

Pencina Michael J. (Orcid ID: 0000-0002-1968-2641)

Cook Nancy R. (Orcid ID: 0000-0002-9705-0842)

Foulkes Andrea (Orcid ID: 0000-0002-9520-0501)

## Statistical Methods for Building Better Biomarkers of Chronic Kidney Disease

Michael J. Pencina<sup>1</sup> Chirag R. Parikh<sup>2</sup> Paul L. Kimmel<sup>3</sup> Nancy R. Cook<sup>4</sup> Joe Coresh<sup>5</sup>  
Harold I. Feldman<sup>6,8</sup> Andrea Foulkes<sup>7</sup> Phyllis A. Gimotty<sup>6,8</sup> Chi-yuan Hsu<sup>9</sup> Kevin Lemley<sup>10</sup>  
Peter Song<sup>11</sup> Kenneth Wilkins<sup>12</sup> Daniel R. Gossett<sup>3</sup> Yining Xie<sup>3</sup> Robert A. Star<sup>3</sup>

### Institution:

<sup>1</sup> Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC

<sup>2</sup> Division of Nephrology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

<sup>3</sup> Division of Kidney, Urology and Hematology, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD

<sup>4</sup> Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

<sup>5</sup> Departments of Epidemiology, Medicine and Biostatistics, Johns Hopkins University, Baltimore, MD

<sup>6</sup> Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

<sup>7</sup> Department of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA

<sup>8</sup> Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

<sup>9</sup> Division of Nephrology, University of California San Francisco, San Francisco, CA

<sup>10</sup> Division of Nephrology, Children's Hospital Los Angeles, Department of Pediatrics, University of Southern California Keck School of Medicine, Los Angeles, CA

<sup>11</sup> Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/sim.8091](https://doi.org/10.1002/sim.8091)

<sup>12</sup> Biostatistics Program, Office of the Director, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, and Department of Preventive Medicine & Biostatistics, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD

**CORRESPONDENCE**

Michael J. Pencina, Ph.D.  
Vice Dean for Data Science & Information Technology  
Professor of Biostatistics & Bioinformatics  
Duke University School of Medicine  
DUMC 2927  
200 Trent Drive  
M144 Davison Bldg.  
Durham, NC 27710  
USA  
Email: michael.pencina@duke.edu

**ABSTRACT**

The last two decades have witnessed an explosion in research focused on development and assessment of novel biomarkers for improved prognosis of disease. As a result, best practice standards guiding biomarker research have undergone extensive development. Currently there is great interest in the promise of biomarkers to enhance research efforts and clinical practice in the setting of chronic kidney disease, acute kidney injury, and glomerular diseases. However, some have questioned whether biomarkers currently add value to the clinical practice of nephrology. The current state of the art pertaining to statistical analyses regarding use of such measures is critical.

In December 2014, the National Institute of Diabetes and Digestive and Kidney Diseases convened a meeting, “Toward Building Better Biomarker Statistical Methodology,” with the goals of summarizing current best practice recommendations and articulating new directions for methodological research. This report summarizes its conclusions and describes areas that need attention. Suggestions are made regarding metrics that should be commonly reported. We outline the methodological issues related to traditional metrics and considerations in prognostic modeling, including discrimination and case mix, calibration, validation, and cost-benefit analysis. We highlight the approach to improved risk communication and the value of graphical displays. Finally, we address some “new frontiers” in prognostic biomarker research, including the competing risk framework, the use of longitudinal biomarkers, and analyses in distributed research networks.

**Keywords**

1. Risk model
2. Discrimination
3. Calibration
4. Cost-benefit
5. Validation
6. Risk Communication

## 1 | INTRODUCTION

This report provides a summary of considerations and discussions that took place during the National Institute of Diabetes and Digestive and Kidney Diseases meeting, “Toward Building Better Biomarker Statistical Methodology” held December 2, 2014. It was motivated by the recognition of need for improved prediction of outcomes and response to therapeutic interventions in patients with acute kidney injury, chronic kidney disease and glomerular disease, to name a few. Here we review research topics that offer meaningful contributions to the field of statistical methodologies for prognostic biomarkers, with a particular focus on kidney diseases.

We review phases of biomarker development, as outlined in a report from the American Heart Association, and expand this review by suggesting some metrics that could be commonly reported.<sup>1</sup> Our scope is also limited to questions related to prognostic biomarkers, and not to biomarker research in general.<sup>2</sup> We start with a summary of best practice recommendations. Subsequently, we outline some methodological issues related to traditional metrics and considerations regarding prognostic modeling, including discrimination and case mix, calibration, validation and cost-benefit analysis. We then focus on the challenges facing improved risk communication and the importance of graphical displays. Finally, we address some “new frontiers” in prognostic biomarker research, including the competing risk framework, the use of longitudinal biomarkers, and analyses in distributed research networks, and end with a brief summary and consideration of other important directions. Our goal is to provide a statistical framework for models in which biomarker analyses permit meaningful clinical interactions between patients and physicians in the outpatient setting or at the bedside.

## 2 | BEST PRACTICE RECOMMENDATIONS

It is essential to consider the clinical context of biomarker applications when working on their discovery and development. Investigators should be guided in part by key clinical questions such as, “Of what use is this biomarker to me or my patient in the clinical setting?” It is also essential that biomarkers studies are motivated by, and address, existing clinical needs that go beyond population level to individual clinical considerations, spanning matters of importance to

the patient and the physician in the consultation room or at the bedside, including risk communication and specific clinical actions.<sup>3</sup>

Several statistical best practices exist. First, it is essential to avoid the loss of information by unnecessary categorization of outcomes and predictors. Instead, models need to incorporate continuous or ordinal outcomes, and flexible modeling techniques (such as fractional polynomials or splines) should be adopted to avoid forcing linearization or dichotomization of predictors. Furthermore, different phases in the process of biomarker development require different statistical methods to summarize the findings. For example, it is premature to assess improvement in model performance during the early proof of concept stage while it is insufficient to rely on tests of association in the later phases. In addition, situations such as competing risk and multi-category outcomes require special consideration.

Risk prediction models should be developed that will be broadly useful in clinical settings. When evaluating performance of biomarker-enriched models, it is essential to keep in mind the impact of case mix. Unbiased internal validation using resampling (e.g., bootstrapping or cross-validation) of biomarker results is necessary. The importance of calibration in the prognostic context was also highlighted. Furthermore, cost-effectiveness and cost-benefit analyses are important for determining the best treatment in certain situations.

### **3 | DIFFERENT PHASES OF EVALUATION OF RISK MARKERS REQUIRE DIFFERENT STATISTICAL METRICS**

A key consideration in evaluation of biomarkers is the stage of their development. The AHA Scientific Statement on the Criteria for Evaluation of Novel Markers of Cardiovascular Risk gives an excellent summary of those stages<sup>2</sup>. The terminology used for presentation of results as well as statistical metrics employed to convey them should differ. During the earliest phase, “Proof of Concept,” investigators establish whether novel marker levels differ between participants who will and will not develop the event of interest. If the biomarker level is not time-varying or largely affected by the occurrence of the event, cross-sectional studies can be employed at this stage. Replication in an independent cohort is desirable. The statistical focus at

this stage should be on the distribution of the biomarker among events and non-events and the observed effect size. Any statistical risk modeling would be premature, and measures of precision of the estimates are preferred over tests of hypotheses.

The second phase, “Prospective Validation,” takes us one step further. Here we attempt to establish whether the new marker predicts development of future events in a prospective cohort, or nested case-control, or case-cohort study. In this phase descriptive plots of the marker’s impact over time, as well as relative risks and hazard ratios with the corresponding confidence intervals, should be reported.

The third phase focuses on the incremental value of the marker on the disease risk model. Does a marker add prognostic information to established, standard risk factors? Does the risk model improve with the inclusion of the marker? A substantial portion of statistical research on prognostic biomarkers has been devoted to improving methodologies for this phase. An emerging consensus supports the use of hierarchical assessment. First, the likelihood ratio test is performed to determine if the new marker(s) remains associated with the outcome after controlling for a large number of already known and used predictors.<sup>4-6</sup> This can be supplemented by the likelihood-based adequacy index described by Harrell.<sup>7</sup> The index quantifies the proportion of the log-likelihood explained by all predictors that can be explained without the new marker(s). Values close to 1 indicate that the new marker(s) offer little added information whereas values close to zero suggest the new marker(s) contain most of the explanatory power. After the added value of the new marker(s) has been established, discrimination, calibration, and reclassification metrics can be presented with corresponding confidence intervals. These three domains are considered in more detail in the next section. Here we stress that p-values should not be reported with measures of discrimination, calibration, or reclassification. There are at least two reasons for this recommendation. Pepe et al have shown that for absolutely continuous risk functions, the hypothesis of association with outcome after controlling for standard risk factors is equivalent to the hypothesis of incremental value as determined by the most commonly used measures of discrimination and reclassification. Since

the likelihood ratio test is the most powerful, there is no need to perform any other tests<sup>5</sup>. Furthermore, the existing asymptotic tests for the hypotheses of incremental value are incorrect due to the lack of normality under the null hypothesis.<sup>4,5</sup>

After an incremental value has been established, the question of clinical utility must be addressed. What would be the impact of routine use of the new biomarker in clinical practice? Would it alter treatment or clinical care? Plots of risk distributions (see Section 6.2) as well as decision analytic measures of net benefit, relative utility, or weighted net reclassification improvement offer simple yet informative summaries that address this question. They can also be expressed as the number of false positives per each additional true positive or as the number needed to test.<sup>8,9</sup> Separate assessment for subjects with and without events might also be instructive, especially as some classification threshold may have improvements in sensitivity with reductions in specificity.<sup>10</sup>

A more extensive version of the above considerations leads to a full cost-effectiveness analysis. This requires meaningful cost or utility estimates assigned to different combinations of clinical decisions and associated outcomes. Some of the inputs might require parameters that are best estimated from randomized controlled trials (RCTs). RCTs may also be necessary to establish whether routine use of the marker improves clinical outcomes. Impact on clinical outcomes and cost-effectiveness form the last two phases in the evaluation of biomarkers.

Table 1 provides a summary of these recommendations. It follows the phases of biomarker evaluation proposed by the AHA scientific statement and expands the considerations by adding statistical metrics that might be relevant for each phase.

## **4 | DISCRIMINATION**

Discrimination of a prognostic model refers to its ability to distinguish between those who will versus those who will not develop the event of interest. It is most commonly quantified using the area under the receiver operating characteristic curve (AUC).<sup>11,12</sup> Recently, the discrimination slope and other R-squared-type measures have gained popularity as discrimination metrics.<sup>13,14</sup>

### **4.1 | Importance of risk factor distributions**



One frequently overlooked feature of discrimination is its dependence on the distributions of the biomarker of interest as well as of other predictors.<sup>15</sup> This affects the discriminatory ability of the baseline model as well as the impact of added biomarkers. This phenomenon can be illustrated using two biomarker studies published in the *New England Journal of Medicine*. Wang et al<sup>16</sup> used Framingham data with baseline age of participants extending approximately from 30 to 70 years (based on the mean and standard deviation reported) to show that a panel of biomarkers added to a cardiovascular risk prediction model with standard risk factors increased the AUC from 0.80 to 0.82. Zethelius et al<sup>17</sup> used data from the Uppsala study of men aged 69 to 74 years at baseline to show that a biomarker panel added to a cardiovascular risk prediction model based on standard risk factors increased the AUC from 0.69 to 0.75. There were a few data characteristics that differed between the two studies, including some of the biomarkers used and men-only sample in the Uppsala study. Here we focus only on the difference in baseline age ranges: 40 years in the Wang paper and 4 years in the Zethelius paper. The difference in baseline AUCs may be explained by the difference in baseline age distributions. A simple simulation confirms this hypothesis. We choose the baseline age of the general population to be distributed uniformly between 35 and 75, roughly corresponding to what is used in the Framingham study. We select the strength of association between age and the outcome to obtain age-only AUCs of 0.6, 0.7, 0.8, and 0.9. Then we add another uncorrelated normal variable with an effect size (standardized mean difference between persons with and without events) of 0.5. We set the sample size at 1,000,000 to limit sampling variability. We repeat the steps for samples generated from the same population, but this time the age range is reduced to 70-73, a selection close to the range in the Uppsala study. The results are presented in Table 2.

First, note the striking impact of baseline age range on the AUC. For example, the AUC decreases from 0.80 in case of the wide age range of 35-75 years to 0.56 for the range of 70-73. The impact of age is substantially reduced in studies with sufficiently narrow age range, an easily predictable but all too often overlooked observation. The change in AUC is a function of the strength of the baseline model – it is large for weak models and small for strong models. This

implies that researchers evaluating the same marker in studies with different distributions of age (wide versus narrow) could reach opposite conclusions about the usefulness of a marker. The results depicted are based on a simple model presented to illustrate the phenomenon. The same results will hold when we replace age with different predictors and their combinations. In studies of chronic kidney disease, for example, in addition to age, the distribution of urine protein excretion and level of renal function might be critical in comparing outcomes in different studies. Thus, the distribution of risk factors in the sample under study has a powerful impact on the results and needs to be considered when drawing conclusions.<sup>18</sup>

Several approaches can be considered. Ideally, we should be able to quantify discrimination on the population for whom the new biomarker is considered. Thus, the incremental impact of the biomarker might be population specific, with the features of the population under investigation clearly described in the scientific report. In addition, investigators could consider having a standard population on which prediction models are assessed to permit comparisons across studies. If this is not possible, micro-simulations could be undertaken to put the results in appropriate context. In some settings, covariate-adjusted measures could be considered.<sup>19</sup> In general, risk prediction models should be developed in broadly representative clinical settings. This concept ties to the issue of case mix, because general predictions cannot be made from a limited study. Journal editors should be sensitive to this limitation.

The issue of risk factor distribution also applies to combinations of biomarkers. The question arises regarding to the extent to which combinations of biomarkers that improve prediction in a mixed-disease population (i.e., substantial case mix) also improve prediction in clinical settings. If the biomarker is evaluated in a heterogeneous population, does this mean that the transfer of the biomarker to another clinical setting is compromised? This concept is related to the notion of biomarkers that reflect pathways that participate in disease progression in all people, compared with combining biomarkers that reflect the diversity of pathways that sometimes cause disease in some people. This reflects the value of the panel of biomarkers at an individual level.

This discussion also relates to the phases of biomarker assessment described in the previous section. Samples used during the proof of concept evaluation will most likely be more narrow and homogenous given their smaller size. Prospective validation and the following phases will hopefully be conducted on broader and more heterogeneous samples. This increases the probability of failure for biomarkers that looked promising during early stages. A couple of safeguards could be considered to limit this risk. First, it is essential to account for the matched case-control design of many early phase studies because by their very nature they might give an over-optimistic impression of the biomarker's discriminatory potential. Matching reduces variability that is due to the predictors on which we match, thus reducing their impact. Matching has an effect similar to the narrowing of the age range in the earlier example. Fortunately, reliable methods have been proposed to account for this problem.<sup>20</sup> Second, before proceeding to assessment of incremental value, micro-simulations should be used to assess the potential value of the marker on a broader and more heterogeneous data set.

#### **4.2 | R-squared-type measures of discrimination**

The AUC is by far the most commonly reported measure of discrimination and model performance.<sup>21</sup> As a result, the AUC has the unquestionable advantage of familiarity but it also has several shortcomings.<sup>22-24</sup> Consequently, several other metrics have been evaluated. Some of them are being rediscovered as potential alternatives to the AUC. Here we consider the R-squared type measures.

In linear regression, R-squared is the prime metric of model performance. Extensions to binary and survival outcomes are possible, but they are not unique. This might partially explain the lack of consistent applications of R-squared measures in prognostic modeling. Furthermore, each extension of R-squared has its own limitations.<sup>14</sup> The linear regression R-squared can be interpreted as a measure of explained variation, a measure of reduction in residual variation, and a measure of correlation of predictors with outcome. Each extends to a slightly different metric in the case of binary outcomes even though they are all asymptotically equivalent, converging to the same quantity for correctly specified models. However, the R-squared based on “model sum

of squares” measures only discrimination while the one based on “error sum of squares” measures both discrimination and calibration and is closely linked to the Brier score. The average of these two, called by Tjur the coefficient of discrimination,<sup>14</sup> is the same as the discrimination slope introduced by Yates<sup>25</sup> and used as a building block for the integrated discrimination improvement (IDI) metric proposed by Pencina et al.<sup>26-28</sup> A related class of R-squared measures can be derived based on the log-likelihood.<sup>7</sup>

While the R-squared measures have many appealing features,<sup>14,26</sup> lack of familiarity and problems with interpretation have limited their application. Further research is thus required, with particular focus on R-square interpretation and practical properties in applied settings. Reference values for different situations may be needed. Given that the R-squared measures belong to the third phase of marker assessment, it would be particularly interesting to see and compare how informative they are in developing a sense of the marker performance in the fourth phase, focused on clinical utility. Thus, further research between the links of AUC and R-squared measures and their links to decision analytic concepts, including but not limited to net benefit, relative utility, and the weighted NRI, appears warranted.<sup>10</sup> In addition, more work needs to be done to develop a better sense for understanding their magnitude as well as the impact of model calibration on the results.

### **4.3 | Effect of correlation structure on model performance**

Recent findings<sup>29,30</sup> underscore the impact of the correlation structure that exists in the data on the ability of markers to improve discrimination. Contrary to the popular belief that uncorrelated predictors might offer the largest gains in model performance, in many settings biomarkers negatively correlated with each other (within event and non-event individuals) and positively correlated with the outcome might lead to highest gains. Even more surprisingly, strong positive correlation might also be beneficial in some settings. For a simple illustration of this phenomenon see Figure 1 in Demler et al.<sup>30</sup> We add a new biomarker of intermediate strength (effect size 0.5) to a model containing a single biomarker with effect size 0.74, yielding a baseline AUC of 0.70. If the correlation between the two biomarkers (conditional on the event

status) is zero, addition of the second biomarker increases the AUC from 0.70 to 0.74. But if the conditional correlation is negative, close to -0.5, AUC increases to 0.80. Finally, large positive conditional correlation (close to 0.75) leads to a negligible increase in the AUC, while very large positive correlation (close to 1) leads to an appreciable increase. In all of the above scenarios, the effect size of the added biomarker is constant and equal to 0.50. Consequently, the approach of pre-selecting biomarkers based on their univariate effect should be discouraged.<sup>29</sup>

## 5 | CALIBRATION, VALIDATION AND COST-BENEFIT

### 5.1 Calibration

Prognostic models evolved from the diagnostic setting where the main focus was on their discriminative ability. The awareness of the importance of calibration when evaluating models predicting future events has been increasing recently.<sup>31,32</sup> Several publications assessing model comparisons have demonstrated that the models can be affected by lack of calibration. However, there remains confusion about the meaning and interpretation of calibration and appropriate ways to quantify it.

First, it is important to note that there are several levels of calibration that can be achieved. Calibration in-the-large requires that the mean of model-based risks equals the event rate, and is automatically satisfied when the model was developed using a regression technique applied to the same data on which it is being evaluated. The simplest reason for miscalibration is when the model is applied in an external setting.

Calibration in-the-large can also be measured by the intercept of the regression model with the event as the dependent variable and the linear predictor as a single independent variable. Over- and under-dispersion refer to model-based probabilities being more or less spread out than the actual risks in the population. Over- and under-dispersion can be measured and corrected using the calibration slope, which is the slope of the regression of event on the linear predictor.

There are “weak” and “strong” calibration metrics that require that the observed risk is equal to the mean estimated risk for those with that model-based risk (“weak”), and for those with any and each covariate pattern (“strong”). No single summary metric does an adequate job

quantifying the last two levels of calibration. Graphic displays are the most appealing. For example, the recent TRIPOD statement on reporting standards for predictive models presents a user-friendly example of an easy to understand calibration plot.<sup>33</sup> Still, more research is needed to help guide investigators in reading these plots, with a particular focus on how we can distinguish between the different levels of calibration. Furthermore, we also need to better understand how these levels relate to each other and which features of the model drive or contribute to different forms of miscalibration.

These considerations help frame questions about the impact on calibration exerted by addition of new biomarkers. To what extent should we expect new biomarkers to improve the different calibration levels described above? It is clear that they will have no impact on calibration in-the-large, but more work is needed to see how they can affect other types of calibration. Also, some consensus needs to be built around the degrees of acceptable miscalibration (if any). In some settings, optimal calibration may not be possible, especially in specific geographic populations or subgroups. It is important to consider how to address concerns that a model might not be calibrated appropriately for a patient belonging to a particular subgroup.

## 5.2 | Validation

The purpose of predictive models is to apply them outside the sample that was used for their development. It is well known that the apparent performance of prognostic models on the data on which they were developed will be better than their performance in different samples.<sup>34</sup> However, the majority of biomarker studies develop and assess models on the same data. Several methods have been proposed to quantify the potential over-optimism due to same sample assessment. These include repeated random split sample techniques,  $x$ -fold cross-validation and different bootstrap techniques.<sup>35</sup> It is less clear, however, to what extent the impact of new biomarkers would be over-estimated when assessed on the same sample on which the model was developed. Empirical evidence shows that results following various corrections for over-optimism remain very close to those without the correction when the number of events is

sufficiently large and the number of biomarkers is sufficiently small.<sup>36</sup> It would be helpful to know to what extent this is to be expected and when meaningful reductions in apparent performance should occur. On the other hand, exploratory analyses studying many genes or proteins at an early research phase require more careful internal validation than studies involving large data sets with known risk factors. Model selection may require 10 to 100 repetitions, assessing linearity, looking at splines, and choosing the most appropriate functional forms. Cross-validation and bootstrapping can help with complex models. Methodological exploration is needed to address these questions.

The importance of external validation is of particular relevance in the context of larger amounts of data and when more computationally-intensive algorithms are used. Regression still remains the most popular strategy, partially due to its easier implementation and interpretation. Algorithms based on regression methods are also known to validate well in external samples.<sup>36</sup> However, adoption of more advanced techniques, including statistical and machine learning methods, might lead to appreciable gains in prognostic ability of individual biomarkers, and especially for combinations of biomarkers.<sup>37</sup> Still, their ability to be well validated must be established. Model validation is critical for measures of discrimination as well as calibration with particular focus on individual-level estimates of absolute risk.

### **5.3 | Cost-benefit considerations**

Cost-benefit analyses need to be undertaken before applying a new marker in practice. A particular biomarker might be better for a specific use, but the cost might outweigh the benefits. An important question in the area of cost-effectiveness and cost-benefit analysis is how best to ascertain costs – not just monetary costs, but also clinical utility, psychological costs, treatment risks, and patient preferences. For example, how willing is a patient to endure a biopsy to improve the assessment of her risk? Furthermore, the frame of reference needs to be defined: should the cost be guideline-based (e.g., uniform costs for everyone), established by individual preferences, or be a combination of both approaches? The frame of reference selected might alter the importance or value of a given marker. Thus, it is essential to understand how the frame

of reference, patient preferences, and societal demands might change the metrics used to quantify cost-benefit. It is also important to learn how best to communicate the costs and benefits to patients and to better elucidate and quantify patient preferences. The lack of agreed-upon inputs for cost-benefit analyses is one of the main impediments to their more frequent and meaningful application. On the other hand, cost-benefit analyses might be more theoretical in areas where no alternative treatments are available. An example is nephrology, where there are few well-defined and accepted preventive treatments for chronic kidney disease (CKD). Angiotensin-converting-enzyme inhibitors and angiotensin receptor blockers halt progression of CKD but the cost-benefit analysis across the spectrum of risk of kidney failure in the future is unknown. More research is needed to develop reliable cost-benefit frameworks specific to each given practical area of application. Ideally, this could be done through a consensus among multi-disciplinary experts.

Even if reliable and generally accepted frameworks are created for cost-benefit analyses corresponding to prognostic models in relevant disciplines, it is unreasonable to expect all biomarker papers in the discipline to perform full cost-benefit or cost-effectiveness assessments. Such analyses do not need to be performed until later stages in biomarker evaluation. Furthermore, even when reliable frameworks are created, differences in interpretation and frames of reference will have persisting problems with heterogeneity and some controversy. Against this backdrop, simplified decision-analytic approaches offer an appealing bridge between a full cost-benefit analysis and purely statistical metrics. Net benefit, relative utility and weighted net reclassification index (NRI) belong to this group.<sup>10,35</sup> In the many instances where classification thresholds are not well established, it is helpful to present these decision analytic measures across a range of plausible thresholds. Graphical representation is particularly attractive, resulting in net benefit, relative utility, or weighted NRI curves.<sup>10,38</sup> A good example of a net benefit curve and its interpretation is given in the TRIPOD statement.<sup>39</sup> There is little disadvantage to routine presentation of these metrics, especially using graphic displays across the entire range of thresholds, as long as their interpretation remains within the confines of what



these metrics can and cannot do. They are not a replacement for a full cost-effectiveness analysis. They also rely on the assumption of rational choice, and link the classification threshold to a ratio of costs or dis-utilities of wrong decisions. Many popularly used classification thresholds are not created in this way. The choices of costs and utilities might depend on the frame of reference, and are not likely to be constant. Finally, they are sensitive to miscalibration.<sup>40</sup> Their relationship to statistical measures of model performance applicable in the third phase of biomarker assessment needs further exploration.

## **6 | RISK COMMUNICATION AND GRAPHICAL DISPLAYS**

### **6.1 | Better risk communication**

Development and application of appropriate statistical methodology is imperative for fruitful biomarker research. However, effective communication of the methods and results among the researchers and the clinical community is also important. Development of such communication is best facilitated by multilevel interaction and collaboration between statisticians, clinicians, and study designers. For example, in cardiology, the Framingham Heart Study serves as an example of such interaction between statisticians and physicians. There is a delicate balance between the ability to explain more difficult concepts in simpler language and diluting these concepts.

Four aspects of these important considerations are highlighted. First, the consistent use of uniform terminology and statistical metrics will enhance their understanding and interpretability. These might need to be specific to each phase of biomarker development as outlined in Section 2. Second, it is important to distinguish between the merits of focus and reporting based on absolute versus relative risk. It is also important to recognize that some of the metrics in current use are more aligned with absolute than relative risk. For example, the change in AUC depends more heavily on relative risks, while the discrimination slope and IDI are also linked to absolute risk as are assessments of calibration. Third, it is important to quantify the variability around predicted risk, and to develop methods that help quantify the heterogeneity of individual predictions based on multiple models created for the same outcome. If several models yield similar (high or low) risk assignments, a conclusion is more credible than when the results are

heterogeneous, even if their central tendencies stay the same. Whether this can be quantified and used for improved risk assessment requires further research.

It is important to distinguish between the incremental value of new biomarkers in predictive models and the amount of risk that is attributed to causal factors. The terms predictor, marker, and risk factor are often used interchangeably, but some differentiation might be desired. Different standards and justifications for use might be applied for markers of risk, compared to causal factors. Generally, a marker will be useful if it can offer added value in predictive model application. A causal factor might add little to an existing model, due to its correlation structure with other predictors, or if the baseline model is already very good, but might still be important because its reduction or elimination can have positive clinical consequences.<sup>41</sup> More research is needed to find ways to best quantify the impact of biomarkers that could be causal factors in mechanistic pathways.<sup>42,43</sup>

## 6.2 | Graphic displays of biomarker effect

Our usual approach of summarization of results as a single number can detract from the interpretation of the complexity of findings in a well-designed study. However, much more information can be conveyed in pictorial displays of a biomarker effect. A calibration plot is a useful graphical display which conveys more information than any single-number calibration metric.<sup>39</sup> Graphical displays can be tailored for different audiences. For example, telling a patient that their risk of disease is X may not give them much information without appropriate context. However, if we tell them that their risk is X and show them where they fall on the risk spectrum that includes their peers, the information becomes much more meaningful.

Consider an example based on Framingham Heart Study data. A sample consisting of 3969 men, aged 30 to 75 years, free of cardiovascular disease, was used to develop predictive models based on age alone, and then adding systolic blood pressure, the presence of antihypertensive treatment, levels of total and HDL cholesterol, and smoking and diabetes mellitus. Figure 1 presents histograms of risk for the age-only (dark gray) and all-risk factors (light gray) models. The wider distribution of risk according to the all-risk factors model demonstrates the improved

discriminatory ability of the second model compared to the one only using age. Still, large portions of the histogram overlap, illustrating that age encompasses a substantial portion of the overall prognostic value. Such presentations also enable individual risks to be considered in the context of risks observed in the larger sample.

A different way to illustrate discriminatory ability is achieved when the distribution of risk is presented separately for those with and without events in a plot called a “discrimination histogram.”<sup>14</sup> Figure 2 presents parallel histograms for the two models depicted in Figure 1. Separation between model-based risks for events versus nonevents (histogram with a mass on the right versus on the left, respectively) indicates the model’s ability to discriminate between events and nonevents. Comparing the amount of separation between the histograms for the reduced and full model allows one to determine how much value the new predictors offer. The histogram for events is more right-shifted for the full model. We note that more extreme risk estimates are attained if we include risk factors beyond age. Still, even with all the risk factors included in the full model, the event histogram is not sufficiently right-shifted, implying more room for improvement of the model. Of note, other metrics of performance can be read quickly from the discrimination histogram. For example, the area to the right of any pre-selected vertical line on the event plot equals sensitivity at the threshold where the line is drawn. Conversely, the area to the left of a vertical line on the nonevent display on the histogram plot equals specificity.

Other useful graphic displays include plots of net benefit or relative utility as a function of classification threshold, as well as change of predicted risks as a function of risk estimated using the baseline model. A more detailed review of different graphs is provided by Steyerberg et al.<sup>44</sup>

## 7 | NEW FRONTIERS OF PROGNOSTIC BIOMARKER RESEARCH

An active interest in the role of biomarkers from epidemiologists, policy-makers, funders of research and pharmaceutical companies, and the availability of data from diverse sources poses new challenges as well as opportunities for the development of novel statistical methodologies. Some of these are addressed below.

### 7.1 | Methodology for biomarker assessment in the competing risk framework

Most commonly, biomarkers are assessed in the context of single binary or time-to-event outcomes. Some people in the sample experience the outcome and others do not. However, settings where more than one outcome is possible are increasingly common. For example, in nephrology, when modeling end stage renal disease (ESRD), it is recognized that the vast majority of patients with CKD do not develop ESRD and, instead, succumb to a competing risk (e.g., cardiovascular mortality). In the kidney transplant setting, graft failure competes with the death of a patient with a functioning graft. The presence of a competing risk will generally affect the estimated absolute risk, and may affect the relative contribution of predictors on each of the end-points.

The competing outcomes are generally approached in two ways. They can either be viewed as a nuisance, an effect we need to account for, but without direct interest in inference about them, or they can be considered multi-category outcomes where each outcome is of interest. The former approach can be illustrated with the kidney transplant example – the primary interest lies in graft failure, but we have to account for those who die with functioning grafts, which are censored events in the analysis. On the other hand, we might be interested in estimating the differential impact of biomarkers measured in patients with CKD or ESRD versus cardiovascular outcomes. Many statistical models can be applied to both settings. These range from more familiar nominal logistic regression based on generalized logits<sup>45</sup> and the Fine and Gray model for survival data<sup>46</sup> to modern machine learning approaches adapted for time-to-event analysis.<sup>47</sup> All models relate the predictor to each of the outcomes. In some extreme cases, such models can produce results that at first might appear counterintuitive. For example, heavy smokers might

die of cancer before developing cardiovascular outcomes, making smoking appear to be unrelated to cardiovascular outcomes or even a protective factor, in effect weakening or nullifying associations between risk factors and outcomes. An alternative approach of Andersen<sup>48</sup> adjusts only the absolute risk for the competing cause, but derives the relative risks from models with a single outcome. In all cases, accounting for competing risks alters the estimated predicted risk. How these different approaches and the unverifiable assumptions inherent to competing risk models affect prediction needs further evaluation.<sup>49</sup>

The vast majority of metrics of prognostic value have been proposed in the context of single binary or time-to-event outcomes. Only more recently, some extension to multi-category or competing risk settings has been proposed.<sup>50-52</sup> However, several important questions remain to be answered. First, we have very little or no intuition or empirical experience with which to interpret the metrics that have been proposed beyond relating them to the most extreme and unrealistic cases. Second, it is not clear how the multi-category outcome metrics compare with and affect each other, and what features of competing risk models drive the performance of these metrics. It is likely that some metrics will favor a strong ability to distinguish between outcomes 1 and 2, even if the model does a poor job distinguishing between outcomes 1 and 3 and 2 and 3, while others will favor models that do a moderate job distinguishing between all pairs of outcomes. Finally, it is not clear what framework is required in settings where the competing risk is a nuisance.

## **7.2 | Longitudinal biomarkers**

Technological advances and reduced costs of measurement with multiplex technologies enable the measurement of several biomarkers over time. In longitudinal studies adjusting for confounding may involve using baseline data, including baseline biomarker measurements. These advances may create promising opportunities for improved inference. The magnitude and rate of change in a biomarker may be informative in enhancing the understanding of the future clinical trajectory of the disease. However, whether rates of change inform prediction more than

static biomarker measurements remains an unresolved issue. Results will likely be context and biomarker-dependent, and might also be related to the adopted modeling approach.<sup>53-553</sup>

Meaningful advances in statistical methodology have been developed over the last decade, preparing the field for consideration of longitudinal measurements.<sup>56</sup> Joint modeling of survival and longitudinal data have been investigated by many researchers. Data quality is essential to use such complex models with both time-to-event and longitudinal parameters in practice. Meaningful clinical interpretation<sup>57</sup> and real applications showing the added value of the more advanced methodologies, such as discovering subgroups exhibiting distinct disease courses, must be demonstrated.<sup>58-60</sup>

An additional epidemiological challenge includes the problem of time-dependent confounding. Appropriate clinical interpretation is closely linked to the timeline of the effect of exposure. The proportional hazards model, with time-dependent covariates where the last observation affects the outcome<sup>61</sup> may be compared with models with “historical” exposures and their trajectories. Exposure accumulation stops at a given time point, often called “baseline,” and then follow-up starts at this point.<sup>62</sup> The former approach might be the most informative for acute exposures that affect the outcome immediately (e.g., increase in blood pressure preceding a cerebrovascular accident). The latter might be best suited for modeling a chronic exposure that leads to damage accumulating over time (e.g., cholesterol depositing in the arterial wall and cardiovascular disease or hypertension and CKD). Joint models can be used in both scenarios, but also allow capturing scenarios that fall in between. Modern longitudinal analysis methods that accommodate both time-dependent confounding and informative censoring can also be used, and have been applied in kidney disease.<sup>63,64</sup> Further work aimed at creating guidance for investigators on which models should be used and when would be of great value.

Joint modeling can be accomplished in either frequentist or Bayesian frameworks. The latter might be appealing in situations where new data is constantly acquired as happens when electronic medical record-based data, including longitudinal measurement of biomarkers, is acquired from health systems. Statistical methodologies which allow updating of models with

new information as it becomes available are necessary to meet rising expectations and data availability.

### **7.3 | Assessment of biomarkers in multiple cohorts and distributed data settings**

The emergence of health systems as sources of data, as well as the proliferation of various consortia and data sharing initiatives undertaken by various industry and public sponsors, increase the need for reliable analytic methods in distributed data settings. Classical meta-analyses and patient-pooled analyses are increasingly common and can be effectively applied in biomarker research. The advantages of this process are obvious: more data from more diverse sources can only help in creating more generalizable knowledge. But many challenges remain, and it would be naïve to assume that increasing sample size will be a panacea. Currently, meta-analysis requires that data be analyzed at the “lowest common denominator” under a critical assumption that parameters of interest are homogeneous across the different underlying subpopulations from which different datasets are collected. Such assumption of homogeneity may be easily violated in practice<sup>65</sup> and therefore potentially important information which is not consistently available might not be used. Statistical methods that allow us to evaluate and accommodate study heterogeneities are needed to provide sensible data harmonization leading to a true gain of statistical power.<sup>66</sup>

Enhancing open collaborative mechanisms should be a priority. Too often, funded consortia become closed groups. In thinking about openness, considering the collaborative platform is important because subject-level data is needed for full assessment and understanding of study heterogeneities, which is essential to come up with good strategies in the analysis of integrated data. The recent push for data sharing and transparency from multiple industry and public sources<sup>67-70</sup> with a definitive set of recommendations contained in the recent report from the Institute of Medicine should help alleviate some of these concerns.<sup>71</sup>

Our list is by no means comprehensive. Several other important topics in biomarker research need dedicated attention which goes beyond this overview. These include the applications of machine learning methods for biomarker discovery, appropriate ways of building biomarker

models for sequential screening as well as methods to appropriately handle missing data in biomarker studies.

## 8 | SUMMARY

A plethora of methodological questions and challenges face prognostic biomarker researchers. Our considerations are far from exhaustive, and do not extend beyond prognostic biomarker research. Other areas of biomarker research are outside the scope of this presentation, but are extremely important and require separate consideration. First, design is of critical importance to the success of a proposed study, but effective tools for designing cohort studies in which biomarkers do not have strong signals are lacking. Second, predictive biomarkers might hold even more promise than prognostic biomarkers in affecting patient care. For example, in oncology, predictive biomarkers and genetic information play key roles in identifying individuals most likely to benefit from new therapies. Furthermore, predictive biomarkers are closely related to the problem of heterogeneity of treatment and risk-benefit considerations associated with many therapies which have side effects that cannot be ignored. The use of biomarkers should help identify those most likely to benefit from a specific therapeutic approach, as well as those with different risks of harm. Several of the topics raised here are relevant to approaches using predictive biomarkers.

Third, biomarkers as mediators of effect play an essential role in enhancing our understanding of pathways of disease. There are many sub-phenotypes of disease that might be related to specific biomarkers (e.g., biomarkers of cardiac injury or kidney injury). Assessing the contribution of biomarkers in mediating disease can help the field progress toward causality and temporal expectations. For example, knowing that 30% of acute kidney injury is mediated through cardiac injury implies that the maximum contribution of non-cardiac kidney injury is 70%. Using such disease models, path analyses and causal mediation methods<sup>72-74</sup> and integrated biomarker outcomes might help us to better understand the underlying disease.



Finally, the use of biomarkers as surrogates requires further research and consideration. This is particularly important as we move away from large outcome trials and towards smaller, more targeted and more cost-effective studies which can be enhanced by the use of biomarkers.

We hope that our formulation of questions and challenges will help direct biostatisticians and other methodological investigators towards problems of most pressing interests and needs. These considerations might also help applied and clinical researchers set their expectations and better articulate their needs and problems.

**REFERENCES**

1. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, Go AS, Harrell FE Jr., Hong Y, Howard BV, Howard VJ, Hsue PY, Kramer CM, McConnell JP, Normand SL, O'Donnell CJ, Smith SC Jr., Wilson PW. American Heart Association Expert Panel on Subclinical Atherosclerotic, D, Emerging Risk, F, the Stroke, C: Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*, 2009;119:2408-2416.
2. U.S. Food and Drug Administration. (2014). Guidance for Industry and FDA Staff: Qualification Process for Drug Development Tools. *Center for Drug Evaluation and Research (CDER)*. Silver Spring, MD: Office of Communications, Division of Drug Information.
3. Sniderman AD, D'Agostino RB, Pencina MJ. The role of physicians in the era of predictive analytics. *JAMA*, 2015;314:25-26.
4. Demler OV, Pencina MJ, D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. *Stat Med*, 2012;31:2577-2587.
5. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med*, 2013;32:1467-1482.
6. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? *Ann Intern Med*, 2014;160:136-137.
7. Harrell FE. *Regression modeling strategies*, New York, Springer, 2015.
8. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 2010;21:128-138.
9. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *J R Stat Soc a Stat*, 2009;172:729-748.

10. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Med Decis Making*, 2013;33:490-501.
11. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med*, 2006;25:3474-3486.
12. Pencina MJ, D'Agostino RB: Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*, 2004;23:2109-2123.
13. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am J Epidemiol*, 2011;174:364-374.
14. Tjur T. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *Am Stat*, 2009;63:366-372.
15. Korn EL, Simon R. Measures of explained variation for survival data. *Stat Med*, 1990;9:487-503.
16. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*, 2006;355:2631-2639.
17. Zethelius B, Berglund L, Sundstrom J, Ingelsson E, Basu S, Larsson A, Venge P, Arnlov J. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *N Engl J Med*, 2008;358:2107-2116.
18. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*, 2010;172:971-980.
19. White IR, Rapsomaniki E. Covariate data. *Biom J*, 2015;57:592-613. - adjusted measure

20. Pepe MS, Fan J, Seymour CW, Li C, Huang Y, Feng Z. Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clin Chem*, 2012;58:1242-1251.
21. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*, 2015;68:25-34.
22. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 2007;115:928-935.
23. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn*, 2009;77:103-123.
24. Hilden J. The area under the Roc curve and its competitors. *Medical Decision Making*, 1991;11:95-101.
25. Yates JF: External Correspondence - Decompositions of the Mean Probability Score. *Organ Behav Hum Perf*, 1982;30:132-156.
26. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*, 2012;176:473-481.
27. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*, 2008;27:157-172; discussion 207-112 .
28. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest*, 2012;42:216-228.
29. Bansal A, Pepe MS. When does combining markers improve classification performance and what are implications for practice? *Stat Med*, 2013;32:1877-1892.
30. Demler OV, Pencina MJ, D'Agostino RB. Impact of correlation on predictive ability of biomarkers. *StatMed*, 2013;32:4196-4210.

31. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. *Stat Med*, 2015;34:1659-1680.
32. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol*, 2011;173:1327-1335.
33. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*, 2015;162:W1-73.
34. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 1996;15:361-387.
35. 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. *CJASN*, 2014;9:1488-1496.
36. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema, JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*, 2001;54:774-781.
37. Pencina MJ, D'Agostino RB Sr. Thoroughly modern risk prediction? *Sci Transl Med*, 2012;4:131fs110.
38. Parikh CR, Thiessen-Philbrook H. Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. *JASN*, 2014;25:1621-1629.
39. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*, 2015;162:55-63.
40. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Dec Making*, 2015;35:162-169.

41. Sniderman AD, Pencina M, Thanassoulis G. Limitations in the conventional assessment of the incremental value of predictors of cardiovascular risk. *Curr Opin Lipidol*, 2015;26:210-214.
42. Holland PW. Causal inference, path analysis, and recursive structural equation models (with discussion). *Sociol Methodol*, 1988;18:449–484.
43. Imai K, Tingley D, Yamamoto T. Experimental designs for identifying causal mechanisms. *J R Stat Soc a Stat*, 2013;176:5-51.
44. Steyerberg EW, Vedder MM, Leening MJ, Postmus D, D'Agostino RB Sr., Van Calster B, Pencina MJ. Graphical assessment of incremental value of novel markers in prediction models: from statistical to decision analytical perspectives. *Biometrical Journal Biometrische Zeitschrift*, 2015;57:556-570.
45. Hedeker DR, Gibbons RD. *Longitudinal data analysis*, Hoboken, N.J., Wiley-Interscience, 2006.
46. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Statist Assoc*, 1999;94:496-509.
47. Chapfuwa P, Tao C, Li C, Page C, Goldstein B, Carin L, Henao R. Adversarial Time-to-Event modelling. International Conference on Machine Learning, 2018.
48. Andersen PK. *Statistical models based on counting processes*, New York, Springer-Verlag, 1993.
49. Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 1975;72:20-22.
50. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the *c*-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat Med*, 2012;31:2610-2626.
51. Li JL, Jiang BY, Fine JP. Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics*, 2013;14:382-394.

52. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, 2009;20:555-561.
53. Hallan SI, Matsushita K, Sang Y, Mahmoodi BK, Black C, Ishani A, Kleefstra N, Naimark D, Roderick P, Tonelli M. Age and association of kidney measures with mortality and end-stage renal disease. *JAMA*, 2012;308:2349-2360.
54. Vedder MM, de Bekker-Grob EW, Lilja HG, Vickers AJ, van Leenders GJ, Steyerberg EW, Roobol MJ. The added value of percentage of free to total prostate-specific antigen, PCA3, and a kallikrein panel to the ERSPC risk calculator for prostate cancer in prescreened men. *Eur Urol*, 2014;66:1109-1115.
55. Vickers AJ, Pencina MJ. Prostate-specific antigen velocity: new methods, same results, still no evidence of clinical utility. *Eur Urol*, 2013;64:394.
56. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *J Stat Softw*, 2010;35:1-33.
57. Hatfield LA, Carlin BP. Clinically relevant graphical predictions from Bayesian joint longitudinal-survival models. *Health Serv Outcomes Res Meth*, 2012;12:169-181.
58. Vickers AJ, Till C, Tangen CM, Lilja H, Thompson IM. An empirical evaluation of guidelines on prostate-specific antigen velocity in prostate cancer detection. *JNCI*, 2011;103:462-469.
59. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncology*, 2010;28:2796-2801.
60. Komarek A, Komarkova L. Clustering for multivariate continuous and discrete longitudinal data. *Ann Appl Stat*, 2013;7:177-200.
61. Arnlov J, Pencina MJ, Amin S, Nam BH, Benjamin EJ, Murabito JM, Wang TJ, Knapp PE, D'Agostino RB, Bhasin S, Vasan RS. Endogenous sex hormones and cardiovascular disease incidence in men. *Annals of Intern Med*, 2006;145:176-184.

62. Navar-Boggan AM, Peterson ED, D'Agostino RB, Neely B, Sniderman AD, Pencina MJ. Hyperlipidemia in early adulthood increases long-term risk of coronary heart disease. *Circulation*, 2015;131: 451-458.
63. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. *Stat Med*, 2013;32:1584-1618.
64. Yang W, Israni RK, Brunelli SM, Joffe MM, Fishbane S, Feldman HI. Hemoglobin variability and mortality in ESRD. *JASN*, 2007;18:3164-3170.
65. Wang F, Wang L, Song PXX. Quadratic inference function approach to merging longitudinal studies: validation and joint estimation. *Biometrika*, 2012;99:755-762.
66. Wang F, Song PX, Wang L. Merging multiple longitudinal studies with study-specific missing covariates: a joint estimating function approach. *Biometrics*, 2015;71(4):929-940.
67. Krumholz HM, Ross JS: A model for dissemination and independent analysis of industry data. *JAMA*, 2011;306:1593-1594.
68. Institute of Medicine. (U.S.) *Sharing clinical trial data: maximizing benefits, minimizing risk*, National Academies Press, 2015.
69. Pencina MJ, Louzao DM, McCour, BJ, Adams MR, Tayyabkhan RH, Ronco P, Peterson, ED. Supporting open access to clinical trial data for researchers: The Duke Clinical Research Institute–Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. *Am Heart J*, 2016;172:64-69.
70. U.S. National Institutes of Health. HHS and NIH take steps to enhance transparency of clinical trial results. 2014.
71. Lo B. Sharing clinical trial data maximizing benefits, minimizing risk. *JAMA*, 2015;313:793-794.
72. Detilleux J, Reginster JY, Chines A, Bruyere O. A Bayesian path analysis to estimate causal effects of bazedoxifene acetate on incidence of vertebral fractures, either directly or through non-linear changes in bone mass density. *Stat MethMed Res*, 2012;25(1):400-412.



73. Roysland K, Gran JM, Ledergerber B, von Wyl V, Young J, Aalen OO. Analyzing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study. *Stat Med*, 2011;30:2947-2958.
74. Winship C, Mare RD. Structural equations and path-analysis for discrete data. *Am J Sociol*, 1983;89:54-110.

## TABLES AND FIGURES

**TABLE 1** Statistical metrics for different phases of biomarker evaluation

<b>Phase of Evaluation</b>	<b>Focus of Assessment</b>	<b>Potential Metrics</b>
Proof of Concept	Do novel marker levels differ between events and nonevents	Distribution of biomarker in events and nonevents Effect Size
Prospective Validation	Does the new marker predict development of future events?	Plots of marker's impact over time Relative Risks / Hazard Ratios
Incremental Value	Is the new marker associated with outcome after controlling for standard risk factors?  If association is established, report impact on: Discrimination  Calibration Reclassification	Likelihood Ratio $\chi^2$ Test Adequacy Index  Change in AUC, Discrimination Slope (IDI), other R-squared measures  Calibration Plot Reclassification tables, event and nonevent NRI
Clinical Utility	Does incorporation of the biomarker lead to better clinical decisions?	Change in Net Benefit (number of false positives per one true positive), Relative Utility (number needed to test), weighted NRI
Clinical Outcomes	Does use of the biomarker improve clinical outcomes?	Reduction in event rates among patients whose management is guided using new biomarker
Cost-effectiveness	Does routine measurement of the new marker affect duration and quality of life, as well as monetary costs?	Quality-adjusted life years (QALY) gained, cost per QALY



**TABLE 2** AUC of four different models based on age range and addition of a new uncorrelated marker with effect size of 0.5

Age Range	Model	Scenario 1	Scenario 2	Scenario 3	Scenario 4
		AUC			
35-75	Age only	0.600	0.700	0.800	0.900
35-75	Age + marker	0.667	0.734	0.815	0.905
70-73	Age only	0.512	0.538	0.561	0.590
70-73	Age + marker	0.639	0.644	0.652	0.663

Author Manuscript

**FIGURE 1** Distribution of predicted risk according to model with limited (age-only) and expanded set of predictors (age plus risk factors)

Caption: Dark grey – Age-only model. Light grey – Age plus risk factors.

**FIGURE 2** Distribution of predicted risk among individuals who do and do not experience events according to model with limited (age-only) and expanded set of predictors (age plus risk factors)

Caption: Top panel – Predicted risk among individuals who did not experience events.  
Bottom panel – Predicted risk among individuals who experienced events.

# Statistical Methods for Building Better Biomarkers of Chronic Kidney Disease

Michael J. Pencina<sup>1</sup> Chirag R. Parikh<sup>2</sup> Paul L. Kimmel<sup>3</sup> Nancy R. Cook<sup>4</sup> Joe Coresh<sup>5</sup>  
Harold I. Feldman<sup>6,8</sup> Andrea Foulkes<sup>7</sup> Phyllis A. Gimotty<sup>6,8</sup> Chi-yuan Hsu<sup>9</sup> Kevin Lemley<sup>10</sup>  
Peter Song<sup>11</sup> Kenneth Wilkins<sup>12</sup> Daniel R. Gossett<sup>3</sup> Yining Xie<sup>3</sup> Robert A. Star<sup>3</sup>

## Institution:

<sup>1</sup> Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC

<sup>2</sup> Division of Nephrology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

<sup>3</sup> Division of Kidney, Urology and Hematology, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD

<sup>4</sup> Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

<sup>5</sup> Departments of Epidemiology, Medicine and Biostatistics, Johns Hopkins University, Baltimore, MD

<sup>6</sup> Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

<sup>7</sup> Department of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA

<sup>8</sup> Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

<sup>9</sup> Division of Nephrology, University of California San Francisco, San Francisco, CA

<sup>10</sup> Division of Nephrology, Children's Hospital Los Angeles, Department of Pediatrics, University of Southern California Keck School of Medicine, Los Angeles, CA

<sup>11</sup> Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI

<sup>12</sup> Biostatistics Program, Office of the Director, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, and Department of Preventive Medicine & Biostatistics, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD

**CORRESPONDENCE**

Michael J. Pencina, Ph.D.  
Vice Dean for Data Science & Information Technology  
Professor of Biostatistics & Bioinformatics  
Duke University School of Medicine  
DUMC 2927  
200 Trent Drive  
M144 Davison Bldg.  
Durham, NC 27710  
USA  
Email: [michael.pencina@duke.edu](mailto:michael.pencina@duke.edu)