Published in final edited form as:

Stat Med. 2015 January 15; 34(1): 179–180. doi:10.1002/sim.6328.

Re: A latent class method for diagnostic tests: the new, reference, gold standard problem

John Collins, Ph.D. and

Department of Rehabilitation Science, College Health and Human Services, George Mason University, 4400 University Drive, 2G7, Fairfax, VA 22030

Minh Huynh, Ph.D.

Epidemiology and Biostatistics Section, RMD, Mark O. Hatfield Clinical Research Center, National Institutes of Health, 6100 Executive Boulevard, 3C01-B, Rockville, MD 20852, minh.huynh@nih.gov

We appreciate the letter from Dr. Baker and the opportunity to comment and contextualize our review within a broader framework of medical testing models. Accuracy without gold standard models represent one end of a continuum of statistical models of medical testing where this continuum is determined by the quality and completeness of the information set available to the researcher. We define the information set of a dataset intended for use in a model of diagnostic test accuracy to be the collection of all factors affecting the suite of available tests and their administration, and all factors affecting the sensitivity and specificity of these tests individually and collectively. In our review we considered a number of the salient features of certain types of imperfect and incomplete information sets with an emphasis upon the modeling issues that arise. Our use of the words imperfect and incomplete is explained next, and from this we hope to clarify a distinction between the situation presented in Dr. Baker's article and those covered in our review. The output of a medical testing model is an estimation of test outcomes conditional on, in our terminology, the data's information set. Conceptually, the model is predicting test outcomes $Y(t)$ conditional on the observed information set $X(t)$ as functions of time t : $E[Y(t) | X(t)]$, where $E[\cdot]$ is the expectation. There is an ideal, full information set $Z(t)$, and if observation is done without error then $X(t) = Z(t)$. Unobserved information is labeled $U(t)$ so that $Z(t) = (X(t); U(t))$. Error may, however, be present as in the case when an imperfect reference test is used as if it were a gold standard. In this situation the observed information set is $(X(t); e(t))$, where $X(t)$ is without error and $e(t)$ is an error term. This suggests four categories of models based on the properties of the information set, each with their own challenges:

1. Perfect and complete, $E[Y(t) | X(t); U(t)]$;
2. Imperfect and complete, $E[Y(t) | X(t); U(t); e(t)]$;
3. Perfect and incomplete, $E[Y(t) | X(t)]$;
4. Imperfect and incomplete $E[Y(t) | X(t); e(t)]$.

The information set captures more than just these four categories, but let us restrict our attention to them for the moment. These categories are determined by the existence of a gold standard or imperfect reference test and the extent of their use in the dataset. In this limited

sense “perfect” reflects the availability of a gold standard test, whether given to all subjects or not, and “complete” for whether the data was fully verified by either the gold standard or an established imperfect reference test. Imperfect data are test results that have not been verified by a gold standard. An established reference test can at best provide an imperfect verification of the data. Specifically, the first category corresponds to data for which all test results are verified by a perfect gold standard while the second category handles data for which all subjects are verified by an imperfect reference test in absence of any gold standard. The situation considered in Dr. Baker’s article best fits within the third category, where an established imperfect reference test whose properties have separately been verified by a gold standard is used to verify the new test. In contrast, our review focuses on the last category of datasets for which no gold standard is available to verify the properties of the imperfect tests used. This includes as a special case datasets in which verification is available only for a subset of subjects that are chosen through a nonrandom process. A common example of this is a screening design where only screening test positives are verified by the gold standard. The result of this decision is that the verification method provides no information about the screening test’s specificity. The information set applies to more than the availability of gold standard and reference tests and so we now provide examples of what other aspects are included in this set. The suite of available tests refers to all tests available to the researchers, whether these are used in the dataset or not. The question of the (un)availability of a gold standard or established imperfect reference test is, however, the most important element to the suite of tests. In choosing the particular tests to use from the suite of available ones the researcher must take into account behaviors or other characteristics of the population(s) under study that may affect test accuracy, the cost and invasiveness of the tests, and the temporal nature of the testing. Temporal aspects include whether testing is performed ad-hoc, ex-ante, or post-hoc as well as expected time lapse between disease onset and testing date. Test administration refers to the pattern of testing used and any decision rules used to determine this pattern, such when using a sequential or parallel screening design. Finally, it should be noted that test sensitivity and specificity are affected by behaviors of the raters or laboratories in which testing is performed. Many of the components of the information set can be handled separately when modeling the accuracy of diagnostic tests in the absence of a gold standard. We have attempted in our review to discuss these choices and to emphasize the tradeoffs between model validity and identifiability that result.