Journal of Clinical Epidemiology 58 (2005) 982–990

Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews

Johannes B. Reitsma[a,*], Afina S. Glas[a], Anne W.S. Rutjes[a], Rob J.P.M. Scholten[b], Patrick M. Bossuyt[a], Aeilko H. Zwinderman[a]

[a]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands

[b]Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, The Netherlands

Accepted 21 February 2005

Abstract

Background and Objectives: Studies of diagnostic accuracy most often report pairs of sensitivity and specificity. We demonstrate the advantage of using bivariate meta-regression models to analyze such data.

Methods: We discuss the methodology of both the summary Receiver Operating Characteristic (sROC) and the bivariate approach by reanalyzing the data of a published meta-analysis.

Results: The sROC approach is the standard method for meta-analyzing diagnostic studies reporting pairs of sensitivity and specificity. This method uses the diagnostic odds ratio as the main outcome measure, which removes the effect of a possible threshold but at the same time loses relevant clinical information about test performance. The bivariate approach preserves the two-dimensional nature of the original data. Pairs of sensitivity and specificity are jointly analyzed, incorporating any correlation that might exist between these two measures using a random effects approach. Explanatory variables can be added to the bivariate model and lead to separate effects on sensitivity and specificity, rather than a net effect on the odds ratio scale as in the sROC approach. The statistical properties of the bivariate model are sound and flexible.

Conclusion: The bivariate model can be seen as an improvement and extension of the traditional sROC approach.

Keywords: Diagnosis; Diagnostic accuracy studies; Sensitivity and specificity; Meta-analysis; Meta-regression; Review

1. Introduction

Diagnostic accuracy studies are a vital step in the evaluation of diagnostic technologies [1–3] Accuracy studies measure the level of agreement between the results of a test under evaluation and that of the reference standard. There are several different measures of diagnostic accuracy [4,5], but the majority of diagnostic accuracy studies present estimates of sensitivity and specificity, either alone or in combination with other measures [6].

Because the majority of diagnostic papers report estimates of sensitivity and specificity, meta-analytic approaches have focused on these measures [6–13]. Pooling pairs of sensitivity and specificity is not straightforward, because these measures are often negatively correlated within studies.

The summary Receiver Operating Characteristic (sROC) approach has become the method of choice for the meta-analysis of studies reporting pairs of sensitivity and specificity [9,12,14–18]. The sROC approach converts each pair of sensitivity and specificity into a single measure of accuracy, the diagnostic odds ratio [19]. The disadvantage of a single measure of diagnostic accuracy is that it does not distinguish between the ability of detecting the sick (sensitivity) and identifying the well (specificity). Discriminating between these abilities is important to determine the optimal use of a test in clinical practice. The bivariate model we propose has the distinct advantage of preserving the two-dimensional nature of the underlying data. It can also produce summary estimates of sensitivity and specificity, acknowledging any possible (negative) correlation between these two measures.

* Corresponding author. Tel.: +31-20-5663273; fax: +31-20-6912683. *E-mail address*: j.reitsma@amc.uva.nl (J.B. Reitsma).

doi: 10.1016/j.jclinepi.2005.02.022

We will discuss both approaches and illustrate their use by reanalyzing the data from a published meta-analysis [20].

2. Pooling pairs of sensitivity and specificity: why simple methods fail

Diagnostic reviews start with a set of individual studies presenting estimates of sensitivity and specificity. One intuitive approach is to do separate pooling of sensitivity and specificity using standard methods for proportions. However, sensitivity and specificity are often negatively correlated within studies, and ignoring this correlation would be inappropriate [7,11,12].

One possible cause for this negative correlation between sensitivity and specificity is that studies may have used different thresholds to define positive and negative test results. In some cases, this may have been done explicitly; for example, studies that used different cutoff points to classify a continuous biochemical measurement as either positive or negative. In other situations there may have been implicit variations in thresholds between studies due to differences in observers, laboratories, or equipment. Unlike other sources of variation, a difference in threshold leads to a particular pattern between sensitivity and specificity. This pattern is well known from studies showing the effect of different cutoffs in case of a biochemical test with a continuous outcome [21–23]. Lowering the cutoff value will then lead to more patients with a positive result, thereby increasing the number of true positives but also the number of false positive results. This means that sensitivity will be higher, but at the expense of specificity. This trade-off between sensitivity and specificity leads to a concave, shoulder-like curve when sensitivity is plotted against 1 minus specificity, the receiver operating characteristic (ROC) curve. In many publications involving ROC plots, sensitivity is referred to as the true positive rate (TPR) and 1 minus specificity as the false positive rate (FPR).

In the next paragraphs, we discuss how the sROC and the bivariate approach deal with estimates of sensitivity and specificity that can show large variability and possible negative correlation.

3. Summary ROC approach

We provide a short description of the sROC approach as outlined by Moses and Littenberg. More details can be found elsewhere [9,12,14–18].

The sROC approach starts with plotting the observed pairs of sensitivity and specificity of each study in ROC space (see Fig. 1). The aim of the sROC approach is to find a smooth curve through these points. The key step is to transform the TPR (sensitivity) and FPR (1 − specificity) scale of the ROC graph so that the relation becomes more linear and a straight line can be fitted to the data points [12].

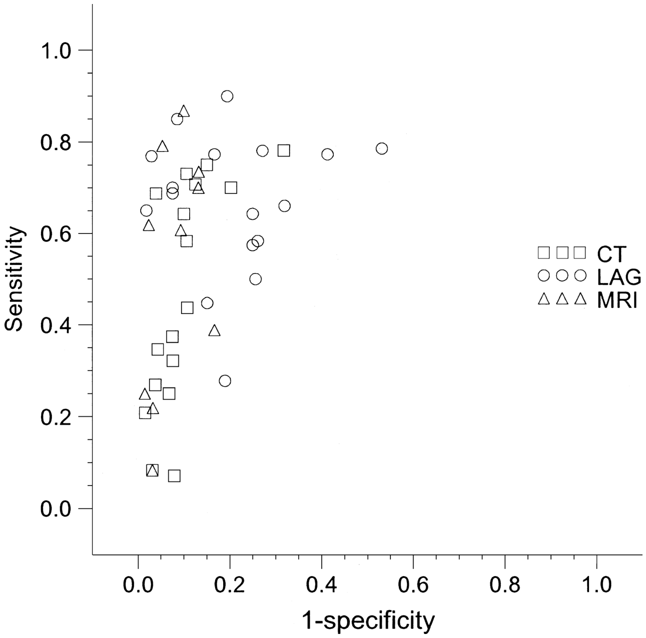

Fig. 1. ROC plot of sensitivity against 1 − specificity for 44 studies comparing the diagnostic accuracy of lymphangiography (LAG), computed tomography (CT), and magnetic resonance imaging (MRI) in the diagnosis of lymph node metastases in women with cervical cancer. Data from a published meta-analysis [20].

The following transformations of TPR and FPR are used. D is defined as the difference in the logit transformed values of TPR and FPR, while S is the sum of these same logits:

$$D = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) - \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right) = \ln(\text{DOR}) \qquad (1)$$

and

$$S = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) + \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right) \qquad (2)$$

D is the log of the diagnostic odds ratio (DOR). The DOR is a single overall indicator of diagnostic accuracy, and it indicates how more often (expressed as odds) a positive test result occurs among patients with the condition of interest compared to patients without the condition. S relates to the test threshold. It has a value of 0 in a study where sensitivity equals specificity, S is positive in studies where sensitivity is higher than specificity, and S is negative when specificity is higher.

A linear regression line is fitted through the transformed points of the ROC graph, showing how the (log) DOR varies with the implicit threshold:

$$D = \alpha + \beta \cdot S \qquad (3)$$

The model is generally fitted using either weighted or unweighted least squares linear regression [12].

3.1. Example using the sROC approach

We illustrate the use of the sROC approach by reanalyzing the data of a published meta-analysis [20]. In this meta-analysis, Scheidler et al. compared three imaging techniques for the diagnosis of lymph node metastasis in women with cervical cancer. Forty-four studies in total were included: 17 studies evaluated lymphangiography, another 17 studies examined computed tomography and the remaining 10 studies focused on magnetic resonance imaging. Diagnosis of metastatic disease by lymphangiography (LAG) is based on the presence of nodal-filling defects, whereas computed tomography (CT) and magnetic resonance imaging (MRI) rely on nodal enlargement.

Similar to the original meta-analysis we fitted three different, unweighted regression lines, one for each imaging modality. The intercepts (α) and the slopes (β) of these three regression lines are given in Table 1. In the final step, these linear regression lines are transformed back to the original axes of the ROC to obtain the sROC curve. Fig. 2 shows the three sROC curves, one for each imaging modality.

The interpretation of the intercept and the slope of the linear regression model of eq. (3) is not straightforward. When the diagnostic odds ratio (DOR) does not depend on the threshold S (e.g., $\beta \approx 0$), the intercept would provide a summary estimate of the DOR. When the DOR does vary with S, the coefficient of the slope (β) has no direct interpretation, but has a considerable effect on the shape of the sROC curve [18].

The disadvantage of the diagnostic odds ratio as the outcome parameter is that summary estimates of sensitivity and specificity are not directly available. It is only possible to obtain an estimate of sensitivity by specifying a value of specificity, or vice versa. Many meta-analyses have reported sensitivity and specificity at the Q-point. The Q-point is the point on the sROC curve where sensitivity equals specificity and is located where the diagonal line running from the top left corner to the lower right corner intersects the sROC-curve (see Fig. 2). Unfortunately, the Q-point may lead to summary values of sensitivity and specificity that are not close, or even outside the range of values from the original studies (see Q-point for MRI in Fig. 2).

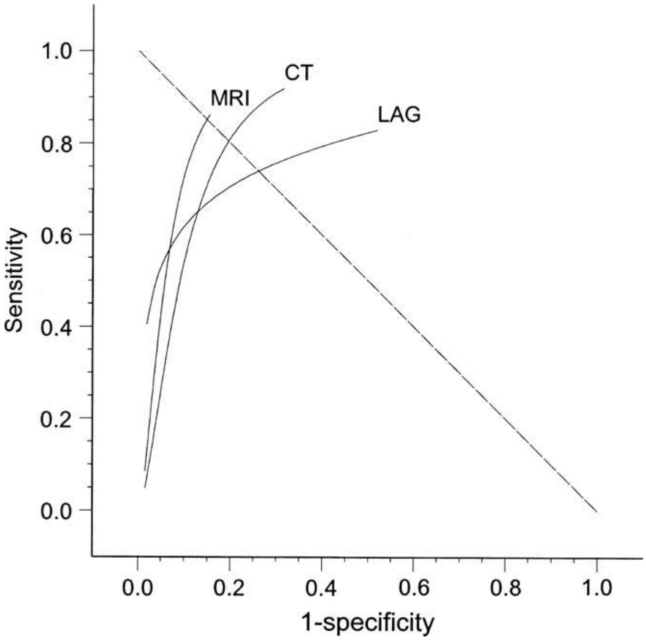

Fig. 2. Estimated summary ROC curves and Q-points for each of the three imaging modalities. Q-point is the point on summary ROC curve where sensitivity equals specificity (intersection of the dashed, diagonal line with the summary ROC curve). See Fig. 1 for primary data.

These Q-points are also used to test for a difference in overall accuracy between diagnostic tests. The rationale is that Q-points remove the effect of possible difference in threshold by comparing the diagnostic odds ratios at a specific value of S, namely zero. However, testing at a different value of S could lead to different conclusions if the diagnostic odds ratio of one or both tests varies with S. In our example there is a statistically significant difference in accuracy (diagnostic odds ratio) at the Q-point between lymphangiography and MRI, but at the overall mean value of S the difference in diagnostic odds ratio is not significant (see Table 1 and Fig. 2).

4. Bivariate model

The bivariate model uses a different starting point for the meta-analysis of pairs of sensitivity and specificity. Rather

Table 1
Intercepts and slopes of the linear regression line underlying the summary ROC approach

Imaging modality	Intercept (SE) α	Coefficient for S (SE) β	DOR at mean of S (95% CI)	Q-point (95% CI)
LAG	2.09 (0.30)	−0.35 (0.20)	16.02 (8.37 to 30.66)	0.74 (0.68 to 0.79)
CT	2.84 (0.44)	0.23 (0.14)	10.90 (6.49 to 18.33)	0.81 (0.73 to 0.87)
MRI	3.51 (0.56)	0.25 (0.17)	20.26 (10.33 to 39.74)	0.85 (0.77 to 0.91)
P-value LAG vs. CT			0.36	0.15
P-value LAG vs. MRI			0.62	0.01
P-value CT vs. MRI			0.15	0.34

Separate lines are fitted for each of the three imaging modalities. Comparison of accuracy (DOR) at the average value of S and at the Q-point, where S equals 0.

Abbreviations: LAG, lymphangiography; CT, computed tomography; MRI, magnetic resonance imaging; DOR, diagnostic odds ratio; Q-point, point on summary ROC curve where sensitivity equals specificity; CI, confidence interval; SE, standard error.

than transforming these two distinct outcome measures into a single indicator of diagnostic accuracy as in the sROC approach, the bivariate model preserves the two-dimensional nature of the data throughout the analysis.

The bivariate model is based on the following line of reasoning [10,24,25]. We assume that the sensitivities from individual studies (after logit transformation) within a meta-analysis are approximately normally distributed around a mean value with a certain amount of variability around this mean. This is a random effect approach, similar to what is used in therapeutic trials, to incorporate unexplained variability in the analysis. This variation in underlying sensitivities between studies can be related to remaining differences in study population, differences in implicit threshold, or unnoticed variations in index test protocol. The same considerations apply to the specificities of these studies. The potential presence of a (negative) correlation between sensitivity and specificity within studies is addressed by explicitly incorporating this correlation into the analysis. The combination of two normally distributed outcomes, the logit transformed sensitivities and specificities, while acknowledging the possible correlation between them, leads to the bivariate normal distribution [25,26].

Besides variability between studies in the true underlying sensitivities and specificities, there is also variation due to sampling. Studies differ in size and variation due to chance is more likely in smaller studies. Therefore, we extended the bivariate model by incorporating the precision by which sensitivity and specificity have been measured in each study using the approach of Van Houwelingen et al. [24]. It means that studies with a more precise estimate of sensitivity are given a higher weight in the analysis of sensitivities. The same is true for studies with more precise estimates of specificity. A more technical description of the bivariate model can be found in Appendix 1.

These bivariate models can be analyzed using linear mixed model techniques that are now widely available in statistical packages. The parameters of the bivariate model are estimated in a single model to incorporate the possible correlation between sensitivities and specificities. To a degree, the bivariate diagnostic model can be viewed as a longitudinal analysis with two measurements (corresponding to sensitivity and specificity) within each person (corresponding to individual studies). The main commands to run the bivariate model using the linear mixed model procedure in SAS are given in Appendix 2 (the complete syntax together with the original data can be found on the journal's website at www.Elsevier.com).

The bivariate model can be seen as an improvement and extension of the simple summary ROC approach, as it can produce the following results:

- In a single step, the bivariate model will estimate the amount of between-study variation in sensitivity and specificity separately, in addition to the degree of correlation between sensitivity and specificity. This provides important background information about the heterogeneity of results between studies and the possibility of an implicit threshold.
- The bivariate model produces summary estimates of sensitivity and specificity and their 95% confidence interval. These intervals take into account the heterogeneity beyond chance between studies (random effects model).
- Using the parameters of the bivariate distribution we can calculate either a confidence ellipse around the mean values of logit sensitivity and specificity or a prediction ellipse for individual values of sensitivity and specificity, taking into account the possible (negative) correlation between sensitivity and specificity.
- The parameters of the bivariate distribution can also be used to obtain a sROC curve. This bivariate sROC curve would be similar to the standard sROC of Moses and Littenberg, if we would correct for two statistical shortcomings of the standard sROC approach (see paragraph on statistical properties).
- Other measures derived from sensitivity and specificity can be calculated, such as the diagnostic odds ratios and likelihood ratios.
- Covariates can be added to the bivariate model and they lead to separate effects on sensitivity and specificity, but net effects on the diagnostic odds ratio are still available. This means that we can explicitly test whether sensitivity or specificity or both are different between two diagnostic technologies.

We will now use the bivariate model to reanalyze the same dataset as we used in the summary ROC approach.

4.1. Example using the bivariate model

The bivariate model directly provides summary estimates of (logit) sensitivity and specificity with corresponding 95% CI for the three imaging modalities (see Table 2). Because of the bivariate nature of the analysis we can either test for differences in sensitivity, or specificity, or both, between the three modalities. The results show that the mean specificity of LAG is significantly lower than that of CT and MRI. However, LAG has the highest sensitivity, which is significantly different from CT, but not from MRI. There are no statistically significant differences in the mean value of sensitivity or specificity between CT and MRI.

The difference between LAG and CT/MRI could be viewed as an implicit threshold effect. It shows that LAG is a more sensitive test, but at the expense of more false positive test results, and hence, a lower specificity. It means that the overall accuracy after "correction" for threshold differences (e.g., diagnostic odds ratio) is similar between the three techniques. This explains the results of the summary ROC approach—no differences among the three techniques. The results of the bivariate model show a more complete picture: a difference in sensitivity and specificity between

Table 2
Summary estimates for sensitivity, specificity, and diagnostic odds ratio from the bivariate model

Imaging modality	Mean sensitivity (95% CI)	Mean specificity (95% CI)	Mean DOR (95% CI)
LAG	0.67 (0.57 to 0.76)	0.80 (0.73 to 0.85)	8.13 (5.16 to 12.82)
CT	0.49 (0.37 to 0.61)	0.92 (0.88 to 0.95)	11.34 (6.66 to 19.30)
MRI	0.56 (0.41 to 0.70)	0.94 (0.90 to 0.97)	21.42 (10.81 to 42.45)
P-value LAG vs. CT	0.023	0.0002	0.35
P-value LAG vs. MRI	0.23	0.0001	0.021
P-value CT vs. MRI	0.47	0.34	0.15

Comparison between three imaging modalities.

Abbreviations: LAG, lymphangiography; CT, computed tomography; MRI, magnetic resonance imaging; DOR, diagnostic odds ratio; CI, confidence interval.

LAG and the other two techniques, but no difference in overall accuracy.

We can still use the visual power of the ROC plot to present the results of the bivariate model. The 95% coverage region of the estimated bivariate distribution of logit sensitivity and specificity can be transformed back to the original ROC axes. We can use the ellipse around the mean estimate of sensitivity and specificity of each modality to show the region containing likely combinations of the mean value of sensitivity and specificity. These ellipses clearly show the differences in sensitivity and specificity of LAG compared to CT and MRI (see Fig. 3). In noncomparative situations and in situations where the between-study variance is large, a 95% prediction ellipse would be useful to show the range of likely values for an individual study.

5. General discussion of methods of meta-analysis

In this section we discuss the statistical methods for meta-analysis of studies of diagnostic accuracy by comparing them on a few key items. These are: (1) the choice of outcome measure; (2) the effect of covariates; (3) the statistical properties of the model.

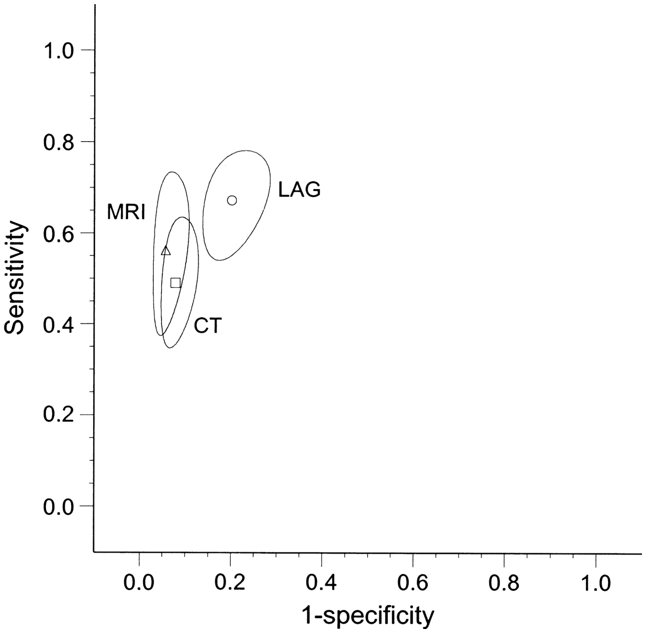

Fig. 3. Bivariate summary estimates of sensitivity and specificity for each of the three imaging modalities and the corresponding 95% confidence ellipse around these mean values. See Fig. 1 for primary data.

5.1. Choice of outcome measure

There are several reasons to prefer sensitivity and specificity as the main outcome measures in the meta-analysis of diagnostic accuracy studies producing dichotomous index test results. First, the vast majority of primary accuracy studies report results in pairs of sensitivity and specificity [6]. Second, our understanding of how sources of bias and variability can affect estimates of diagnostic accuracy is largely based on these measures [27,28]. Third, sensitivity and specificity are at the heart of diagnostic theory and teaching, and therefore the most familiar measures to clinicians. Fourth, alternative measures, such as likelihood ratios and diagnostic odds ratio, can be derived from them.

The bivariate model directly analyses these well-known measures of diagnostic accuracy in a straightforward manner, while acknowledging the possibility of an implicit threshold effect. The model uses a random effects approach in the estimation of summary estimates of sensitivity and specificity and their corresponding 95% confidence intervals. Because the bivariate model estimates the strength and the shape of the correlation between sensitivity and specificity, we can either draw a 95% confidence ellipse around their mean values or draw a 95% prediction ellipse for individual values of sensitivity and specificity. This will enhance our understanding of the heterogeneity of results between studies and the correlation within studies.

In contrast, the sROC approach focuses on the diagnostic odds ratio, and therefore it does not yield unique summary estimates of sensitivity and specificity. This greatly limits its clinical application. Summary values of sensitivity and specificity at the so-called Q-point are just an arbitrary choice of possible values, and may not reflect the sensitivity and specificity reported in primary studies. In our example, Q-values of MRI were misleading (sensitivity = specificity = 0.85), because they are far away from the pooled estimates of sensitivity and specificity from the bivariate model (0.56 and 0.94, respectively). The sROC-curve describes how sensitivity and specificity interact within studies, but the interpretation is not the same as in the normal ROC curve documenting the effect of varying the cutoff value of a continuous measurement. There are several other factors in meta-analysis that can lead to opposing changes in sensitivity and specificity, thereby mimicking the effect of differences in threshold, such as partial verification and various selection mechanisms of patients. Drawing a ROC curve might suggest that all variation is related to a threshold effect, while it is only one of the possible explanations. Another difficulty in the sROC approach is choosing the range of values over which to draw the sROC curve, because a complete curve (from the southwest corner to the northeast corner) is misleading. The boundaries of either the confidence or the prediction ellipse from the bivariate model are, however, well defined.

5.2. *Comparing index tests and examining the effect of covariates*

The sROC approach uses the odds ratio to compare diagnostic accuracy between tests. This removes the effect of a possible difference in threshold, but at the same time it can mask important clinical differences in test performance. In our example, LAG was found to be more sensitive and less specific compared to the other two techniques, which might indicate that it can be an important modality for ruling out lymphadenopathy. In the bivariate model we can specifically test whether there is a difference in sensitivity, specificity, or both.

Examining, quantifying and explaining sources of bias and variability in meta-analysis is a major issue, in particular for diagnostic studies, as there are many possible differences in design, in the selection of patients, and in test protocol between studies [8,24,28–34]. Again, the effect of these differences in design and conduct will be estimated in the sROC approach as changes on the diagnostic odds ratio scale [8,29]. An unchanged odds ratio, however, may obscure the effect of a design feature that increases sensitivity but at the same time lowers specificity, or vice versa. The bivariate model enables an analysis of the effects on sensitivity and specificity separately, whereas a net effect on the odds ratio scale is still available.

5.3. *Statistical properties*

The statistical properties of the bivariate model for performing diagnostic meta-analyses are sound. First, it incorporates and estimates the correlation that might exist between estimates of sensitivity and specificity within studies. This means that the bivariate model will produce valid results whether or not this correlation is high, medium, or absent. This simplifies the overall approach to meta-analysis as outlined in recent guidelines [17,35]. The examination of the degree of correlation between sensitivity and specificity is a cornerstone in these guidelines. If a “moderate” correlation is present, the sROC approach is advocated. If the correlation is small, separate pooling of sensitivity and specificity is promoted. Our bivariate model will automatically deal with both situations. Second, the bivariate model is a direct extension of the methods used for meta-analysis of data from therapeutic trials [24,25]. It takes into account the differences in precision by which sensitivity and specificity have been measured within and across studies, and it incorporates and estimates the amount of between-study variability in both sensitivity and specificity (random effects model).

The sROC approach is based on the linear regression of D on S and has two statistical shortcomings. First, there is measurement error in both the dependent (D) and the independent variable (S). There are well-known methods that take into account measurement error in both the dependent and independent variables, but these are hardly used in the sROC approach. However, a more serious problem is ignoring the covariance that might exist between D and S, because they are defined as the difference and sum of the same two measures. Correcting both errors will lead to a bivariate regression model of D and S, which would be the same as our bivariate model, only with a less transparent definition of the coefficients. Therefore, our bivariate model can be seen as an improvement and an extension of the sROC model of Moses and Littenberg [12]. It is important to note that a linear regression line is also available from the bivariate model to produce the equivalent of a sROC curve. This curve runs through the center of the bivariate confidence ellipse of either the mean values or the individual values. But for reasons stated before, we prefer the use of the confidence or prediction ellipse.

The statistical shortcomings of the sROC approach have been described before [11,13,36] Rutter et al. [11,37] have developed a hierarchical sROC approach. Their aim was also to obtain meaningful summary estimates of sensitivity and specificity, and to improve the handling of within- and between-study variability. The specification of their model is rather complex, and they apply a Bayesian approach for estimating the parameters. Because we reanalyzed the same meta-analysis that was used in their article, we can directly compare the results. Despite the differences in models, the summary estimates and 95% CI of sensitivity and specificity are almost identical. However, their Bayesian approach comes at a price. It is based on Markov Chain Monte Carlo simulations, which requires programming, simulations, an evaluation of model convergence and adequacy, and a synthesis of simulation results. As a result, few if any application of this method can be found in the medical literature, although recently a less complicated way of fitting the hierarchical sROC model has been described [38].

In contrast, the setup of the bivariate model is straightforward, as it directly models sensitivity and specificity. The development of mixed model technology in commercial software means that bivariate models can now be analyzed using standard procedures in statistical packages. Several examples of the bivariate model have already been published [39–42]. The full SAS syntax to replicate our analysis can be found on the journal’s website at www.Elsevier.com.

5.4. Diagnostic accuracy studies: quality of design and reporting

Despite the availability of flexible and advanced statistical models, performing a meta-analysis of diagnostic studies will remain difficult. We see three main reasons [9,15]. First, a systematic review of diagnostic accuracy studies, like any other review, is threatened by publication bias [43]. Second, many reports of studies of diagnostic accuracy lack information on key elements of design and conduct [44]. Without complete and accurate reporting we cannot correctly identify potential sources of bias and variability [45]. This hampers a statistical analysis of these sources of genuine heterogeneity. Third, many studies on diagnostic accuracy have major shortcomings in design or conduct [44,46]. Health care workers need evidence from well-designed studies to make informed choices, but synthesis of study results remains useless in the absence of premium quality primary studies. In all, there is a strong need to improve the methodological quality of diagnostic studies in addition to better standards of reporting.

Appendix 1. Technical description of the bivariate model

We are dealing with individual studies ($I = 1,\ldots, k$) that have reported sensitivity ($p_{A,i}$) determined in N_A individuals with the condition of interest and specificity ($p_{B,i}$) measured in N_B subjects without this condition. We define $\theta_{A,i}$ as the logit-transformed sensitivity in study i, and $\theta_{B,i}$ as the logit-transformed specificity.

We assume that the true logit sensitivities of these individual studies ($\theta_{A,i}$) are normally distributed around some common mean value θ_A with a between-study variability of σ_A^2. The same random effect assumption is used for the specificities of the studies, where we use θ_B to denote the mean value of logit specificity and σ_B^2 as the between-study variance in logit specificity.

We explicitly incorporate the possibility of correlation between (logit) sensitivity and specificity within studies. Combining two normal distributions that can be correlated leads to the following bivariate normal model:

$$\begin{pmatrix} \theta_{A,i} \\ \theta_{B,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_A \\ \theta_B \end{pmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

where σ_{AB} is the covariance between logit sensitivity and specificity.

We extend this bivariate model by incorporating the precision by which sensitivity and specificity have been measured in each study. When N_A and N_B are large and $0 < p_{A,i}; p_{B,i} < 1$, the corresponding variance of the estimated logit transformed sensitivity and specificity in each study are given by

$$s_{A,i}^2 = \frac{1}{n_{A,i} \cdot p_{A,i} \cdot (1 - p_{A,i})} \text{ and } s_{B,i}^2 = \frac{1}{n_{B,i} \cdot p_{B,i} \cdot (1 - p_{B,i})}$$

If we treat the observed variance of logit sensitivity and specificity as fixed quantities, a standard approach in meta-analysis, we can write:

$$\begin{pmatrix} \hat\theta_{A,i} \\ \hat\theta_{B,i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{A,i} \\ \theta_{B,i} \end{pmatrix}, C_i\right) \text{ with } C_i = \begin{pmatrix} s_{A,i}^2 & 0 \\ 0 & s_{B,i}^2 \end{pmatrix}$$

The final model then becomes

$$\begin{pmatrix} \hat\theta_{A,i} \\ \hat\theta_{B,i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_A \\ \theta_B \end{pmatrix}, \Sigma + C_i\right)$$

with C_i being a diagonal matrix holding the s_i^2's.

This model can be fitted by likelihood-based methods, in particular the SAS proc mixed procedure because it allows the user to fix the within trial variance at specific values per study [24].

The standard output of the bivariate model includes: mean logit sensitivity (θ_A) and specificity (θ_B) with their standard errors and 95% confidence intervals; and estimates of the between-study variability in logit sensitivity (σ_A^2) and specificity (σ_B^2) and the covariance between them (σ_{AB}).
Based on these parameters, we can calculate other measures of interest:

- The likelihood ratio for positive and negative test results:

$$LR+ = \frac{e^{\theta_A}/(1 + e^{\theta_A})}{1 - \{e^{\theta_B}/(1 + e^{\theta_B})\}} \quad \text{and}$$

$$LR- = \frac{1 - \{e^{\theta_A}/(1 + e^{\theta_A})\}}{e^{\theta_B}/(1 + e^{\theta_B})}$$

- The diagnostic odds ratio defined by: DOR $= e^{(\theta_A + \theta_B)}$
- The correlation between logit sensitivity and specificity: $\frac{\sigma_{AB}}{\sqrt{\sigma_A^2} \cdot \sqrt{\sigma_B^2}}$
- Several summary ROC linear regression lines based on either the regression of logit sensitivity on specificity given by: $\hat\theta_{A|\theta_B} = \bar\theta_A + \frac{\sigma_{AB}}{\sigma_B^2}(\theta_B - \bar\theta_B)$ or the regression of logit specificity on sensitivity, or a orthogonal regression line by minimizing the perpendicular distances. These lines can be transformed back to the original ROC scale to obtain a summary ROC curve.

Appendix 2. Fitting the bivariate model using Proc Mixed

Example of main syntax. A copy of the full SAS program is available on request from the first author.

/* The dataset (*bi_meta*) has one outcome variable (*logit*), but each study has two records. One record contains logit

```
sensitivity, the other logit specificty. We use two indicator
variables to distinguish these measures: dis and non_dis
*/
proc mixed data=bi_meta method=reml cl;
/* study_id and modality are categorical variables */
   class study_id modality;

/* model statement: asking for different estimates of mean
sensitivity and specificity for each modality, provide large
value for degrees of freedom to obtain p-values based on
normal distribution rather than the t-distribution (=default
in SAS) */
   model logit = dis*modality non_dis*modality / noint
   cl df=1000, 1000, 1000, 1000, 1000, 1000;

/* random effects for logit sensitivity and specificity with
possible correlation (UN=unstructured covariance struc-
ture) */
   random dis non_dis / subject=study_id type=un;

/* use the repeat statement to define different within-study
variances for sens and spec in each study */
   repeated / group=rec;

/* name the file holding the all the (co)variances parameters,
keep the within-study variance constant */
   parms / parmsdata=cov hold=4 to 91;

/* use contrast statement for testing specific hypotheses */
/* testing for differences in sensitivities */
   contrast 'CT_sens vs LAG_sens' dis*modality 1 -1 0 /
   df=1000;
   contrast 'CT_sens vs MRI_sens' dis*modality 1 0 -1/
   df=1000;
   contrast 'LAG_sens vs MRI_sens' dis*modality 0 1 -
   1/ df=1000;

/* testing for differences in specificities */
   contrast 'CT_spec vs LAG_spec' non_dis*modality
   1 -1 0 / df=1000;
   contrast 'CT_spec vs MRI_spec' non_dis*modality 1
   0 -1/ df=1000;
   contrast 'LAG_spec vs MRI_spec' non_dis*modality 0
   1 -1/ df=1000;
run;
```

References

[1] Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Publishing Group; 2002. p. 19–38.

[2] Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134(6):587–94.

[3] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003;56(11): 1118–28.

[4] Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981;94(4 Pt 2):557–92.

[5] Habbema JDF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Publishing Group; 2002. p. 117–44.

[6] Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Serv Res 2002;2(1):4.

[7] Walter SD, Jadad AR. Meta-analysis of screening data: a survey of the literature. Stat Med 1999;18(24):3409–24.

[8] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061–6.

[9] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48(1):119–30 [discussion 131–2].

[10] Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk hernia-tion. Methods Inf Med 1990;29(1):12–22.

[11] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001; 20(19):2865–84.

[12] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12(14):1293–316.

[13] Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323(7305): 157–62.

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

[15] Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120(8):667–76.

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

[17] Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making 1993; 13(3):253–7.

[18] Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002;21(9):1237–56.

[19] Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56(11):1129–35.

[20] Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. JAMA 1997;278(13):1096–101.

[21] Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11(2):95–101.

[22] Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. Radiology 1988;167(2):565–9.

[23] Sorribas A, March J, Trujillano J. A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests. Stat Med 2002;21(9):1213–35.

[24] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21(4):589–624.

[25] van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Stat Med 1993;12(24):2273–84.

[26] Kotz S, Balakrishnan N, Johnson NL. Bivariate and trivariate normal distributions. In: Continuous multivariate distributions. New York: Wiley; 2000. p. 251–348.

[27] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138(1):W1–12.

[28] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140(3):189–202.

[29] Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002; 21(11):1525–37.

[30] Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999;18(3):321–59.

[31] Song F. Exploring heterogeneity in meta-analysis: is the L'Abbe plot useful? J Clin Epidemiol 1999;52(8):725–30.

[32] Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999;18(20):2693–708.

[33] van Houwelingen H, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 1999;18(1):110–5.

[34] Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? Stat Med 1987;6(3):351–60.

[35] Deville WL, Buntinx F, Bouter LM, Montori VM, De Vet HC, Van Der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2(1):9.

[36] Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. Acad Radiol 1995;2(Suppl 1): S37–47 [discussion S65–9, S83].

[37] Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2(Suppl 1):S48–56 [discussion S65–7, S70].

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

[39] Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. J Urol 2003;169(6):1975–82.

[40] Bipat S, Glas AS, van der Velden J, Zwinderman AH, Bossuyt PM, Stoker J. Computed tomography and magnetic resonance imaging in staging of uterine cervical carcinoma: a systematic review. Gynecol Oncol 2003;91(1):59–66.

[41] Scholten RJ, Opstelten W, van der Plas CG, Bijl D, Deville WL, Bouter LM. Accuracy of physical diagnostic tests for assessing ruptures of the anterior cruciate ligament: a meta-analysis. J Fam Pract 2003;52(9):689–94.

[42] Koelemay MJ, Nederkoorn PJ, Reitsma JB, Majoie CB. Systematic review of computed tomographic angiography for assessment of carotid artery disease. Stroke 2004;35(10):2306–12.

[43] Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol 2002;31(1):88–95.

[44] Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274(8):645–51.

[45] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003;138(1):40–4.

[46] Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984;252(17):2418–22.

Annotated SAS syntax

Fitting the bivariate model of (logit) sensitivity and specificity using SAS software. The SAS proc mixed procedure is a convenient mixed model application because it allows the user to fix the within trial variance at specific values per trial. Further details can be found here (van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21(4):589-624.)

```
/*
First step: read in the original dataset of 44 studies included in the meta-analysis of Scheidler et al. Each study is represented
by one record. Key variables: study_id=identification number, modality=type of test (1=CT, 2=LAG, 3=MRI), author=sur-
name of first author, and 4 variables holding the number of patients in each of the 4 cells of the 2×2 table: tp=true positive,
fp=false positive, fn=false negative, tn=true negative
*/

data meta ;
    length author $20;
    input study_id author $ year modality tp fp fn tn;

/* correction if one of the 4 cells is empty */
    if tp eq 0 or fp eq 0 or fn eq 0 or tn eq 0 then do;
        tp = tp+0.5; fp = fp+0.5; fn = fn+0.5; tn = tn+0.5;
    end;

/* calculation of sensitivity and specificity */
    sens = tp/(tp+fn); spec = tn /(tn+fp);

/* logit transformation with corresponding variance */
    log_sens = log(sens/(1-sens)); var_log_sens = 1/(sens*(1-sens)*(tp+fn));
    log_spec = log(spec/(1-spec)); var_log_spec = 1/(spec*(1-spec)*(tn+fp));
datalines;
1       Grumbine          1981    1     0     1     6    17
2       Walsh             1981    1    12     3     3     7
3       Brenner           1982    1     4     1     2    13
4       Villasanta        1983    1    10     4     3    25
5       vanEngelshoven    1984    1     3     1     4    12
6       Bandy             1985    1     9     3     3    29
7       Vas               1985    1    20     4     8    31
8       King              1986    1    17     5     7    21
9       Feigen            1987    1     2     0     9    32
10      Camilien          1988    1     3     1     9    38
11      Janus             1989    1     1     1     2    18
12      Matsukuma         1989    1     5     2     2    61
13      Heller            1990    1    21     8    40   184
14      Kim               1990    1     4     3     9    42
15      Ho                1992    1     0     0     5    15
16      Kim               1993    1     7    11    22   158
17      Subak             1995    1     3     3     2    29
18      Kindermann        1970    2    19     1    10    81
19      Lecart            1971    2     8     9     2    13
20      Piver             1971    2    41     1    12    49
21      Piver             1973    2     5     1     2    18
22      Kolbenstvedt      1975    2    45    58    32   165
23      LemanJr           1975    2     8     6     2    32
24      Brown             1979    2     5     8     1     7
25      Lagasse           1979    2    15    17    11    52
26      Kjorstad          1980    2    16    11     8    24
```

```
27    Ashraf      1982    2     4     8     2     25
28    deMuylder   1984    2     8    12    10     70
29    Smales      1986    2    10     4     4     55
30    Feigen      1987    2     2     5     6     23
31    Swart       1989    2     7    10     7     30
32    Heller      1990    2    44    50    12    135
33    Lafianza    1990    2     8     3     1     37
34    Stellato    1992    2     4     3     0     14
35    Hricak      1988    3     9     2     2     44
36    Greco       1989    3     3     6     5     32
37    Janus       1989    3     3     2     1     16
38    Kim         1990    3     3     1    12     44
39    Ho          1992    3     0     0     5     15
40    Kim         1993    3     7     2    22    167
41    Hawnaur     1994    3    12     4     4     29
42    Kim         1994    3    23     5    14    230
43    Subak       1995    3     8     5     5     53
44    Heuck       1997    3    16     2     2     22
;
run;

/*
In the next step we rebuild the original dataset so that each study will now have two records: one record holding (logit)
sensitivity and the other (logit) specificity. Define two indicator variables: dis and non_dis. When a record holds logit sensitivity
the value for dis=1 and for non_dis=0, when a record holds specificity dis=0 and non_dis=1
*/

data bi_meta;
    set meta;

/* creating the record for logit sensitivity */
    dis = 1; non_dis = 0;
    logit = log_sens; var_logit = var_log_sens;
    rec+1;                              /* variable rec uniquely identifies each record */
    output;                             /* writes the record holding sensitivity to the new dataset */

/* creating the record for logit sensitivity */
    dis = 0; non_dis = 1;
    logit = log_spec; var_logit = var_log_spec;
    rec+1; output;
    run;

/*
In the next step we create a special dataset that contains the 44 calculated variances of logit sensitivity and the 44 variances
of specificity in each study. Three additional records are created which contain the starting values for the 3 additional variance
parameters of the bivariate model: record 1 holds the starting value for the between-study variance in logit sensitivity, record
2 the covariance between logit sensitivity and specificity, and record 3 the between-study variance in logit specificity
*/

data cov;
/* start the file with three starting values for (co)variance parameters of the random effects (zero works well)
    SAS expects the name est for this variable */
if _n_ eq 1 then do;
    est = 0; output; est =0; output; est = 0; output;
end;
```

```
/* followed by the 88 calculated variances of sensitivity and specificity of each study */
    set bi_meta;
    est = var_logit; output;
    keep est;
    run;

/*
Use the Proc Mixed module in SAS is used to set up the bivariate model. We apply the approach of Van Houwelingen et al
to incorporate both the within and between study variance, more details and explanations can be found there.{van Houwelingen,
2002 #2} Use the restricted maximum likelihood estimation (REML) method in estimating the model
*/

proc mixed data=bi_meta method=reml cl ;                                  /* option cl will give confidence intervals */
/* study_id and modality are categorical variables */
    class study_id modality;

/* model statement: asking for different estimates of mean sensitivity and specificity for each modality, provide large value
for degrees of freedom to obtain p-values based on normal distribution rather than the t-distribution (=default in SAS) */
    model logit = dis*modality non_dis*modality / noint cl df=1000, 1000, 1000, 1000, 1000, 1000;

/* random effects for logit sensitivity and specificity with possible correlation (UN=unstructured covariance structure) */
    random dis non_dis / subject=study_id type=un ;

/* use the repeat statement to define different within-study variances for sens and spec in each study */
    repeated / group=rec;

/* name the file holding the all the (co)variances parameters, keep the within-study variance constant */
    parms / parmsdata=cov hold=4 to 91;

/* use contrast statement for testing specific hypotheses */
/* testing for differences in sensitivities */
    contrast ’CT_sens vs LAG_sens’ dis*modality 1 -1 0 / df=1000 ;
    contrast ’CT_sens vs MRI_sens’ dis*modality 1 0 -1/ df=1000 ;
    contrast ’LAG_sens vs MRI_sens’ dis*modality 0 1 -1/ df=1000 ;

/* testing for differences in specificities */
    contrast ’CT_spec vs LAG_spec’ non_dis*modality 1 -1 0 / df=1000 ;
    contrast ’CT_spec vs MRI_spec’ non_dis*modality 1 0 -1/ df=1000 ;
    contrast ’LAG_spec vs MRI_spec’ non_dis*modality 0 1 -1/ df=1000 ;

/* testing for differences in DOR */
    contrast ’CT_odds vs LAG_odds ’ dis*modality 1 -1 0 non_dis* modality 1 -1 0 / df=1000 ;
    contrast ’CT_odds vs MRI_odds ’ dis* modality 1 0 -1 non_dis* modality 1 0 -1 / df=1000 ;
    contrast ’LAG_odds vs MRI_odds ’ dis* modality 0 1 -1 non_dis* modality 0 1 -1 / df=1000 ;
run;
```