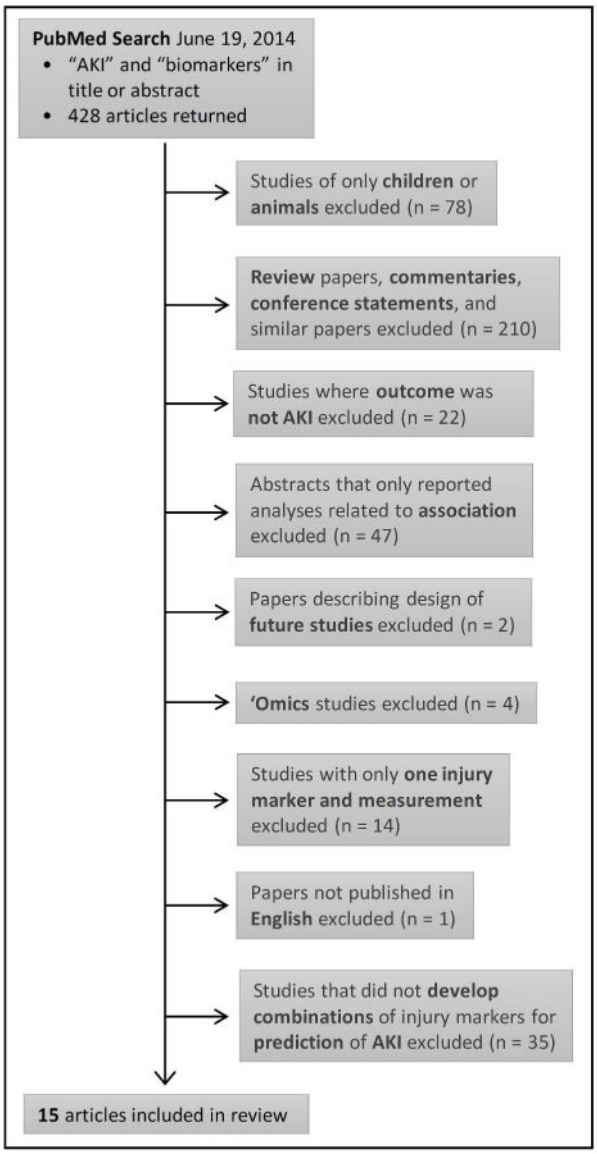

Figure 1. Overview of Study Selection

Abbreviations: AKI = acute kidney injury.

Kidney Int. Author manuscript; available in PMC 2016 August 01.

Table 1

Characteristics of Included Studies.

First Author, Year	Journal	Biomarkers*	Clinical Setting	AKI Outcome Definition	Sample Size		Method for Combining Biomarkers
					Cases	Controls	
Vaidya, 2008	Clinical and Translational Science	uKIM1, uNGAL, uIL18, uHGF, uCysC, uNAG, uVEGF, uIP10, total protein	Inpatient nephrology consultation service	Peak sCr > 50% increase over admission or known baseline	102	102	Logic regression
Han, 2009	CJASN	uKIM1, uNAG, uNGAL	Cardiac surgery	≥ 0.3 mg/dl increase in sCr from baseline or increase 2- to 3-fold within 72h	36	54	Logistic regression
Liangos, 2009	Biomarkers	uKIM1, uNAG, uNGAL, uIL18, uCysC, u(α-1 microglobulin)	Cardiac surgery with CPB	≥ 50% increase in sCr within 72h of CPB	13	90	Logistic regression
Che, 2010	Nephron Clinical Practice	pCysC, uNGAL, uIL18, uRBP, uNAG	Cardiac surgery	≥ 50% increase in SCr from baseline in 72h	14	15	Logistic regression
de Geus, 2011	Nephron Extra	uNGAL, pNGAL, uCysC, pCysC	ICU admissions	≥ 50% increase in SCr occurring and persisting for >24h after admission¥	47	444	Logistic regression
Parikh, 2011	JASN	uIL18, uNGAL, pNGAL	Cardiac surgery	Dialysis or ≥ 100% increase in sCr	60	1159	Logistic regression
Katagiri, 2012	Annals of Thoracic Surgery	uLFABP, uNAG	Cardiac surgery	≥ 0.3 mg/dL or 50% increase in sCr from baseline within 3 days	28	49	Logistic regression
Kokkoris, 2012	Renal Failure	pNGAL, uNGAL, pCysC, sCr	ICU admissions	Any AKI by RIFLE in 7 days	36	64	Logistic regression
Cho, 2013	Journal of Korean Medical Science	uNGAL, uLFABP	ICU admissions	≥ 0.3 mg/dL or ≥ 50% increase of sCr in 5 days	54	91	Unclear
Kambhampati, 2013	Journal of Cardiovascular Surgery	Fluid balance, uNGAL, uIL18, pMCP-1, pTNFalpha	Cardiac surgery	≥ 0.3 mg/dl increase in sCr in 48h	27	73	Logistic regression
Liu, 2013	Biomarkers	uNGAL, uLFABP	Cardiac surgery	≥ 0.3 mg/dL or ≥ 50% increase of sCr in 72h	26	83	Logistic regression

First Author, Year	Journal	Biomarkers*	Clinical Setting	AKI Outcome Definition	Sample Size		Method for Combining Biomarkers
					Cases	Controls	
Luo, 2013	Clinical Nephrology	uKIM1, uNGAL, uIL18	Percutaneous coronary intervention	≥ 0.5 mg/dL or ≥ 25% increase of sCr at 48h	12	30	Logistic regression
Parikh, 2013	CJASN	uKIM1, uLFABP, uIL18, uNGAL, pNGAL	Cardiac surgery	Dialysis or ≥ 100% increase in sCr	60	1159	Logistic regression
Siew, 2013	Kidney International	uNGAL, uLFABP, uCysC	ICU admissions	≥ 0.3 mg/dL or ≥ 50% increase of sCr within 48h of biomarker measurement	127	245	Logistic regression
Zeng, 2014	Clinical Chemistry and Laboratory Medicine	uNGAL, uLFABP	Admitted for major surgery	≥ 0.3 mg/dL or ≥ 50% increase of SCr within 48h	37	160	Logistic regression

Abbreviations: CJASN – Clinical Journal of the American Society of Nephrology; uKIM1 – urine kidney injury marker-1; uNGAL – urine neutrophil gelatinase-associated lipocalin; pNGAL – serum neutrophil gelatinase-associated lipocalin; uIL18 – urine interleukin 18; uCysC – urine cystatin C; pCysC – serum cystatin C; sCr – serum creatinine; uLFABP – urine liver-type fatty acid binding protein; uHGF – urine hepatocyte growth factor; uNAG – urine N-acetyl-β-D-glucosaminidase; uVEGF – urine vascular endothelial growth factor; uIP10 – urine chemokine interferon-inducible protein 10; u(α-1 microglobulin) – urine α-1 microglobulin; uRBP – urine retinol-binding protein; pMCP-1 – serum monocyte chemoattractant protein 1; pTNF-alpha – serum tumor necrosis factor-alpha; CPB – cardiopulmonary bypass; ICU – intensive care unit.

*All biomarkers considered for combinations, including injury and functional markers.

¥Paper considered sustained and transient AKI; here we report analyses related only to sustained vs. no AKI (combinations only reported for sustained AKI).

Table 2

Sources of Bias.

	Possible Sources of Bias				
First Author, Year	**Resubstitution bias**	**Few events**	**Fit many models***	**Multiple centers**	**Reported AUC**
Vaidya, 2008	No	No (n_{ev} = 102, n_m = 4)	Yes (up to 2^9)	No	0.75–0.78
Han, 2009	Yes	No (n_{ev} = 36, n_m = 3)	No (5)	No	0.75–0.78
Liangos, 2009	Yes	Yes (n_{ev} = 13, n_m = 3)	No (7)	Yes	0.78
Che, 2010	Yes	Yes (n_{ev} = 14, n_m = 5)	Yes (29)	No	0.98
de Geus, 2011	Yes	No (n_{ev} = 47, n_m = 2)‡	No (7)	No	0.83
Parikh, 2011	No^	No (n_{ev} = 60, n_m = 3)	Yes (165)	Yes	0.77
Katagiri, 2012	Yes	No (n_{ev} = 28, n_m = 2)	No (9)	No	0.81
Kokkoris, 2012	Yes	No (n_{ev} = 36, n_m = 2–3)	Yes (11)	No	0.823–0.835
Cho, 2013	Yes	No (n_{ev} = 54, n_m = 2)	No (1)	No	0.800
Kambhampati, 2013	Yes	Yes (n_{ev} = 27, n_m = 5)	No (1)	No	0.80
Liu, 2013	Yes	No (n_{ev} = 26, n_m = 2)	No (2)	No	0.911–0.927
Luo, 2013	Yes	Yes (n_{ev} = 12, n_m = 3)	No (2)	No	0.99
Parikh, 2013	No^	No (n_{ev} = 60, n_m = 3)	Yes (455)	Yes	0.78
Siew, 2013	Yes	No (n_{ev} = 127, n_m = 2)	No (1)	No	0.59
Zeng, 2014	Yes	No (n_{ev} = 37, n_m = 2)	Yes (64)	No	0.91–0.94

Dark red indicates a high likelihood of bias, dark pink indicates possible bias, and pale pink indicates low likelihood of bias. In the “Resubstitution bias” column, “no” indicated some attempt was made to address resubstitution bias. Under “Few events”, n_{ev} is the number of events and n_m is the number of markers in the main reported combination; we considered n_{ev}:n_m < 10 to be few events. The number of events reported may include individuals with missing marker values. We considered more than 10 models to be “many”. We calculated the number of models by hand (i.e., these were not explicitly reported by the authors). Under “Multiple centers”, “yes” indicates that multiple centers were involved and possible differences were not considered, though this does not mean there was bias due to center differences. Abbreviations: AUC – area under the receiver operating characteristic curve.

*Includes search over univariate models (if best individual markers were chosen for combination).

^Resubstitution bias was addressed, but may not be totally removed.

‡Reported 10 individuals with sustained AKI also had missing marker values.

Table 3

Replication in TRIBE.

Study Information			Reported AUC (95% CI)	AUC in TRIBE (95% CI)			
First Author, Year	**Provided Combination**	**Biomarkers**		**Apparent**	**Center-Adjusted**	**Optimism-Corrected**	**Center-Adjusted and Optimism-Corrected**
de Geus, 2011	No	uNGAL, pNGAL at admission	0.83 (0.75–0.91)	0.665 (0.585, 0.739)	0.655 (0.573, 0.736)	0.663 (0.583, 0.737)	0.647 (0.565, 0.728)
Kokkoris, 2012	Yes	pNGAL, sCr at admission	0.823 (0.73–0.89)	0.702 (0.655, 0.745)	0.691 (0.647, 0.733)[¥]		
		pNGAL, uNGAL, sCr at admission	0.835 (0.75–0.90)	0.704 (0.660, 0.746)[¥]	0.690 (0.649, 0.733)[¥]		
Cho, 2013	No	uNGAL, uLFABP at admission	0.800 (0.727–0.872)	0.598 (0.564, 0.630)	0.585 (0.549, 0.618)	0.597 (0.563, 0.629)	0.583 (0.547, 0.616)
Liu, 2013	Yes	CuNGAL, CuLFABP at 0h	0.927 (0.868, 0.986)	0.587 (0.553, 0.622)[¥,#]	0.570 (0.533, 0.605)[¥,#]		
		CuNGAL, CuLFABP at 2h	0.911 (0.836, 0.987)	0.587 (0.551, 0.618)[¥,#]	0.569 (0.535, 0.606)[¥,#]		
Luo, 2013	No	uKIM1, uNGAL, uIL18 at 24h	0.99 (0.90–1.00)	0.654 (0.607,0.700)[+]	0.588 (0.544, 0.639)[+]	0.650 (0.603, 0.696)[+]	0.580 (0.535, 0.631)[+]
Siew, 2013	No	CuNGAL, CuLFABP at 0h and 48h	0.59 (0.56–0.69)	0.622 (0.568, 0.676)[#]	0.610 (0.550, 0.661)[#]	0.615 (0.560, 0.669)[#]	0.602 (0.542, 0.653)[#]
Zeng, 2014	Yes	CuNGAL at 12h, CuLFABP at 4h	0.94 (0.89–0.98)	0.620 (0.564, 0.675)[¥,1]	0.622 (0.561, 0.674)[¥,1]		
				0.614 (0.561, 0.673)[¥,2]	0.609 (0.553, 0.668)[¥,2]		
		CuNGAL at 12h,	0.91 (0.85–0.97)	0.639 (0.591, 0.686)[¥,3]	0.619 (0.561, 0.674)[¥,3]		

Study Information			Reported AUC (95% CI)	AUC in TRIBE (95% CI)			
First Author, Year	Provided Combination	Biomarkers		Apparent	Center-Adjusted	Optimism-Corrected	Center-Adjusted and Optimism-Corrected
		CuLFABP at 12h					
				0.642 (0.584, 0.696)[¥,4]	0.630 (0.572, 0.687)[¥,4]		

We present an overview of the studies replicated in TRIBE, including whether the paper reported the estimated combination, the biomarkers involved in the combination, the reported AUC, and the AUC in TRIBE. For the AUC in TRIBE, we considered the apparent, center-adjusted, optimism-corrected, and center-adjusted and optimism-corrected AUCs for those combinations re-estimated in TRIBE. The optimism-corrected AUC adjusts for resubstitution bias. For the combinations provided in the article, we considered the apparent and center-adjusted AUCs. The 95% confidence intervals in TRIBE were estimated by bootstrapping. Abbreviations: AUC – area under the receiver operating characteristic curve; uKIM1 – urine kidney injury marker-1; uNGAL – urine neutrophil gelatinase-associated lipocalin; CuNGAL – corrected (for urine creatinine) urine neutrophil gelatinase-associated lipocalin; pNGAL – serum neutrophil gelatinase-associated lipocalin; uIL18 – urine interleukin 18; sCr – serum creatinine; uLFABP – urine liver-type fatty acid binding protein; CuLFABP – corrected (for urine creatinine) urine liver-type fatty acid binding protein; CI – confidence interval.

[¥] Based on coefficients from paper.

[#] Based on markers measured at 0–6h in TRIBE.

[+] Used KIM1, IL18 and NGAL measured at day 2.

[1] Used uNGAL at 6–12h, uLFABP at 0–6h.

[2] Used uNGAL at 12–18h, uLFABP at 0–6h.

[3] Used uNGAL at 6–12h, uLFABP at 6–12h.

[4] Used uNGAL at 12–18h, uLFABP at 6–12h.

Table 4

Recommendations Regarding the Design and Analysis of Biomarker Combination Studies.

Study Design	
Sample size	For binary outcomes, the effective sample size is the minimum of the number of events and the number of non-events ([39–42]). Consider events per variable (EPV), where the number of variables includes transformations and interactions ([42, 43]).
Enrollment and follow-up	Data from a carefully designed and conducted study are preferable to convenience samples ([37, 39, 44, 45]). Can recruit from multiple sites ([17]); special considerations may be needed ([37]). Inclusion/exclusion criteria and referral pattern can affect generalizability and interpretation ([37, 39, 42, 46, 47]). Treatment may modify risk ([37, 43]) and/or predictive accuracy ([37, 43, 48]) and should be addressed in analysis/interpretation ([42, 49]). Prospective cohort studies are preferable: full control over the sample and data collection ([37, 43, 50, 51]). Avoid loss to follow-up and follow for an adequate length of time ([39, 41]). Can use case-cohort or nested case-control studies ([37, 43, 48]). Specialized study designs exist ([45, 52]). Case-control studies cannot be used to estimate risks ([37, 43, 48]) without external data. Matched designs require additional analytic considerations ([17, 52]).
Measuring biomarkers	Clearly define (including blood/urine and methods of preservation/storage) and measure biomarkers in a uniform, standardized way ([16, 41, 48–50]). Use assays intended for general use ([52]) and beware of batch effects ([17]), and the effect of storage and handling of specimens ([53]). Assays should be standardized, valid and reproducible ([41, 48, 54]). Blind measurement of biomarkers to other variables (including the outcome) as appropriate ([17, 37, 42, 43, 45, 50, 55])
Measuring outcomes	Outcome should be relevant to patients and decision-making ([37, 43, 48]). Measure outcome blinded to other variables as appropriate ([17, 37, 43, 48, 55]). Measure outcome carefully ([17, 39, 41, 56]) and uniformly ([43]) using a well-established method for establishing presence or absence ([37, 48, 50, 52]).
Timing of measurements	Timing must be carefully defined ([17, 43, 47, 48]) and relevant to patients and clinical decision-making ([37]). Patients may receive treatment in the interval between measurement of biomarkers and outcome, which may modify risk ([37])
Other design issues	Determine minimally acceptable values of performance measures at the design stage ([52]). Develop a rigorous study protocol with a sound analysis plan ([45, 51, 57, 58]). Study objectives and research question should be clear ([41, 42, 50, 59])
Model Development	
Choosing candidate biomarkers	A candidate biomarker is any biomarker associated with the outcome; the association need not be causal ([16, 48]). The number of candidate biomarkers increases with transformations and interactions ([16, 17, 60]). Use subject-matter knowledge to choose candidate biomarkers ([40, 48, 56]). Interaction terms rarely add predictive ability; should restrict to a small number of interactions with prior rationale ([37, 60]).
Handling continuous predictors	Categorizing continuous biomarkers results in a loss of information ([16, 42, 44, 46, 48, 49, 51, 60]). Can model biomarkers linearly as a starting point, and consider systematically testing simple transformations ([40, 46, 48]).
Missing data	Complete case analysis can lead to bias and increased variance ([44, 48, 60, 61]) depending upon the extent and mechanism of missingness ([17, 48, 60, 61]). Multiple imputation is recommended ([41, 43, 46, 61])
Predictor selection	Smaller and simpler models may have practical advantages ([16, 51, 56, 62]).

Study Design	
	Stepwise methods ([16, 42, 43, 46, 60]) and univariate screening can be problematic ([37, 48]). Predictive role in isolation no guarantee of performance in combination ([16, 48]). Use clinical knowledge, previous studies and practical considerations to reduce the number of candidate biomarkers ([37, 43, 46]). No consensus on model selection ([48]).
Methods for combination	Multiple methods can be used ([17, 37, 50]) though regression methods are common ([37]). Logistic regression is common for binary outcomes with no loss to follow-up ([37, 40, 43]). Usually use multivariable techniques ([51, 63]). No method can be shown to perform best on every real dataset ([60]). Consider model assumptions ([51]).
Evaluation	
Performance metrics	Predictive accuracy, not measures of association or p-values, is what matters ([42–44, 64]). How a model was derived is of little importance if it performs well ([40, 51, 61]) in terms of validity (accuracy of risk estimates) ([46, 65]). Duration of follow-up is critical in interpreting performance ([37]). Calibration ([16, 37, 43, 48, 61]) and discrimination ([37, 43, 48, 61]) are commonly assessed and presented with bootstrap-based confidence intervals ([17, 42, 66]). Discrimination is particularly important for model development; both are important for model validation ([37]). Can evaluate discrimination by assessing AUC, mean risk difference, true/false positive rates, ROC curves and histograms of predicted risks ([16, 43, 46]). AUC is not affected by miscalibration ([46]). Multi-center studies require specialized techniques ([17]). Assess calibration graphically; can use Cox's method to estimate the calibration intercept and slope ([37, 43, 46, 61]).
Internal validation	A necessary step in model development ([37, 41, 46, 51, 64]). Can help to avoid external validation failure by uncovering problems that may make models misleading or invalid ([60]). The apparent performance of many prediction models tends to be optimistic ([37]) due to resubstitution bias (especially with low EPV) and model selection bias ([37, 43]), and must be interpreted with extreme caution ([60]). Bootstrapping is preferred as it uses all of the available data ([43, 46, 48]). If model-building can be fully specified in advance, it can (and should) be incorporated into the bootstrap ([37, 48, 52, 65]). If model selection cannot be automated, test data are needed ([60]). Bootstrapping gives an honest estimate of internal validity, penalized for optimism due to resubstitution bias and model selection bias ([43, 48, 60]).
External validation	Strongly recommended ([37, 41, 43, 51, 52, 55, 58, 60, 61, 64, 67, 68]) for promising prediction tools (those with a rigorous derivation process) ([55]). Internal validation is not a sufficient substitute ([64]), though it is generally an advisable step ([65]). Consider the prospective broad clinical application of the model ([64, 67]). The completely specified model should be applied ([64, 68]) to assess the generalizability and applicability of the model ([47, 61]). May fail for many reasons: different measurement methods and/or definitions, selection bias, inclusion/exclusion criteria, subject source, settings/location (may affect case mix), recruitment, and clinical and demographic characteristics of the population ([17, 37, 54, 61]). Performance is often worse in new samples ([37, 44]), though this does not necessarily imply a validation failure: diminished accuracy is not the same as inaccuracy.
Existing models	Newly developed models should be quantitatively compared to existing models ([37]). Developing a different prediction model per setting makes research localized ([37]); instead, if models already exist in the same or a related setting, investigators should consider evaluating or comparing, and perhaps updating or recalibrating, existing models ([43, 44, 51, 61]) as part of an ongoing validation process ([46]). Recalibration can be used to improve calibration without needing more data ([43, 60]). Discrimination cannot be improved in this way, and will not be affected by recalibration ([60]).
Reporting	Adhere to existing guidelines ([17, 63]).