Imperfect Gold Standards for Kidney Injury Biomarker Evaluation

Sushrut S. Waikar,* Rebecca A. Betensky,† Sarah C. Emerson,‡† and Joseph V. Bonventre*§

*Renal Division, Brigham and Women’s Hospital, Department of Medicine, Harvard Medical School, Boston, Massachusetts; †Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; ‡Department of Statistics, Oregon State University, Corvallis, Oregon; and §Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts

ABSTRACT
Clinicians have used serum creatinine in diagnostic testing for acute kidney injury for decades, despite its imperfect sensitivity and specificity. Novel tubular injury biomarkers may revolutionize the diagnosis of acute kidney injury; however, even if a novel tubular injury biomarker is 100% sensitive and 100% specific, it may appear inaccurate when using serum creatinine as the gold standard. Acute kidney injury, as defined by serum creatinine, may not reflect tubular injury, and the absence of changes in serum creatinine does not assure the absence of tubular injury. In general, the apparent diagnostic performance of a biomarker depends not only on its ability to detect injury, but also on disease prevalence and the sensitivity and specificity of the imperfect gold standard. Assuming that, at a certain cutoff value, serum creatinine is 80% sensitive and 90% specific and disease prevalence is 10%, a new perfect biomarker with a true 100% sensitivity may seem to have only 47% sensitivity compared with serum creatinine as the gold standard. Minimizing misclassification by using more strict criteria to diagnose acute kidney injury will reduce the error when evaluating the performance of a biomarker under investigation. Apparent diagnostic errors using a new biomarker may be a reflection of errors in the imperfect gold standard itself, rather than poor performance of the biomarker. The results of this study suggest that small changes in serum creatinine alone should not be used to define acute kidney injury in biomarker or interventional studies.


Diagnostic tests are judged on the basis of their ability to classify individuals according to disease status. The actual disease status of an individual is often not known with certainty in clinical medicine. Myocardial infarction, for example, is ultimately a pathologic diagnosis; however, in practice, the diagnosis must be made premortem on the basis of blood biochemical markers of myocardial necrosis along with electrocardiographic changes or ischemic symptoms. Cardiac troponins (I and T) are widely considered to be adequate biomarkers for the diagnosis of acute myocardial infarction on the basis of their myocardial tissue specificity and their association with important clinical outcomes. Few diagnostic tests, however, enjoy such acceptance as biomarkers.

Serum creatinine (SCR) concentration is the surrogate test used to diagnose AKI. Importantly, SCR is acknowledged as an inadequate gold standard for several reasons. SCR has poor specificity in the settings of prerenal azotemia, changes in dietary intake, and drug-induced changes in tubular secretion of creatinine, all of which may lead to changes in SCr without actual injury to the kidney. SCR has poor sensitivity in the setting of adequate renal reserve, when SCr may not change despite acute tubular injury because of compensatory increases in function by other nephrons. The use of SCr may also lead to delays in diagnosis because of the relatively slow kinetics of the rise in SCr after injury. A recent study of nephrotoxicity in rodents showed that the sensitivity of SCr is poor, especially when histologic injury to the kidney is mild. Because of the limitations of SCr, there has been considerable interest recently in identifying a troponin for the kidney. Despite widespread acknowledgment of the limitations, definitions of AKI continue to rely on SCr as a diagnostic standard, perhaps because of the historical absence of validated primary biomarkers of injury. New biomarkers of tubular injury have been sought because the kidney tubule is the most metabolically active segment of the nephron and is uniquely susceptible to ischemic and nephrotoxic insults. Animal and human studies have resulted in a number...
of promising biomarkers that may revolutionize the diagnosis of AKI, enabling more accurate and earlier diagnosis of tubular injury, and clinical studies of these biomarkers in humans are increasing.

In clinical studies, researchers have used changes in SCr as the gold standard against which to test novel tubular injury biomarkers. This method is less than ideal, however, because the inadequacies of SCr are the raison d’être for kidney injury biomarker discovery and qualification studies in the first place. Furthermore, SCr is an approximate and imperfect measure of glomerular filtration and does not directly reflect tubular function or injury, which many kidney injury biomarkers identify. Change in SCr is a continuous variable, but it is dichotomized to define a binary outcome (AKI present or absent). The choice of a cutoff value will directly affect the true sensitivity and specificity of SCr as the gold standard. Using small changes in SCr to define AKI will lead to relatively higher sensitivity but lower specificity, whereas using larger changes in SCr—or the need for renal replacement therapy in severe AKI—will result in lower sensitivity but higher specificity for true tubular injury. As was recently discussed and now addressed in detail in this study, even minor imperfections in the diagnostic performance of a gold standard test such as SCr can result in significant misinterpretations of the diagnostic performance of a novel biomarker under investigation. The conceptual framework for this exercise is that true AKI—the actual disease or clinical condition that our diagnostic tests intend to identify—is not synonymous with the clinical conditions identified by changes in SCr. Using changes in SCr as the gold standard may therefore lead to substantial distortions of the apparent diagnostic performance characteristics of a novel biomarker, as discussed later in this article. For the purposes of this exercise, we will assume that the tubular injury process we are attempting to identify with a biomarker can be unequivocally known and dichotomized. It may be reasonable to use histopathology as a dichotomized marker, although even histopathology may not be adequate due to a time delay in the development of pathologic lesions and insensitivity of pathology when injury is subtle. Furthermore, histopathology is not practical for clinical studies, in which biopsies are infrequent.

**Table 1. The effect of an imperfect gold standard on the sensitivity and specificity of a new biomarker that is in fact 100% sensitive and specific for AKI**

<table>
<thead>
<tr>
<th>True AKI (i.e., tubular injury)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AKI</strong></td>
<td><strong>No AKI</strong></td>
</tr>
<tr>
<td>AKI according to SCr</td>
<td>180</td>
</tr>
<tr>
<td>No AKI according to SCr</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>200</td>
</tr>
<tr>
<td>Sensitivity = 90%</td>
<td>Specificity = 90%</td>
</tr>
</tbody>
</table>

**AKI According to SCr**

<table>
<thead>
<tr>
<th></th>
<th><strong>Total</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AKI</strong></td>
<td><strong>No AKI</strong></td>
</tr>
<tr>
<td>New biomarker positive</td>
<td>180</td>
</tr>
<tr>
<td>New biomarker negative</td>
<td>80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>260</td>
</tr>
</tbody>
</table>

Apparent sensitivity = 69% Apparent specificity = 97%

**MODELING THE EFFECTS OF USING AN IMPERFECT GOLD STANDARD ON EVALUATION OF THE PERFORMANCE OF AN IDEAL BIOMARKER**

To understand how an imperfect gold standard can distort the apparent diagnostic performance of a new test, consider the following scenario. In a study of 1000 individuals, assume that 200 truly have AKI with tubular injury, with the diagnosis based not on changes in SCr, but on another ideal diagnostic test that has 100% sensitivity and 100% specificity. The imperfect gold standard test, SCr, would have its own sensitivity and specificity for the true diagnosis of AKI: SCr would have neither perfect sensitivity due to renal reserve in some patients nor perfect specificity due to prerenal azotemia. Even if the sensitivity and specificity of SCr are each 90% (likely to be overestimates), a 2 × 2 contingency table (Table 1) can be constructed that shows how many individuals are correctly and incorrectly classified by SCr as having AKI.

Of the 800 individuals without true AKI as defined by tubular injury, SCr would falsely identify 80 as having AKI. Of the 200 individuals with true AKI, SCr would falsely identify 20 as not having AKI. Now imagine that a new biomarker is studied in this cohort of patients and that the new biomarker is in fact perfect compared with the true gold standard. How would such a perfect biomarker seem to perform compared with SCr? Table 1 shows the results: The apparent sensitivity of the perfect biomarker is only 69%, and the apparent specificity is 97%. Figure 1 illustrates the phenomenon graphically for sensitivity and specificity of 80%.

Assuming that the results of the gold standard and the novel biomarker are independent given disease status (an assumption termed conditional independence), the equations that describe the apparent sensitivity (Equation 1) and apparent specificity (Equation 2) of a novel biomarker are as follows:

\[
\text{Apparent sensitivity} = \frac{(\text{prevalence})(\text{sensitivity}_A)(\text{sensitivity}_B)}{\text{prevalence}(\text{sensitivity}_A)(\text{sensitivity}_B) + (1 - \text{prevalence})(1 - \text{specificity}_A)(1 - \text{specificity}_B)}
\]

\[
\text{Apparent specificity} = \frac{(\text{prevalence})(\text{sensitivity}_C)(\text{sensitivity}_D) + (1 - \text{prevalence})(1 - \text{specificity}_C)(\text{specificity}_D)}{(1 - \text{specificity}_D) + (1 - \text{prevalence})(1 - \text{specificity}_C)(\text{specificity}_D) + (\text{prevalence})(\text{sensitivity}_C)(1 - \text{specificity}_D)}
\]
where the subscripts $G$ and $B$ refer to the imperfect gold standard and the novel biomarker, respectively. Receiver operating characteristic (ROC) curves are graphical plots of sensitivity versus 1 - specificity; the area under the ROC curve (AUC-ROC) is a summary statistic widely used to assess diagnostic test performance characteristics. Because ROC curves are monotonic, the upper and lower bounds of the AUC-ROC are calculated for a given sensitivity and specificity value as follows:

Lower bound of AUC-ROC
\[
= \text{sensitivity} \times \text{specificity}
\]

Upper bound of AUC-ROC
\[
= \text{sensitivity} + ([1 - \text{sensitivity}] \times \text{specificity})
\]

The lower and upper bounds for AUC-ROC curves are derived by plotting the point (1 - specificity, sensitivity) and finding monotone curves through the given point that have minimal and maximal AUC-ROCs, respectively. These curves will be step functions with a vertical jump at $X = 1 - \text{specificity}$.

We now consider the implications for diagnostic studies of novel biomarkers using the imperfect gold standards likely used in diagnostic studies of AKI. For simplicity, we consider the special case in which the novel biomarker is in fact a perfect test.

Defining AKI by Small Changes in SCr
A recently proposed consensus definition of AKI is an increase in SCr of as little as 0.3 mg/dl over 48 h.\(^9\) The motivation for this definition (known as the Acute Kidney Injury Network [AKIN] stage 1) was the finding that small changes in SCr herald a significantly increased risk of death in hospitalized individuals.\(^13,14\) Lowering the threshold for a diagnostic test usually increases sensitivity for disease detection at the expense of specificity.

The diagnostic performance characteristics of SCr have not been adequately assessed in humans, but results from animal models of nephrotoxicity have shown poor sensitivity when using histopathology as the gold standard.\(^7\) We generously assume that AKIN stage 1 has 80% sensitivity (i.e., 20% of individuals with true AKI have no increase in SCr, perhaps due to renal reserve) and 90% specificity (i.e., 10% of individuals with no actual parenchymal kidney injury have a $\geq 0.3$ mg/dl increase in SCr, perhaps due to prerenal azotemia). At a true disease prevalence of 20%, the apparent sensitivity of a perfect biomarker (compared with AKIN stage 1) is 67%, apparent specificity is 95%, and the

Figure 1. Apparent diagnostic performance characteristics of a perfect biomarker compared with an imperfect gold standard. (A) Actual disease prevalence = 20%. Red squares represent true disease positives. Blue squares represent true disease negatives. (B) Imperfect gold standard sensitivity = 80%. Of the 20 true disease positives, 4 are classified as disease negative by the imperfect gold standard (darker blue). (C) Imperfect gold standard specificity = 80%. Of the 80 true disease negatives, 16 are classified as disease positive by the imperfect gold standard (darker red). (D) The imperfect gold standard has classified 32 individuals as disease positive (red) and 68 individuals as disease negative (blue). A perfect biomarker will correctly identify only 16 of the original true positives (lighter red) and fail to identify the 16 false positives (darker red), leading to an apparent sensitivity of 16/32 = 50%. A perfect biomarker will identify as disease negative only 64 of the original true negatives (lighter blue) and fail to identify the 4 false negatives (darker blue), leading to an apparent specificity of 64/68 = 94%.

\[
\left(1 - \text{prevalence}\right)\left(\text{specificity}_G\right)\left(\text{specificity}_B\right) + \left(\text{prevalence}\right)
\]

\[
\left(1 - \text{prevalence}\right)\left(\text{specificity}_G\right)\left(1 - \text{specificity}_B\right) + \left(1 - \text{specificity}_G\right)\left(\text{prevalence}\right)\left(1 - \text{specificity}_B\right)
\]

\[
\left(1 - \text{specificity}_G\right)\left(\text{prevalence}\right)\left(\text{specificity}_B\right) + \left(1 - \text{specificity}_B\right)\left(\text{prevalence}\right)\left(1 - \text{specificity}_G\right)
\]

\[
\left(1 - \text{specificity}_G\right)\left(1 - \text{specificity}_B\right)
\]

where the subscripts $G$ and $B$ refer to the imperfect gold standard and the novel biomarker, respectively. Receiver operating characteristic (ROC) curves are graphical plots of sensitivity versus 1 - specificity; the area under the ROC curve (AUC-ROC) is a summary statistic widely used to assess diagnostic test performance characteristics. Because ROC curves are monotonic, the upper and lower bounds of the AUC-ROC are calculated for a given sensitivity and specificity value as follows:
lower and upper bounds of the AUC-ROC are 0.63 and 0.98, respectively (0.81 under the assumption of conditional independence).

Figure 2 illustrates how differences in true disease prevalence affect these estimates. At low disease prevalence, the dominant effect is on a perfect biomarker’s apparent sensitivity. In contrast, apparent specificity remains high at low disease prevalence. The apparent AUC-ROC exhibits a wide range

---

**Figure 2.** Apparent diagnostic performance characteristics of a perfect biomarker when an imperfect gold standard has 80% sensitivity and 90% specificity. (A) The apparent sensitivity and specificity. (B) The apparent AUC-ROC curve (solid line, under the assumption of conditional independence) with lower and upper bounds (dotted lines). Results are plotted for a range of true disease prevalence estimates.

**Table 2.** Apparent diagnostic performance characteristics of a perfect biomarker as a function of disease prevalence and sensitivity and specificity of an imperfect gold standard

<table>
<thead>
<tr>
<th>Imperfect Gold Standard Diagnostic Performance</th>
<th>True Disease Prevalence, %</th>
<th>Apparent Sensitivity, %</th>
<th>Apparent Specificity, %</th>
<th>Lower Bound of AUC-ROC</th>
<th>Upper Bound of AUC-ROC</th>
<th>AUC-ROC Assuming Conditional Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% sensitive, 90% specific</td>
<td>5</td>
<td>30</td>
<td>99</td>
<td>0.29</td>
<td>0.99</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>47</td>
<td>98</td>
<td>0.46</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>67</td>
<td>95</td>
<td>0.63</td>
<td>0.98</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>84</td>
<td>87</td>
<td>0.73</td>
<td>0.98</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>92</td>
<td>75</td>
<td>0.69</td>
<td>0.98</td>
<td>0.84</td>
</tr>
<tr>
<td>80% sensitive, 80% specific</td>
<td>5</td>
<td>17</td>
<td>99</td>
<td>0.17</td>
<td>0.99</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>31</td>
<td>97</td>
<td>0.30</td>
<td>0.98</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>50</td>
<td>94</td>
<td>0.47</td>
<td>0.97</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>73</td>
<td>86</td>
<td>0.62</td>
<td>0.96</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>86</td>
<td>73</td>
<td>0.64</td>
<td>0.96</td>
<td>0.79</td>
</tr>
<tr>
<td>25% sensitive, 100% specific</td>
<td>5</td>
<td>100</td>
<td>96</td>
<td>0.96</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>100</td>
<td>92</td>
<td>0.92</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>100</td>
<td>84</td>
<td>0.84</td>
<td>1.00</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>100</td>
<td>67</td>
<td>0.67</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>100</td>
<td>47</td>
<td>0.47</td>
<td>1.00</td>
<td>0.74</td>
</tr>
<tr>
<td>25% sensitive, 99% specific</td>
<td>5</td>
<td>57</td>
<td>96</td>
<td>0.55</td>
<td>0.98</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>74</td>
<td>92</td>
<td>0.68</td>
<td>0.98</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>86</td>
<td>84</td>
<td>0.72</td>
<td>0.98</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>94</td>
<td>66</td>
<td>0.63</td>
<td>0.98</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>97</td>
<td>47</td>
<td>0.46</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td>100% sensitive, 25% specific</td>
<td>5</td>
<td>7</td>
<td>100</td>
<td>0.07</td>
<td>1.00</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>13</td>
<td>100</td>
<td>0.13</td>
<td>1.00</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>25</td>
<td>100</td>
<td>0.25</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>47</td>
<td>100</td>
<td>0.47</td>
<td>1.00</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>67</td>
<td>100</td>
<td>0.67</td>
<td>1.00</td>
<td>0.83</td>
</tr>
</tbody>
</table>
of possible values, depending on the nature of the concordance between the imperfect gold standard and the perfect biomarker (Table 2).

Contrast nephropathy is one example in which the risk of significant AKI is low.\textsuperscript{15} The majority of cases of contrast nephropathy are diagnosed on the basis of small changes in SCr alone,\textsuperscript{16} which may very well represent misclassifications of conditions such as prerenal azotemia. If 100 patients are studied and 5 have true AKI (i.e., tubular injury from contrast media) and 5 have prerenal azotemia (i.e., no tubular injury but SCr changes), a perfect biomarker of tubular injury may seem to have only 50% sensitivity using the small increase in SCr as the gold standard.

DEFINING AKI ACCORDING TO THE NEED FOR DIALYSIS

The need for renal replacement therapy after AKI usually reflects severe parenchymal kidney injury. Rare exceptions may include dialysis initiation in patients with advanced chronic kidney disease or dialysis for volume overload, electrolyte abnormalities, or toxic ingestions. Assume that the gold standard in this case, the need for acute dialysis, has a specificity of 100% (i.e., among patients without true AKI or tubular injury, no one requires acute dialysis) and a sensitivity of 25% (i.e., of all individuals with true AKI, only 25% require dialysis). At a true disease prevalence of 20%, the apparent sensitivity of a perfect biomarker is 100%, apparent specificity is 84%, and the lower and upper bounds of the apparent AUC-ROC are 0.84 and 1.00, respectively (0.92 under the assumption of conditional independence).

Figure 3 illustrates how differences in true disease prevalence affect these estimates. Note that, in this case, the apparent sensitivity and upper bound of the AUC-ROC remain 100% only when the gold standard is in fact perfectly specific. Even rare false positives (specificity of 99% of the imperfect gold standard) lead to an apparent sensitivity of 86% and lower and upper bounds of the apparent AUC-ROC of 0.72 and 0.98, respectively (0.85 under the assumption of conditional independence; Table 2).

PREVIOUS WORK ON THE IMPERFECT GOLD STANDARD

The effect of imperfect reference standards has generally been neglected in the expanding clinical literature on diagnostic test accuracy. In the biostatistical literature, several approaches have been proposed based on the assumption of conditional independence. If the gold standard has a known false positive and false negative rate, and the true disease prevalence is known, the apparent sensitivity and specificity of a new diagnostic test can be calculated.\textsuperscript{17–19} Unfortunately, the required parameters are not usually known with certainty. Hui and Walter\textsuperscript{20} proposed a method to estimate the error rate of a diagnostic test even when the error rates of the gold standard are unknown by applying both tests simultaneously in two populations with different disease prevalence. Walter and Irwig\textsuperscript{21} reviewed latent class models for use when no gold standard exists; the approach requires a minimum of three (imperfect) diagnostic tests and the use of maximum likelihood techniques to yield estimates of disease prevalence and test accuracy. All of these approaches make the assumption of conditional independence of the new diagnostic test and the gold standard, which may not be a reasonable assumption in many clinical settings. Analytical approaches that incorporate conditional dependence have been described by Vacek\textsuperscript{22} and Phelps and Hutson.\textsuperscript{19}

Figure 3. Apparent diagnostic performance characteristics of a perfect biomarker when an imperfect gold standard has 25% sensitivity and 100% specificity. (A) The apparent sensitivity and specificity. (B) The apparent AUC-ROC curve (solid line, under the assumption of conditional independence) with lower and upper bounds (dotted lines). Results are plotted for a range of true disease prevalence estimates.
**EXAMPLES ASIDE FROM NEPHROLOGY: IRON DEFICIENCY ANEMIA**

AKI is not the only condition in clinical medicine that is classified on the basis of an imperfect and continuous value gold standard. Iron deficiency anemia, for example, is diagnosed in clinical practice on the basis of serum ferritin concentration because the gold standard—examination of a bone marrow aspirate for stainable iron—is invasive, expensive, and time-consuming, and carries its own set of diagnostic uncertainties. Another potential gold standard examination—assessing the therapeutic response to a trial of iron supplementation—is not practical for making an initial diagnosis. Soluble transferrin receptor (sTfR) concentration was examined as an alternative index of iron status in anemia. Wians et al.\(^{23}\) for example, reported an AUC-ROC of 0.958 for sTfR as a biomarker to distinguish iron deficiency from anemia of chronic disease: ferritin $<$ 20 ng/ml was used to define iron deficiency, and ferritin $>$ 1.5 $\times$ the sex-specific upper limit of the normal reference range defined anemia of chronic disease. In this study, exclusion of intermediate values for ferritin may have led to overestimation of the accuracy of sTfR. Furthermore, the authors did not address misclassification of disease status by the imperfect gold standard, ferritin. Although ferritin has become the de facto gold standard for the assessment of iron stores, ferritin does not perfectly correlate with the result of the true gold standard: examination of stainable iron in bone marrow aspirates. Reports of the sensitivity and specificity of ferritin have ranged from 52% to 71% and 93% to 100%, respectively.\(^{24,25}\)

**DEFINING AKI FOR BIOMARKER STUDIES**

Despite consensus definitions of AKI\(^ {9,26}\) recent clinical studies of novel biomarkers used a variety of Scr criteria to define AKI (e.g., a 50% increase over baseline,\(^ {27}\) 0.3 mg/dl or 50% increase over baseline,\(^ {28}\) and 25% increase over baseline\(^ {29}\) ). As discussed in this study, varying the choice of definition for AKI will affect the apparent diagnostic performance of a novel biomarker. This has, in fact, been shown for the biomarker neutrophil gelatinase-associated lipocalin (NGAL). Haase et al.\(^ {30}\) reported that the AUC-ROC for NGAL was 0.75 when using the AKIN stage 1 definition (0.3 mg/dl or 50% increase within 48 h) and 0.83 when defining AKI as the need for dialysis.

The finding of low sensitivity of a promising kidney injury biomarker in a setting such as contrast nephropathy, in which the expected true prevalence of disease is low, should raise the question of disease misclassification by the gold standard. In fact, this has been reported in a meta-analysis\(^ {30}\) of NGAL’s performance as a biomarker of AKI in various clinical settings. Sensitivity was only 77.8%, whereas specificity was 96.3% for three pooled studies involving contrast nephropathy defined by using small changes in Scr (25% or 0.5 mg/dl in two adult studies\(^ {31,32}\) and 50% in a pediatric study with mean baseline Scr values $<$ 1.0 mg/dl\(^ {33}\) ).

In clinical settings with a higher expected prevalence of true AKI, apparent sensitivity is less affected by false positive designations by the gold standard, but apparent specificity may be reduced by false negative designations. Higher true AKI prevalence may occur in biomarker studies performed in the intensive care unit (ICU), in which the frequency of hypotension and sepsis may lead to frequent tubular injury. The prediction of lower specificity for a biomarker in high-risk clinical settings has been borne out for NGAL, which had a pooled specificity of 75.5% for the diagnosis of AKI in critically ill patients.\(^ {30}\)

Mishra et al.\(^ {27}\) reported an AUC-ROC of 0.998 for NGAL as a biomarker in pediatric cardiac surgery using a 50% rise in Scr as the definition of AKI. The inability to replicate this level of performance in subsequent larger studies\(^ {34–36}\) suggests that the findings were either specific to the unique pediatric population studied or reflect conditional dependence between Scr and NGAL. The latter point merits emphasis: an AUC-ROC of 1.0 may in fact indicate a biomarker that is subject to the same limitations as Scr.

**NON–CREATININE-BASED ENDPOINTS FOR BIOMARKER STUDIES**

Longer-term outcomes such as mortality, or the eventual need for renal replacement therapy, may be used to compare a new biomarker against an imperfect gold standard. Indeed, the association of troponin with mortality,\(^ {2,3}\) in conjunction with its known tissue specificity,\(^ {1}\) contributed to its adoption for the diagnosis of myocardial infarction.\(^ {37}\) One difficulty with extrapolating this approach to AKI biomarker studies may be the large sample sizes required for statistical power, the long latency between an episode of kidney injury and outcomes such as progressive chronic kidney disease, and confounding by other risk factors and clinical events.

A biomarker may also be associated with mortality or another long-term outcome because of an association with sepsis or inflammation, without being reflective of actual kidney injury. The observation that even a 0.1 mg/dl increase in Scr is associated with mortality does not assure that this small change in Scr is a reliable predictor of kidney injury.\(^ {38}\) This may have important implications for using a biomarker as a surrogate endpoint in an interventional study because a surrogate biomarker should be in the causal pathway of the disease process.

Another possible study design involves using exposure status to test a biomarker’s accuracy. Consider, for example, a study in which biomarkers are tested after exposure to a drug with known nephrotoxic potential, such as cisplatin. If biomarkers are measured in well matched patients who did and did not receive cisplatin, exposure status could be used as the criterion against which biomarkers are compared, assuming that there is a high correlation between exposure status and kidney injury. In this type of design, Scr does not need to be used as a gold standard. The risk of such a study design, however, is the identification of biomarkers...
that are too sensitive to be of clinical use. Tubular enzymes such as\( N\)-acetyl-\( \beta\)-(D)-glucosaminidase as well as \( \alpha\)-glutathione and \( \pi\)-glutathione \( S\)-transferases, for example, are known to be elevated in the urine after cardiac surgery\(^{39–41}\) but have not been adopted (perhaps inappropriately) into clinical practice because of concerns regarding possible nonspecificity of the appearance of tubular enzymes in the urine. Nevertheless, these types of studies may be useful to identify biomarkers that fulfill the vision put forth by the US Food and Drug Administration regarding qualification and use of biomarkers in drug development, dose regulation, and clinical monitoring of nephrotoxic drug exposure.\(^{42}\)

The ultimate validation of a biomarker’s utility would be to demonstrate, ideally in a randomized controlled trial, that biomarker measurement actually alters clinical management and improves clinical outcomes. For example, results might show that clinical decisions aided by knowledge of AKI status as inferred by biomarker elevations lead to reductions in length of stay, ICU-related complications, need for renal replacement therapy, long-term renal function decline, or mortality.

**BIOMARKERS OF TUBULAR INJURY VERSUS BIOMARKERS OF GFR**

It should be noted that tubular injury—the focus of many biomarkers—may not always couple with reductions in GFR, thereby leading to apparent false positive results. Conversely, reductions in GFR from prerenal azotemia may not always reflect tubular injury, thereby leading to apparent false negative results with the biomarker. In this regard, one unresolved question is which pathophysiological process is more likely to be clinically relevant and important to monitor: changes in GFR or tubular injury? The answer to this question is clear in some contexts, such as preclinical nephrotoxicity studies, in which SCr underperforms tubular injury biomarkers when histology is used as the arbiter.\(^7\)

In hospital-acquired AKI, the measure of a biomarker’s performance may be the additional clinical value afforded by its measurement. In this context, validation of a biomarker may come from a trial in which patients at risk for AKI (e.g., after cardiac surgery, in the ICU, or with acute decompensated heart failure) are randomized to a diagnostic strategy involving conventional biomarkers alone (BUN, SCr, or urine output) or conventional plus novel biomarkers of tubular injury. Fluid or hemodynamic management may be adjusted on the basis of information on kidney injury. Rational endpoints for such a trial could include: net fluid balance, duration of mechanical ventilation, length of stay, or need for renal replacement therapy. Such an approach will inform the assessment of biomarkers relative to functional changes reflected by changes in SCr.

**SUGGESTED SCr-BASED DEFINITIONS FOR BIOMARKER STUDIES**

Using a highly sensitive but poorly specific definition for AKI will lead to a relatively large number of false positive cases compared with true positive cases, thereby significantly reducing the apparent sensitivity of a novel biomarker. In contrast nephropathy, for example, one currently accepted definition\(^43\) of an increase of \( \geq 0.5 \text{ mg/dl} \) or 25% in SCr may be adequate for epidemiology studies but is inappropriate for a diagnostic biomarker study. Figure 4 shows that, for the low prevalence estimates we may expect in many AKI studies, maintaining high specificity of an AKI definition is more important than high sensitivity. If the choice is between a highly sensitive definition of AKI with some false positives versus a perfectly specific AKI definition with some false negatives, the latter leads to far less distortion of biomarker performance. Accordingly, we suggest at a minimum using RIFLE I, AKI stage 2, or stage 2 of the definition based on creatinine kinetics that we previously proposed.\(^6\) Higher specificity will be achieved with definitions that require larger changes in SCr. Studies should be adequately powered ideally for hard endpoints, such as the need for renal replacement therapy for uremia to reduce ambiguity with respect to the presence of...
true kidney injury, because smaller studies with liberal endpoints may lead to gross misrepresentations of biomarker performance. Consideration should also be given to approaches suggested in the biostatistical literature when a reliable gold standard test does not exist; for example, measuring two biomarkers that are assumed to be conditionally independent in individuals drawn from two populations with different disease prevalence. As described in Table 3, several study designs should be considered to prevent researchers from prematurely discarding novel biomarkers.

CONCLUSION

Biomarker development in nephrology is crucial in the development of therapeutics strategies for AKI prevention and treatment. Underpowered studies using small changes in SCr as endpoints may have the unintended and perverse effect of underestimating the utility of novel biomarkers that actually outperform SCr itself. When using a nonideal gold standard to evaluate novel biomarkers, appropriate study design considerations become critical to avoid misleading conclusions that would preclude the acceptance into clinical medicine of new useful biomarkers that have the chance to revolutionize the approach to AKI diagnosis and therapeutics.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grants R33DK074099 (S.S.W., R.A.B., and J.V.B.), K23DK075941 (S.S.W.), and T32NS048005 (S.C.E.). This work was conducted with support from Harvard Catalyst! The Harvard Clinical and Translational Science Center (NIH Award UL1 RR 025758 and financial contributions from Harvard University and its affiliated academic health care centers).

The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, the National Center for Research Resources, or the National Institutes of Health.

REFERENCES


Table 3. Suggested study designs of biomarker studies in AKI

1. Use an AKI definition that minimizes false positive misclassifications of disease status, particularly in clinical settings with relatively low expected prevalence of true AKI. Examples of appropriate AKI definitions include the following: Acute Kidney Injury Network (AKIN) stage 2 or 3, RIFLE stage “I” or “F,” or stage 2 or 3 of a definition based on SCr kinetics.
2. Examine hard endpoints (e.g., in-hospital or 28-d mortality, need for renal replacement therapy (RRT), or substantial and sustained reduction in kidney function at 28 d) instead of short-term changes in SCr—recognizing, however, that the “injury” marker is then being tested as a prognostic marker and not necessarily an injury marker.
3. Test whether biomarkers outperform conventional measures of kidney function in the accurate identification of individuals exposed to nephrotoxic agents versus well matched, nonexposed individuals.
4. Test whether clinical outcomes (most likely intermediate endpoints such as length of stay or postoperative fluid balance, but ideally hard endpoints such as renal replacement therapy or mortality) are improved in patients who do versus do not undergo kidney injury biomarker testing with subsequent clinical decision making on the basis of biomarker results.
5. Use the approaches suggested in the biostatistical literature when a reliable gold standard does not exist; for example, measurement of two biomarkers that are assumed to be conditionally independent (e.g., SCr and a novel tubular injury biomarker) in individuals drawn from two populations with different disease prevalence.