ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 566−574

Journal of Clinical Epidemiology

Statistical methods for the meta-analysis of diagnostic tests must take into account the use of surrogate standards

Jian Kang[a,*], Rollin Brant[b], William A. Ghali[c]

[a]*Sports Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, 2500 University Drive, NW, Calgary, Alberta, Canada T2N1N4*
[b]*Department of Statistics, University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, British Columbia, Canada V6T1Z4*
[c]*Department of Community Health Sciences and Family Medicine, Faculty of Medicine, University of Calgary, TRW building, 3280 Hospital Drive, NW, Calgary, Alberta, Canada T2N4Z6*

Accepted 23 December 2012; Published online 1 March 2013

Abstract

Background: Evaluating the performance of a new diagnostic test presents a challenge if the conventional ''gold'' standard is invasive, hazardous, or expensive, especially if that test has been supplanted in usual clinical practice by a ''silver'' standard test that is more acceptable and perhaps only slightly suboptimal. In such a case, a systematic literature review will typically uncover a mix of study types, some using the gold and some the silver.

Objective: We sought to develop and compare statistical methods to account for this kind of heterogeneity in performing a meta-analysis.

Study Design and Setting: We compared the performance of estimation methods based on generalized mixed models which incorporate heterogeneity, especially choice of reference test, and random between-study variation in sensitivity and specificity with more conventional methods which neglect the differences in reference tests. Computer simulations were conducted to assess bias and root mean square error of point estimates and coverage of interval estimates.

Results: Methods ignoring the difference in reference tests severely underestimated sensitivity and specificity under the assumption of conditional independence. Bias was substantial even for references with small departure from the standard and persisted with increasing sample size. Coverage of interval estimates was far from nominal level.

Conclusion: In the presence of varying reference tests, avoidance of bias and invalid confidence intervals for diagnostic performance requires applying a model that accounts for differences in reference test and heterogeneity among studies.

Keywords: Surrogate measurement; Diagnostic accuracy; Misclassification; Meta-analysis; Generalized linear mixed model; Gibbs sampling

1. Introduction

In an ideal study evaluating a new diagnostic test, the conventional standard is applied to each patient to confirm disease status. This may, however, be impractical because of invasiveness, high cost, or technical challenges. Our motivating example is the diagnosis of deep vein thrombosis (DVT), in which venography, an invasive procedure, is the accepted standard test. Ultrasonography, which is not risky to patients and is well-known diagnostic characteristics, is a commonly applied surrogate. A recent meta-analysis [1] of the performance of D-dimer testing (an inexpensive screening blood test) included studies using both venography and ultrasonography as reference tests, but none where patients underwent all three tests. This article proposes models and methods appropriate for a diagnostic test meta-analysis combining studies, which evaluate the candidate test using either the conventional standard or clinically accepted surrogate reference.

In the meta-analysis of diagnostic tests, the clinical importance of accounting for dissimilarities among studies has been widely acknowledged [2−4]. The use of different references in meta-analysis of diagnostic test was recognized as the major concern when pooling results from different studies [5,6]. Ignoring this information produced biased conclusions on the accuracy of the diagnostic test [5,6]. Improvements were also discussed by several authors [7−10]. Bivariate random effects model [32,33] was also proposed for the meta-analysis of diagnostic test, which is a generalization of the summary receiver operating characteristic (ROC) curve [26]. However,

Conflict of interest statement: No conflict of interest/financial disclosure is declared.

* Corresponding author. Tel.: (403) 220-2221; fax: (403) 284-2451.
E-mail address: kang@ucalgary.ca (J. Kang).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2012.12.008

What is new?

- Failure to account for the use of different reference standards produces biased estimates and flawed confidence intervals.
- Modern Bayesian methods facilitate the implementation of models, which appropriately account for heterogeneity.

all these approaches aimed at corrections either when only results of the test of interest vs. the surrogate were available or in the situation in which each individual received all tests. De Bock et al. [11] suggested a solution when at least two types of reference tests were applied under the assumption that sensitivity and specificity of the candidate test did not change across studies. This assumption often fails for tests, such as the D-dimer, which admit different cutoffs for positivity.

Although the relationship between two or more diagnostic tests is most fully described by a multinomial distribution, the preponderance of research in this area has focused on binary or ordinal data using logit and probit links for cumulative probabilities [12−15]. Daniels and Gatsonis [16] applied the baseline-category logit in a hierarchical Bayesian model for cluster multinomial data. Hartzel presented a general approach for logit random effects modeling on clustered multinomial responses [12]. Log-linear models provide a convenient framework for analysis of multinomial data [17−20]. The application of these methods in this context is complicated by incompleteness of the data because subjects are never evaluated by all three modalities. Recent work by Walter et al. [36] emphasizes the need for careful examination of assumptions concerning assumed patterns of association between candidate tests and imperfect reference standard.

In this article, we propose methods for analyzing incomplete multinomial data in a model in which heterogeneity between studies is taken into account via an underlying log-linear model with random effects. Following the work of other authors, construct models to accommodate variation in

Disease prevalence: inevitable as recognized in all studies of diagnostic tests [1,3,21,22].

Test yield: a likelihood if the classification process depends on thresholds for positivity [23,24].

Diagnostic odds ratio (DOR): ratio of the odds for positive tests between diseased and nondiseased populations needs to be considered [32].

We furthermore extend these models to accommodate varying choice of reference test, applying modern Bayesian methods of statistic estimation to facilitate technical development. We follow by application to a previous published data set and conclude with simulations to demonstrating the deleterious effects of ignoring differences between reference tests.

2. The mixed-effects log-linear model for three-way classifications

Our underlying framework is best expressed in terms of a hypothetical study in which N subjects receive all three (candidate, surrogate, and standard) tests, and test results are tabulated in a $2 \times 2 \times 2$ contingency table, with elements y_{ijk}, recording the number of individuals with specific patterns of test results. Denoting test results as 0 for negative and 1 for positive and consider our motivating example, letting D, U, and V denote D-dimer (candidate), ultrasonography (surrogate), and venography (standard), respectively, then y_{ijk} (where i, j, and $k = 0$ or 1) denotes the observed frequency of subjects for whom $D = i$, $U = j$, and $V = k$ (e.g., y_{000} is the number of patients testing negative on all three tests). We will denote the expected frequency of such cases as m_{ijk}, which equals $N \cdot p_{ijk}$,, where p_{ijk} is the joint probability of such an outcome at the individual level.

Log-linear models for multinomial models provide a convenient framework by specifying the logarithm of the probability vector $p = (p_{000}, p_{001}, \ldots p_{111})$ as $\log(p) = X\beta$, where X is a matrix of known values and β is a parameter vector. Because of the underlying factorial structure, analysis of variance (ANOVA) type notation is commonly used to characterize relationships, whereby the most general model takes the form

$$\log p_{ijk} = \beta_0 + \beta_D \cdot i + \beta_U \cdot j + \beta_V \cdot k + \beta_{DU} \cdot i \cdot j + \beta_{DV} \cdot i \cdot k + \beta_{UV} \cdot j \cdot k + \beta_{DUV} \cdot i \cdot j \cdot k. \quad (1)$$

Heuristically, the parameters describe patterns of association between the subscript variables; for example, $\beta_{DUV} \neq 0$ indicates that the pattern of association cannot be described in terms of simple two-way associations. More specific interpretations apply depending on assumptions. For example, if $\beta_{DV} = \beta_{DUV} = 0$, then D and U are independent conditional on V and

β_D log odds of the D-dimer false-positive rate (FPR),

β_U log odds of the ultrasound FPR,

β_V log odds of positive venography when both other tests are negative,

β_{DV} log diagnostic odds ratio (DOR) for D-dimer, and

β_{UV} log DOR for ultrasound.

The log-linear model is applicable in the present context as joint probabilities for results from studies involving only two tests can be obtained by summing over the possible values of the unobserved test, for example, the probability of a subject testing positive on both D-dimer and ultrasound is the marginal probability $p_{110} + p_{111}$. Similarly, joint marginal distributions of D vs. V and U vs. V can be calculated, and a joint likelihood for the entire set of study results can be formed as product of contributions from each study.

Note that inference is not possible for all parameters of the general model because estimation of the parameter β_{DUV} requires observing a complete $2 \times 2 \times 2$ table.