Journal of Clinical Epidemiology 57 (2004) 698–711

Journal of Clinical Epidemiology

Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies

Mir Said Siadaty[a,*], John T. Philbrick[b], Steven W. Heim[c], Joel M. Schectman[b]

[a]*Division of Biostatistics and Epidemiology, University of Virginia School of Medicine, Box 800717, Charlottesville, VA 22908, USA*
[b]*Department of Internal Medicine, University of Virginia School of Medicine, Box 800744, Charlottesville, VA, 22908, USA*
[c]*Department of Family Medicine, University of Virginia School of Medicine, Box 800729, Charlottesville, VA, 22908, USA*

Accepted 16 December 2003

Abstract

Objective: Current methods for meta-analysis of diagnostic tests do not allow utilizing all the information from papers in which several tests have been studied on the same patient sample. We demonstrate how to combine several studies of diagnostic tests, where each study reports on more than one test and some tests (but not necessarily all of them) are shared with other papers selected for the meta-analysis. We adopt statistical methodology for repeated measurements for the purpose of meta-analysis of diagnostic tests.

Study Design and Setting: The method allows for missing values of some tests for some papers, takes into account different sample sizes of papers, adjusts for background and confounding factors including test-specific covariates and paper-specific covariates, and accounts for correlations of the repeated measurements within each paper. It does not need individual-level data, although it can be modified to use them, and uses the two-by-two table of test results vs. gold standard.

Results: The results are translated from diagnostic odds ratios (DOR) to more clinically useful measures such as predictive values, post-test probabilities, and likelihood ratios. Models to capture between-study variation are introduced. The fit and influence of specific studies on the regression can be evaluated. Furthermore, model-based tests for homogeneity of DORs across papers are presented.

Conclusion: The use of this new method is illustrated using a recent meta-analysis of the D-dimer test for the diagnosis of deep venous thrombosis.

Keywords: Meta-analysis; Diagnostic test; Repeated measurement; Marginal model; DOR

1. Introduction

To combine papers on performance of screening/diagnostic tests in a meta-analysis one can choose from several probabilistic measures. In his recent review, Deeks [1] summarized three: pooling sensitivities and specificities, pooling likelihood ratios, and constructing diagnostic odds ratios (DOR), with summary receiver operating characteristic (SROC) curves. Hasselblad and Hedges [2] suggested that the most practically useful general method of combining evidence to estimate an ROC curve is that of Moses et al. [3] who assume that the studies of the tests that one wants to compare are statistically independent. In other words, if each study used the same or overlapping subjects to give performance measures for each test, the standard errors given in Moses et al. would be incorrect. Walter [4], in a recent article, derives properties of the SROC, and gives standard errors for area under curve (AUC) and Q^*. (Q is a point of indifference on the SROC curve between the false positive and false negative diagnosis errors.) However, he too assumes that the summary measures for each test are independent of other tests. Practically, this means no one paper can provide performance measures for two tests on the same or overlapping subject populations. In reality, there are many papers where more than one competing diagnostic test is studied simultaneously. It therefore is important to seek a method whereby all the information each paper provides can be utilized in a meta-analysis. Walter mentions that methods to take such dependencies into account have been proposed for therapeutic studies, and suggests that extensions or alternative approaches for diagnostic test comparisons would be useful.

We demonstrate how to extend and improve on the method by Moses et al. [3] to combine several studies of diagnostic tests, where each study reports on one or more tests and some tests (but not necessarily all of them) are shared with other papers selected for the meta-analysis. The method is suitable for meta-analysis of several competing tests across studies that report the two-by-two table of test result vs. gold standard.

[*] Corresponding author. Tel.: 434-982-4436; fax: 434-924-8437.
E-mail address: mirSiadaty@virginia.edu (M.S. Siadaty).

0895-4356/04/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2003.12.007

The results are translated from DOR to more clinically useful measures such as predictive values, post-test probabilities, and likelihood ratios. Models to capture between-study variation are introduced. The fit and influence of specific studies on the regression can be evaluated. Furthermore, model-based tests for homogeneity of DORs across papers are presented.

2. Motivation

A cursory search of literature for studies of diagnostic tests in almost any field shows an abundance of papers where two or more tests have been evaluated, and fewer papers where only one diagnostic test is studied. For example, in a recent meta-analysis of diagnostic tests for deep vein thrombosis (DVT) Heim et al. [5] selected 23 papers. Each paper studied 1 to 13 different tests. Table 1 shows the distribution of studied tests per paper.

In these 23 papers, 21 different tests were studied. The papers overlap partially in the types of tests they studied. In other words, not every test has been studied in every paper. In addition, the number of studied tests per paper varies. There were papers that studied only one test. Statistical methods for matched pairs/groups may fail to utilize groups where one (or more) of the members of the matched group is missing. Furthermore, methods that treat multiple tests within a paper as independent observations are ignoring the fact that the tests have been performed on the same patient sample, and hence, are dependent.

Because studies usually have different sample sizes, one may want to adjust for this in the analysis. Also, there may be test-specific covariates (within each paper) and paper-specific covariates for which one wants to adjust.

Papers usually report only study-level summary measures such as sensitivity and specificity. None of the papers in the above example report individual data. Additionally, as Littenberg and Moses [6] point out, full ROC curves are rarely published. Hence, analysis methods should be able to work with study-level data and possibly be extensible to the patient level.

The method of choice should be able to test hypotheses of interest, give *P*-values, and estimate magnitude of effect/difference. Evaluating effect of covariates and adjusting for potential confounders are sometimes needed. A clear interpretation of the results with straightforward clinical meaning is desirable.

3. Statistical modeling

In the setting shown in Table 1, each paper includes performance measures for one or more diagnostic tests. The results for several tests reported from a single paper define a cluster of repeated measurements that are potentially correlated, because all the tests have been measured in the same (or overlapping) sample of patients. We take the view of a repeated-measurement situation. We adopt statistical methodology for repeated measurements for the purpose of meta-analysis of diagnostic tests. There are several choices for repeated measurement analysis. They include random effects models (including random coefficient models [7]), transition models, marginal models, and marginalized versions of nonmarginal models. One may use any of the mentioned approaches to solve the question at hand. We are interested in population averages, and the interdependence of the studies is not of primary interest, and is mainly treated as a nuisance. A marginal model is a good candidate [8].

This method can utilize papers that have not studied all of the diagnostic tests of interest (missing values of some tests in some papers). This is usually not the case for methods of analysis of matched groups. The method adjusts for background and confounding factors, accounts for correlations of the repeated measurements within each paper, and gives the results in a format relatively easy to interpret and understand. Fitting a marginal model is relatively easy, and there are several statistical software packages that have implemented the method, including SAS, S-Plus, R, and Stata.

In the model

$$logit(E(y_{pt})) = \boldsymbol{x}_{pt}\, \boldsymbol{\beta}$$

y_{pt} is the result of diagnostic test t in paper p in a binary format (diseased, healthy). p is an index number for the papers selected in the meta-analysis. Within each p index, there is one or more tests that are indicated by index t. There are m papers that we want to include in the meta analysis, hence $p = 1,2, \ldots, m$. Besides, for each paper p there are n_p tests measured. In other words, the number of studied tests is allowed to vary for each paper. $\boldsymbol{x}_{pt}$ is a vector of predictors within paper p for test t, and $\boldsymbol{\beta}$ is the vector of regression coefficients. Because our outcome is a dichotomous one (diseased, healthy), y_{pt} is distributed binomially, and we use a logit link function. The correlation structure between repeated measurements is modeled separately in the marginal model. Using an independence structure (equivalent to Huber-White sandwich estimator [9]) is a common practice, and several other choices are available.

For each test one extracts a two-by-two table of test result vs. gold standard from the corresponding paper. This means that the model is fitted to grouped binary data. The method expands each two-by-two table to the original sample size; hence, different sample sizes of different papers are accounted for. The primary study units are persons, not papers or tests. Therefore, the effective sample size will be larger than the number of papers selected for the meta-analysis. This will allow more covariates to be included in the model without overfitting. Also, it makes the transition from aggregated data currently presented in published papers to patient-level data quite simple.

Consider the following model, where Disease is an indicator variable for results of gold standard and Result is an

Table 1
Tests studied in each paper

		Diagnostic test																					
		Asserachrom	Auto Dimertest	BC D-Dimer	D-Dimer test	Dimertest	Dimertest EIA	Dimertest GOLD EIA	Dimertest II	Enzygnost	Fibrinostika	IL Test	Instant I.A.	Liatest	LPIA	Minutex	Nephelotex	NycoCard	SimpliRED	Tinaquant	Turbiquant	VIDAS	
	Paper	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total
1	2Brenner, B. 1995 (86)						X		X										X				3
2	4D'Angelo, A. 1996 (103)																					X	1
3	5Elias, A. 1996 (171)	X			X					X	X		X					X				X	7
4	6Escoffre-Barbe, M. 1998 (464)													X									1
5	7Farrell, S. 2000 (48)																		X				1
6	8Fiessinger, J. 1997 (30)																		X				1
7	12Janssen, M. 1997 (132)					X					X					X			X	X		X	6
8	19Legnani, C. 1997 (81)							X					X		X	X	X	X				X	7
9	20Legnani, C. 1999 (99)			X				X														X	3
10	21Lennox, A. 1999 (200)																		X				1
11	22Leroyer, C. 1997 (448)	X											X										2
12	26Scarano, L. 1997 (126)									X			X					X					3
13	29van der Graaf, F. 2000 (99)	X		X						X	X	X	X	X		X		X	X	X	X	X	13
14	30Wells, P. 1999 (150)																		X				1
15	31Wells, P. 1995 (214)																		X				1
16	33Funfsinn, N. 2001 (106)	X	X																	X		X	4
17	37Harper, P. 2001 (235)											X							X				2
18	57Carter, C. 1999 (199)																		X				1
19	61Permpikul, C. 2000 (65)																		X				1
20	63Sadouk, M. 2000 (177)																			X	X		2
21	68Wijns, W. 1998 (74)	X											X									X	3
22	91Kharia, HS. 1998 (79)																	X					1
23	92Perrier, A. 1999 (474)																					X	1
	TOTAL	5	1	2	1	1	1	2	1	3	3	2	6	2	1	3	1	5	11	4	2	9	66

Numbers in parentheses in front of the paper names are sample sizes for the corresponding paper.

indicator for results of diagnostic test (note that we extract a two-by-two table for each test from each paper, so the results of gold standard and diagnostic test are already dichotomized based on the threshold value the authors of that paper have chosen). (In model 1, the Result of the diagnostic Test is modeled while the actual status of Disease is considered as a predictor. If we reversed their role, the interpretation of the regression coefficient for this model would be the same.)

$$logit(Result_{pt}) = \beta_0 + \beta_1 {*} Disease_{pt} \qquad (1)$$

In (1), β_1 is the ratio of the odds of event (positive Test) when disease is present (Disease = 1) vs. when disease is absent (Disease = 0) in the log scale; hence, the log odds ratio (LOR). Equivalently, it is ratio of odds of no event (negative Test) when Disease = 0 vs. when Disease = 1. In other words, it is ratio of TN/FP over FN/TP, or TN*TP / FP*FN. This is the DOR built from Table 2.

The better the test, the more it matches the gold standard; hence, the bigger the counts in cells TP and TN and the smaller the counts in cells FP and FN. The relationship of interest translates to maximizing the DOR. (There is an additive ratio that could be maximized, [TP + TN]/[FP + FN]. However, the OR is more common to use and more statistical tools are available for it.) In a log scale, LDOR of zero means the test is as good as pure chance (flipping a coin) and the magnitude of the positive LDOR reflects the test's diagnostic value beyond chance. (A negative LOR means the test is performing worse than by simply flipping a coin to guess the diagnosis of a patient.)

Model (1) is useful for explaining the meaning and interpretation of the regression coefficient. However, it does not

Table 2
Classifying patients based on test results and actual disease status (by gold standard)

	Disease (by gold standard Gj)	
	D+	D−
Result (by test Ti)		
T+	TP	FP
T−	FN	TN

TP = true positives.
FP = false positives.
FN = false negatives.
TN = true negatives.

differentiate between performances of different types of diagnostic tests. Now we use the following marginal logistic model.

$$logit(Result_{pt}) = \beta_0 + \beta_1 * Disease_{pt} + \boldsymbol{\beta_2} * Test_{pt} + \boldsymbol{\beta_3} * Disease_{pt} * Test_{pt} \quad (2)$$

In (2), $\boldsymbol{\beta_3}$ is the (log) ratio of two DORs from two different tests. In other words, it is a ratio of performance of one test vs. the other. If this ratio is statistically different from zero, the two tests are performing differently in diagnosing the disease, with the coefficient sign indicating which test performed better (a positive coefficient means the test in the numerator is better than the denominator, which is the reference category). Note that the variable Test is a categoric variable, showing types of tests studied in each paper, which should be represented in the model by a suitable number ($t - 1$) of indicator variables. Hence, coefficients $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ are vectors of coefficients.

One can extend model (2) to include other covariates and factors. For example, the percentage of diseased cases in the sampled population (observed prevalence) may exert an effect on the accuracy of a test. The following model has an extra variable Prvlnc. Each additional covariate is entered as both a main effect and an interaction term with Disease, the latter coefficient (β_5) reflecting the impact of that covariate on the LDOR.

$$logit(Result_{pt}) = \beta_0 + \beta_1 * Disease_{pt} + \boldsymbol{\beta_2} * Test_{pt} + \beta_3 * Prvlnc_{pt} + \boldsymbol{\beta_4} * Disease_{pt} * Test_{pt} + \beta_5 * Disease_{pt} * Prvlnc_{pt} \quad (3)$$

Note that the added covariate can be paper-specific, meaning it can have the same value for all the diagnostic tests studied within a single paper. Additionally, it can be test-specific, meaning it can have different values for each test within the same paper. Moreover, the added covariate can have the same value for a specific test across all the papers. Therefore, the model can accommodate for both test-specific covariates and paper-specific covariates. For the example of prevalence, all the tests within the same paper may have been performed on exactly the same sample of patients, and hence, have the same observed prevalence for all the tests within that paper (paper-specific). Or some of the tests within that paper may have been performed on a subsample where the observed prevalence is dramatically different from the total sample (test-specific).

To verify enough sample size for a model with a certain complexity, one calculates the number of parameters to be estimated for the model. Then, using a rule of thumb of 10 samples per parameter, multiply the number of parameters by 10. This gives a rough estimate of the minimum sample size needed for that specific model. The effective sample size (calculated based on the nominal sample size, and the correlation between the outcome measurements) is usually bigger than the sum of sample sizes of the papers included in the meta-analysis.

It is good practice to measure homogeneity of performance of a test across the papers, then attempt to pool the results from several papers into a summary measure for that test, because (some) summary performance measures may behave unexpectedly when different papers report very heterogeneous results for the same test. To test homogeneity of DOR for a specific test across the papers one can use the following model.

$$logit(Result) = \beta_0 + \beta_1 * Disease + \boldsymbol{\beta_2} * Paper + \boldsymbol{\beta_3} * Disease * Paper \quad (4)$$

Note that there is no p and t index; so it is an ordinary logistic model (a logistic model under the family of generalized linear models) not a marginal one. One selects rows of data related to a specific test, then runs the model (4) on this subsetted data. Variable Paper represents an ID number for the papers that studied that specific diagnostic test, and is entered into the model as a categoric variable. A significant $\boldsymbol{\beta_3}$ means a significant heterogeneity of DORs for that test across the papers. $\boldsymbol{\beta_3}$ is a vector and has $k - 1$ components, where k is the number of papers that included that specific test in their study. Therefore, an overall test (type III) of the interaction term is equivalent to a Breslow-Day test [10]. If one is interested in locating the papers that account for the heterogeneity, the components of the $\boldsymbol{\beta_3}$ vector may be helpful. One should know that the *P*-values of the components of the $\boldsymbol{\beta_3}$ are sensitive to the choice of the reference category. One may prefer to use a "deviation" contrast for the *Disease*Paper* term, where the effect for each category of *Paper* is compared to the overall effect.

If one wants to relax the assumption of independence of DOR and test threshold (equivalent to homogeneity of DORs), starting from model (1), one enters a covariate (call it *Z*) into the model, where *Z* is a function of axes of the ROC graph, true positive rate (TPR) and false positive rate (FPR).

$$logit(Result) = \beta_0 + \beta_1 * Disease + \beta_2 * Z + \beta_3 * Disease * Z \quad (5)$$

One can show that LDOR in model (5) is equal to $\beta_1 + \beta_3 * Z$, hence, dependent on values of covariate *Z*. (Document showing the relationships can be found at http://www.people.virginia.edu/~mss4x/meta.html.) Thus, *Z* may (partially) account for systematic between-study variation, if present. If one defines *Z* to be $\log\{TPR*FPR/[(1 - TPR)*(1 - FPR)]\}$, model (5) is equivalent to the method proposed by the paper of Moses et al. [3]. However, this definition of *Z* constrains the SROC curve. The ROC curve has to cross the equivalent homogeneous ROC curve on the antidiagonal line (where $TPR = 1 - FPR$), it cannot have more than one inflection point, and estimates of β_3 bigger than 1 or smaller than -1 produce some "unintuitive" ROC curves. One may prefer to use a smooth function of FPR and TPR as *Z*, for instance, a restricted cubic spline of FPR. If such *Z* covariate turns out to be insignificant, one may return to the simpler model.

Under the assumption of no relationship between DOR and the test threshold, one can show that each DOR is on a one-to-one correspondence with a ROC curve. Because the estimated DOR is a summary measure, the corresponding ROC curve is a SROC curve. (To make a one-to-one correspondence between DOR and SROC, one needs the assumption that the ratio of two cells in the diseased column before and after a threshold change changes the same amount as for the nondiseased column.) Hence, one can compute and draw the SROC for each test (using the estimated coefficient of the test and populating the model equation with average or modal values for the covariates), compare with other tests, and calculate the AUCs. Also, SROC curves can be constructed for different values of a covariate within the same test. This gives a graphic representation of the effect of a variable on the performance of a test.

Although there is no unique best indicator for performance of a test, from a clinical point of view, predictive values and post-test probabilities may have more straightforward diagnostic meaning. One can get the estimates of DORs from the model and use them to calculate such measures. (The relationship between different performance measures, and the codes to convert them to each other, are illustrated by documents that are in URL http://www.people.virginia.edu/~mss4x/meta.html.)

The influence of specific studies on the regression can be evaluated by examining the residuals. These results plus the ones from test of homogeneity may constitute additional exclusion criteria for papers included in a meta-analysis.

If there is a multilayer cluster structure (nested clusters), a random-effects model for repeated measures may be easier to implement. (Although marginal model theory allows for this, currently available software does not readily support such functionality [unless all submatrices are assumed to share the same correlation parameter]. An alternative approach for implementing nested clusters is utilizing software for "survey analysis." For example, package "survey" in *R* supports nested clusters.) For example, a paper may report several two-by-two tables for the same test by using different threshold values (to dichotomize test result to positive and negative). These tables for the same test in a paper constitute a cluster of repeated measurements themselves. This cluster is nested within that paper, along with other tests studied by that paper. Also, if some of the published papers are not completely independent of each other, one may prefer to include them as members of a cluster of potentially correlated papers. This relaxes the assumption of independence of the studies. Please note that if there is a natural ordering among the members of each cluster, a marginal model is preferred over the random-effects model.

When patient-level data are available, besides using a threshold to dichotomize the test result, one can directly enter the test value into the model. Then, instead of logit, one chooses a link function that matches the level of measurement of the test result (i.e., an identity link for a continuous outcome).

4. Implementation through a case study

Continuing the example of DVT, one needs to restructure the data extracted from the papers as shown below. (The 21 D-dimer diagnostic tests may be viewed as different manufacturers' version of the same test. However, the methodology presented here is equally applicable to tests that are remotely similar.) This is a grouped binary data structure. We assume individual-level data are not available, and that two-by-two tables of test result vs. gold standard are reported for each test per paper.

In Table 3, the first column, Paper, is an index of papers (the *p* index), indicating the paper on which each row of information is based. The second column, Test, is an index of tests (the *t* index), indicating the diagnostic test used for that row of data. The third column, Disease, shows the actual status of disease (present or absent, 1 or 0) based on the gold standard. Column "N" is the "column total" from Table 2. When Disease = 0, *n* is the sum of FP + TN. Conversely, when Disease = 1, *n* equals TP + FN. The column "Result" contains the number of subjects with negative diagnostic test results. Hence, when Disease equals 0, Result refers to TN and when Disease equals 1, Result contains FN [11]. (This is due to the SAS proc Genmod default of using the last category as the reference category [and option "descending" has effect in multinomial case only]. Otherwise, one can equally model the positive cells for column Result. For S-PLUS or R, one may need to replace columns "n" and Result with other numbers, such as TP and FP [refer to the appropriate software manual, under "grouped binary data"].)

One can add extra columns for other potential confounders or covariates to be included in the analysis. In this example, column Prvlnc is observed prevalence of DVT. Setting describes the populations from which subjects were recruited (representing patient mix). Gold is the type of method used as gold standard because it varies among papers.

We therefore use the following marginal logistic model.

$$\begin{aligned} logit(Result_{pt}) = \ & \beta_0 + \beta_1{*}Disease_{pt} + \boldsymbol{\beta_2}{*}\ Test_{pt} \\ & + \beta_3{*}\ Prvlnc_{pt} + \beta_4{*}\ Gold_{pt} \\ & + \beta_5{*}\ Setting_{pt} + \boldsymbol{\beta_6}{*}Disease_{pt}{*}\ Test_{pt} \\ & + \beta_7{*}Disease_{pt}{*}\ Prvlnc_{pt} \\ & + \beta_8{*}Disease_{pt}{*}\ Gold_{pt} \\ & + \beta_9{*}Disease_{pt}{*}Setting_{pt} \qquad (6) \end{aligned}$$

Model (6) is simpler than an all two-way interaction model [for explanation of the variables in (6) look at Table

Table 3
Data structure

Paper	Test	Disease	*N*	Result	Prvlnc	Setting	Gold
2Brenner, B. 1995 (86)	Dimertest EIA	0	36	17	58	Outpatient	US-V
2Brenner, B. 1995 (86)	Dimertest EIA	1	50	6	58	Outpatient	US-V
2Brenner, B. 1995 (86)	Dimertest II	0	36	24	58	Outpatient	US-V
2Brenner, B. 1995 (86)	Dimertest II	1	50	10	58	Outpatient	US-V
2Brenner, B. 1995 (86)	SimpliRED	0	36	22	58	Outpatient	US-V
2Brenner, B. 1995 (86)	SimpliRED	1	50	3	58	Outpatient	US-V
4D'Angelo, A. 1996 (103)	VIDAS	0	81	36	21	In-Mix	US-V
4D'Angelo, A. 1996 (103)	VIDAS	1	22	1	21	In-Mix	US-V
5Elias, A. 1996 (171)	Asserachrom	0	96	21	44	In-Mix	US-V
5Elias, A. 1996 (171)	Asserachrom	1	75	2	44	In-Mix	US-V
5Elias, A. 1996 (171)	D-Dimer test	0	96	60	44	In-Mix	US-V
5Elias, A. 1996 (171)	D-Dimer test	1	75	20	44	In-Mix	US-V
5Elias, A. 1996 (171)	Enzygnost	0	96	29	44	In-Mix	US-V
5Elias, A. 1996 (171)	Enzygnost	1	75	5	44	In-Mix	US-V
5Elias, A. 1996 (171)	Fibrinostika	0	96	36	44	In-Mix	US-V
5Elias, A. 1996 (171)	Fibrinostika	1	75	5	44	In-Mix	US-V
5Elias, A. 1996 (171)	Instant I.A.	0	96	18	44	In-Mix	US-V
5Elias, A. 1996 (171)	Instant I.A.	1	75	5	44	In-Mix	US-V
5Elias, A. 1996 (171)	NycoCard	0	96	36	44	In-Mix	US-V
5Elias, A. 1996 (171)	NycoCard	1	75	15	44	In-Mix	US-V
5Elias, A. 1996 (171)	VIDAS	0	96	25	44	In-Mix	US-V
5Elias, A. 1996 (171)	VIDAS	1	75	2	44	In-Mix	US-V
6Escoffre-Barbe, M. 1998 (464)	Liatest	0	188	66	59	In-Mix	US-V
.. (lines of data skipped)	..	..	..	..	..	..	..
68Wijns, W. 1998 (74)	VIDAS	1	32	4	43	In-Mix	V
91Kharia, HS. 1998 (79)	NycoCard	0	50	20	37	Outpatient	US-V
91Kharia, HS. 1998 (79)	NycoCard	1	29	1	37	Outpatient	US-V
92Perrier, A. 1999 (474)	VIDAS	0	363	125	23	Outpatient	US-V
92Perrier, A. 1999 (474)	VIDAS	1	111	2	23	Outpatient	US-V

This is a sample of data. Only a few rows have been shown here.

2 and the description of models (1) to (3)]. It only keeps the two-way interactions where one of the variables is Disease. Following a hierarchical form, all main-effect terms are included. (The grouped binary data structure necessitates the Disease*covariate interactions to evaluate covariate effects on the LDOR.) In this case, only Prvlnc is a continuous variable (between 0 and 1). (One may prefer to use a logit transform of prevalence in the model.) The rest of the variables are categorical, and therefore, should be represented by appropriate indicator variables. The model assumes effect of covariates is the same for all test types. We have presented codes that implement model (6) in SAS (Appendix).

Table 4 shows the results of the modeling, plus some postmodel calculations. We have provided codes that, starting from the model estimates, calculate positive predictive values (PPV), negative predictive value (NPV), likelihood ratios, and post-test probabilities and odds. (For the codes, plus other supplementary material, please see appendix, and the following URL: http://www.people.virginia.edu/~mss4x/meta.html.)

For example, test "Dimertest" (line 5) has a DOR that is 8% that of test VIDAS (the reference category in this analysis, the last line), a significant difference ($P < .0001$). One can conclude that test VIDAS is significantly better than Dimertest, and that it diagnoses patients correctly almost 13 times better (where patients are a mixture of disease-positive and -negative people). One can use the negative predictive values (NPV) to see if any of the tests are good for "ruling out" purposes. Note that calculation of NPV (and PPV) is based on assumptions of 39% prevalence and 90% sensitivity. These are the average observed values across the papers. A model-based approach enables us to calculate the performance measures for different tests, while controlling for variables that potentially confound the performance comparison.

The results can be presented graphically, for ranges of parameters. For instance, Fig. 1 shows the effect of sensitivity (or specificity) on PPV and NPV of test VIDAS for a specific estimate of disease prevalence (the left plot). Also, it shows the effect of different prevalences on the PPV and NPV, when assuming a specific value for sensitivity or specificity (the right plot). Utilizing all the point estimates of VIDAS DOR, prevalence, and sensitivity, the graph calculates and shows the point estimates of PPV and NPV. We have shown, in the Appendix, how to calculate DOR of each test based on the model estimates.

Although all of these 23 studies met methodologic standards designed to limit bias, wide variations in the test characteristics were observed among them. This emphasizes the role of a model where some of the variation can be diminished through model-based adjustments. Overall, the multivariate analysis identified three assays with DORs that

Table 4
Estimates of test performance, adjusted for repeated measures, different sample size, prevalence, gold standard, and patient mix

Diagnostic test	Relative DOR[a]	*P*-value	DOR[a]	AUC[a]	PPV[b]	NPV[b]	Pab[b]	Pnr[b]	LRab[b]	LRnr[b]
1 Asserachrom	0.63	0.2758	32.66	0.918	0.73	0.92	0.73	0.08	4.166	0.1276
2 Auto Dimertest	1.67	0.2194	86.85	0.959	0.86	0.93	0.86	0.07	9.585	0.1104
3 BC D-dimer	0.38	0.0929	19.85	0.886	0.65	0.91	0.65	0.09	2.885	0.1453
4 D-dimer test	0.37	0.0751	18.99	0.883	0.64	0.91	0.64	0.09	2.799	0.1474
5 Dimertest	0.08	<.0001	4.12	0.721	0.46	0.83	0.46	0.17	1.312	0.3184
6 Dimertest EIA	0.44	0.0584	22.82	0.896	0.67	0.92	0.67	0.08	3.182	0.1394
7 Dimertest GOLD EIA	1.31	0.6902	67.86	0.951	0.83	0.93	0.83	0.07	7.686	0.1133
8 Dimertest II	0.53	0.1483	27.67	0.908	0.70	0.92	0.70	0.08	3.667	0.1325
9 Enzygnost	0.76	0.6806	39.50	0.928	0.76	0.93	0.76	0.07	4.85	0.1228
10 Fibrinostika	1.03	0.9529	53.30	0.942	0.80	0.93	0.80	0.07	6.23	0.1169
11 IL Test	0.43	0.0811	22.15	0.894	0.67	0.92	0.67	0.08	3.115	0.1406
12 Instant I.A.	0.67	0.4596	34.70	0.921	0.74	0.93	0.74	0.07	4.37	0.1259
13 LPIA	1.01	0.9878	52.35	0.941	0.80	0.93	0.80	0.07	6.135	0.1172
14 Liatest	0.88	0.8265	45.47	0.935	0.78	0.93	0.78	0.07	5.447	0.1198
15 Minutex	0.45	0.0694	23.19	0.897	0.67	0.92	0.67	0.08	3.219	0.1388
16 Nephelotex	1.68	0.3526	87.37	0.959	0.86	0.93	0.86	0.07	9.637	0.1103
17 NycoCard	0.28	0.0327	14.52	0.861	0.60	0.91	0.60	0.09	2.352	0.1620
18 SimpliRED	0.46	0.1755	23.92	0.899	0.68	0.92	0.68	0.08	3.292	0.1376
19 Tinaquant	1.07	0.922	55.43	0.943	0.80	0.93	0.80	0.07	6.443	0.1162
20Turbiquant	0.20	0.0009	10.46	0.831	0.55	0.89	0.55	0.11	1.946	0.1860
21 VIDAS	1.00	[c]	51.90	0.941	0.80	0.93	0.80	0.07	6.09	0.1173

Relative DOR = ratio of test's DOR to that of reference category (VIDAS).
DOR = Diagnostic odds ratio.
AUC = Area under curve (assuming homogeneous DOR).
PPV = Positive predictive value.
NPV = Negative predictive value.
Pab = Post test probability of abnormal test.
Pnr = Post test probability of normal test.
LRab = Likelihood ratio for abnormal test.
LRnr = Likelihood ratio for normal test.

[a] For prevalence of 39%, gold standard of Venography, and outpatient setting.

[b] For sensitivity of 90% (besides the assumptions of [a]).

[c] Reference category.

were significantly ($P < .05$) different (all lower) from the VIDAS assay. The DORs of the other 17 assays were not significantly different from the VIDAS assay DOR, although most trended toward lower discriminant ability.

We included factors associated with study design and patient population in the model (Table 5). Increasing prevalence of DVT in the study population was independently associated with poorer assay performance ($P = .01$), while the choice of venography as the reference standard was associated with better assay performance ($P < .005$). This shows that the choice of reference standard confounds D-dimer study results. The model found no significant effect on assay performance of patient mix (outpatient or other).

We have provided a code to convert the estimated DORs into SROC curves (see Appendix). One can use the SROC graphs to compare different assays, or different values of a covariate (Fig. 2). One should be aware that the plotted SROC curves might extend beyond the observed data.

Using the SROC curve one can compute the AUC. The AUC for VIDAS is 0.94 and that of Dimertest is 0.72. The AUC of VIDAS decreases from 0.94 to 0.88 when venography, as the gold standard, is replaced by ultrasonography. Using the covariance matrix of coefficients one can compute the CI for each SROC curve.

Table 6 shows the result of test for homogeneity of odds ratio for each test. When only one paper has studied a test, no *P*-value can be computed (and testing the homogeneity is meaningless). Out of the 14 tests where multiple papers have studied them, half of them show significant heterogeneity across the corresponding papers (not adjusted for multiple comparisons). This necessitates further study of heterogeneity within each test. For example, for test Instant I.A., row 12, the six papers reported sensitivities and specificities that translate to ORs of 3.12 by Elias, 60.46 by Legnani, 6.75 by Leroyer, 208.14 by Scarano, 26.68 by van der Graaf, and 8.66 by Wijns. Or for test Enzygnost, row 9, the three papers reported ORs of 5.69 by Elias, 25.56 by van der Graaf, and a virtually infinite OR by Scarano. It is a bit reassuring that all the ORs are in the same direction, but the between-study variation is quite large. One may want to omit the outliers and study the effect on the estimates. An alternative is to introduce a covariate *Z*, as described in model (5), to capture between-study variation.

The residuals can help locate observations that are poorly accounted for by the model. Fig. 3 shows the residuals of each paper in a boxplot. One expects the median of the residuals for each paper (the solid horizontal line inside the box) to be around the horizontal zero line, with no

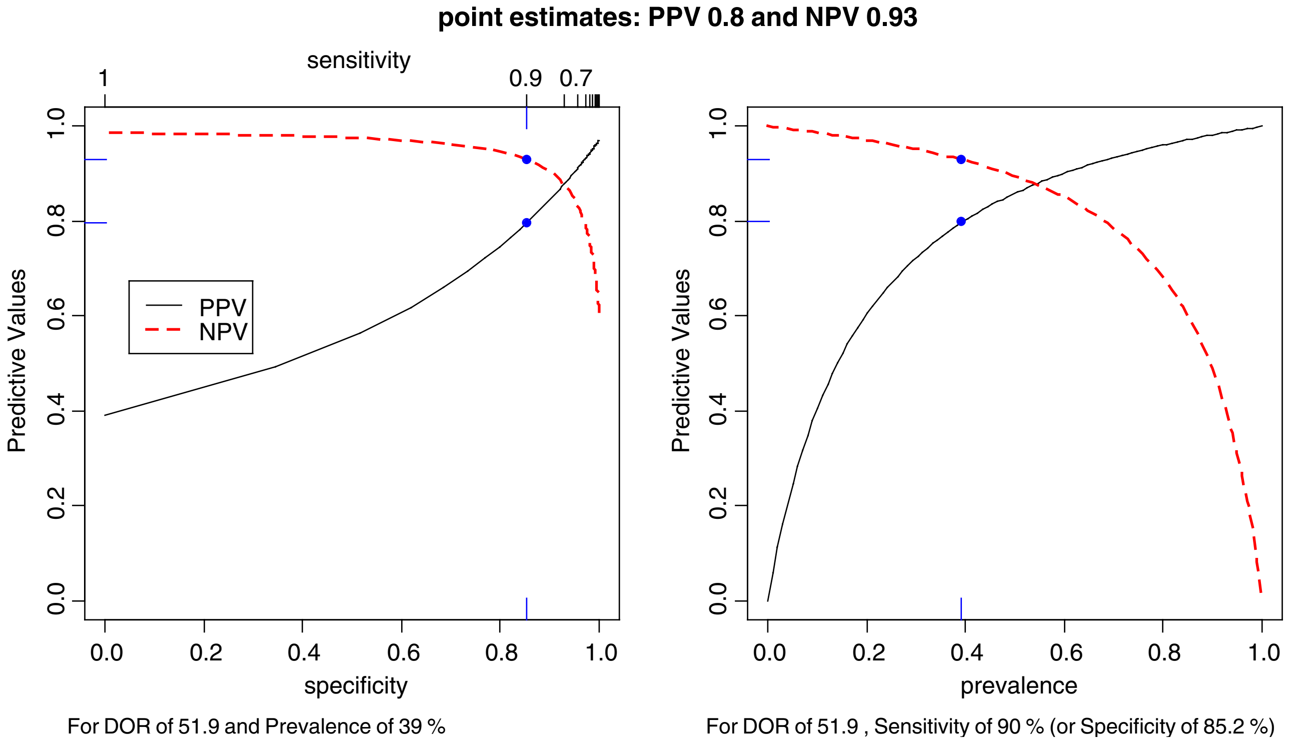


Fig. 1. Performance graphs for test VIDAS.

outliers. Paper #5 (Elias, 1996) has a big negative residual and a lower outlier. It is for test Instant IA. Additionally, papers #30 and 31 have negative residuals also. They are from two papers by Wells both about test SimpliRED. Table 7 shows the five biggest and five smallest Pearson's residuals across all papers.

Comparing the above table with the one for homogeneity of ORs shows consistent papers/tests that have both heterogeneous OR and big residuals. Again, a sensitivity analysis by omitting these observations and refitting the models may be worthwhile.

5. Discussion

Deeks mentions that, when a threshold effect exists, study results may be best summarized as an SROC. (If the observed heterogeneity between the studies arises due to variation in the diagnostic threshold, a threshold effect exists.) Even in this case, Deeks believes that the SROC is difficult to interpret and apply to practice. Lee [12] believes that, using a Lorenz curve, Pietra and Gini indexes have a closer tie with real-world medical diagnosis.

Although the method proposed by Moses et al. [3] is commonly used to build a summary ROC curve, it has several drawbacks.

1. It introduces a new measure of test efficacy (the Q^*, where the SROC crosses the line of sensitivity = specificity), which is not utilizing the whole SROC curve. Q^* is a point of indifference on the SROC curve between false positive and false negative diagnostic errors. It assumes implicitly that the two errors are of equal value. However, "one must weigh the two to balance the overall performance of the test in a population; the optimal diagnostic threshold need not then correspond to the Q^* point …(also) it conveys no additional statistical information beyond the odds ratio " [4].
2. It relies on asymptotic normality in the dependent variable D, ignores errors in the independent variable S, and assumes that the test measurement vs. the threshold value follows a logistic distribution to reach the conclusion that the line in the (U, V) space is straight.
3. It requires adding an arbitrary number to the cells (continuity correction), introducing downward bias.
4. It requires arbitrarily eliminating some of the observations (points outside the ROC left upper region), introducing a possibly upward bias.

Table 5
Effect of covariates on test performance

Covariate	Relative DOR[a]	*P*-value
Reference standard		
US or US/Venog	0.37	0.0045
Venog	1	[b]
Patient mix		
Mixed	0.61	0.2068
Outpatient	1	[b]
Prevalence (for 10% increase)	0.8	0.01

[a] Relative DOR = effect on DOR compared to that of reference category.
[b] Reference category.

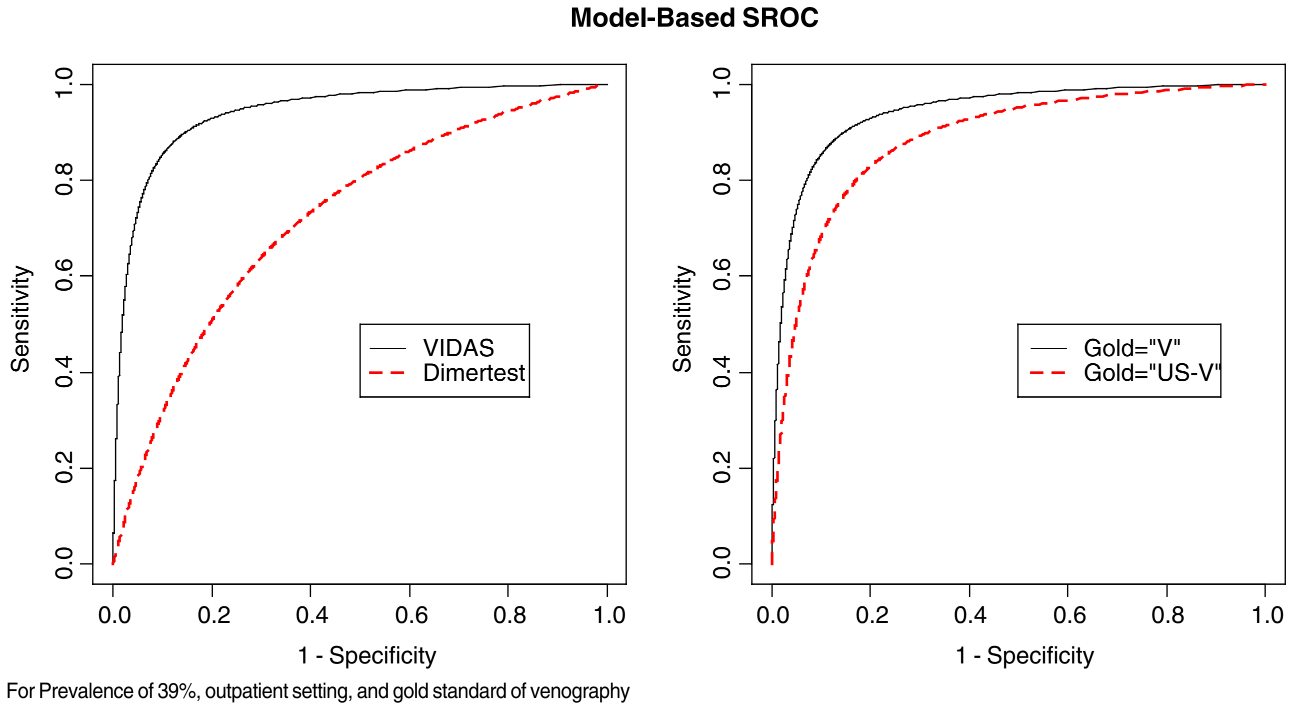


Fig. 2. Comparison of two DVT tests (left), and two gold standards within test VIDAS (right).

5. It does not take into account sample size of each study.
6. It does not address the issue of correlated measurements. If overlapping subjects were used to give performance measures for different tests, the standard errors would be incorrect.

There are similarities between the method proposed here and the one of Moses et al. [3]; however, it does not have the disadvantages enumerated above. (In generalized estimating equations [GEE] regression coefficients remain consistent even when the correlation structure is mis-specified. However, the linear predictor should be specified correctly. Additionally, the missing values should be completely at random.)

Table 6
Homogeneity of ORs

Diagnistic test	Number of papers	*P*-value
1 Asserachrom	5	.0727
2 Auto Dimertest	1	—
3 BC D-dimer	2	.673
4 D-dimer test	1	—
5 Dimertest	1	—
6 Dimertest EIA	1	—
7 Dimertest GOLD EIA	2	.168
8 Dimertest II	1	—
9 Enzygnost	3	.001
10 Fibrinostika	3	.0803
11 IL Test	2	.5013
12 Instant I.A.	6	<.0001
13 Liatest	2	.299
14 LPIA	1	—
15 Minutex	3	.019
16 Nephelotex	1	—
17 NycoCard	5	.0007
18 SimpliRED	11	.0005
19 Tinaquant	4	.0284
20 Turbiquant	2	.4894
21 VIDAS	9	.0064

Each DOR is on a one-to-one correspondence with a ROC curve (under the “independence of test threshold and OR” assumption). Hence, one can use the estimated DOR out of the model to build a SROC curve for each test type. Also, the method by Moses et al. [3] is equivalent to fitting a logistic regression model with interaction terms, where the primary study unit is paper. Because the test statistic introduced by Moses et al. [3] (the Q^*) is built on the computed SROC curve rather than the coefficients of the logistic model, it may inherit some of the flaws from it.

Estimating summary DOR implicitly assumes that the studies are homogeneous (in terms of estimated OR across studies). This may not be the case. However, Walter [4] shows that the AUC calculated from the SROC is a reasonable approximation with heterogeneous studies. In the method by Moses et al. [3], even though a nonzero slope could capture/show the heterogeneity, their test statistic Q^* is invariant to heterogeneity.

The method for capturing between-study variation presented here (by generalizing the way the covariate Z is defined) can capture a broader class of heterogeneity, while the approach by Moses et al. [3] can capture only certain kinds. They assume their D and S transforms have linear relationship. When the slope coefficient of their model, the B, is unequal to zero (that is there is heterogeneity), Walter shows that their SROC curve is constrained to have one of the two general S shapes he presented.

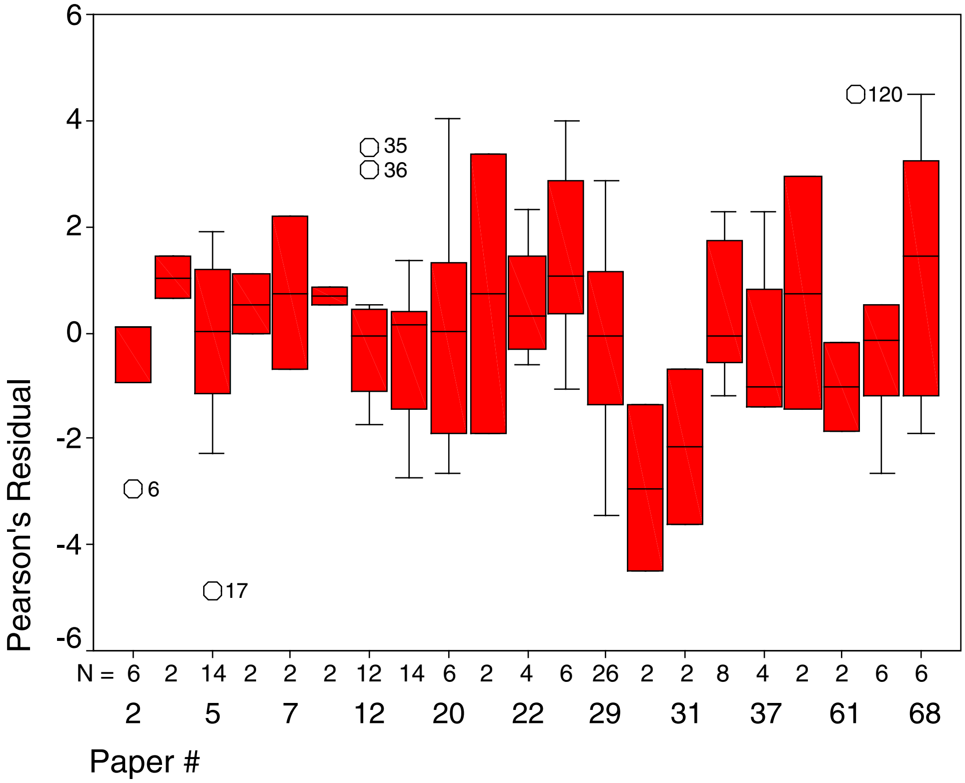


Fig. 3. Residuals.

Although currently the majority of meta-analytic efforts have no access to patient-level data, it is predictable that this may change in the near future, considering numerous recommendations by advisory bodies. The method presented here can be extended to utilize patient level data with no difficulty. The model has been tailored for grouped binary data. It takes only a minor change to run it on “ungrouped” patient-level data. A related point is that because the proposed method expands each paper to its original sample size, it does not have the flaws caused by too few papers studying a test (limited number of studies), because the primary study units are persons not papers. Even a test with only a single published study can be entered into the meta-analysis. Of course, it is still sensitive to small numbers of subjects per study.

Table 7
Extreme residuals

Paper	Test	Pearson’s
63Sadouk, M. 2000 (177)	Tinaquant	4.49
68Wijns, W. 1998 (74)	VIDAS	4.48
20Legnani, C. 1999 (99)	VIDAS	4.03
26Scarano, L. 1997 (126)	Instant I.A.	4.01
12Janssen, M. 1997 (132)	SimpliRED	3.51
5Elias, A. 1996 (171)	Instant I.A.	−4.87
30Wells, P. 1999 (150)	SimpliRED	−4.49
31Wells, P. 1995 (214)	SimpliRED	−3.62
29van der Graaf, F. 2000 (99)	NycoCard	−3.46
2Brenner, B. 1995 (86)	SimpliRED	−2.96

To fit the model proposed here one requires software supporting GEE. Such software is readily available in SAS (genmod procedure), R (function geese),[16] and STATA (command xtgee), with R being freely available open source software [13]. We doubt the method can be hand calculated. (Function geese{} in package geepack currently does not handle grouped binary data. Function gee() in package gee does not support the *t* index.)

In summary, the proposed method handles papers that overlap partially in the types and numbers of tests they studied. It allows for missing values of some tests for some papers, takes into account different sample sizes of papers, adjusts for background and confounding factors including test-specific covariates and paper-specific covariates, and accounts for correlations of the repeated measurements within each paper. It does not need (but can accommodate) individual-level data, and uses a two-by-two table of test result vs. the gold standard. It is capable of testing hypotheses of interest as well as providing estimates of the magnitude of effect/difference. Additionally one can translate the model estimates into measures that are more practical, such as predictive values and post-test probabilities. The proposed method not only gives more accurate results than the alternative methodologies, but extends the meta-analytic approach to clustered data not appropriate for previous methods.

Methods for estimation/visualization of influence of individual studies on the estimated parameters and the overall model fit that are available for repeated-measures models are directly applicable here.

Correspondence between the suggested method and recent measures of performance like area swept out by the curve, projected length of the curve, or the Lorenz curve indexes Pietra and Gini may be worth investigating [12].

Acknowledgment

We thank Frank E. Harrell, Jr for useful comments.

References

[1] Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323:157–62.

[2] Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995;117(1):167–78.

[3] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;30:1293–316.

[4] Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002;21:1237–56.

[5] Heim SW, Philbrick JT. D-dimer assays for deep venous thrombosis: a systematic review. J Gen Intern Med 2002;17(Suppl 1):112.

[6] Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13(4):313–21.

[7] Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS system for mixed models. Cary (NC): SAS Institute Inc.; 1996.

[8] Diggle P, Heagerty P, Liang K-Y, Zeger S. Analysis of longitudinal data. 2nd edition. New York: Oxford University Press; 2002. p. 142–7.

[9] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Proc Fifth Berkeley Symp Math Stat 1967; 1:221–33.

[10] Davis W, Breslow NE, Day NE. Statistical methods in cancer research: the analysis of case–control studies (Iarc Scientific Publications, No. 32, reprint edition). New York: Oxford University Press; 1993. p. 142.

[11] SAS Institute Inc. SAS/STAT user's guide, version 8, vol. 2. Cary (NC): SAS Institute Inc.; 1999.

[12] Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Stat Med 1999;18(4):455–71.

[13] Ihaka R, Gentleman R. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.

Appendix

This appendix contains the SAS code to implement some of the models of the paper. Also, it includes the SAS output for the models. It has a section on calculating DOR for each diagnostic test starting from the model estimates.

We have written several functions (in R) that implement postmodel calculations, convert model-based DOR into other performance measures, plus producing several types of plots, some of which have been presented in this paper. The address http://www.people.virginia.edu/~mss4x/meta.html has the supplementary material. The Web site includes documents showing the relationship between different performance measures.

Implementing model (6) in SAS

The following exhibit is model (6) implemented in SAS [11]. Note the grouped binary format in the model statement.

Option subject in the repeated statement defines the p index of papers. Option within defines the t index of tests. Because not every paper studied the same tests, defining the t index by "within" is advisable.

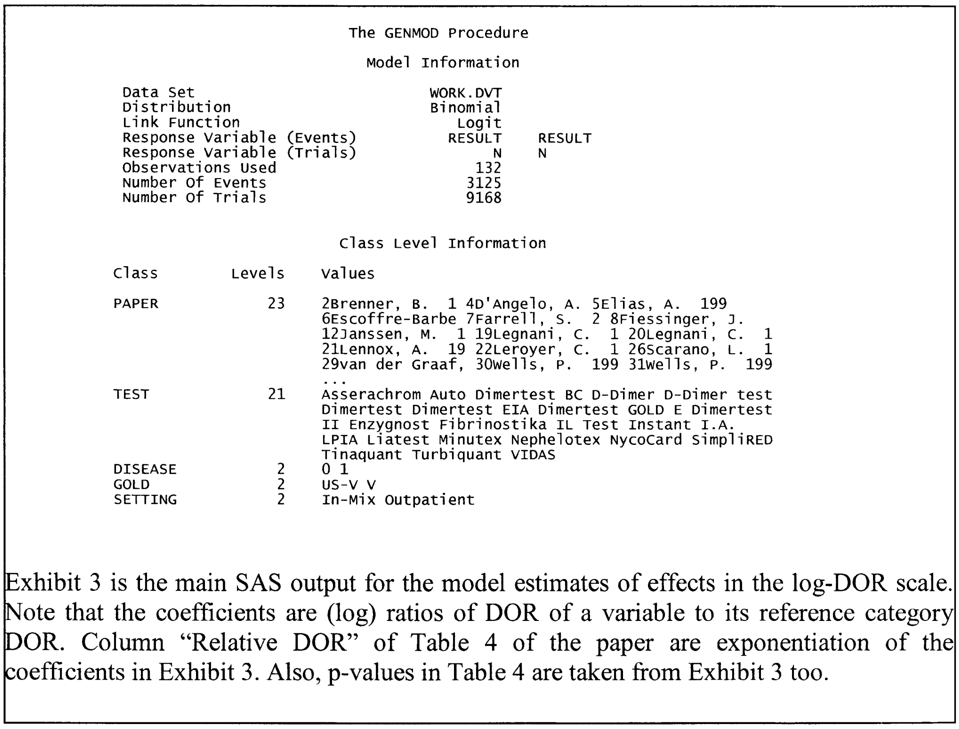
Option type defines the type of correlation matrix (printed through option corrw). We have used an exchangeable correlation matrix here. By rerunning the command with a few choices of correlation matrix one can evaluate sensitivity of the final inference to the specification of the correlation matrix. Furthermore, one may use an unstructured correlation matrix to get an idea how the correlations between observations are. Option covb enables one to build confidence intervals for contrasts of interest.

Exhibit 2 shows there were 9,168 measurements ("Number of Trials") cumulatively in the 23 papers (the sum of column "N"). This includes repeated measurements from the same sample for several tests within the same paper. This is the result of expanding the grouped binary data shown in Table 3. Note that the primary study unit is each person, not the papers or tests. The effective sample size is smaller than the nominal sample size of 9,168 (due to correlation between the measurements), but it is bigger than the sum of the sample sizes of the papers included in the meta-analysis (here it is 3,860). To calculate the effective sample

Exhibit 1. Computer Command for Fitting the Marginal Logistic Model

```
proc genmod data=dvt ;
class paper test disease gold setting;
model result/n = disease test prvlnc setting gold disease*test
disease*prvlnc disease*setting disease*gold
/ dist=bin link=logit;
repeated subject=paper / within=test*disease type=exch covb corrw;
output out=dvtrsdl resraw=rawr reschi=pearsonr;
run;
```

Exhibit 2. Computer (SAS) Output, Summary Report



```
                              The GENMOD Procedure

                                Model Information

          Data Set                          WORK.DVT
          Distribution                      Binomial
          Link Function                        Logit
          Response Variable (Events)          RESULT    RESULT
          Response Variable (Trials)               N    N
          Observations Used                      132
          Number Of Events                      3125
          Number Of Trials                      9168


                             Class Level Information

         Class        Levels    Values

         PAPER            23    2Brenner, B.  1 4D'Angelo, A. 5Elias, A.  199
                                6Escoffre-Barbe 7Farrell, S.  2 8Fiessinger, J.
                                12Janssen, M.  1 19Legnani, C.  1 20Legnani, C.  1
                                21Lennox, A.  19 22Leroyer, C.  1 26Scarano, L.  1
                                29van der Graaf, 30Wells, P.  199 31Wells, P.  199
                                ...
         TEST             21    Asserachrom Auto Dimertest BC D-Dimer D-Dimer test
                                Dimertest Dimertest EIA Dimertest GOLD E Dimertest
                                II Enzygnost Fibrinostika IL Test Instant I.A.
                                LPIA Liatest Minutex Nephelotex NycoCard SimpliRED
                                Tinaquant Turbiquant VIDAS
         DISEASE           2    0 1
         GOLD              2    US-V V
         SETTING           2    In-Mix Outpatient
```

Exhibit 3 is the main SAS output for the model estimates of effects in the log-DOR scale. Note that the coefficients are (log) ratios of DOR of a variable to its reference category DOR. Column "Relative DOR" of Table 4 of the paper are exponentiation of the coefficients in Exhibit 3. Also, p-values in Table 4 are taken from Exhibit 3 too.

size one needs to take into account the correlation of measurements. Stronger correlation results in a smaller effective sample size.

Using a rule of thumb of spending one degree of freedom (*df*) per 10 primary study units, one can decide how elaborate a model can be fitted. Because our effective sample size is at least 3,860, the number of *df*s one can spend is roughly 380. In the SAS output for this DVT example, SAS reports 84 *df* have been spent for the five covariates of the model, or equivalently 84 parameters have been estimated for the model (not shown in the exhibit). (Number of parameters of a model is based on number of covariates in the model, number of categories of the categorical covariates, and the presence of interaction terms.) Therefore, we have enough sample size for a model of this complexity, without fearing overfitting.

The number of events shown in the exhibit is the sum of TN and FN (column "Result"). It is the number of measurements the tests considered as not having disease.

Exhibit 3 is the main SAS output for the model estimates of effects in the log-DOR scale. Note that the coefficients are (log) ratios of DOR of a variable to its reference category DOR. Column "Relative DOR" of Table 4 of the paper are exponentiation of the coefficients in Exhibit 3. Also, *P*-values in Table 4 are taken from Exhibit 3.

Exhibit 4 is SAS code for implementing model (4). More accurately, it implements a version of Model (4) that needs a single run for estimating all the *P*-values for test of homogeneity of all the diagnostic tests. The model is

$$\begin{aligned}\text{logist(Result)} = \; & \beta_0 + \beta_1\text{*Test} + \beta_2\text{* Disease * Test}\\ & + \beta_3\text{* Paper * Test}\\ & + \beta_4\text{* Disease * Paper * T1}\\ & + \beta_5\text{* Disease * Paper * T2}\\ & + \beta_6\text{*Disease* Paper* T3}\\ & + \ldots + \beta_{k+3}\text{*Disease * Paper *T}k\end{aligned}$$

where variable Test is a categorical standing for different types of diagnostic tests studied across all papers. The variables T1 to T*k* are indicators for each test; hence, *k* is equal to the total number of unique tests studied in all papers included in the meta-analysis. A type III global test of each of the coefficient β_4 to β_{k+3} is equivalent to the Breslow-Day test of OR homogeneity for each diagnostic test.

Calculating DORs from the model output

To calculate DOR for each test, using model (6), we start with the reference category test VIDAS. The value for variable Test in the model therefore is 0. This gives us

$$\begin{aligned}&\beta_0 + \beta_1\text{* Disease}_{\text{pt}} + \beta_3\text{* Prvlnc}_{\text{pt}} + \beta_4\text{* Gold}_{\text{pt}}\\ &\quad + \beta_5\text{* Setting}_{\text{pt}} + \beta_7\text{* Disease}_{\text{pt}}\text{ * Prvlnc}_{\text{pt}}\\ &\quad + \beta_8\text{* Disease}_{\text{pt}}\text{* Gold}_{\text{pt}} + \beta_9\text{* Disease}_{\text{pt}}\text{* Setting}_{\text{pt}}\end{aligned}$$

for the Disease = 1, and $\beta_0 + \beta_3*Prvlnc_{pt} + \beta_4*Gold_{pt} + \beta_5*Setting_{pt}$ for the Disease = 0. Subtracting these two

Exhibit 3. Ratio of Diagnostic ORs (in log scale), SAS outrput

```
                                  The GENMOD Procedure

                             Analysis Of GEE Parameter Estimates
                              Empirical Standard Error Estimates

                                                  Standard   95% Confidence
Parameter                               Estimate     Error       Limits            Z Pr > |Z|

TEST*DISEASE     Asserachrom       0     -0.4632    0.4251  -1.2964   0.3699   -1.09   0.2758
TEST*DISEASE     Asserachrom       1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Auto Dimertest    0      0.5148    0.4192  -0.3067   1.3364    1.23   0.2194
TEST*DISEASE     Auto Dimertest    1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     BC D-Dimer        0     -0.9612    0.5720  -2.0824   0.1599   -1.68   0.0929
TEST*DISEASE     BC D-Dimer        1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     D-Dimer test      0     -1.0053    0.5648  -2.1122   0.1016   -1.78   0.0751
TEST*DISEASE     D-Dimer test      1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Dimertest         0     -2.5344    0.4693  -3.4542  -1.6147   -5.40   <.0001
TEST*DISEASE     Dimertest         1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Dimertest EIA     0     -0.8218    0.4341  -1.6726   0.0291   -1.89   0.0584
TEST*DISEASE     Dimertest EIA     1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Dimertest GOLD E  0      0.2680    0.6724  -1.0498   1.5858    0.40   0.6902
TEST*DISEASE     Dimertest GOLD E  1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Dimertest II      0     -0.6292    0.4352  -1.4821   0.2238   -1.45   0.1483
TEST*DISEASE     Dimertest II      1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Enzygnost         0     -0.2732    0.6636  -1.5738   1.0274   -0.41   0.6806
TEST*DISEASE     Enzygnost         1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Fibrinostika      0      0.0266    0.4500  -0.8554   0.9086    0.06   0.9529
TEST*DISEASE     Fibrinostika      1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     IL Test           0     -0.8515    0.4881  -1.8081   0.1051   -1.74   0.0811
TEST*DISEASE     IL Test           1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Instant I.A.      0     -0.4028    0.5447  -1.4704   0.6648   -0.74   0.4596
TEST*DISEASE     Instant I.A.      1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     LPIA              0      0.0086    0.5593  -1.0875   1.1047    0.02   0.9878
TEST*DISEASE     LPIA              1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Liatest           0     -0.1323    0.6032  -1.3145   1.0500   -0.22   0.8265
TEST*DISEASE     Liatest           1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Minutex           0     -0.8056    0.4437  -1.6752   0.0640   -1.82   0.0694
TEST*DISEASE     Minutex           1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Nephelotex        0      0.5207    0.5602  -0.5773   1.6187    0.93   0.3526
TEST*DISEASE     Nephelotex        1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     NycoCard          0     -1.2738    0.5963  -2.4425  -0.1050   -2.14   0.0327
TEST*DISEASE     NycoCard          1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     SimpliRED         0     -0.7745    0.5716  -1.8948   0.3459   -1.35   0.1755
TEST*DISEASE     SimpliRED         1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Tinaquant         0      0.0658    0.6722  -1.2517   1.3832    0.10   0.9220
TEST*DISEASE     Tinaquant         1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     Turbiquant        0     -1.6014    0.4826  -2.5473  -0.6555   -3.32   0.0009
TEST*DISEASE     Turbiquant        1      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     VIDAS             0      0.0000    0.0000   0.0000   0.0000     .       .
TEST*DISEASE     VIDAS             1      0.0000    0.0000   0.0000   0.0000     .       .
PRVLNC*DISEASE   0                       -0.0221    0.0088  -0.0394  -0.0048   -2.50   0.0125
PRVLNC*DISEASE   1                        0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*SETTING  0                 In-Mix     -0.4875    0.3862  -1.2444   0.2693   -1.26   0.2068
DISEASE*SETTING  0                 Outpatient  0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*SETTING  1                 In-Mix      0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*SETTING  1                 Outpatient  0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*GOLD     0                 US-V       -0.9888    0.3482  -1.6713  -0.3063   -2.84   0.0045
DISEASE*GOLD     0                 V           0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*GOLD     1                 US-V        0.0000    0.0000   0.0000   0.0000     .       .
DISEASE*GOLD     1                 V           0.0000    0.0000   0.0000   0.0000     .       .
```

Exhibit 4 is SAS code for implementing model (4). More accurately, it implements a version of Model (4) that needs a single run for estimating all the p-values for test of homogeneity of all the diagnostic tests. The model is

$$logit(Result) = \beta_0 + \beta_1 *Test + \beta_2 *Disease*Test + \beta_3 *Paper*Test + \beta_4 *Disease*Paper*T1 + \beta_5 *Disease*Paper*T2 + \beta_6 *Disease*Paper*T3 + \ldots + \beta_{k+3} *Disease*Paper*Tk$$

where variable Test is a categorical standing for different types of diagnostic tests studied across all papers. The variables T1 to Tk are indicators for each test, hence k is equal to the total number of unique tests studied in all papers included in the meta-analysis. A type III global test of each of the coefficient β_4 to β_{k+3} is equivalent to the Breslow-Day test of OR homogeneity for each diagnostic test

Exhibit 4. Computer Command (for SAS) for Testing Homogeneity of ORs

```
proc genmod data=dvt ;
class disease paper test;
model result/n = test disease*test paper*test paper*disease*t1
paper*disease*t2 paper*disease*t3 paper*disease*t4 paper*disease*t5
paper*disease*t6 paper*disease*t7 paper*disease*t8 paper*disease*t9
paper*disease*t10 paper*disease*t11 paper*disease*t12 paper*disease*t13
paper*disease*t14 paper*disease*t15 paper*disease*t16 paper*disease*t17
paper*disease*t18 paper*disease*t19 paper*disease*t20 paper*disease*t21
/ dist=bin link=logit type3;
run;
```

formulas gives the formula for LDOR of VIDAS. It is $\beta_1 + \beta_7 * Prvlnc_{pt} + \beta_8 * Gold_{pt} + \beta_9 * Setting_{pt}$. Now we replace variable Prvlnc by its mean value 39.0, Setting is replaced by the more frequent category "outpatient," and Gold is replaced by category "V." Because category V for Gold, and outpatient for Setting are the reference categories, this effectively removes terms involving β_8 and β_9 from the formula above; hence, $\beta_1 + \beta_7 * Prvlnc_{pt}$. β_1 is 4.9792, and β_7 is -0.0271, thus giving log-DOR of 3.9494 for VIDAS. By exponentiating the log DOR we get the DOR of VIDAS [exp(3.9494) = 51.90421]. The next step is to pick a test that we want its DOR, say "Dimertest." The ratio of DOR of test Dimertest to the VIDAS is 0.079309 [which is exp(-2.5344)]; hence, its DOR is 4.116486.