Methods Research Report

Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy: Methods for the Joint Meta-Analysis of Multiple Tests

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy: Methods for the Joint Meta-Analysis of Multiple Tests

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10055-I

Prepared by:
Tufts Evidence-based Practice Center, Tufts Medical Center
Boston, MA

Investigators
Thomas A. Trikalinos, M.D.
David C. Hoaglin, Ph.D.
Kevin M. Small, Ph.D.
Christopher H. Schmid, Ph.D.

AHRQ Publication No. 12(13)-EHC151-EF
January 2013

This report is based on research conducted by Tufts Evidence-based Practice Center, Tufts Medical Center, under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, (Contract No. 290-2007-10055-I). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical research and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact effectivehealthcare@ahrq.hhs.gov.

Authors TAT and CHS are involved in developing open source software for meta-analysis, but this software is available at no charge and the authors receive no financial benefit. None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Trikalinos TA, Hoaglin DC, Small KM, Schmid CH. Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy: Methods for the Joint Meta-Analysis of Multiple Tests. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center, under Contract No. 290-2007-10055-I.) AHRQ Publication No. 12(13)-EHC151-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. www.effectivehealthcare.ahrq.gov.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below, at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D.
Director, EPC Program
Agency for Healthcare Research and Quality

Elisabeth U. Kato, M.D., M.R.P.
Task Order Officer
Agency for Healthcare Research and Quality

Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy: Methods for the Joint Meta-Analysis of Multiple Tests

Structured Abstract

Background: Existing methods for meta-analysis of diagnostic test accuracy focus primarily on a single index test rather than comparing two or more tests that have been applied to the same patients in paired designs.

Objectives: We develop novel methods for the joint meta-analysis of studies of diagnostic accuracy that compare two or more tests on the same participants.

Development of methods: We extend the bivariate meta-analysis method proposed by Reitsma et al. (J Clin Epidemiol. 2005; 58[10]:982-90) and modified by others to simultaneously meta-analyze $M \geq 2$ index tests. We derive and present formulas for calculating the within-study correlations between the true-positive rates (TPR, sensitivity) and between the false-positive rates (FPR, one minus specificity) of each test under study using data reported in the studies themselves. The proposed methods respect the natural grouping of data by studies, account for the within-study correlation between the TPR and the FPR of the tests (induced because tests are applied to the same participants), allow for between-study correlations between TPRs and FPRs (such as those induced by threshold effects), and calculate asymptotically correct confidence intervals for summary estimates and for differences between summary estimates. We develop algorithms in the frequentist and Bayesian settings, using approximate and discrete likelihoods to model testing data.

Application: Published meta-analysis of 11 studies on the screening accuracy of detecting trisomy 21 (Down syndrome) in liveborn infants using two tests: shortened humerus (arm bone), and shortened femur (thigh bone). Secondary analyses included an additional 19 studies on shortened femur only.

Findings: In the application, separate and joint meta-analyses yielded very similar estimates. For example, in models using the discrete likelihood, the summary TPR for a shortened humerus was 35.3 percent (95% credible interval [CrI]: 26.9, 41.8%) with the novel method, and 37.9 percent (27.7 to 50.3%) when shortened humerus was analyzed on its own. The corresponding numbers for the summary FPR were 4.8 percent (2.8 to 7.5%) and 4.8 percent (3.0 to 7.4%).

However, when calculating comparative accuracy, joint meta-analyses resulted in shorter confidence intervals compared with separate meta-analyses for each test. In analyses using the discrete likelihood, the difference in the summary TPRs is 0 percent (-8.9, 9.5%; TPR higher for shortened humerus) with the novel method versus 2.6 percent (-14.7, 19.8%) with separate meta-analyses. The standard deviation of the posterior distribution of the difference in TPR with joint meta-analyses is half of that with separate meta-analyses.

Conclusions: The joint meta-analysis of multiple tests is feasible. It may be preferable over separate analyses for estimating measures of comparative accuracy of diagnostic tests. Simulation and empirical analyses are needed to better define the role of the proposed methodology.

Contents

Tables

Figures

Appendix

Background

The value of diagnostic testing for patient management ultimately derives from its effect on patient-relevant outcomes. Testing primarily affects outcomes by indirect means, influencing downstream patient management decisions, including decisions for further testing, and administration or choice of treatment. Thus, the relative accuracy of diagnostic tests has considerable importance. The ideal study compares the effectiveness and safety of complete test-and-treat strategies, but such studies are exceedingly rare.[1] Instead, one has to assemble data on effectiveness and safety from various sources,[2] including studies of test performance (test accuracy) and studies of the effects of interventions. For this reason, meta-analysis of accuracy plays a key role in evaluations of medical tests,[3] although it does not directly determine the comparative effectiveness and safety of test-and-treat strategies.[4-7]

Two key characteristics of a diagnostic test, sensitivity and specificity, depend on its threshold for classifying an outcome as positive. If high values are "positive", lowering the threshold increases sensitivity at the expense of specificity; increasing the threshold moves the measures in the opposite directions. Recent statistical work has focused on developing bivariate analytic techniques that jointly model sensitivity and specificity in order to account for this inherent negative correlation.[8,9] Such models have focused on a single test.

To compare of the accuracy of two (or more) tests, analysts typically take one of two approaches: (a) perform a separate meta-analysis for each test and compare the meta-analytic summaries; or (b) perform a meta-regression using the type of test as a categorical predictor. Both methods assume that all individuals studied are independently sampled (i.e., that the tests study two distinct sets of individuals, each containing individuals with disease and individuals without disease). But when the same individuals receive both tests, their results are correlated. A valid statistical model must account for this dependence. Because the two common approaches do not take these correlations into account, neither is a valid method for comparing tests performed on the same individuals.

In broad terms, diagnostic accuracy studies can be noncomparative, when they assess one index test at a given setting, or comparative, when they assess the performance of two or more tests. Estimates of comparative test accuracy can be obtained from either category of studies. Estimates from the former group are confounded by study setting, whereas estimates from the latter group are not. Further, comparative studies can have paired (or "crossover") designs (where each test is applied to the same patients) or parallel designs (where each test is applied in disjoint sets of patients, e.g., using random allocation or other means). A paired design is statistically much more efficient, in that one needs much smaller sample sizes to detect a given difference in test accuracy, compared with a parallel design.

Reflecting this, data from studies that evaluate two tests can take more complicated forms. A "crossover" study, which obtains results on both tests for each individual, may report the sensitivity and specificity (e.g., as a 2×2 table) for each test separately, and not the full cross-classifications. Or one set of individuals may receive both tests, a second subset only the first test, and a third subset only the second test. Yet another study design randomly assigns individuals to the two tests or, without randomization, administers each test to separate sets of individuals (including those with disease and those without disease). This work focuses primarily on the case in which the same individuals receive both tests and the results are cross-classified. It briefly discusses some of the other forms, which represent minor modifications to the models for assessing studies with paired designs.

Jointly analyzing the true- and false-positive rate (equivalently, sensitivity and one minus specificity) of two diagnostic tests made on the same individuals requires multivariate models, to account for the relation between the outcomes of the two tests in patients with disease and in patients without disease within studies and variation in those outcomes among studies. In this work, we derive a model that accounts for such relations. The model can be fitted to data that report totals of true and false positive and negative results for each test and for each combination of tests. We provide algorithms for maximizing the resulting likelihood and for calculating asymptotically correct confidence intervals for summary estimates and for differences between summary estimates. We also provide algorithms for fitting models with Markov Chain Monte Carlo (MCMC) methods in a Bayesian framework.

The remainder of the work is organized around a motivating example from a meta-analysis of two second-trimester ultrasonographic tests that screen for trisomy 21 (Down syndrome). We first describe the example, and then we present the model for two tests and extensions to more than two tests. We then proceed with a description of estimation and inference, and a comment on some patterns of missing data (at the study level). We then present an analysis of the data, and we conclude with a discussion of the method and its applications.

Illustrative Example

We examine the screening accuracy of second-trimester ultrasound markers in detecting liveborn infants with trisomy 21 (Down syndrome). Briefly, trisomy 21 is a clinically important chromosomal abnormality whose frequency is strongly associated with maternal age (more common in older mothers). Infants with trisomy 21 manifest mental retardation and have high risk of congenital structural defects.[10] Until the mid-1980s maternal age was the only predictor for trisomy 21. Since then, biochemical measurements in maternal plasma (alpha-fetoprotein, human chorionic gonadotrophin, inhibin A and estriol), as well as many ultrasonographic markers in the first and second trimester, have been examined as screening tests. Mothers and fetuses identified by a positive screening test result are typically offered a definitive diagnosis via amniocentesis, an invasive diagnostic test.

As an illustration, we use information from 11 studies combined in a published meta-analysis that examined the screening accuracy of seven ultrasonographic markers or their combination in detecting trisomy 21 in liveborn infants.[11] We focus on two such markers, shortened humerus (arm bone), and shortened femur (thigh bone) of the fetus. Two of the authors (TAT and DCH) extracted information from the full text of the pertinent articles on the provenance of the paper (first author, year of publication), the definition of positive screening tests, and cross-classification counts for the presence or absence of each marker among infants with and without trisomy 21. The extractors discussed and reconciled all results and also discussed them with a third author (CHS).

We emphasize that this example is for illustration only, and the analyses presented here should not be used for decisionmaking. Current prenatal screening programs do not rely on isolated markers, but combine sequential biochemical and ultrasound testing. Further, prenatal screening programs also aim to detect trisomy 13 (Patau syndrome), trisomy 18 (Edwards syndrome) and neural tube defects.[10] In addition, new screening methods sequence cell-free fetal DNA circulating in maternal plasma and are summarized by Palomaki et al.[12]

Table 1 and Table 2 show the extracted data. Table 1 shows 11 studies that reported counts of the results of each test.[13-23] Five studies reported sufficient information to calculate counts of combinations of test results among infants with trisomy 21 and among healthy infants,[13-16,20] 1 study reported sufficient information to calculate counts of combinations of test results in trisomy 21 infants only,[23] and the remaining 5 reported only separate counts for the two tests.[17-19,21,22] Three studies measured humerus in a subset of healthy infants with femur measurements.[17-19] Table 2 shows 19 additional studies (reported in 18 publications[24-41]) that measured shortened femur only. No studies measured shortened humerus only.

Table 1. Cross-classification of counts of shortened femur or shortened humerus among infants with trisomy 21 and healthy infants (studies reporting both tests)

Study Year [reference]	Design	Positive Test (ratio of observed to expected length)	Trisomy 21							Healthy						
			N	H+	F+	H+/F+	H+/F-	H-/F+	H-/F-	N	H+	F+	H+/F+	H+/F-	H-/F+	H-/F-
Benacerraf 1991[13]	Case-control	H: <0.90; F: <0.91	24	12	10	9	3	1	11	400	25	40	19	6	21	354
Benacerraf 1992[14]	Case-control	H: <0.90; F: <0.91	32	17	23	17	0	6	9	588	34	63	23	11	40	514
Benacerraf 1994[15]	Case-control	H: <0.90; F: <0.91	45	20	22	20	0	2	23	106	3	4	1	2	3	100
Biagiotti 1994[16]	Case-control	H: <0.90; F: <0.91	27	10	13	10	0	3	14	500	60	60	31	29	29	411
Bromley 1997[17]	Case-control	H: <0.90; F: ≤0.91	53	19	25	–	–	–	–	177 (149)*	5	14	–	–	–	–
Johnson 1995[18]	Case-control	H: ≤0.90; F: ≤0.90	36	8	15	–	–	–	–	794 (486) †	25	127	–	–	–	–
Lockwood 1993[19]	Prospective	[Observed minus expected, mm] H <−3.6; F <−3.4	42	6	6	–	–	–	–	4874 (2775)‡	111	161	–	–	–	–
Nyberg 1993[20]	Case-control	H: ≤0.89; F: ≤0.91	45	11	11	8	3	3	31	942	42	44	15	27	29	871
Nyberg 1998[21]	Case-control	H: ≤0.89; F: ≤0.91	142	27	30	–	–	–	–	930	11	43	–	–	–	–
Rodis 1991[22]	Case-control	[Percentile of measurement] H <5th; F <5th	11	7	2	–	–	–	–	1470	74	74	–	–	–	–
Vintzileos 1996[23]	Prospective	H: <0.89; F: <0.88	22	10	5	4	6	1	11	493	49	50	–	–	–	–

F[+|-] = shortened femur [present | absent]; H[+|-] = shortened humerus [present | absent]; N = total (per disease category)

* Total 177 for femur, and 149 for humerus. All those tested for humerus were also tested for femur.

† Total 794 for femur, and 486 for humerus. All those tested for humerus were also tested for femur.

‡ Total 4874 for femur, and 2775 for humerus. All those tested for humerus were also tested for femur.

Table 2. Number of counts of shortened femur among infants with trisomy 21 and healthy infants (studies of shortened femur only)

Study year [reference]	Design	Positive test (ratio of observed to expected length)	Trisomy 21		Healthy	
			N	F+	N	F+
Benacerraf 198928	Case-control	<0.91	20	7	3480 (709)*	28
Brumfield 198929	Case-control	[BPD/FL≥1.80]	15	6	45	1
Campbell 199424	Prospective	[BPD/FL≥1.5 SD in controls]	6	3	264	20
Cuckle 198926	Case-control	≤0.90	83	20	1340	84
Dicke 198927	Case-control	<0.91	33	5	177	18
Ginsberg 199025	Case-control	[BPD/FL>1.5 SD in controls]	12 (11)†	5	212	14
Grandjean 199530	Prospective	<0.91	34	15	2763	495
Grist 199032	Prospective	≤0.90	6	3	428	28
Hill 198934	Case-control	≤0.91	22	11	286	43
Johnson 199331	Prospective	[FL/Foot length≤0.90]	14	10	331	31
LaFollette 198933	Case-control	≤0.91	30	4	229	27
Lockwood 1987 (New Haven)35	Case-control	[BPD/FL>1.5 SD in controls]	35	18	349	26
Lockwood 1987 (Boston)35	Case-control	[BPD/FL>1.5 SD in controls]	20	14	195	9
Lynch 198941	Case-control	[BPD/FL>1.5 SD in controls]	9	5	9	5
Marquette 199038	Case-control	[BPD/FL>1.5 SD in controls]	31	3	155	14
Nyberg 199039	Case-control	≤0.91	49	7	572	35
Nyberg 199537	Prospective	≤0.91	18	5	232	14
Shah 199040	Case-control	[BPD/FL; threshold not stated]	17	3	17	1
Verdin 199836	Case-control	[>97.5 percentile of BPD/FL in controls]	11	6		

BPD = biparietal diameter; F+ = shortened femur; FL = observed femoral length; SD = standard deviation.
*3,480 were included, 709 were analyzed for shortened femur.
†12 were included, 11 were analyzed for shortened femur.

Models and Estimation

Over several decades, many methods have been proposed for meta-analyzing data on performance of medical tests. The most theoretically motivated methodologies respect the multivariate nature of performance metrics, allow the true-positive rate (TPR, sensitivity) and false-positive rate (FPR, one minus specificity) to vary together across studies (because of a threshold effect or other reasons), and allow for between-study heterogeneity. The meta-analysis methods by Reitsma et al.[8] (further augmented elsewhere[42,43]) and by Rutter and Gatsonis[9] meet these desiderata. Reitsma's method summarizes data as a "summary point", that is, a summary TPR and a summary FPR. Rutter's approach summarizes data by a "summary line" (equivalent to the hierarchical summary receiver operating characteristic curve when transformed into [ROC] space), which describes how the average TPR changes with the average FPR. Choosing the most helpful summary is largely subjective and application-dependent, and in some situations the two summaries provide meaningful and complementary information.[44]

We focus on meta-analyses in which a summary point can be considered a helpful summary of test performance. We briefly review the Reitsma et al. random effects bivariate model and its augmentations[42,43] for the meta-analysis of TPR and FPR of a single test, and describe its extension to two and three or more tests. We use M to denote the number of tests. In the following, we first describe models using the normal approximation to within-study variance, as historically these were developed first, and then present models that more accurately capture the discrete nature of the events (binomial distributions for one test and multinomial distributions for more than one test).[45,46]

The Case of a Single Test

Consider a meta-analysis of K studies (indexed by k) evaluating a single test among individuals with and without a condition of interest ("disease"). Table 3 shows the true (unobserved, population) probabilities (denoted by π's) of positive and negative test results in those with disease (D) and those without disease ($\bar{D}$) in study k. By definition, the TPR of the test is equal to $\pi_{k,1}^{D}$, and the FPR of the test is equal to $\pi_{k,1}^{\bar{D}}$.

Table 3. Notation for the probability of results for a single test in those with (D) and without disease ($\bar{D}$) in study k

Test Results	Disease	No Disease
Negative (–)	$\pi_{k,0}^{D}$	$\pi_{k,0}^{\bar{D}}$
Positive (+)	$\pi_{k,1}^{D} = TPR_k$	$\pi_{k,1}^{\bar{D}} = FPR_k$

Structural Model

We assume a bivariate random effects model.[8] One can approximate the between-study variance using the normal distribution on logit-transformed data (approximate likelihood).[a] We write the logistic transformation of TPR and FPR in study *k*:

[a] One can also use the normal approximation with the arcsine or probit transformation, or with untransformed data. Reitsma et al. used the canonical (logistic or logit) transformation, and this is the choice we follow here.

$$\eta_k = \text{logit}(TPR_k) = \text{logit}(\pi^D_{k,1}) \quad (1)$$

$$\xi_k = \text{logit}(FPR_k) = \text{logit}(\pi^{\bar{D}}_{k,1}) \quad (2)$$

Across studies, the random effects for the true logit-transformed TPR and FPR are likely correlated, e.g., because of a threshold effect. We model this dependency using a bivariate normal distribution:

$$\begin{pmatrix} \eta_k \\ \xi_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} H \\ \Xi \end{pmatrix}, \mathbf{T} \right). \quad (3)$$

The means H and Ξ are the overall meta-analytic summaries of the logit-TPR and logit-FPR, respectively, and $\mathbf{T}$ is the between-study covariance matrix:

$$\mathbf{T} = \begin{pmatrix} \tau^2_\eta & \rho_{\eta\xi}\tau_\eta\tau_\xi \\ & \tau^2_\xi \end{pmatrix}, \quad (4)$$

where the variances τ^2_η and τ^2_ξ represent the between-study heterogeneity in logit-transformed TPR and FPR, respectively, and $\rho_{\eta\xi}$ is the corresponding between-study correlation.

Observational Model: Normal (Approximate) Likelihood

In study k we obtain sample estimates of η_k and ξ_k based on the observed counts of test results in those with and without disease.

Table 4 gives the notation for the cross-classification of the results of a medical test in patients with and without disease in a single study. The p's in

Table 4 estimate the corresponding probabilities (π's) in Table 3. For example, the estimate of the probability of a true positive result in study k is $\hat{\pi}^D_{k,1} = x^D_{k,1} \,/\, N^D_k$, and correspondingly for the other probabilities. We estimate TPR by $\widehat{TPR}_k = \hat{\pi}^D_{k,1}$, and FPR by $\widehat{FPR}_k = \hat{\pi}^{\bar{D}}_{k,1}$. Finally, the logit-transformed TPR and FPR in (1) and (2) are estimated by

$$\hat{\eta}_k = \text{logit}(\hat{\pi}^D_{k,1}) \quad (5)$$

and

$$\hat{\xi}_k = \text{logit}(\hat{\pi}^{\bar{D}}_{k,1}). \quad (6)$$

Table 4. Notation for observed counts and estimated probabilities for a single test in those with disease (D) and without disease ($\bar{D}$) in study k

Test Results	Disease, Counts	Disease, Estimated Probabilities	No Disease, Counts	No Disease, Estimated Probabilities
Negative (–)	$x^D_{k,0}$	$\hat{\pi}^D_{k,0}$	$x^{\bar{D}}_{k,0}$	$\hat{\pi}^{\bar{D}}_{k,0}$
Positive (+)	$x^D_{k,1}$	$\hat{\pi}^D_{k,1} = \widehat{TPR}_k$	$x^{\bar{D}}_{k,1}$	$\hat{\pi}^{\bar{D}}_{k,1} = \widehat{FPR}_k$
Total	N^D_k	1	$N^{\bar{D}}_k$	1

As mentioned above, TPR and FPR are proportions, so the binomial distribution is most suitable for modeling their within-study behavior (see below). However, it has not been uncommon to approximate the within-study distribution of the logit transformed proportions using normal distributions. Because TPR and FPR refer to disjoint sets of patients, $\hat{\eta}_k$ and $\hat{\xi}_k$ are independent conditional on study *k*:

$$\hat{\eta}_k \sim \mathcal{N}\left(\eta_k,\ \sigma^2_{k\eta}\right), \text{ and} \tag{7}$$

$$\hat{\xi}_k \sim \mathcal{N}\left(\xi_k,\ \sigma^2_{k\xi}\right). \tag{8}$$

To facilitate later descriptions, we rewrite (7) and (8) as a single bivariate normal distribution

$$\begin{pmatrix} \hat{\eta}_k \\ \hat{\xi}_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \eta_k \\ \xi_k \end{pmatrix},\ \boldsymbol{\Sigma}_k\right), \tag{9}$$

where the within-study covariance matrix $\boldsymbol{\Sigma}_k$ has zero off-diagonal elements:

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma^2_{k\eta} & 0 \\ & \sigma^2_{k\xi} \end{pmatrix}. \tag{10}$$

The elements of $\boldsymbol{\Sigma}_k$ are considered known and are calculated from the data using the formulas in the Appendix. (This assumption is typically made, but without formal justification.)

Observational Model—Binomial Likelihood

In place of (7) and (8) –or equivalently, (9) and (10)– we can use two independent binomials to model counts of test results in study k:

$$x^D_{k,1} \sim \text{Bin}\left(TPR_k,\ N^D_k\right) \tag{11}$$

$$x^{\bar{D}}_{k,1} \sim \text{Bin}\left(FPR_k,\ N^{\bar{D}}_k\right), \tag{12}$$

while retaining the structural model specified by (3) and (4).

The Case of Two Tests

Consider a meta-analysis of two tests (indexed by m) and data from K studies. The outcome of each test is either negative (–) or positive (+). Table 5 gives the notation for the probability of each combination of test results in diseased patients and nondiseased participants in study k. Each of the $2^M = 2^2 = 4$ rows shows the corresponding proportion in the population.

Table 5. Notation for the probability of each combination of results in those with disease (D) and without disease ($\bar{D}$) in study k – the case of two tests

Test 1	Test 2	Disease	No Disease
–	–	$\pi^D_{k,00}$	$\pi^{\bar{D}}_{k,00}$
–	+	$\pi^D_{k,01}$	$\pi^{\bar{D}}_{k,01}$
+	–	$\pi^D_{k,10}$	$\pi^{\bar{D}}_{k,10}$
+	+	$\pi^D_{k,11}$	$\pi^{\bar{D}}_{k,11}$

The marginal TPR and FPR in each test are:

$$TPR_{k,1} = \pi^{D}_{k,1\bullet} = \pi^{D}_{k,10} + \pi^{D}_{k,11}, \quad (13)$$

$$TPR_{k,2} = \pi^{D}_{k,\bullet 1} = \pi^{D}_{k,01} + \pi^{D}_{k,11}, \quad (14)$$

$$FPR_{k,1} = \pi^{\bar{D}}_{k,1\bullet} = \pi^{\bar{D}}_{k,10} + \pi^{\bar{D}}_{k,11}, \quad (15)$$

and

$$FPR_{k,2} = \pi^{\bar{D}}_{k,\bullet 1} = \pi^{\bar{D}}_{k,01} + \pi^{\bar{D}}_{k,11}. \quad (16)$$

In the equations, $\bullet$ indicates summation over the respective subscript. In addition, we define the "jointly true positive rate" (JTPR), and "jointly false positive rate" (JFPR):

$$JTPR_k = \pi^{D}_{k,11} \quad (17)$$

and

$$JFPR_k = \pi^{\bar{D}}_{k,11}. \quad (18)$$

Structural Model: TPR and FPR Parameterization

The sets of probabilities in the Disease and No Disease columns of Table 5 are each mutually exclusive and exhaustive and could be modeled with a multinomial distribution. Alternatively, we could re-express these parameters as functions of the sensitivities and specificities of each test and approximate the multinomial distribution using a multivariate normal distribution. As in the previous section, we work with logit-transformed probabilities. For study k and test m equations (1) and (2) become:

$$\eta_{km} = \text{logit}(TPR_{k,m}) \quad (19)$$

$$\xi_{km} = \text{logit}(FPR_{k,m}). \quad (20)$$

We also model the logit-JTPR, η_{k^*}, and logit-JFPR, ξ_{k^*}, which capture information on the agreement between the two tests in those with and without disease, respectively:

$$\eta_{k^*} = \text{logit}(\pi^{D}_{k,11}), \text{ and} \quad (21)$$

$$\xi_{k^*} = \text{logit}(\pi^{\bar{D}}_{k,11}). \quad (22)$$

By arranging quantities for the two tests in column vectors, we write

$$\boldsymbol{\eta}_k = (\eta_{k1}, \eta_{k2}, \eta_{k^*})', \text{ and}$$

$$\boldsymbol{\xi}_k = (\xi_{k1}, \xi_{k2}, \xi_{k^*})'.$$

with the prime (') denoting transpose, bold symbols denoting vectors or matrices, and italics denoting scalars.

Across studies, the joint distribution of the random effects for the true logit-transformed TPRs, FPRs, JTPR and JFPR may involve correlation. We model this dependency using a six-dimensional normal distribution analogous to (3):

$$\begin{pmatrix} \eta_{k1} \\ \eta_{k2} \\ \hline \eta_{k*} \\ \hline \xi_{k1} \\ \xi_{k2} \\ \hline \xi_{k*} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_k \\ \boldsymbol{\xi}_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{H} \\ \boldsymbol{\Xi} \end{pmatrix}, \mathbf{T} \right), \tag{23}$$

where the means $\mathbf{H} = (H_1, H_2, H_*)'$ and $\boldsymbol{\Xi} = (\Xi_1, \Xi_2, \Xi_*)'$ are column vectors of the overall means of the respective quantities for the two tests, and $\mathbf{T}$ is a between-study covariance matrix, analogous to that in (4).

We propose an unstructured and a simplified structured specification for $\mathbf{T}$:

Unstructured variant, $\mathbf{T}_A$

$$\mathbf{T}_A = \left[\begin{array}{cc:c:cc:c} \tau^2_{\eta 1} & \rho_{\eta 1\eta 2}\tau_{\eta 1}\tau_{\eta 2} & \rho_{\eta 1\eta *}\tau_{\eta 1}\tau_{\eta *} & \rho_{\eta 1\xi 1}\tau_{\eta 1}\tau_{\xi 1} & \rho_{\eta 1\xi 2}\tau_{\eta 1}\tau_{\xi 2} & \rho_{\eta 1\xi *}\tau_{\eta 1}\tau_{\xi *} \\ & \tau^2_{\eta 2} & \rho_{\eta 2\eta *}\tau_{\eta 2}\tau_{\eta *} & \rho_{\eta 2\xi 1}\tau_{\eta 2}\tau_{\xi 1} & \rho_{\eta 2\xi 2}\tau_{\eta 2}\tau_{\xi 2} & \rho_{\eta 2\xi *}\tau_{\eta 2}\tau_{\xi *} \\ \hdashline & & \tau^2_{\eta *} & \rho_{\eta *\xi 1}\tau_{\eta *}\tau_{\xi 1} & \rho_{\eta *\xi 2}\tau_{\eta *}\tau_{\xi 2} & \rho_{\eta *\xi *}\tau_{\eta *}\tau_{\xi *} \\ \hline & & & \tau^2_{\xi 1} & \rho_{\xi 1\xi 2}\tau_{\xi 1}\tau_{\xi 2} & \rho_{\xi 1\xi *}\tau_{\xi 1}\tau_{\xi *} \\ & & & & \tau^2_{\xi 2} & \rho_{\xi 2\xi *}\tau_{\xi 2}\tau_{\xi *} \\ \hdashline & & & & & \tau^2_{\xi *} \end{array}\right] \tag{24}$$

In (24), $\mathbf{T}_A$ has 21 parameters that have to be estimated.

Structured variant, $\mathbf{T}_B$

For the case of two tests, we can impose structure (and reduce the number of parameters to be estimated to 13) by the following requirements:

a. All logit-TPRs have the same variance, τ^2_η
b. All logit-FPRs have the same variance, τ^2_ξ
c. The logit-JTPRs have the same variance, $\tau^2_{\eta *}$
d. The logit-JFPRs have the same variance, $\tau^2_{\xi *}$
e. The correlation between logit-TPRs in different tests is $\rho_{\eta\eta}$
f. The correlation between logit-FPRs in different tests is $\rho_{\xi\xi}$
g. The correlation between logit-JTPR and logit-JFPR is $\rho_{\eta *\xi *}$
h. The correlation between the logit-TPR and logit-FPR in the same test is $\rho_{\eta\xi}$
i. The correlation between the logit-TPR and logit-FPR in different tests is $\varrho_{\eta\xi}$
j. The correlation between the logit-TPR for each test and the logit-JTPR is $\rho_{\eta\eta *}$
k. The correlation between the logit-FPR for each test and the logit-JFPR is $\rho_{\xi\xi *}$
l. The correlation between the logit-TPR for each test and the logit-JFPR is $\rho_{\eta\xi *}$

m. The correlation between the logit-FPR for each test and the logit-JTPR is $\rho_{\eta^*\xi}$

$$\mathbf{T}_B = \left[\begin{array}{cc|c|cc|c} \tau_\eta^2 & \rho_{\eta\eta}\tau_\eta^2 & \rho_{\eta\eta^*}\tau_\eta\tau_{\eta^*} & \rho_{\eta\xi}\tau_\eta\tau_\xi & \varrho_{\eta\xi}\tau_\eta\tau_\xi & \rho_{\eta\xi^*}\tau_\eta\tau_{\xi^*} \\ & \tau_\eta^2 & \rho_{\eta\eta^*}\tau_\eta\tau_{\eta^*} & \varrho_{\eta\xi}\tau_\eta\tau_\xi & \rho_{\eta\xi}\tau_\eta\tau_\xi & \rho_{\eta\xi^*}\tau_\eta\tau_{\xi^*} \\ \hline & & \tau_{\eta^*}^2 & \rho_{\eta^*\xi}\tau_{\eta^*}\tau_\xi & \rho_{\eta^*\xi}\tau_{\eta^*}\tau_\xi & \rho_{\eta^*\xi^*}\tau_{\eta^*}\tau_{\xi^*} \\ \hline & & & \tau_\xi^2 & \rho_{\xi\xi}\tau_\xi^2 & \rho_{\xi\xi^*}\tau_\xi\tau_{\xi^*} \\ & & & & \tau_\xi^2 & \rho_{\xi\xi^*}\tau_\xi\tau_{\xi^*} \\ \hline & & & & & \tau_{\xi^*}^2 \end{array}\right]. \qquad (25)$$

Structural Model: Probability Parameterization

Another, perhaps more direct approach, models the probabilities $\boldsymbol{\pi}_k^D = (\pi_{k,00}^D, \pi_{k,01}^D, \pi_{k,10}^D, \pi_{k,11}^D)$ and $\boldsymbol{\pi}_k^{\bar{D}} = (\pi_{k,00}^{\bar{D}}, \pi_{k,01}^{\bar{D}}, \pi_{k,10}^{\bar{D}}, \pi_{k,11}^{\bar{D}})$ directly, rather than re-expressed in terms of sensitivity, specificity and joint positivity. One could work with the logits of these probabilities and reformulate the multivariate normal model in terms of them or one could proceed to model the probabilities themselves in the form of a Dirichlet distribution. The resulting parameter estimates could then be re-expressed in terms of TPRs and FPRs, if desired, either directly through transformations of MCMC simulations from Bayesian analyses or using the delta method with likelihood analyses. We leave this for future work.

Observational Model: Multivariate Normal (Approximate) Likelihood

We calculate sample estimates for the quantities in the structural model using the observed counts in Table 6. For each combination of test results, the estimate of the population proportion is

$$\hat{\pi}_k^D = x_k^D / N_k^D \qquad (26)$$

(we have suppressed the indices corresponding to the rows in Table 6); N_k^D is the total number of those with disease. (For notational simplicity we assume that both tests are applied to all patients, and thus $N_{k,1}^D = N_{k,2}^D = N_k^D$.) The corresponding notation for those without disease replaces D with $\bar{D}$. We estimate the sensitivity and specificity of test m in study k by the replacing true probabilities (π's) with the respective sample estimates ($\hat{\pi}$'s) in (13) through (16); and the logit-transformed TPR, FPR, JTPR and JFPR by

$$\hat{\eta}_{km} = \text{logit}(\widehat{TPR}_{k,m}), \qquad (27)$$

$$\hat{\xi}_{km} = \text{logit}(\widehat{FPR}_{k,m}), \qquad (28)$$

$$\hat{\eta}^* = \text{logit}(\hat{\pi}_{k,11}^D), \text{ and} \qquad (29)$$

$$\hat{\xi}_{k^*} = \text{logit}(\hat{\pi}_{k,11}^{\bar{D}}). \qquad (30)$$

Table 6. Notation for observed counts and estimated probabilities in those with disease (D) and without disease ($\bar{D}$) in study k – the case of two tests

Test 1	Test 2	Disease, Counts	Disease, Estimated Probabilities	No Disease, Counts	No Disease, Estimated Probabilities
–	–	$x_{k,00}^{D}$	$\hat{\pi}_{k,00}^{D}$	$x_{k,00}^{\bar{D}}$	$\hat{\pi}_{k,00}^{\bar{D}}$
–	+	$x_{k,01}^{D}$	$\hat{\pi}_{k,01}^{D}$	$x_{k,01}^{\bar{D}}$	$\hat{\pi}_{k,01}^{\bar{D}}$
+	–	$x_{k,10}^{D}$	$\hat{\pi}_{k,10}^{D}$	$x_{k,10}^{\bar{D}}$	$\hat{\pi}_{k,10}^{\bar{D}}$
+	+	$x_{k,11}^{D}$	$\hat{\pi}_{k,11}^{D}$	$x_{k,11}^{\bar{D}}$	$\hat{\pi}_{k,11}^{\bar{D}}$
Totals		N_k^D	1	$N_k^{\bar{D}}$	1

We assume that both tests are applied to all patients, and thus $N_{k,1}^{D} = N_{k,2}^{D} = N_k^{D}$.

We model within-study variability in study k using a multivariate normal distribution in a manner analogous to (9). More explicitly,

$$\begin{pmatrix} \hat{\eta}_{k1} \\ \hat{\eta}_{k2} \\ \hline \hat{\eta}_{k*} \\ \hline \hat{\xi}_{k1} \\ \hat{\xi}_{k2} \\ \hline \hat{\xi}_{k*} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_k \\ \hat{\boldsymbol{\xi}}_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\eta}_k \\ \boldsymbol{\xi}_k \end{pmatrix}, \boldsymbol{\Sigma}_k \right). \tag{31}$$

The within-study covariance matrices $\boldsymbol{\Sigma}_k$ in (31) include the within-study correlations between the logit-transformed TPRs, FPRs, JTPR and JFPR. The elements of $\hat{\boldsymbol{\eta}}_k$ are correlated because they pertain to the same patients (those with disease). The elements of $\hat{\boldsymbol{\xi}}_k$ are correlated because they are calculated in the same patients (those without disease). However, $\hat{\boldsymbol{\eta}}_k$ is not correlated with $\hat{\boldsymbol{\xi}}_k$ because they are calculated in disjoint samples. Thus, $\boldsymbol{\Sigma}_k$ is a block-diagonal matrix:

$$\boldsymbol{\Sigma}_k = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{k\eta} & \mathbf{0} \\ \hline & \boldsymbol{\Sigma}_{k\xi} \end{array} \right] = \left[\begin{array}{cc:c|ccc} \sigma_{k\eta 1}^2 & \sigma_{k\eta 12} & \sigma_{k\eta 1*} & & & \\ & \sigma_{k\eta 2}^2 & \sigma_{k\eta 2*} & & \mathbf{0} & \\ \hdashline & & \sigma_{k\eta **}^2 & & & \\ \hline & & & \sigma_{k\xi 1}^2 & \sigma_{k\xi 12} & \sigma_{k\xi 1*} \\ & \mathbf{0} & & & \sigma_{k\xi 2}^2 & \sigma_{k\xi 2*} \\ \hdashline & & & & & \sigma_{k\xi **}^2 \end{array} \right], \tag{32}$$

where $\boldsymbol{\Sigma}_{k\eta}$ and $\boldsymbol{\Sigma}_{k\xi}$ are the 3×3 covariance matrices of the logit-transformed metrics in those with disease and those without disease, respectively, and $\mathbf{0}$ a 3×3 matrix of zeros. The

elements of $\mathbf{\Sigma}_k$ are considered known and are calculated from the data using the formulas in the Appendix. The analogy between (9) and (31), and between (10) and (32) is clear.

If the tests have not been applied in the same patients (i.e., the comparative studies used a parallel design) (32) becomes a diagonal matrix, i.e., a matrix with all off-diagonal elements equal to zero.

Observational Model: Multinomial Likelihood

Alternatively, one can use the multinomial distribution to model the cross-classification of the results of two or more tests. Specifically, the column vectors of counts $\mathbf{x}^D = (x^D_{k0\cdots0},\ldots,x^D_{k1\cdots1})'$ and $\mathbf{x}^{\bar{D}} = (x^{\bar{D}}_{k0\cdots0},\ldots,x^{\bar{D}}_{k1\cdots1})'$ follow multinomial distributions:

$$\mathbf{x}^D \sim \mathcal{M}\left(N^D_k,\ \boldsymbol{\pi}^D_k\right), \tag{33}$$

and

$$\mathbf{x}^{\bar{D}} \sim \mathcal{M}\left(N^{\bar{D}}_k,\ \boldsymbol{\pi}^{\bar{D}}_k\right). \tag{34}$$

The estimates of the true probabilities $\boldsymbol{\pi}^D_k = (\pi^D_{k0\cdots0},\ \ldots,\ \pi^D_{k1\cdots1})'$ and $\boldsymbol{\pi}^{\bar{D}}_k = (\pi^{\bar{D}}_{k0\cdots0},\ \ldots,\ \pi^{\bar{D}}_{k1\cdots1})'$ are $\hat{\pi}^D$ and $\hat{\pi}^{\bar{D}}$, respectively, with elements defined in (26) (and its counterpart for those without disease).

As above, if the tests have not been applied in the same patients (i.e., the comparative studies used a parallel design) (33) and (34) become sets of independent binomials.

The Case of Three or More Tests

One can extend the models to $M > 2$ tests that have been applied in the same patients, but some care is required to ensure that the resulting model has enough parameters to represent the probabilities of all the possible outcomes in cross-classified data. The number of parameters is not the only challenge, as we discuss below.

Number of Parameters

Because the number of parameters needed to fully specify the test results among both diseased and nondiseased individuals is $2^M - 1$, the total number of parameters for a fully specified model increases very rapidly. We need:

- $2^M - 1$ parameters to model the probabilities (or functions of probabilities) in those with disease
- $2^M - 1$ parameters in those without the disease
- $(2^{M+1} - 2)(2^{M+1} - 1)/2$ parameters to model between-study variances and covariances, assuming the unstructured covariance matrix $\mathbf{T}_A$ in (24), for a total of $(2^{M+1} - 2) + [(2^{M+1} - 2)(2^{M+1} - 1)/2]$ parameters.

As per Sections 3.1 and 3.2, one needs 5 and 27 parameters for one and two tests, respectively. For three and four tests one needs 119 and 495 parameters, respectively! Numerical difficulties arise both in unstructured parameterizations because of the large number of parameters and in structured parameterizations in which it may be difficult to enforce positive definiteness of the covariance matrices during optimization.

As mentioned already, several parameterizations of the probabilities are possible. One model uses logits of the $2^M - 1$ probabilities of individual test combinations in both diseased and nondiseased individuals. Another uses the M logit-TPRs together with the logits of other suitable functions of the individual probabilities, as we did for the case of two tests. For illustration, when $M = 3$, $2^M - 1 = 7$, and the additional four functions could be the probability that Test 1 and Test 2 are positive, the probability that Test 1 and Test 3 are positive, the probability that Test 2 and Test 3 are positive, and the probability that all three tests are positive. (In a particular application, the choice of such functions should take into account the likely probabilities, so as to minimize difficulties with small probabilities and small cell counts.)

For the case of three or more tests, the practical consequences of such representations are unclear. In practice, it is unlikely that a systematic review would find completely cross-classified data on more than two tests.

Handling the observational model is less complex presuming the multinomial likelihood is used. Within each group of diseased or nondiseased individuals, the counts of combinations of positive and negative test results on all the tests forms a cross-classification that can be represented by a multinomial distribution. The approximate normal likelihood becomes too complex when $M>3$.

Estimation and Inference

Separate (one test at a time) and joint meta-analysis models using the normal approximation can be fit using (restricted) maximum likelihood.

Separate meta-analyses models that use the binomial distribution can be fit in the generalized linear mixed models framework using routines readily available in general statistical packages such as xtmelogit in Stata or lmer in R. However, the joint meta-analysis models using the multinomial likelihood cannot be fit in these general routines. The available generalized linear mixed model (GLMM) packages in R, Stata and SAS do not allow the user to specify the random effects distribution in (24), where the random effects pertain to sums of the probabilities in Table 5. Optimizing the likelihood for joint meta-analysis using the multinomial likelihood outside a GLMM package is nontrivial, because it involves calculating complicated integrals numerically. Thus we did not develop routines for fitting this model. Instead we fitted the model using Markov Chain Monte Carlo (MCMC) methods in the Bayesian framework, as described later in this section.

Maximum Likelihood Estimation (Model Using the Normal Approximation)

To fit the normal approximation model, optimize the log likelihood

$$LogL = \frac{1}{2}\sum_{k=1}^{K}\left(\log(|\mathbf{W}_k|) - \mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k\right), \tag{35}$$

where $\mathbf{W}_k = (\boldsymbol{\Sigma}_k + \mathbf{T})^{-1}$ and $\mathbf{D}_k = \begin{pmatrix} \hat{\boldsymbol{\eta}}_k - \mathbf{H} \\ \hat{\boldsymbol{\xi}}_k - \boldsymbol{\Xi} \end{pmatrix}$; $|\mathbf{W}_k|$ denotes the determinant of $\mathbf{W}_k$. The parameters to be estimated are the summary effects $\mathbf{H}$ and $\boldsymbol{\Xi}$ and the elements of the between-study covariance matrix $\mathbf{T}$. Alternatively, one can optimize the restricted likelihood, which was the approach we used in the applied example:

$$LogL^* = \frac{1}{2}\sum_{k=1}^{K}\left(\log(|\mathbf{W}_k|) - \mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k\right) + \frac{1}{2}\log\left(\left|\sum_{k=1}^{K}\mathbf{W}_k\right|\right) \tag{36}$$

As mentioned previously, it is typical meta-analytic practice to consider the elements of $\boldsymbol{\Sigma}_k$ known, but calculate them from the data. Appendix A provides formulas for these calculations. The matrix equations for the log likelihood remain the same for bivariate meta-analysis of one test and for the joint meta-analysis of two or more tests.

By optimizing (35) or (36) we obtain the (restricted) maximum likelihood estimators $\hat{\mathbf{H}}$, $\hat{\boldsymbol{\Xi}}$ and $\hat{\mathbf{T}}$. We also obtain the $(2^{M+1}-2)\times(2^{M+1}-2)$ estimated covariance matrix $\mathbf{C} = (c_{ij})$ of the $(\hat{\mathbf{H}}', \hat{\boldsymbol{\Xi}}')'$ as the inverse Hessian matrix.

Confidence Intervals

Confidence Intervals for the Summary Estimates H_m and Ξ_m

Confidence intervals for summary estimates are obtained in a similar manner for bivariate analyses of one test and for joint meta-analyses of two or more tests. Therefore, the formulas below are for *M* tests.

The $100(1-\alpha)\%$ simultaneous confidence interval (usually a 95% confidence interval) for H_m (the summary logit-TPR in test *m*) is given by:

$$\left(\hat{H}_m - q_\alpha\sqrt{c_{mm}},\ \hat{H}_m + q_\alpha\sqrt{c_{mm}}\right), \tag{37}$$

where c_{mm} is the variance of $\hat{H}_m$, and q_α is the square root of the $100(1-\alpha)$ percentile of the chi-squared distribution with $2^{M+1}-2$ degrees of freedom. This simultaneous confidence interval is a special case of Scheffé's F-projections for multiple comparisons; it controls type I error for the family of all possible linear combinations of the estimated parameters.[47] The simultaneous confidence interval for Ξ_m (the summary logit-FPR in test *m*) is given by:

$$\left(\hat{\Xi}_m - q_\alpha\sqrt{c_{m+2^M-1,\,m+2^M-1}},\ \hat{\Xi}_m + q_\alpha\sqrt{c_{m+2^M-1,\,m+2^M-1}}\right), \tag{38}$$

where $c_{m+2^M-1,\,m+2^M-1}$ is the variance of $\hat{\Xi}_m$.

Confidence Intervals for Differences $H_i - H_j$ and $\Xi_i - \Xi_j$ Between Summary Estimates of Two Tests

For two tests that have been applied to the same patients, one can either perform a meta-analysis for Test 1 and a separate one for Test 2, or a joint meta-analysis for the two tests. In either case, one can compare the diagnostic accuracy of the tests by calculating the difference between the logit-TPRs $H_1 - H_2$ and the difference between the logit-FPRs $\Xi_1 - \Xi_2$. The confidence intervals for such differences are calculated in different ways for separate versus joint meta-analyses of the two tests.

Confidence Intervals for Differences Based on Separate Meta-Analyses Per Test

Separate bivariate meta-analyses of the two tests ignore within-study correlations and treat the summary estimates of the two tests as independent. The resulting asymptotic confidence interval for the difference in logit TPR of tests i and j is

$$\left(\hat{H}_i - \hat{H}_j - z_{\alpha/2}\sqrt{\mathrm{var}(\hat{H}_i) + \mathrm{var}(\hat{H}_j)},\ \hat{H}_i - \hat{H}_j + z_{\alpha/2}\sqrt{\mathrm{var}(\hat{H}_i) + \mathrm{var}(\hat{H}_j)}\right), \tag{39}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. Because the confidence intervals in (39) ignore within-study correlations, their coverage differs from the nominal $100(1-\alpha)\%$. Bonferroni's inequality offers a simple adjustment to control the type I error. One substitutes $z_{\alpha/(2f)}$ for $z_{\alpha/2}$ in (39), where f is the number of comparisons of interest. It may be reasonable to consider $f = 2M + M(M-1)$, which equals the number of estimated mean logit-TPRs and mean logit-FPRs plus the total number of pairwise differences among the

mean logit-TPRs and plus the total number of pairwise differences among the mean logit-FPRs. (The above considers all other modeled quantities, such as the logit-JTPR and the logit-JFPR, as nuisance parameters that are not of interest.)

Confidence Intervals for Differences Based on Joint Meta-Analyses of All Tests

For joint multivariate meta-analyses of all tests, differences and simultaneous confidence intervals are obtained as follows. For convenience, write $\boldsymbol{\beta} = \begin{pmatrix} \mathbf{H} \\ \boldsymbol{\Xi} \end{pmatrix}$; that is, arrange the true summary logit-transformed quantities in a column vector. For a vector $\mathbf{a} = (a_1, \ \ldots \ , a_{2(2^M-1)})'$ let $L(\mathbf{a}, \boldsymbol{\beta}) = \mathbf{a}'\boldsymbol{\beta}$ be a linear combination of the true summaries, and $L(\mathbf{a}, \hat{\boldsymbol{\beta}}) = \mathbf{a}' \begin{pmatrix} \hat{\mathbf{H}} \\ \hat{\boldsymbol{\Xi}} \end{pmatrix}$ its estimate. Then $100(1-\alpha)\%$ simultaneous confidence intervals for all possible linear combinations are given by

$$\left(L(\mathbf{a}, \hat{\boldsymbol{\beta}}) - q_\alpha \sqrt{(\mathbf{a}'\mathbf{C}\mathbf{a})} \ , \ L(\mathbf{a}, \hat{\boldsymbol{\beta}}) + q_\alpha \sqrt{(\mathbf{a}'\mathbf{C}\mathbf{a})} \right). \tag{40}$$

In particular, to estimate differences between the summary logit-TPRs of tests i and j, set $a_i = 1$, $a_j = -1$, and all other elements of $\mathbf{a}$ to 0. Then $L(\mathbf{a}, \hat{\boldsymbol{\beta}}) = \hat{H}_i - \hat{H}_j$, and the confidence interval in (40) becomes

$$\left(\hat{H}_i - \hat{H}_j - q_\alpha \sqrt{c_{ii} + c_{jj} - 2c_{ij}} \ , \ \hat{H}_i - \hat{H}_j + q_\alpha \sqrt{c_{ii} + c_{jj} - 2c_{ij}} \right). \tag{41}$$

In an analogous manner, to estimate differences between summary logit-FPRs for tests i and j, set $a_{i+2^M-1} = 1$, $a_{j+2^M-1} = -1$, and all other elements of $\mathbf{a}$ to 0, and proceed as in (40) to obtain

$$\begin{aligned} &(\hat{\Xi}_i - \hat{\Xi}_j - q_\alpha \sqrt{c_{i+2^M-1, i+2^M-1} + c_{j+2^M-1, j+2^M-1} - 2c_{i+2^M-1, j+2^M-1}} \ , \\ &\hat{\Xi}_i - \hat{\Xi}_j + q_\alpha \sqrt{c_{i+2^M-1, i+2^M-1} + c_{j+2^M-1, j+2^M-1} - 2c_{i+2^M-1, j+2^M-1}}). \end{aligned} \tag{42}$$

MCMC Estimation and Credible Intervals for Models Using Discrete Likelihoods

We fit models using the binomial and multinomial distributions at the within-study level with MCMC methods. To this end, and in addition to equations in the Models and Estimation chapter, we specified vague prior distributions for the following modeled parameters.

The true means were assigned independent vague normal priors:

$$\begin{pmatrix} \mathbf{H} \\ \boldsymbol{\Xi} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \ 10^6 \cdot \mathbf{I}_6 \right),$$

where $\mathbf{I}_6$ is the 6×6 identity matrix.

To assign priors for the covariance matrix $\mathbf{T}$ we use the factorization $\mathbf{T} = \text{diag}(\boldsymbol{\tau}) \, \mathbf{R} \, \text{diag}(\boldsymbol{\tau})$ where $\text{diag}(\boldsymbol{\tau})$ is the diagonal matrix whose diagonal elements are the square roots of the variances of the η_k and ξ_k and $\mathbf{R}$ is the correlation matrix corresponding to the covariances of

the η_k and ξ_k. We assign independent uniform priors to the elements of $\boldsymbol{\tau}$, i.e., the standard deviations of the random effects):

$$\tau_m \sim \mathcal{U}(10^{-4}, 5)$$

The priors for $\mathbf{R}$ must guarantee that the matrix is positive definite with elements between -1 and 1. We follow Lu and Ades[48] in factorizing $\mathbf{R}$ using the Cholesky decomposition for square symmetric matrices $\mathbf{R} = \mathbf{L}\mathbf{L}'$, and in assigning specially constructed priors to the elements of the lower triangular matrix $\mathbf{L}$ (this is the spherical parameterization of Pinheiro and Bates[49]):

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \cos(\phi_{21}) & \sin(\phi_{21}) & 0 & 0 & 0 & 0 \\ \cos(\phi_{31}) & \cos(\phi_{32})\cdot\sin(\phi_{31}) & \sin(\phi_{32})\cdot\sin(\phi_{31}) & 0 & 0 & 0 \\ \cos(\phi_{41}) & \cos(\phi_{42})\cdot\sin(\phi_{41}) & \cos(\phi_{43})\cdot\sin(\phi_{42})\cdot\sin(\phi_{41}) & \sin(\phi_{43})\sin(\phi_{42})\cdot\sin(\phi_{41}) & 0 & 0 \\ \cos(\phi_{51}) & \cos(\phi_{52})\cdot\sin(\phi_{51}) & \cos(\phi_{53})\cdot\sin(\phi_{52})\cdot\sin(\phi_{51}) & \cos(\phi_{54})\sin(\phi_{53})\cdot\sin(\phi_{52})\sin(\phi_{41}) & \sin(\phi_{54})\sin(\phi_{53})\cdot\sin(\phi_{52})\sin(\phi_{41}) & 0 \\ \cos(\phi_{61}) & \cos(\phi_{62})\cdot\sin(\phi_{61}) & \cos(\phi_{63})\cdot\sin(\phi_{62})\cdot\sin(\phi_{61}) & \cos(\phi_{64})\sin(\phi_{63})\cdot\sin(\phi_{62})\sin(\phi_{64}) & \cos(\phi_{65})\sin(\phi_{64})\cdot\sin(\phi_{63})\sin(\phi_{62})\cdot\sin(\phi_{61}) & \sin(\phi_{65})\sin(\phi_{64})\cdot\sin(\phi_{63})\sin(\phi_{62})\cdot\sin(\phi_{61}) \end{bmatrix}$$

Setting uniform independent priors for ϕ's in the interval 0 to $\pi = 3.14159...$ yields a prior for $\mathbf{R}$ in which all elements are between -1 and 1 and positive definiteness is guaranteed

$$\phi_{ij} \sim \mathcal{U}(0, \pi) \ .$$

See Lu and Ades for a short discussion on the density of the elements of $\mathbf{R}$ using the priors above.[48] See Pinheiro and Bates for a discussion of additional parameterizations.[49]

95% Credible Intervals

With MCMC it is straighfrorward to obtain credible intervals for any quantity or any function of quantities explicitly, by simulation. In particular, we used 95% central credible intervals as the 2.5 and 97.5 percentile of the MCMC simulations.

Software and Computation

For the normal approximation models, the log likelihood in (35) and (36) for the unstructured variant of the $\mathbf{T}$ matrix can be optimized using routines such as mvmeta in Stata. We have developed our own Stata routines to optimize both the structured and the unstructured variant of $\mathbf{T}$. (*mvmeta* uses a simple imputation of zero point estimates and large variances or covariances to simplify programming when handling studies with missing data; our routines do not need such imputations.) For convergence, starting values from fixed effect meta-analysis estimates appear to suffice. Note however, that the routine for the structured covariance matrix is not as robust: it failed to converge in the dataset used in this example (but does converge in other datasets). The optimization uses a modified Newton-Raphson algorithm. The routines are available from the authors upon request (see also www.cebm.brown.edu).

We ran MCMC analyses using JAGS version 3.1.0 through the R package rjags. We used three chains with a burn-in of at least 100,000 iterations and between 100,000 and 800,000 iterations for recording results. We monitored convergence with the Gelman-Rubin diagnostic for stochastic nodes corresponding to the meta-analysis means and the elements of their between-study covariance matrices. We declared convergence when the 97.5 percentile of the diagnostic was 1.10 or less for all monitored stochastic nodes, and provided that on visual inspection the traceplots of the MCMC chains were suggestive of good mixing.

Incomplete or Missing Data

In practice, not all studies in a meta-analysis report counts on all combinations of test results in diseased and nondiseased participants. In the example, even with only two tests, the level of detail varies substantially. For some studies we can construct the full cross-classification for both TPR and FPR, for some we can construct the cross-classification for TPR but not FPR or vice versa; some studies report only the estimated TPRs and FPRs of the tests, and others report an intermediate level of detail. In principle, it is desirable to incorporate all available suitable evidence, perhaps including results from studies that evaluated only a single test. And if the total number of tests to be analyzed is greater than two, it is likely that many studies will have evaluated only two of the tests. We can describe many patterns of missingness, but we comment on the following three:

Information To Extract Cross-Classification Counts Is Not Reported

Often studies report information on the TPR or the FPR of each test, but do not provide data to reconstruct the cross-classification of test results. Assuming that data are missing at random, one could proceed with imputation of the missing within-study correlations between the quantities of interest, using information from the observed correlations from studies with complete data. In performing the imputation, one must ensure the positive definiteness of the covariance matrix. Another option imputes the missing counts directly. This is the approach taken in the analysis of the example in this work.

A Subset of Studies Reports Only on Diseased or Only on Nondiseased Individuals

If information is missing at random, one obtains unbiased estimates by optimizing the appropriate likelihood omitting contributions from the missing subsets.

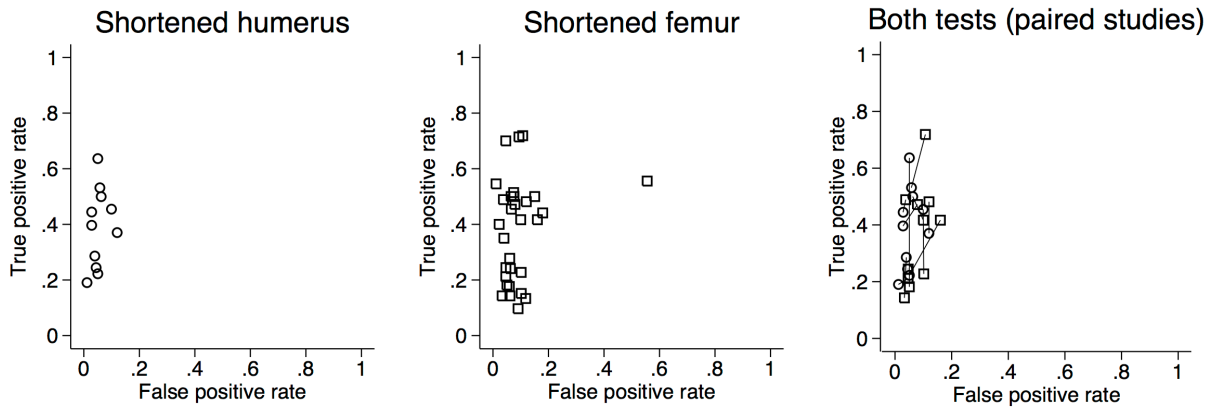
Some Studies Do Not Report Results From Each Test

If information is missing at random, one again optimizes the likelihood of the observed data omitting the missing tests.

Analysis of the Example

As is typical for screening tests, individual studies have chosen thresholds to attain primarily low FPR (ranging from less than 1% to 18% for both tests, except for a single study with shortened femur with an FPR of 55%), with TPRs that range from 10 percent to 72 percent for both tests (Figure 1).

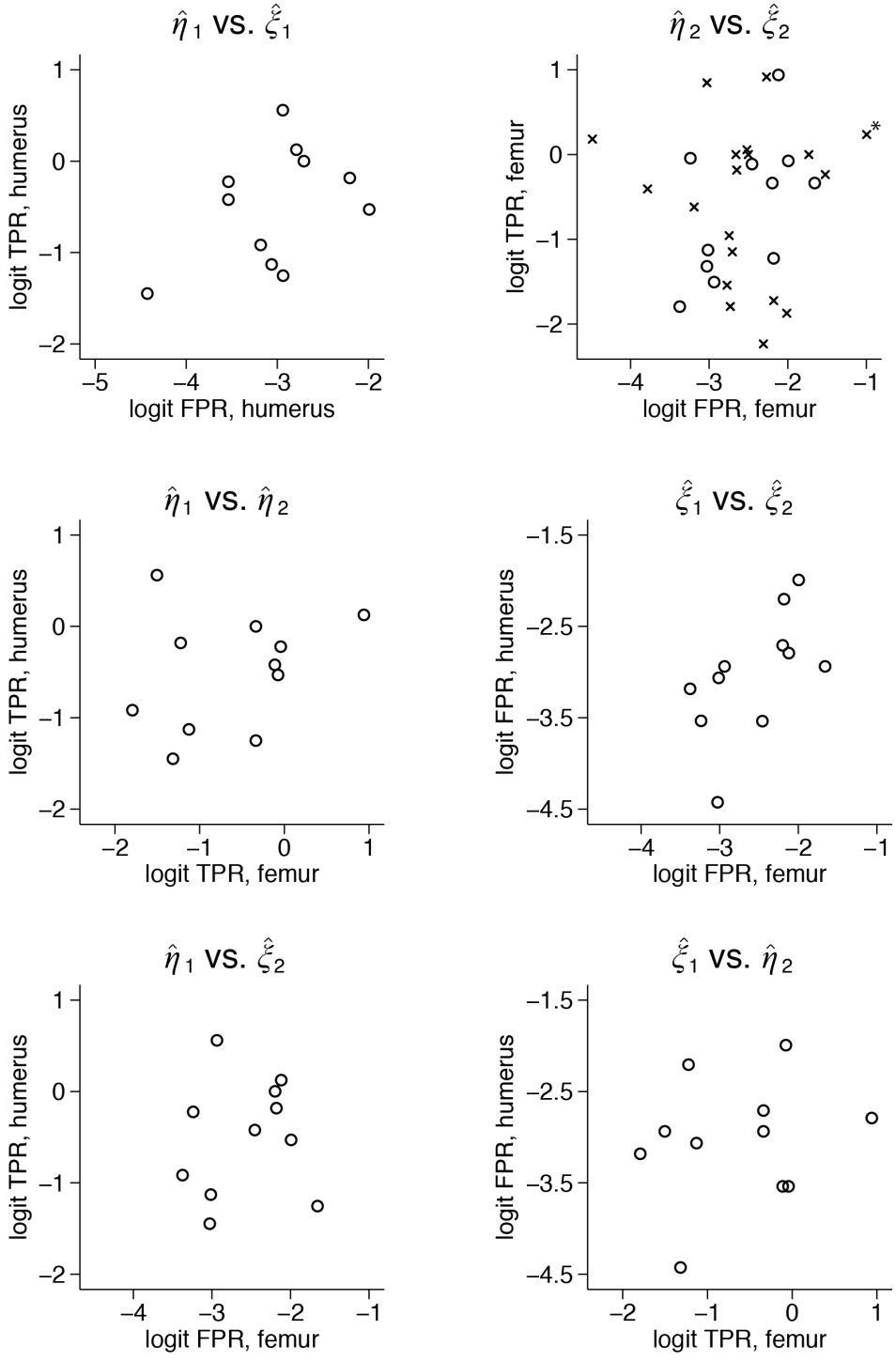
Figure 1. Observed sensitivities and false-positive rates for the two tests in the example



Note: Markers denote studies (circles: shortened humerus; squares: shortened femur). In the rightmost plot, lines connect test results in the same patients (from paired studies).

Figure 2 shows scatter plots of logit-transformed TPR and FPR for the studies in the example. Most plots give the impression of positive correlation between the estimated quantities. For descriptive purposes we calculated empirical Spearman correlation coefficients, which corroborate the visual impressions Table 7). This informal exploration lends support to the notion that the joint multivariate meta-analysis of the two tests will take into account information that separate bivariate meta-analyses would ignore.

Figure 2. Scatter plots of $\hat{\eta}$'s and $\hat{\xi}$'s in the example



Note: Circles denote paired test studies. In the $\hat{\eta}_2$ vs. $\hat{\xi}_2$ plot, "x" marks denote studies of shortened femur only. The asterisk ("*") marks a study where the FPR is 55 percent (logit-FPR is 0.20) but has been relocated to lie in the plotted region. The range of the axes in the scatter plots is approximately 3 units in the logit scale, and the aspect ratio has been kept to 1 to facilitate visual assessments. FPR: false-positive rate (one minus specificity); TPR: true-positive rate (sensitivity).

Table 7. Empirical correlations between $\hat{\eta}$'s and $\hat{\xi}$'s in the example

	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_*$	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\xi}_*$
$\hat{\eta}_1$	1	0.22	0.57	0.41	0.15	0.31
$\hat{\eta}_2$		1	0.89	0.23	0.46 [0.16]*	0.10
$\hat{\eta}_*$			1	0.39	0.45	0.26
$\hat{\xi}_1$				1	0.74	0.93
$\hat{\xi}_2$					1	0.79
$\hat{\xi}_*$						1

*0.46 for the 11 paired studies; 0.16 for all 30 studies of shortened femur.

Note: Figure 2 shows scatterplots between $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{\xi}_1$ and $\hat{\xi}_2$.

Subscript 1 refers to shortened humerus and subscript 2 to shortened femur. Subscript * refers to JTPR or JFPR.

Estimates of Diagnostic Accuracy

Table 8 shows the results of the meta-analyses. The upper part of the table (nonshaded rows a through f) shows results in the 11 paired studies. The lower part (shaded rows a through f) shows the corresponding results including the additional 19 studies on shortened femur. The table includes results from separate meta-analyses for each test using the normal approximation (row a) and the binomial likelihood (row b), and the following joint meta-analyses: using the normal approximation and accounting for within study correlation (row c); using the normal approximation and ignoring within-study correlation (i.e., setting the off-diagonal elements of $\mathbf{\Sigma}_k$ to zero; row d); using the multinomial likelihood (which models within study correlation; row e); and using independent binomials for each test in both diseased and nondiseased individuals (i.e., a discrete likelihood but ignoring the within-study correlations; row f).

In general, the analyses give similar point estimates for the true-positive rates and similar point estimates for the false-positive rates. The 95% credible intervals in Bayesian models using discrete likelihoods (rows b, e, f) tend to be longer than the respective 95% confidence intervals in non-Bayesian models using the normal approximation. Overall, the lengths of the confidence or credible intervals in Table 8 differ only slightly between the separate and joint meta-analyses. To facilitate comparisons of the length of the confidence intervals between separate meta-analyses of each test and joint analyses, we report 95% confidence intervals that are not corrected for multiple comparisons. Specifically, we used $z_{0.025} = 1.96$ instead of q_α in (37), (38), (40), (41) and (42).

Including all 30 studies does not change the point estimates and confidence/credible intervals for shortened humerus appreciably. It results in lower uncertainty for shortened femur, as all 19 additional studies were on shortened femur.

Table 8. Point estimates and individual 95% confidence intervals with alternative meta-analysis methods

	Summary TPR (%) Short humerus	Summary TPR (%) Short femur	Summary JTPR (%)	Summary FPR (%) Short humerus	Summary FPR (%) Short femur	Summary JFPR (%)
Analyses in the 11 paired studies						
Separate meta-analyses of the two tests						
a. Bivariate, normal likelihood	36.9 (29.3, 45.1)	36.0 (26.0, 47.3)	NA	5.0 (3.5, 7.1)	7.6 (5.5, 10.3)	NA
b. Bivariate, binomial likelihood	37.9 (27.7, 50.3)	35.4 (23.1, 49.5)	NA	4.8 (3.0, 7.4)	7.4 (5.0, 10.7)	NA
Joint meta-analyses of the two tests						
c. Normal likelihood, using within-study correlation	37.3 (29.9, 45.5)	37.5 (27.9, 48.1)	30.9 (23.0, 40.2)	4.9 (3.4, 7.0)	7.6 (5.6, 10.2)	3.0 (2.2, 4.1)
d. Normal likelihood, ignoring within-study correlation	37.5 (29.4, 46.5)	36.4 (26.1, 48.0)	27.5 (19.5, 37.2)	4.9 (3.4, 7.0)	7.3 (5.3, 10.0)	2.8 (2.0, 4.0)
e. Multinomial likelihood (uses within-study correlations)	35.3 (26.9, 41.8)	35.0 (22.4, 46.2)	26.1 (16.6, 34.0)	4.9 (2.8, 7.5)	7.3 (4.6, 10.5)	2.7 (1.6, 4.2)
f. Binomial likelihood (ignores within-study correlation)	34.6 (20.3, 44.2)	35.9 (20.5, 50.4)	26.8 (11.3, 39.2)	4.8 (2.9, 7.7)	7.3 (4.6, 11.5)	2.8 (1.7, 4.4)
Analyses in all 30 studies						
Separate meta-analyses of the two tests						
a. Bivariate, normal likelihood	36.9 (29.3, 45.1)	35.8 (29.1, 43.1)	NA	5.0 (3.5, 7.1)	7.6 (6.1, 9.4)	NA
b. Bivariate, binomial likelihood	38.0 (27.8, 50.3)	35.2 (27.8, 43.3)	NA	4.8 (3.0, 7.4)	7.4 (5.8, 9.3)	NA
Joint meta-analyses of the two tests						
c. Normal likelihood, using within-study correlation	34.2 (27.2, 42.0)	37.1 (30.6, 44.2)	29.8 (23.1, 37.5)	4.9 (3.5, 6.7)	7.6 (6.2, 9.4)	3.0 (2.3, 3.8)
d. Normal likelihood, ignoring within-study correlation	35.9 (28.9, 43.6)	35.8 (29.0, 43.3)	26.7 (20.4, 34.1)	5.0 (3.6, 7.0)	7.5 (6.1, 9.3)	2.9 (2.2, 3.9)
e. Multinomial likelihood (uses within-study correlations)	34.7 (27.9, 42.5)	34.8 (28.7, 41.9)	25.2 (18.9, 32.4)	4.9 (3.1, 7.4)	7.4 (6.0, 9.2)	2.8 (1.9, 3.9)
f. Binomial likelihood (ignores within-study correlation)	35.9 (27.7, 45.1)	34.9 (28.0, 42.7)	27.5 (19.2, 37.2)	4.9 (3.1, 7.4)	7.4 (5.8, 9.3)	2.8 (1.8, 4.1)

FPR = false-positive rate (1–specificity); REML = restricted maximum likelihood; TPR = true-positive rate (sensitivity); JFPR/JTPR = joint-[false|true]-positive rate. The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24).

Table 9 shows the standard errors and the standard deviations of the posteriors of the summary estimates in the logit scale for maximum likelihood and Bayesian analyses, respectively. Meta-analyses that model the within-study correlations between tests generally gives more precise estimates. Standard errors based on the Bayesian models with discrete likelihoods are larger than those from the non-Bayesian normal likelihoods. When all 30 studies are considered, differences in the standard errors or standard deviations between separate and joint meta-analyses are attenuated primarily for shortened femur, to which the additional data pertain.

In sum, for estimating summaries of TPRs and FPRs, the payoff of joint meta-analyses versus separate meta-analyses is at best modest.

Table 9. Standard errors or posterior standard deviations of logit-transformed summary effects with alternative meta-analysis methods

	SE/Posterior SD			
	logit-TPR (shortened humerus)	**logit-TPR (shortened femur)**	**logit-FPR (shortened humerus)**	**logit-FPR (shortened femur)**
Analyses in the 11 paired studies				
Separate meta-analyses of the two tests				
a. Bivariate, normal likelihood	0.1745	0.2397	0.1934	0.1714
b. Bivariate, binomial likelihood	0.2430*	0.2957*	0.2399*	0.2072*
Joint meta-analyses of the two tests				
c. Normal likelihood, using within-study correlation	0.1715	0.2230	0.1947	0.1646
d. Normal likelihood, ignoring within-study correlation	0.1871	0.2446	0.1910	0.1753
e. Multinomial likelihood (uses within-study correlations)	0.1686*	0.2691*	0.2514*	0.2134*
f. Binomial likelihood (ignores within-study correlation)	0.2581*	0.3405*	0.2659*	0.2456*
Analyses in all 30 studies				
Separate meta-analyses of the two tests				
a. Bivariate, normal likelihood	0.1745	0.1560	0.1934	0.1173
b. Bivariate, binomial likelihood	0.2423*	0.1730*	0.2397*	0.1306*
Joint meta-analyses of the two tests				
c. Normal likelihood, using within-study correlation	0.1692	0.1492	0.1720	0.1162
d. Normal likelihood, ignoring within-study correlation	0.1646	0.1594	0.1764	0.1187
e. Multinomial likelihood (uses within-study correlations)	0.1658*	0.1473*	0.2257*	0.1195*
f. Binomial likelihood (ignores within-study correlation)	0.1919*	0.1651*	0.2308*	0.1257*

FPR = false-positive rate (1–specificity); SE = standard error; TPR = true-positive rate (sensitivity). The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24).
*Standard deviation of the posterior.

Estimates of Comparative Diagnostic Accuracy

We can compare the TPRs and FPRs of shortened humerus and shortened femur using meta-analysis. A convenient metric is the difference between the summary TPRs or the summary FPRs of the tests. Another metric is the difference in the summary logit-TPRs (and the same for the logit-FPRs). The exponential of the latter is an odds ratio: it expresses how many times higher the odds of a true positive (or false positive) are when shortened humerus is the test, versus when shortened femur is the test. For both metrics, a positive difference in, e.g., the TPR favors shortened humerus, in that its average TPR is higher than that of shortened femur. A difference of zero favors neither test, and a negative difference favors shortened femur. For FPR the direction is reversed: a negative difference favors shortened humerus, and so on.

Separate Meta-Analyses Versus Joint Analyses Accounting for Within-Study Correlations

Table 10 shows differences in the absolute scale, and Table 11 shows relative differences (differences in the logit scale and odds ratios). In each table, the point estimates are generally similar across analyses. This is congruent with Table 8, where the point estimates for the summary TPRs and FPRs were similar across analyses. However, the confidence or credible intervals are shorter for the joint meta-analyses compared with the corresponding separate meta-analyses. For example, consider the analyses using the discrete likelihood. Among the 11 paired studies, the length of the confidence interval for the difference in TPRs in row e (joint meta-analyses accounting for within-study correlations) is approximately half of that in row b (separate meta-analyses; Table 10). The corresponding confidence intervals for FPRs are approximately 20 percent shorter. It so happens that in analyses using binomial and multinomial likelihoods, the shortening of the credible intervals is more evident for the difference in TPRs and logit-TPRs. In analyses using the normal approximations, the gain (in terms of shortening confidence interval lengths) is maximal for the difference in FPRs and logit-FPRs.

The sensitivities do not differ beyond what is expected by chance. The corresponding odds ratios in are close to one, between 1.03 and 1.13. All analyses suggest that shortened femur has smaller summary false-positive rate than shortened humerus, corresponding to odds ratios of 0.64 or 0.65. The differences in the summary false-positive rates are statistically significant in the joint meta-analyses (albeit uncorrected for multiple comparisons).

Separate Meta-Analyses Versus Joint Analyses Not Accounting for Within-Study Correlations

As shown in Table 10 and Table 11, there is gain in terms of shorter confidence or credible interval lengths from joint analyses compared with separate meta-analyses, even if the within-study correlations are ignored (assumed to be zero; compare rows a and d for analyses with normal approximations; and rows b and f for analyses using the discrete likelihood). However, the gain is less than that obtained when within-study correlations are accounted for.

Finally, when all 30 studies are included, the differences between separate meta-analyses, joint meta-analyses accounting for correlation and joint meta-analyses not accounting for correlation remain in the same direction, but are attenuated.

Table 10. Comparative test performance: Differences in the summary TPRs or FPRs and corresponding standard errors or standard deviations of the posterior distributions

	Difference in TPR	post. SD_{Diff} or SE_{Diff}	Difference in FPR	post. SD_{Diff} or SE_{Diff}
Analyses in the 11 paired studies				
Separate meta-analyses of the two tests				
a. Bivariate, normal likelihood	0.9 (-12.5, 14.3)	6.85	-2.6 (-5.5, 0.4)	1.51
b. Bivariate, binomial likelihood	2.6 (-14.7, 19.8)	8.70*	-2.5 (-6.3, 1.1)	1.83*
Joint meta-analyses of the two tests				
c. Normal likelihood, using within-study correlation	-0.2 (-8.9, 8.6)	4.45	***-2.8 (-4.7, -0.8)***	1.00
d. Normal likelihood, ignoring within-study correlation	1.2 (-9.6, 12.0)	5.51	***-2.4 (-4.4, -0.4)***	1.02
e. Multinomial likelihood (uses within-study correlations)	0.0 (-8.9, 9.5)	4.62*	-2.5 (-5.4, 0.3)	1.43*
f. Binomial likelihood (ignores within-study correlation)	-1.4 (-10.2, 8.8)	4.81*	-2.5 (-6.4, 0.5)	1.68*
Analyses in all 30 studies				
Separate meta-analyses of the two tests				
a. Bivariate, normal likelihood	1.0 (-9.6, 11.7)	5.42	***-2.6 (-5.0, -0.2)***	1.24
b. Bivariate, binomial likelihood	2.9 (-10.1, 17.0)	6.87*	-2.5 (-5.2, 0.5)	1.44*
Joint meta-analyses of the two tests				
c. Normal likelihood, using within-study correlation	-2.9 (-8.5, 2.7)	2.86	***-2.7 (-4.6, -0.9)***	0.94
d. Normal likelihood, ignoring within-study correlation	0.1 (-8.3, 8.4)	4.26	***-2.5 (-4.2, -0.8)***	0.89
e. Multinomial likelihood (uses within-study correlations)	-0.2 (-8.9, 8.8)	4.50*	-2.5 (-4.9, 0.2)	1.29*
f. Binomial likelihood (ignores within-study correlation)	0.9 (-9.2, 11.6)	5.29*	-2.5 (-4.9, 0.2)	1.30*

FPR = false-positive rate (1–specificity); SE = standard error; TPR = true-positive rate (sensitivity). The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24). For models fit with REML (rows a, c and d) the differences were calculated using the delta method on the fitted model parameters.

*Standard deviation of the posterior.

Bold italic font: Statistically significant difference ($p<0.05$) or posterior probability >0.975 that the difference is less than 0.

Table 11. Comparative test performance: Differences in the summary logit-TPRs or logit-FPRs and corresponding standard errors or standard deviations of the posterior distributions

Analysis	TPR			FPR		
	Difference in logits	post. SD_{Diff} or SE_{Diff}	Odds ratio	Difference in logits	post. SD_{Diff} or SE_{Diff}	Odds ratio
Analyses in the 11 paired studies						
Separate meta-analyses of the two tests						
a. Bivariate, normal likelihood	0.038 (-0.543, 0.619)	0.2965	1.04 (0.58, 1.86)	-0.440 (-0.947, 0.066)	0.2584	0.64 (0.39, 1.07)
b. Bivariate, binomial likelihood	0.112 (-0.632, 0.889)	0.3827*	1.12 (0.53, 2.43)	-0.452 (-1.086, 0.168)	0.3166*	0.64 (0.34, 1.18)
Joint meta-analyses of the two tests						
c. Normal likelihood, using within-study correlation	-0.006 (-0.309, 0.296)	0.1544	0.99 (0.73, 1.34)	***-0.480 (-0.810, -0.150)***	0.1684	***0.62 (0.44, 0.86)***
d. Normal likelihood, ignoring within-study correlation	0.051 (-0.374, 0.476)	0.2169	1.05 (0.69, 1.61)	***-0.427 (-0.761, -0.093)***	0.1704	***0.65 (0.47, 0.91)***
e. Multinomial likelihood (uses within-study correlations)	0.000 (-0.374, 0.447)	0.2087*	1.00 (0.69, 1.56)	-0.443 (-0.956, 0.057)	0.2533*	0.64 (0.38, 1.06)
f. Binomial likelihood (ignores within-study correlation)	-0.061 (-0.432, 0.408)	0.2161*	0.94 (0.65, 1.50)	-0.444 (-1.058, 0.094)	0.2807*	0.64 (0.35, 1.10)
Analyses in all 30 studies						
Separate meta-analyses of the two tests						
a. Bivariate, normal likelihood	0.045 (-0.414, 0.504)	0.2341	1.05 (0.66, 1.66)	***-0.446 (-0.889, -0.003***)	0.2261	***0.64 (0.41, 1.00)***
b. Bivariate, binomial likelihood	0.124 (-0.452, 0.724)	0.2976*	1.13 (0.64, 2.06)	-0.451 (-1.006, 0.076)	0.2730*	0.64 (0.37, 1.08)
Joint meta-analyses of the two tests						
c. Normal likelihood, using within-study correlation	-0.127 (-0.316, 0.062)	0.0965	0.88 (0.73, 1.06)	***-0.474 (-0.810, -0.138)***	0.1733	***0.62 (0.44, 0.87)***
d. Normal likelihood, ignoring within-study correlation	0.004 (-0.345, 0.354)	0.1783	1.00 (0.71, 1.42)	***-0.432 (-0.751, -0.112)***	0.1631	***0.65 (0.47, 0.89)***
e. Multinomial likelihood (uses within-study correlations)	-0.007 (-0.398, 0.385)	0.1989*	0.99 (0.67, 1.47)	-0.446 (-0.939, 0.026)	0.2415*	0.64 (0.39, 1.03)
f. Binomial likelihood (ignores within-study correlation)	0.042 (-0.411, 0.505)	0.2323*	1.04 (0.66, 1.66)	-0.444 (-0.953, 0.028)	0.2468*	0.64 (0.39, 1.03)

FPR = false-positive rate (1–specificity); SE = standard error; TPR = true-positive rate (sensitivity). The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24). *Standard deviation of the posterior.

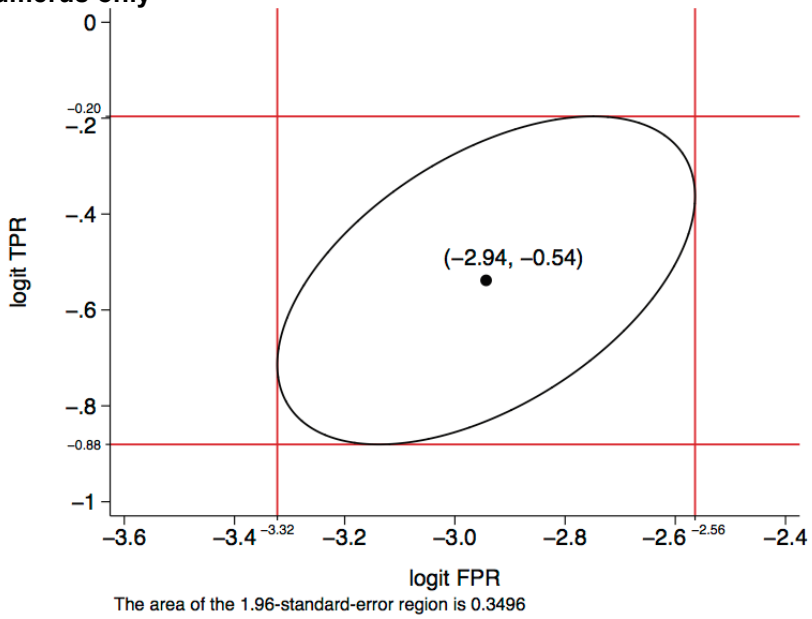
Bold italic font: Statistically significant difference ($p<0.05$) or posterior probability >0.975 that the difference is less than 0.

Four-Dimensional 1.96-Standard-Error Volumes for Separate and Joint Meta-Analyses

In the example, accounting for within-study correlation has at best moderate impact on the length of the confidence intervals for meta-analysis means, but can result in tighter confidence intervals for differences of means (for assessing comparative test accuracy). These findings are not discordant. The explanation involves the concept of a "confidence volume," which generalizes the confidence interval in more than one dimension. The same concept applies to credible intervals in Bayesian analyses in an analogous manner.

To introduce the concept, focus on the meta-analysis of the sensitivity and specificity of a single test, say shortened humerus. The bivariate meta-analysis model yields estimates of means of logit-transformed sensitivities and false-positive rates and their covariance. The geometric representation of the simultaneous confidence interval for the estimated means is a two-dimensional ellipse (Figure 3), whose shape is determined by the estimated covariance, and whose center is determined by the estimated means. The confidence intervals of the estimated means are projections of the elliptical contour corresponding to 1.96 standard errors (95% confidence ellipse). A fundamental metric is the area of the 95% confidence ellipse, in that it characterizes the distribution of the estimated means according to the model. The area can be calculated analytically (see Appendix).

Figure 3. 1.96-standard-error confidence region (ellipse) for a bivariate meta-analysis of shortened humerus only

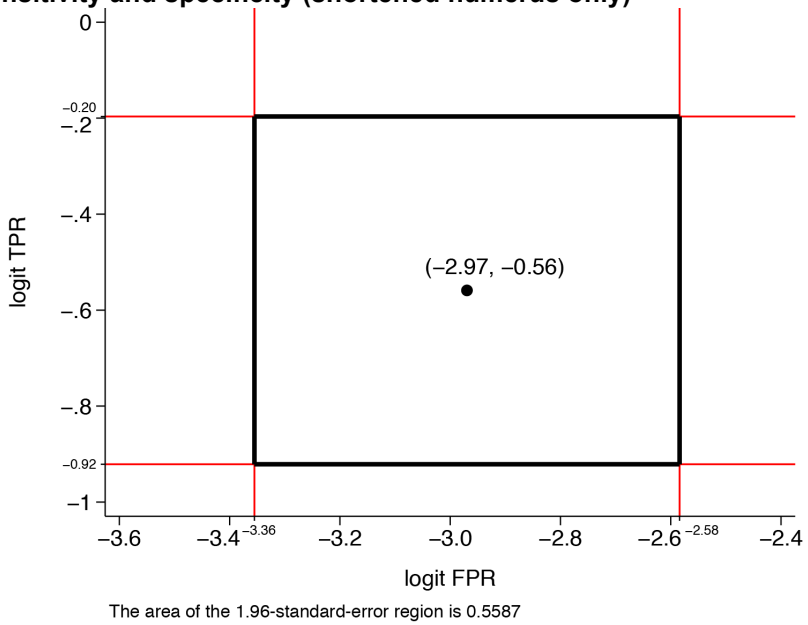


FPR = false-positive rate (1-specificity); TPR = true-positive rate (sensitivity).
Note: The plot is on the logit scale. The black dot represents the summary point, i.e., the mean logit true-positive and false-positive rates for the 11 meta-analyzed studies of shortened humerus. The red horizontal lines represent the 95% confidence interval on the margin of the estimated mean logit true-positive rate, and the red vertical lines represent the corresponding 95% confidence intervals of the estimated mean logit false-positive rate. The ellipse is the 1.96-standard-error region of the joint distribution of the two estimated means.

By contrast, separate univariate meta-analyses would effectively consider the estimated means as independent; the geometric representation of the estimated means with separate univariate meta-analyses is a rectangle (Figure 4). The 1.96-standard-error region of the rectangle is larger than that of the ellipse in Figure 3. (The type I error rate for the rectangle

confidence region in Figure 4 is $1-0.95^2=0.0975$, so the rectangle does not have 95% coverage.) The coordinates of the summary point in Figure 4 come from separate meta-analyses, and are slightly different from Figure 3.

Figure 4. 1.96-standard-error confidence region (rectangle) for independent meta-analyses of sensitivity and specificity (shortened humerus only)



FPR = false-positive rate (1-specificity); TPR = true-positive rate (sensitivity).
Note: The plot is on the logit scale. The black dot represents the coordinates of the mean of logit-transformed sensitivity and false-positive rate for shortened humerus. The red horizontal lines represent the 95% confidence interval on the margin of the summary logit true-positive rate, and the red vertical lines represent the corresponding 95% confidence intervals of the summary logit false-positive rate. The rectangle is the 1.96-standard-error confidence region of the joint distribution of the two means, which have been obtained from separate univariate random effects meta-analyses with REML. The coordinates of the summary point in this figure come from separate meta-analyses, and are slightly different from Figure 3.

The concept of the confidence region extends to more dimensions. When one considers both tests, the summary logit sensitivities and false-positive rates have a joint distribution over four dimensions. Table 12 shows the calculated four-dimensional 1.96-standard-error confidence volumes for the alternative meta-analyses performed here. The confidence volume for joint meta-analyses is much smaller than that from separate meta-analyses for the two tests. The reason is that the joint analyses incorporate more information in the form of within-study correlations.

When considering estimated means and the confidence intervals around them, we observe no great differences between separate meta-analyses of the two tests and joint analyses, in the same way as we see no great differences between the confidence intervals in Figure 3 and those in Figure 4. However, when we calculate differences of means we choose a projection that results in appreciable differences.

Table 12. Four-dimensional volumes within the 1.96-standard error hull for the summary estimates in alternative analyses (normal approximation modeling only)

	Volume (×10-3)	Ranking (small to large)
Analyses in the 11 paired studies		
Separate meta-analyses of the two tests		
a. Bivariate, normal likelihood	13.4	3
b. Bivariate, binomial likelihood	Not calculated	–
Joint meta-analyses of the two tests		
c. Normal likelihood, using within-study correlation	4.9	1
d. Normal likelihood, ignoring within-study correlation	12.0	2
e. Multinomial likelihood (uses within-study correlations)	Not calculated	–
f. Binomial likelihood (ignores within-study correlation)	Not calculated	–
Analyses in all 30 studies		
Separate meta-analyses of the two tests		
a. Bivariate, normal likelihood	7.6	3
b. Bivariate, binomial likelihood	Not calculated	–
Joint meta-analyses of the two tests		
c. Normal likelihood, using within-study correlation	1.8	1
d. Normal likelihood, ignoring within-study correlation	5.7	2
e. Multinomial likelihood (uses within-study correlations)	Not calculated	–
f. Binomial likelihood (ignores within-study correlation)	Not calculated	–

Note: The volumes do not have units and are on the logit scale. The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24).

Estimates of Between-Study Variance and Comparison of Structural Variants of T

Table 13 shows the estimated between-study variance from analyses using the normal approximation, and Table 14 shows posterior medians from analyses using discrete likelihoods. For ease of comparison, we have arranged results from separate meta-analyses in the same format as for results from joint meta-analyses.

All between-study correlations are almost always positive, and appear to be mostly congruent with each other. One should not overinterpret the differences in these estimates across models, because variance components are often not well estimated.

Table 13. Estimates of between-study standard errors and correlations (analyses of the 11 paired studies)

Model	Normal approximation, estimates of $\boldsymbol{\tau}', \mathbf{R}$	Discrete likelihood, estimates of $\boldsymbol{\tau}', \mathbf{R}$
Separate meta-analyses	$\begin{bmatrix} 0.448 \\ 0.696 \\ - \\ 0.590 \\ 0.536 \\ - \end{bmatrix}$, $\begin{bmatrix} 1 & 0 & - & 0.73 & 0 & - \\ & 1 & - & 0 & 0.79 & - \\ & & - & - & - & - \\ & & & 1 & 0 & - \\ & & & & 1 & - \\ & & & & & - \end{bmatrix}$	$\begin{bmatrix} 0.628 \\ 0.807 \\ - \\ 0.684 \\ 0.602 \\ - \end{bmatrix}$, $\begin{bmatrix} 1 & 0 & - & 0.53 & 0 & - \\ & 1 & - & 0 & 0.64 & - \\ & & - & - & - & - \\ & & & 1 & 0 & - \\ & & & & 1 & - \\ & & & & & - \end{bmatrix}$
Joint meta-analyses (accounting for within-study correlation)	$\begin{bmatrix} 0.446 \\ 0.637 \\ 0.578 \\ 0.609 \\ 0.520 \\ 0.513 \end{bmatrix}$, $\begin{bmatrix} 1 & 0.69 & 0.90 & 0.66 & 0.34 & 0.61 \\ & 1 & 0.92 & 0.41 & 0.79 & 0.69 \\ & & 1 & 0.48 & 0.56 & 0.62 \\ & & & 1 & 0.57 & 0.90 \\ & & & & 1 & 0.88 \\ & & & & & 1 \end{bmatrix}$	$\begin{bmatrix} 0.432 \\ 0.717 \\ 0.612 \\ 0.739 \\ 0.611 \\ 0.630 \end{bmatrix}$, $\begin{bmatrix} 1 & 0.74 & 0.84 & 0.88 & 0.78 & 0.94 \\ & 1 & 0.97 & 0.40 & 0.90 & 0.73 \\ & & 1 & 0.54 & 0.85 & 0.78 \\ & & & 1 & 0.48 & 0.88 \\ & & & & 1 & 0.82 \\ & & & & & 1 \end{bmatrix}$
Joint meta-analyses (not accounting for within-study correlation)	$\begin{bmatrix} 0.485 \\ 0.699 \\ 0.626 \\ 0.595 \\ 0.554 \\ 0.572 \end{bmatrix}$, $\begin{bmatrix} 1 & 0.72 & 0.86 & 0.65 & 0.34 & 0.59 \\ & 1 & 0.96 & 0.30 & 0.63 & 0.51 \\ & & 1 & 0.32 & 0.44 & 0.44 \\ & & & 1 & 0.62 & 0.92 \\ & & & & 1 & 0.87 \\ & & & & & 1 \end{bmatrix}$	$\begin{bmatrix} 0.561 \\ 0.879 \\ 0.852 \\ 0.731 \\ 0.673 \\ 0.667 \end{bmatrix}$, $\begin{bmatrix} 1 & 0.94 & 0.96 & 0.58 & 0.57 & 0.72 \\ & 1 & 0.96 & 0.31 & 0.55 & 0.55 \\ & & 1 & 0.35 & 0.43 & 0.52 \\ & & & 1 & 0.48 & 0.84 \\ & & & & 1 & 0.78 \\ & & & & & 1 \end{bmatrix}$

Note: The table shows the factorization of the estimated covariance matrices into vectors of between-study standard deviations and correlation matrices, Note that $\mathbf{T} = \text{diag}(\boldsymbol{\tau})\mathbf{R}\,\text{diag}(\boldsymbol{\tau})$, where $\text{diag}(\boldsymbol{\tau})$ is a matrix with the estimated standard deviations on the diagonal and with all off-diagonal elements 0, and $\mathbf{R}$ is a correlation matrix. The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24).

Table 14. Estimates of between-study standard errors and correlations (analyses of all 30 studies)

Model	Normal approximation, estimates of $\boldsymbol{\tau}^{\prime}, \mathbf{R}$	Discrete likelihood, estimates of $\boldsymbol{\tau}^{\prime}, \mathbf{R}$
Separate meta-analyses	$\begin{bmatrix} 0.448 \\ 0.709 \\ - \\ \hline 0.590 \\ 0.582 \\ - \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0 & - & 0.73 & 0 & - \\ & 1 & - & 0 & 0.28 & - \\ \hdashline & & - & - & - & - \\ \hline & & & 1 & 0 & - \\ & & & & 1 & - \\ \hdashline & & & & & - \end{array}\right]$	$\begin{bmatrix} 0.627 \\ 0.793 \\ - \\ \hline 0.683 \\ 0.639 \\ - \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0 & - & 0.53 & 0 & - \\ & 1 & - & 0 & 0.22 & - \\ \hdashline & & - & - & - & - \\ \hline & & & 1 & 0 & - \\ & & & & 1 & - \\ \hdashline & & & & & - \end{array}\right]$
Joint meta-analyses (accounting for within-study correlation)	$\begin{bmatrix} 0.740 \\ 0.668 \\ 0.780 \\ \hline 0.590 \\ 0.578 \\ 0.519 \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0.83 & 0.97 & 0.28 & -0.27 & 0.04 \\ & 1 & 0.93 & 0.43 & 0.29 & 0.45 \\ \hdashline & & 1 & 0.41 & -0.05 & 0.24 \\ \hline & & & 1 & 0.5 & 0.87 \\ & & & & 1 & 0.86 \\ \hdashline & & & & & 1 \end{array}\right]$	$\begin{bmatrix} 0.470 \\ 0.971 \\ 0.662 \\ \hline 0.745 \\ 0.796 \\ 0.651 \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0.46 & 0.74 & 0.91 & 0.42 & 0.90 \\ & 1 & 0.72 & 0.30 & 0.41 & 0.50 \\ \hdashline & & 1 & 0.53 & 0.47 & 0.71 \\ \hline & & & 1 & 0.33 & 0.84 \\ & & & & 1 & 0.53 \\ \hdashline & & & & & 1 \end{array}\right]$
Joint meta-analyses (not accounting for within-study correlation)	$\begin{bmatrix} 0.475 \\ 0.721 \\ 0.592 \\ \hline 0.628 \\ 0.592 \\ 0.603 \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0.78 & 0.89 & 0.41 & 0.07 & 0.31 \\ & 1 & 0.98 & 0.08 & 0.27 & 0.21 \\ \hdashline & & 1 & 0.22 & 0.24 & 0.28 \\ \hline & & & 1 & 0.67 & 0.93 \\ & & & & 1 & 0.89 \\ \hdashline & & & & & 1 \end{array}\right]$	$\begin{bmatrix} 0.575 \\ 1.149 \\ 0.765 \\ \hline 0.751 \\ 0.833 \\ 0.703 \end{bmatrix}$, $\left[\begin{array}{cc:c:cc:c} 1 & 0.52 & 0.90 & 0.71 & 0.37 & 0.77 \\ & 1 & 0.64 & 0.19 & 0.30 & 0.39 \\ \hdashline & & 1 & 0.43 & 0.31 & 0.57 \\ \hline & & & 1 & 0.36 & 0.83 \\ & & & & 1 & 0.52 \\ \hdashline & & & & & 1 \end{array}\right]$

Note: The table shows the factorization of the estimated covariance matrices into vectors of between-study standard deviations and correlation matrices, Note that $\mathbf{T} = \mathrm{diag}(\boldsymbol{\tau})\mathbf{R}\,\mathrm{diag}(\boldsymbol{\tau})$, where $\mathrm{diag}(\boldsymbol{\tau})$ is a matrix with the estimated standard deviations on the diagonal and with all off-diagonal elements 0, and $\mathbf{R}$ is a correlation matrix. The joint meta-analyses of the two tests use the unstructured variant of the between-study covariance matrix as per equation (24).

Discussion

We propose models for the joint (multivariate) meta-analysis of $M \geq 2$ diagnostic tests. The models are applicable when the tests are applied in the same patients and a substantial number of studies report data on the cross-classification of results from several tests, and for examples in which it is helpful to summarize study accuracy with a "summary point" (such as summary sensitivity and specificity), rather than a "summary line" (such as a hierarchical summary receiver operating characteristic curve). We derive formulas for calculating within-study covariances from data reported in the studies themselves. We show in an applied example that the developed methods can result in tighter confidence intervals for comparisons between sensitivities or false-positive rates of different tests than those from separate meta-analyses for each test.

True comparative accuracy studies involve a network of different diagnostic accuracy studies that include fully cross-classified data from crossover studies, data from crossover studies reported as separate 2×2 tables, and data from parallel test studies that examine all or a subset of the potential index tests. The herein described models can form the basis for network meta-analysis of test accuracy. For example, the case of three or four tests is essentially a case of network meta-analysis. Nevertheless, for a more general solution to the meta-analysis of networks of diagnostic accuracy studies, the herein presented work should be extended to incorporate all of the data structures, respect their constraints and simultaneously estimate all parameters, while respecting consistency equations that arise when the network contains three or more tests that are not all compared in each study.

We hypothesize that most studies of diagnostic accuracy that assess two or more tests in the same patients do not report sufficient data to extract the cross-tabulations of test results as in Table 5. If this is true, only separate meta-analyses are possible. Authors of primary studies of diagnostic accuracy should be encouraged to report such information clearly. In our work with the illustrative example, the extraction of counts for cross-classified test results was challenging. For example, Biagiotti et al.[16] reported cumulative counts of test positives for shortened humerus, shortened femur and for both tests combined using several thresholds. To reconstruct the equivalent of Table 5 for the thresholds of interest (ratio of observed to expected length <0.90 for humerus and <0.91 for femur), we had to employ mixed integer linear programming (details available from the authors on request). Even so, we could not extract cross-tabulation counts from five studies in the trisomy 21 group, and from six studies in the healthy group.

In the application, when the focus is on estimates of the summary TPRs or summary FPRs, the payoff of multivariate analyses was modest at best, in that the differences in the summary estimates and confidence/credible intervals from separate meta-analyses for each test were not that pronounced. For the models presented here, the summary point estimates from separate meta-analyses will be the same as from joint meta-analyses if (a) the within study covariances are zero, or (b) if the within-study covariance matrices are all equal (that is $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k).[50] The former is unlikely for diagnostic tests, because we can reasonably expect that results obtained from tests applied to the same patients will be positively correlated. Therefore, for meta-analysis of diagnostic tests the point estimates will differ when the within-study covariance matrices are most dissimilar across studies. However, the variances of the summary estimates are always affected.

The most pronounced differences between separate and joint meta-analyses are observed when one estimates comparative diagnostic accuracy. We see this as a manifestation of the

theoretical advantages of the joint meta-analyses. First, the joint meta-analyses utilize all the information in the cross-tabulation of the test results, and this translates into smaller standard errors compared with separate meta-analyses. Second, with joint meta-analyses one can obtain simultaneous confidence intervals that control the type I error for all possible linear combinations of summary estimates, as described in (40).[47] Third, when calculating comparative test accuracy, the joint meta-analyses respect the grouping of data by study. In contrast, separate meta-analyses rely on differences of marginal estimates; that is, they first average across studies and then compare the averages. As in meta-analyses of treatment effects, the latter approach can result in biased estimates (in some way analogous to Simpson's paradox).

Already with two tests, transforming the parameters to the logit scale introduces a complication: the sum of the probabilities recovered from the logit-scale parameters is not guaranteed to be 1. One can avoid this complication by constrained optimization or by using parameters in the probability scale and a Dirichlet distribution for the random effects. Further research will be necessary to determine whether a Dirichlet distribution can accommodate sums of probabilities such as TPRs and FPRs as parameters, or only the individual probabilities.

Nevertheless, our work demonstrates complexities introduced by simultaneous consideration of multiple diagnostic tests that are not apparent in treatments of diagnostic test models which focus on a single test. First, while sensitivity and specificity completely describe the model parameters for a single test, they fail to do so once a second test is introduced. Then, it becomes necessary to consider the joint probability of the two tests in addition to the marginal probabilities of each. With additional tests, the need for extra parameters grows rapidly and it is not clear at all which are the most clinically important parameterizations.

Second, the number of cross-classifications grows rapidly with more tests, thus increasing the chance of combinations of test results with small counts. When some counts are small, the normal likelihood becomes suboptimal[46,51] and multinomial observational distributions are preferred. Moreover, not only are the observed counts small but the cell probabilities also may become extremely small and so the asymptotic normality of the random effects distribution may no longer hold either. Random effects will also be imprecisely estimated when the number of participants is small and outliers will become more of an issue, inflating between-study variances and inflating prediction ellipses.

Third, the exponentially exploding number of parameters will require structured correlation matrices for model identifiability and numerical convergence. We have suggested one such structure, but even its simple form did not always permit our models to be estimable. More experience with these models is needed to resolve both numerical and conceptual issues for appropriate structures.

Fourth, as the number of tests increases, the likelihood that each patient will receive each test and have a recorded result becomes unlikely. Missing data will then become the norm, rather than the exception. Such missing data may be random or structural when certain tests are not ordered. Assumptions about the missing data mechanism become crucial with large amounts of missing data and with nearly empty cells. Our assumption that data are missing at random will hold only in special situations. These certainly do not include sequences of tests where test ordering is related to prior knowledge about health status.

Fifth, while comparative accuracy studies should rightly be set up so that patients receive each test in order to reduce biases and ensure the clinical relevance of the resulting inferences, such designs require strict protocols to ensure validity. Many of these issues are covered by items in the QUADAS study quality and STARD reporting quality instruments including prospective

and consecutive data collection, avoidance of verification bias, blinding of readers to test and gold standard results and control of the time between test administrations. As the number of tests grows, quality control becomes more difficult and differences may begin to influence test results.

Finally, in the field of diagnostic accuracy it is hard to imagine a topic where exploration of heterogeneity with covariates is not needed. In principle, adding covariates to the current models is straightforward: one simply has to substitute linear equations of the covariates in the place of the means at the structural model. This will increase the number of parameters further. An additional complication stems from the different parameterizations of the probabilities in the structural model. For example, the interpretation of the coefficients for covariates will be very different for parameterizations based on sums of probabilities (e.g., TPRs and FPRs) versus a direct parameterization of the probabilities themselves. It is not clear how one should judge the appropriateness of the various parameterizations.

So should one use separate or joint meta-analysis for analyzing paired diagnostic accuracy data? This work introduces one approach and is not sufficient for making general methodological recommendations. In theory, the decision for performing separate versus joint meta-analyses depends on the underlying assumptions that the researcher is prepared to make about the data. Ideally, these decisions should be made early in the analysis, and not after an examination of the data. In some instances alternative analyses will yield similar results; in other instances they may not. The fact that we observed differences in comparative test accuracy in a single example is not a strong basis for recommending general use.

More generally but in the same vein, methodological recommendations address the problem of choosing between alternative methodologies, and developing them should be approached as a decision problem.[52] In brief, one must define the decision context, which includes specifying: (a) the perspective from which the problem is approached (e.g., recommendations for specialists who perform publicly funded research may have to set a much higher bar than recommendations for the meta-analysis community, which may want to set pragmatic minimum standards); (b) all reasonable alternative choices; (c) the type of problems to which the recommendation applies; (d) the characteristics or quantities (utilities) that will be used in making a decision, and their relative weights if more than one exist. These represent not only the scientific rigor of each alternative but also issues such as the feasibility of performing an analysis without access to statistical expertise or specialized software, and other concerns. Subsequently, one must examine theoretical arguments, results from simulation analyses, or empirical data, as applicable, and be transparent on how the recommendation is reached. Pending an adequate exploration of this decision problem in their setting, our opinion is that meta-analysts facing problems such the one analyzed here should consider performing joint meta-analysis, if only as a sensitivity analysis.

References

1. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. Jun 6 2006;144(11):850-5.
2. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations.
3. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. Apr 15 1994;120(8):667-76.
4. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. Jan 10 1998;351(9096):123-7.
5. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making. Sep-Oct 2009;29(5):E13-21.
6. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Sep-Oct 2009;29(5):E1-E12.
7. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med. Jun 21 2005;142(12 Pt 2):1048-55.
8. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-90.
9. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol. 1995;2 Suppl 1:S48-S56.
10. Kliegman RM, Behrman RE, Jenson HB, et al. Nelson textboook of pediatrics. Philadelphia, PA: Saunders; 2007.
11. Smith-Bindman R, Hosmer W, Feldstein VA, et al. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. JAMA. Feb 28 2001;285(8):1044-55.
12. Palomaki GE, Deciu C, Kloza EM, et al. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med. Mar 2012;14(3):296-305.
13. Benacerraf BR, Neuberg D, Frigoletto FD, Jr. Humeral shortening in second-trimester fetuses with Down syndrome. Obstet Gynecol. Feb 1991;77(2):223-7.
14. Benacerraf BR, Neuberg D, Bromley B, et al. Sonographic scoring index for prenatal detection of chromosomal abnormalities. J Ultrasound Med. Sep 1992;11(9):449-58.
15. Benacerraf BR, Nadel A, Bromley B. Identification of second-trimester fetuses with autosomal trisomy by use of a sonographic scoring index. Radiology. Oct 1994;193(1):135-40.
16. Biagiotti R, Periti E, Cariati E. Humerus and femur length in fetuses with Down syndrome. Prenat Diagn. Jun 1994;14(6):429-34.
17. Bromley B, Lieberman E, Benacerraf BR. The incorporation of maternal age into the sonographic scoring index for the detection at 14-20 weeks of fetuses with Down's syndrome. Ultrasound Obstet Gynecol. Nov 1997;10(5):321-4.
18. Johnson MP, Michaelson JE, Barr M Jr., et al. Combining humerus and femur length for improved ultrasonographic identification of pregnancies at increased risk for trisomy 21. Am J Obstet Gynecol. Apr 1995;172(4 Pt 1):1229-35.
19. Lockwood CJ, Lynch L, Ghidini A, et al. The effect of fetal gender on the prediction of Down syndrome by means of maternal serum alpha-fetoprotein and ultrasonographic parameters. Am J Obstet Gynecol. Nov 1993;169(5):1190-7.
20. Nyberg DA, Resta RG, Luthy DA, et al. Humerus and femur length shortening in the detection of Down's syndrome. Am J Obstet Gynecol. Feb 1993;168(2):534-8.

21. Nyberg DA, Luthy DA, Resta RG, et al. Age-adjusted ultrasound risk assessment for fetal Down's syndrome during the second trimester: description of the method and analysis of 142 cases. Ultrasound Obstet Gynecol. Jul 1998;12(1):8-14.

22. Rodis JF, Vintzileos AM, Fleming AD, et al. Comparison of humerus length with femur length in fetuses with Down syndrome. Am J Obstet Gynecol. Oct 1991;165(4 Pt 1):1051-6.

23. Vintzileos AM, Egan JF, Smulian JC, et al. Adjusting the risk for trisomy 21 by a simple ultrasound method using fetal long-bone biometry. Obstet Gynecol. Jun 1996;87(6):953-8.

24. Campbell WA, Vintzileos AM, Rodis JF, et al. Efficacy of the biparietal diameter/femur length ratio to detect Down syndrome in patients with an abnormal biochemical screen. Fetal Diagn Ther. May-Jun 1994;9(3):175-82.

25. Ginsberg N, Cadkin A, Pergament E, et al. Ultrasonographic detection of the second-trimester fetus with trisomy 18 and trisomy 21. Am J Obstet Gynecol. Oct 1990;163(4 Pt 1):1186-90.

26. Cuckle H, Wald N, Quinn J, et al. Ultrasound fetal femur length measurement in the screening for Down's syndrome. BJOG. Dec 1989;96(12):1373-8.

27. Dicke JM, Gray DL, Songster GS, et al. Fetal biometry as a screening tool for the detection of chromosomally abnormal pregnancies. Obstet Gynecol. Nov 1989;74(5):726-9.

28. Benacerraf BR, Cnann A, Gelman R, et al. Can sonographers reliably identify anatomic features associated with Down syndrome in fetuses? Radiology. Nov 1989;173(2):377-80.

29. Brumfield CG, Hauth JC, Cloud GA, et al. Sonographic measurements and ratios in fetuses with Down syndrome. Obstet Gynecol. Apr 1989;73(4):644-6.

30. Grandjean H, Sarramon MF. Femur/foot length ratio for detection of Down syndrome: results of a multicenter prospective study. The Association Francaise pour le Depistage et la Prevention des Handicaps de l'Enfant Study Group. Am J Obstet Gynecol. Jul 1995;173(1):16-9.

31. Johnson MP, Barr M, Jr., Treadwell MC, et al. Fetal leg and femur/foot length ratio: a marker for trisomy 21. Am J Obstet Gynecol. Sep 1993;169(3):557-63.

32. Grist TM, Fuller RW, Albiez KL, et al. Femur length in the US prediction of trisomy 21 and other chromosomal abnormalities. Radiology. Mar 1990;174(3 Pt 1):837-9.

33. LaFollette L, Filly RA, Anderson R, et al. Fetal femur length to detect trisomy 21. A reappraisal. J Ultrasound Med. Dec 1989;8(12):657-60.

34. Hill LM, Guzick D, Belfar HL, et al. The current role of sonography in the detection of Down syndrome. Obstet Gynecol. Oct 1989;74(4):620-3.

35. Lockwood C, Benacerraf B, Krinsky A, et al. A sonographic screening method for Down syndrome. Am J Obstet Gynecol. Oct 1987;157(4 Pt 1):803-8.

36. Verdin SM, Economides DL. The role of ultrasonographic markers for trisomy 21 in women with positive serum biochemistry. BJOG. Jan 1998;105(1):63-7.

37. Nyberg DA, Luthy DA, Cheng EY, et al. Role of prenatal ultrasonography in women with positive screen for Down syndrome on the basis of maternal serum markers. Am J Obstet Gynecol. Oct 1995;173(4):1030-5.

38. Marquette GP, Boucher M, Desrochers M, et al. Screening for trisomy 21 with ultrasonographic determination of biparietal diameter/femur length ratio. Am J Obstet Gynecol. Nov 1990;163(5 Pt 1):1604-5.

39. Nyberg DA, Resta RG, Hickok DE, et al. Femur length shortening in the detection of Down syndrome: is prenatal screening feasible? Am J Obstet Gynecol. May 1990;162(5):1247-52.

40. Shah YG, Eckl CJ, Stinson SK, et al. Biparietal diameter/femur length ratio, cephalic index, and femur length measurements: not reliable screening techniques for Down syndrome. Obstet Gynecol. Feb 1990;75(2):186-8.

41. Lynch L, Berkowitz GS, Chitkara U, et al. Ultrasound detection of Down syndrome: is it really possible? Obstet Gynecol. Feb 1989;73(2):267-70.

42. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. Dec 2006;59(12):1331-1332; author reply 1332-3.

43. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol. Sep 2004;57(9):925-32.

44. Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: Meta-analysis of Test Performance When There Is A "Gold Standard".

.

45. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med. Dec 20 2010;29(29):3046-67.

46. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol. Jan 2008;61(1):41-51.

47. Miller R. Simultaneous statistical inference. 2nd ed. New York: Springer Verlag; 1981.

48. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics. Oct 2009;10(4):792-805.

49. Pinheiro JC, Bates DM. Unconstrained parametrizations for variance-covariance matrices. Stat Comput. Sep 1996;6(3):289-96.

50. Riley RD, Abrams KR, Lambert PC, et al. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med. Jan 15 2007;26(1):78-97.

51. Trikalinos TA, Trow P, Schmid CH. Simulation-based comparison of methods for the meta-analysis of proportions and rates [draft report]. Rockville, MD: Agency for Healthcare Research and Quality; 2012.

52. Trikalinos TA, Dahabreh IJ, Wallace BC, et al. A framework for rating the strength of methodological recommendations for systematic reviews and meta-analyses [draft report]. Rockville, MD: Agency for Healthcare Research and Quality; 2012

Appendix A. Formulas for Within-Study Covariance Matrices and for 1.96-Standard Error Volumes

Variances for logit-transformed sensitivities and false positive rates for bivariate meta-analysis (single test)

See Table 3 in the main report for notation. The within-study covariance $\mathbf{\Sigma}_k$ in study k has zero off-diagonal elements, because sensitivity and specificity are calculated in independent groups.

$$\mathbf{\Sigma}_k = \begin{pmatrix} \sigma^2_{k\eta} & 0 \\ & \sigma^2_{k\xi} \end{pmatrix},$$

with

$$\sigma^2_{k\eta} = \frac{1}{N^D_k p^D_{k1}(1-p^D_{k1})},$$

and

$$\sigma^2_{k\xi} = \frac{1}{N^{\bar{D}}_k p^{\bar{D}}_{k1}(1-p^{\bar{D}}_{k1})},$$

Variances for logit-transformed sensitivities and false positive rates for the joint multivariate meta-analysis of two tests

Application of the multivariate delta method yields the following formulas. (The notation [m:1] indicates the sum over all patterns in which the outcome on test m is 1).

Variance of logit TPR in study k for test m:

$$\sigma^2_{k,\eta m} = \frac{1}{N^D_k \hat{\pi}^D_{k[m:1]}(1-\hat{\pi}^D_{k[m:1]})}$$

Variance of logit JTPR in study k:

$$\sigma^2_{k,\eta^*} = \frac{1}{N^D_k \hat{\pi}^D_{k,11}(1-\hat{\pi}^D_{k,11})}$$

Variance for logit FPR in study k for test m:

$$\sigma^2_{k,\xi m} = \frac{1}{N^{\bar{D}}_k \hat{\pi}^{\bar{D}}_{k[m:1]}(1-\hat{\pi}^{\bar{D}}_{k[m:1]})}$$

Variance for logit JFPR in study k:

$$\sigma^2_{k,\xi^*} = \frac{1}{N^{\bar{D}}_k \hat{\pi}^{\bar{D}}_{k,11}(1-\hat{\pi}^{\bar{D}}_{k,11})}$$

Covariance between logit TPRs of tests m and t in study k:

$$\sigma_{k,\eta m\eta t} = \frac{\hat{\pi}^{D}_{k[m:1,t:1]} - \hat{\pi}^{D}_{k[m:1]}\hat{\pi}^{D}_{k[t:1]}}{N^{D}_{k}\hat{\pi}^{D}_{k[m:1]}(1-\hat{\pi}^{D}_{k[m:1]})\hat{\pi}^{D}_{k[t:1]}(1-\hat{\pi}^{D}_{k[t:1]})}$$

Covariance between logit FPRs of tests *m* and *t* in study *k*:

$$\sigma_{k,\xi m\xi t} = \frac{\hat{\pi}^{\bar{D}}_{k[m:1,t:1]} - \hat{\pi}^{\bar{D}}_{k[m:1]}\hat{\pi}^{\bar{D}}_{k[t:1]}}{N^{\bar{D}}_{k}\hat{\pi}^{\bar{D}}_{k[m:1]}(1-\hat{\pi}^{\bar{D}}_{k[m:1]})\hat{\pi}^{\bar{D}}_{k[t:1]}(1-\hat{\pi}^{\bar{D}}_{k[t:1]})}$$

Covariance between logit TPR of test *m* and logit-JTPR in study *k*:

$$\sigma_{k,\eta m\eta^{*}} = \frac{1}{N^{D}_{k}\hat{\pi}^{D}_{k[m:1]}(1-\hat{\pi}^{D}_{k,11})}$$

Covariance between logit FPR of test *m* and logit-JFPR in study *k*:

$$\sigma_{k,\xi m\xi^{*}} = \frac{1}{N^{\bar{D}}_{k}\hat{\pi}^{\bar{D}}_{k[m:1]}(1-\hat{\pi}^{\bar{D}}_{k,11})}$$

Formulas for calculating 1.96-standard-error volumes

Let $t_1,\ldots,t_M$ be the lengths of the half axes of an ellipsoid of dimension *M* that corresponds to the contour surface of one standard error. The volume V_M included in this one-standard-error surface is calculated by integration. We calculated the first three integrals. (In the following three formulas $\pi = 3.14159\ldots$)

$$V_2 = \pi t_1 t_2$$

$$V_3 = \frac{4}{3}\pi t_1 t_2 t_3$$

$$V_4 = \frac{1}{2}\pi^2 t_1 t_2 t_3 t_4$$

For a covariance matrix $\mathbf{C}$ we have to calculate the lengths of the half axes for the one-standard-error contour ellipsoid. Rotation to an orthonormal basis automatically provides the lengths of the half axes; these are the square roots of the eigenvalues $\lambda_1,\ldots,\lambda_M$ of $\mathbf{C}$. So set $t_m = \sqrt{\lambda_m}$ in the formulas above. The 1.96-standard-error volume uncorrected for multiple comparisons is

$$(z_{0.025})^M V_M,$$

With $z_{\alpha/2} = 1.96$ denoting the upper $\alpha/2$ percentile of the standard normal distribution. For example in the main report, Table 11 the volumes in rows (b) and (c) pertain to four-dimensional models and were calculated using the formula above, for $M = 4$. The confidence volume in row (a) corresponds to two independent bivariate models. It is calculated as the product of the confidence volumes of dimension 2, one for each bivariate model.