Published in final edited form as:
Clin Trials. 2013 October ; 10(5): 696–700. doi:10.1177/1740774513497540.

Imperfect Gold Standards for 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

[†]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

[‡]Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts

Abstract

Background—Serum creatinine has been used as the diagnostic test for acute kidney injury (AKI) for decades despite having 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.

Conclusions—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. Apparent errors in diagnosis using a new biomarker may be a reflection of errors in the imperfect gold standard itself rather than poor performance of the biomarker.

Introduction

Diagnostic tests are judged on the basis of their ability to classify individuals according to disease status. Serum creatinine concentration (SCr) is the surrogate test for diagnosis of acute kidney injury (AKI). SCr is acknowledged to be an inadequate gold standard as it has poor specificity in some settings and poor sensitivity others. Change in SCr is a continuous variable, but is dichotomized to define a binary outcome (AKI present or absent). The choice of a cutoff 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; 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. Even minor imperfections in the diagnostic performance of a gold standard test like SCr can result in significant misinterpretations of the diagnostic performance of a novel biomarker

Correspondence to: Rebecca A. Betensky.

This Short Communication summarizes a presentation from the University of Pennsylvania's fifth conference on statistical issues in clinical trials: Emerging statistical issues in Biomarker validation for clinical trials. The material relies heavily on a recent published paper[1] and technical details of the material presented here are available therein.

Disclosure: JVB had a significant equity interest in Genzyme, which, for one year during the study, created a conflict of interest. Following review by the institution, JVB relinquished this interest to be consistent with Harvard and Partners Healthcare System policies, and Partners put a management plan into place.

under investigation. 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.

The Effects of Using an Imperfect Gold Standard

To understand how an imperfect gold standard can distort the apparent diagnostic performance of a new test, consider a study of 1,000 individuals; assume 200 truly have AKI with tubular injury with the diagnosis based not on changes in SCr, but another ideal diagnostic test that has 100% sensitivity and 100% specificity. The imperfect gold standard test, SCr, would then have its own sensitivity and specificity for the true diagnosis of AKI: SCr would not have perfect sensitivity due to renal reserve in some patients or perfect specificity due to pre-renal azotemia. Even if the sensitivity and specificity of SCr are each 90% (likely to be overestimates), then 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* when compared to the true gold standard. How would such a perfect biomarker appear to perform when compared to SCr? Table 1 shows the results: the apparent sensitivity of the perfect biomarker is only 69%, and the apparent specificity is 97%.

The equations that describe the apparent sensitivity and specificity of a novel biomarker, assuming the results of the gold standard and the novel biomarker are independent given disease status, an assumption termed conditional independence, are as follows:

1. *Apparent sensitivity*:

$$\frac{(\text{prevalence})(\text{sensitivity}_{\text{G}})(\text{sensitivity}_{\text{B}})+(1-\text{prevalence})(1-\text{specificity}_{\text{G}})(1-\text{specificity}_{\text{B}})}{(\text{prevalence})(\text{sensitivity}_{\text{G}})+(1-\text{prevalence})(1-\text{specificity}_{\text{G}})}$$

2. *Apparent specificity*:

$$\frac{(1-\text{prevalence})(\text{specificity}_{\text{G}})(\text{specificity}_{\text{B}})+(\text{prevalence})(1-\text{sensitivity}_{\text{G}})(1-\text{sensitivity}_{\text{B}})}{(1-\text{prevalence})(\text{specificity}_{\text{G}})+(\text{prevalence})(1-\text{sensitivity}_{\text{G}})}$$

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 receiver operating characteristic 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 can be calculated for a given sensitivity and specificity value (defined jointly relative to the same cutoff) as follows:

Lower bound of AUC-ROC = sensitivity × specificity

Upper bound of AUC-ROC = sensitivity + (1 − sensitivity) × 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-ROC's respectively. These curves will be step functions with a vertical jump at X = 1 − specificity. These bounds are for the empirical ROC curve; they are not confidence bounds for the true ROC curve. Corresponding bounds for the apparent AUC-ROC assume conditional independence and replace sensitivity and specificity with their "apparent" counterparts.

Defining AKI according to the need for dialysis

The need for renal replacement therapy following AKI almost always 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. As seen in Table 2, if we assume that the gold standard in this case, need for acute dialysis, has specificity of 100% and sensitivity of 25%, then 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. Even rare false positives (specificity of 99% for 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.

Previous work on the imperfect gold standard

The effect of imperfect reference standards has in general been neglected in the expanding clinical literature on diagnostic test accuracy. In the biostatistical literature, several approaches based on the assumption of conditional independence have been proposed. If the gold standard has a known false positive and false negative rate, and the true disease prevalence is known, then the apparent sensitivity and specificity of a new diagnostic test can be calculated [2–4]. Unfortunately, the required parameters are not usually known with certainty. Hui and Walter have 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 prevalences of disease [5]. Walter and Irwig [6] have reviewed latent class models for use when no gold standard exists at all; 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[7] and Phelps *et al*.[4] Alonzo and Pepe [8] have proposed a composite reference standard test to overcome deficiencies with other methods.

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, troponin's association with mortality [9–11], in conjunction with its known tissue specificity[12], contributed to its adoption for the diagnosis of myocardial infarction[13]. 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 CKD, 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. This highlights that a surrogate biomarker should be in the causal pathway of the disease process.

Another possible study design involves the use of exposure status to test a biomarker's accuracy. Consider, for example, a study in which biomarkers are tested following 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 identification of biomarkers that are too sensitive to be of use clinically. These types of studies may be useful to identify biomarkers that fulfill the vision elaborated by the United States FDA regarding qualification and use of biomarkers in drug development, dose regulation, and clinical monitoring of nephrotoxic drug exposure [14].

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

Conclusion

Biomarker development in nephrology is crucial if we hope to develop therapeutic 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 outperform SCr itself. When using a non-ideal 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 article is excerpted from the original publication by Waikar et al. [1]. Republished with permission of Journal of the American Society of Nephrology, from Waikar SS, Betensky RA, Emerson SC, Bonventre JV. Imperfect gold standards for kidney injury biomarker evaluation. J Am Soc Nephrol. 2012 Jan; 23(1):13–21.; permission conveyed through Copyright Clearance Center, Inc."

Supported by R33DK074099, K23DK075941, R01CA075971 and T32NS048005. This paper discusses biomarkers. One such biomarker is KIM-1. JVB is a co-inventor on KIM-1 patents that have been licensed by the Partners Office for Research Ventures & Licensing (RVL) to a number of companies, including Johnson and Johnson, BiogenIdec, R and D systems, Bioassayworks, Rules Based Medicine and Genzyme.

References

1. Waikar SS, Betensky RA, Emerson SC, Bonventre JV. Imperfect gold standards for kidney injury biomarker evaluation. J Am Soc Nephrol. 2012 Jan; 23(1):13–21. [PubMed: 22021710]
2. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies Indices of agreement and their relation to prevalence. Am J Epidemiol. 1966; 83:586–592. [PubMed: 5932702]
3. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis. 1981; 34:599–610. [PubMed: 6458624]
4. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a "fuzzy gold standard". Med Decis Making. 1995; 15:44–57. [PubMed: 7898298]
5. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980; 36:167–171. [PubMed: 7370371]
6. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988; 41:923–937. [PubMed: 3054000]
7. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statistics in Medicine. 18:2987–3003. [PubMed: 10544302]
8. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 1985; 41:959–968. [PubMed: 3830260]
9. Antman EM, Tanasijevic MJ, Thompson B, Schactman M, McCabe CH, Cannon CP, Fischer GA, Fung AY, Thompson C, Wybenga D, Braunwald E. Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes. N Engl J Med. 1996; 335:1342–1349. [PubMed: 8857017]
10. Hamm CW, Ravkilde J, Gerhardt W, Jorgensen P, Peheim E, Ljungdahl L, Goldmann B, Katus HA. The prognostic value of serum troponin T in unstable angina. N Engl J Med. 1992; 327:146–150. [PubMed: 1290492]
11. Wade R, Eddy R, Shows TB, Kedes L. cDNA sequence, tissue-specific expression, and chromosomal mapping of the human slow-twitch skeletal muscle isoform of troponin I. Genomics. 1990; 7:346–357. [PubMed: 2365354]
12. Thygesen K, Alpert JS, White HD, Jaffe AS, Apple FS, Galvani M, Katus HA, Newby LK, Ravkilde J, Chaitman B, Clemmensen PM, Dellborg M, Hod H, Porela P, Underwood R, Bax JJ, Beller GA, Bonow R, Van der Wall EE, Bassand JP, Wijns W, Ferguson TB, Steg PG, Uretsky BF, Williams DO, Armstrong PW, Antman EM, Fox KA, Hamm CW, Ohman EM, Simoons ML, Poole-Wilson PA, Gurfinkel EP, Lopez-Sendon JL, Pais P, Mendis S, Zhu JR, Wallentin LC, Fernandez-Aviles F, Fox KM, Parkhomenko AN, Priori SG, Tendera M, Voipio-Pulkki LM, Vahanian A, Camm AJ, De Caterina R, Dean V, Dickstein K, Filippatos G, Funck-Brentano C, Hellemans I, Kristensen SD, McGregor K, Sechtem U, Silber S, Widimsky P, Zamorano JL, Morais J, Brener S, Harrington R, Morrow D, Lim M, Martinez-Rios MA, Steinhubl S, Levine GN, Gibler WB, Goff D, Tubaro M, Dudek D, Al-Attar N. Universal definition of myocardial infarction. Circulation. 2007; 116:2634–2653. [PubMed: 17951284]
13. Boldt J, Brenner T, Lang J, Kumle B, Isgro F. Kidney-specific proteins in elderly patients undergoing cardiac surgery with cardiopulmonary bypass. Anesthesia and analgesia. 2003; 97:1582–1589. [PubMed: 14633524]

14. Amur S, Frueh FW, Lesko LJ, Huang SM. Integration and use of biomarkers in drug development, regulation and clinical practice: a US regulatory perspective. Biomark Med. 2008; 2:305–311. [PubMed: 20477416]

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 acute kidney injury.

	True AKI (i.e., tubular injury)		
	AKI	**No AKI**	**Total**
AKI according to SCr	180	80	260
No AKI according to SCr	20	720	740
Total	200	800	1000
	sensitivity = 90%	*specificity* = 90%	

	AKI according to SCr		
	AKI	**No AKI**	**Total**
New biomarker positive	180	20	200
New biomarker negative	80	720	800
Total	260	740	1000
	apparent sensitivity = 69%	*apparent specificity* = 97%	

Abbreviations: SCr, serum creatinine; AKI, acute kidney injury

Table 2

The apparent diagnostic performance characteristics of a perfect biomarker as a function of disease prevalence and sensitivity and specificity of an imperfect gold standard.

Imperfect gold standard diagnostic performance	True disease prevalence	Apparent sensitivity	Apparent specificity	Lower bound of Apparent AUC-ROC	Upper bound of Apparent AUC-ROC
	5%	30%	99%	0.29	0.99
	10%	47%	98%	0.46	0.99
80% sensitive, 90% specific	20%	67%	95%	0.63	0.98
	40%	84%	87%	0.73	0.98
	60%	92%	75%	0.69	0.98
	5%	17%	99%	0.17	0.99
	10%	31%	97%	0.30	0.98
80% sensitive, 80% specific	20%	50%	94%	0.47	0.97
	40%;	73%	86%	0.62	0.96
	60%	86%	73%	0.64	0.96
	5%	100%	96%	0.96	1.00
	10%	100%	92%	0.92	1.00
25% sensitive, 100% specific	20%	100%	84%	0.84	1.00
	40%	100%	67%	0.67	1.00
	60%	100%	47%	0.47	1.00
	5%	57%	96%	0.55	0.98
	10%	74%	92%	0.68	0.98
25% sensitive, 99% specific	20%	86%	84%	0.72	0.98
	40%	94%	66%	0.63	0.98
	60%	97%	47%	0.46	0.99
	5%	7%	100%	0.07	1.00
	10%	13%	100%	0.13	1.00
100% sensitive, 25% specific	20%	25%	100%	0.25	1.00
	40%	47%	100%	0.47	1.00
	60%	67%	100%	0.67	1.00

Abbreviations: AUC, area under the receiver operating characteristic curve