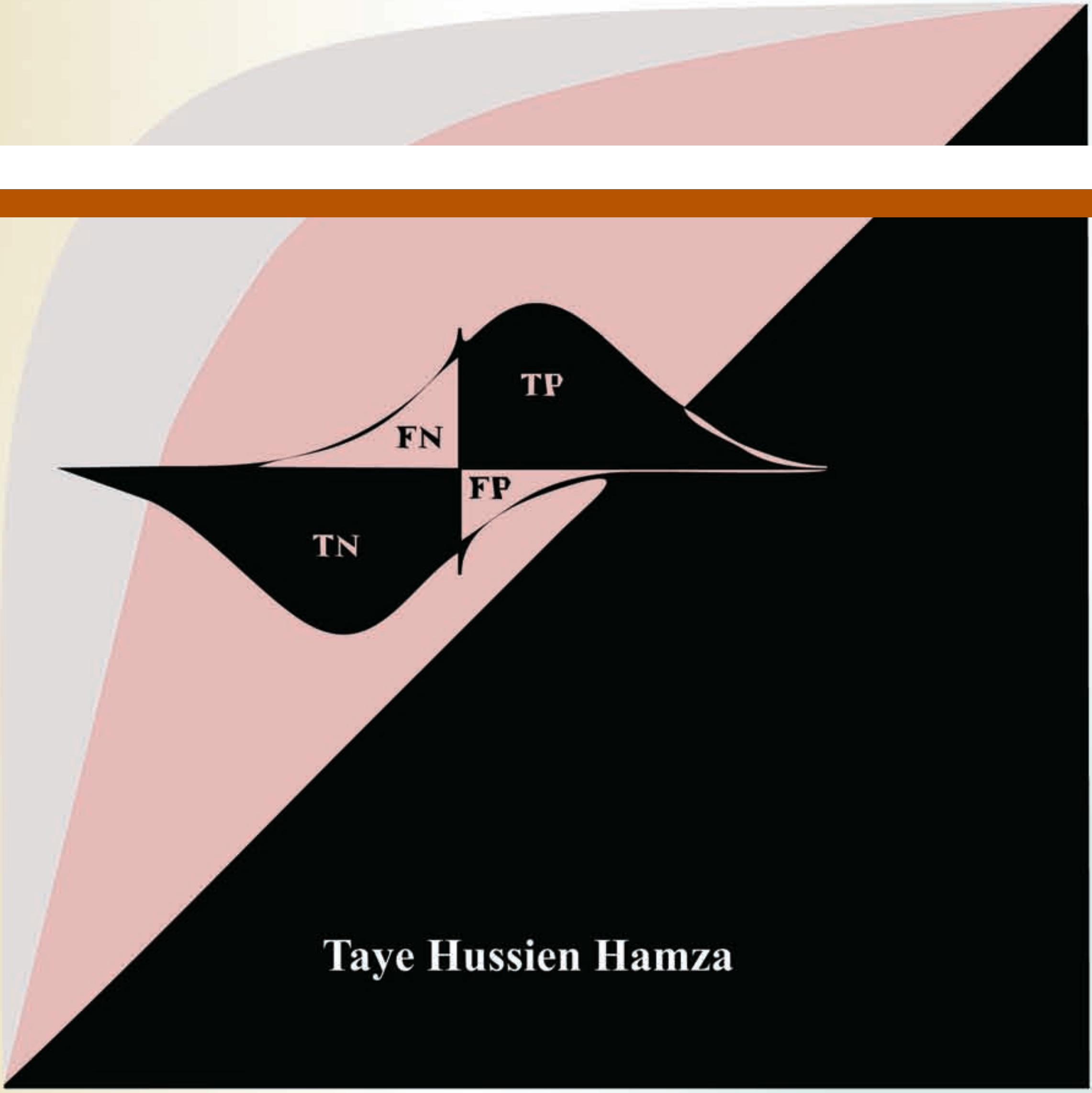
Meta-analysis of Diagnostic Test Evaluation Data: Random Effects Approaches



Taye Hussien Hamza

Meta-analysis of Diagnostic Test Evaluation Data: Random Effects Approaches

Taye Hussien Hamza

Acknowledgment

The publication of this thesis was financially supported by the Department of Epidemiology & Biostatistics of the Erasmus MC, Erasmus University Rotterdam.

Cover design: Dr. Wendimagegn Ghidey
Printed by: Haveka BV, Alblasserdam

ISBN 978-90-9023002-3

Meta-analysis of Diagnostic Test Evaluation Data: Random Effects Approaches

Meta-analyse van diagnostische test evaluatie data: random effecten modellering

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 28 mei 2008 om 15:45 uur

door

Taye Hussien Hamza

geboren te Worailu, Ethiopia

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotoren :	Prof.dr. T. Stijnen
	Prof.dr. J.C. van Houwelingen
Overige leden :	Prof.dr. M.G.M. Hunink
	Prof.dr. E.W. Steyerberg
	Prof.dr. G. Knapp

Contents

Manuscripts based on Studies described in this thesis

Chapter 2

Taye H. Hamza, Hans C. van Houwelingen, Theo Stijnen
The binomial distribution of meta-analysis was preferred to model within-study variability. Journal of Clinical Epidemiology. 2008; 61(1):41-51.

Chapter 3

L.R.Arends, T.H.Hamza, J.C. van Houwelingen, N.H.Heijenbrok-kal, M.G.M. Hunink, T.Stijnen
Bivariate Random Effects Meta-analysis of ROC curves. Medical Decision Making. (to appear)

Chapter 4

Taye H. Hamza, Johannes R. Reitsma, Theo Stijnen
Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal and binomial-normal bivariate summary ROC approaches. Medical Decision Making. (accepted)

Chapter 5

Taye H. Hamza, Hans C. van Houwelingen, Majanka H. Heijenbrok-Kal, Theo Stijnen
Bivariate random-effects meta-regression of diagnostic tests: an application. (submitted)

Chapter 6

Taye H. Hamza, Hans C. van Houwelingen, Theo Stijnen
Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. (submitted)

To Serkalem Abebe
and
Hussien Hamza

CHAPTER 1

Introduction

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies [1]. When several studies have heterogeneous or even conflicting conclusions, a meta-analysis can be used to estimate an average effect or identify a subset of studies associated with a beneficial effect [1]. Although meta-analysis is widely used in epidemiology and evidence-based medicine today, a meta-analysis regarding a medical treatment was not published until 1955. The term 'meta-analysis' was introduced in 1976 [2], and since then its application has rapidly grown. In the medical world the increase began in the 1980's [3]. The foundation of the Cochrane Collaboration (in 1993), an international network of health care professionals who prepare and regularly update systematic reviews, facilitated the conduct of meta-analyses in all areas of health care. Most meta-analyses within the field of medical research have been conducted on randomized controlled trials, epidemiologic studies or diagnostic tests [4]. In this thesis we focus on the meta-analysis of diagnostic tests.

In recent years, the need for systematic reviews and synthesis of published evidence on the accuracy of diagnostic tests has increased [5, 6]. The main objective of a diagnostic review is to summarize the evidence regarding the accuracy of a test or instrument [7]. Many studies may have too small numbers of patients to give precise estimates or are too selected to allow a general applicability. This means that the conclusion regarding the diagnostic accuracy of test may differ among studies. The meta-analytic approach may overcome this problem by combining studies evaluating the same diagnostic technology. The information from such reviews is a key element in clinical and health policy decision making regarding the use of diagnostic tests; it is also essential for guiding the process of technology development and evaluation in diagnostic medicine [6].

1.1 Measures of Diagnostic test accuracy

In diagnostic accuracy studies the true disease status of patients is determined using a reference standard. Often it is assumed that the disease status is classified without error, the reference being a gold standard. Test errors can be characterized by various measures [8]. The two most commonly reported measures are sensitivity, the probability that a test result is positive in patients with the disease of interest, and specificity, the probability that a test result is negative in patients without the disease of interest. For a quantitative test, an alternative to reporting a single pair of sensitivity and specificity estimates is to report a range of pairs obtained by varying the threshold criterion [8]. Such a range of pairs is often presented as a receiver operating characteristic (ROC) curve (see section 1.2 for more discussion on ROC curves). Other measures of diagnostic accuracy are the diagnostic likelihood ratio, defined as the ratio of the probability of a particular test result in people with the disease to the probability of the same test result in people without the disease; positive predictive value, defined as the proportion of patients with positive test results who are correctly diagnosed; and negative predictive value, defined as the proportion of patients with negative test results who are correctly diagnosed.

1.2 Summary Receiver Operating Characteristic (ROC) curve

In studies of quantitative diagnostic tests and a dichotomous reference standard (disease present / absent), the relation between sensitivity and specificity can be represented by the receiver operating characteristic (ROC) curve. The ROC curve is a monotone increasing function on the unit square. It starts at the point (0,0), corresponding to a sensitivity of 0% and a specificity of 100%, and ends at the point (1,1), corresponding to a sensitivity of 100% and a specificity of 0%. It is made by calculating the pairs of sensitivity and specificity corresponding to all possible positivity thresholds of the diagnostic test. In study reports the whole ROC curve is almost never presented, and the usual situation is that per study in the meta-analysis only one or sometimes a few pairs of sensitivity and specificity, corresponding with one or more points on the study specific ROC curve, are available. A key goal in the synthesis of such studies is to derive summary measures of test performance. In particular the goal is often to produce a summary ROC curve that shows the trade-off between sensitivity and specificity as the threshold for positivity varies along some explicit or latent scale. For the last ten or fifteen years, the standard method to obtain such a summary ROC curve is the SROC method of Littenberg and Mosses [9, 10].

1.3 Meta-analytic approaches for diagnostic accuracy studies

Statistical methods for meta-analysis of diagnostic test evaluations depend on the type of data available from the different studies. The test results in the primary studies can be presented in two categories, in more than two categories or on a continuous scale. In this thesis we focused on the case where test results are presented in two categories (one threshold) or a few categories (more than one threshold).

In the last decades statistical methods were developed to meta-analyse such kind of data. When the objective is to pool sensitivity and/or specificity, the easiest and most straightforward approach is taking the average across studies, possibly with a weight depending on the within-study sample sizes or standard errors. This approach assumes that the only source of variation is sampling error. In such cases the homogeneity assumption across studies is satisfied, leading to what is called a fixed effect analysis. However, as many authors pointed out (for example [8, 11]) this approach is almost never appropriate, for many reasons. Even with the more sophisticated random effects approaches (see for example Chapter 2), pooling sensitivities and specificities separately is appropriate if and only if the correlation between sensitivity and specificity is negligible, which is usually not the case. One of the reasons is that different studies may use different test cutoff points (thresholds), implicitly or explicitly, and this needs to be taken into account. The effect of shifting the threshold can be more conveniently shown by a summary ROC curve [9, 10]. Littenberg and Moses, and Moses *et al.* [9, 10] were the first to introduce a summary ROC curve approach, which is usually referred as the SROC method. They accounted for sampling error and threshold values; however in practice other sources of variation can contribute as well to the between-studies variability, for example the difference in patient selection and clinical setting, the type of test used, problems in the verification of test results, or any combination of these factors [8, 11]. Possibly some of the heterogeneity can be explained by using meta-regression: an extension of meta-analysis which examines the relationship with one or more study level characteristics (covariates) through regression modeling. But mostly not all heterogeneity can be explained by covariates and hence a random effects model should be used that allows between-studies heterogeneity beyond what is explained by covariates [12, 13]. The SROC method has several drawbacks as pointed out by different authors (for example [5, 14]). In this thesis we introduce meta-analytic approaches that address these shortcomings. We follow the general approach described in van Houwelingen *et al.* [15] and van Houwelingen *et al.* [16]. Our models can be fitted using General(ized) Linear Mixed Model programs of widely used statistical packages, for example SAS (NL)MIXED, R/S-Plus (n)lme or STATA GLLAMM. We also consider the extension of the models with study level covariates and the extension to the case where we have more than one point per study.

1.4 Aims and outline of the thesis

The aim of this thesis is to develop new statistical methods, improve existing ones and compare their performances through simulation. In the simulation studies we consider different scenarios by varying different parameters such as number of studies included in the meta-analysis, within-study size, and probability of test positivity. Chapters 2 to 5 consider meta-analysis where one pair of sensitivity and specificity is available, and in chapter 6 we focus on meta-analyses with more than one point per study.

In the second chapter we study pooling sensitivities or specificities separately, or in general a proportion, using a random effects meta-analysis approach. Approximate normal and exact binomial likelihood approaches are considered for the within-study distribution of the observed data. The two approaches are compared through a simulation study, in terms of bias, mean-squared error, and coverage probabilities (Wald type and profile likelihood).

In chapter three, a bivariate random effects approach to meta-analysis of sensitivity and specificity is discussed for the case where one point per study is available. Several results can be derived from the fitted model. Noticeably, with different assumptions, it leads to different possible ROC curves. Also study specific ROC curves can be derived through a random intercept model using the empirical Bayes estimates. For the within-study distribution of the observed data, both approximate normal and exact binomial likelihood approaches are discussed.

In the fourth chapter, we consider three random effects approaches that can be seen as extensions of the SROC method of Littenberg and Moses [9, 10]: random intercept, normal-normal bivariate and normal-binomial bivariate random effects approach. We compare the three methods through an extensive simulation in order to give guidelines for practitioners.

Chapter five discusses the extension of the bivariate approach with covariates at the hand of a case study. We show how the bivariate approach can be used to study the effect of study level characteristics on different outcome measures, such as sensitivity and specificity and SROC curve. We also show how to compare differences of the study level characteristics on diagnostic test accuracy, for any given outcome measure.

The sixth chapter extends the bivariate approach to the situation where there is more than one point per study. We propose a multivariate random effects model to analyze meta-analysis data with a fixed number of thresholds across studies. The method can be fitted in standard statistical packages such as SAS.

A general discussion of the thesis is given in chapter seven. We discuss the advantages of the approaches over the standard (SROC curve) approaches and their limitations.

References

[1] Normand SL. Meta-analysis: Formulating, evaluating, combining and reporting. Statistics in Medicine. 1999; 18:32-359.

[2] Glass GV. Primary, Secondary and Meta-analysis of research. Educational Research. 1976; 5:3-8.

[3] Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? International Epidemilogical Association. 2002; 31:1-5.

[4] Whitehead A. *Meta-analysis of Controlled Clinical Trials*. John Wiley & Sons Ltd; 2003.

[5] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 2001; 20:2865-2884.

[6] Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screeening test accuracy evaluations: Methodological Primer. AJR. 2006; 187:271-281.

[7] Pai M, McCulloch, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: What's behind the scenes? ACP Journal Club. 2004; 141: A-11.

[8] Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses Evaluating Diagnostic Tests. Annals of Internal Medicine. 1994; 120(8): 667-676.

[9] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conflicting reports - a new meta-analytic method. Medical Decision Making. 1993; 13:313-321.

[10] 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. Statistics in Medicine. 1993; 12:1293-1316.

[11] Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Statistics in Medicine. 2002; 21:1525-37.

[12] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted. Statistics in Medicine. 2002; 21:1559-73.

[13] Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine. 1998; 17:841-56.

[14] Arends LR, Hamza TH, van Houwelingen, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate Random Effects Meta-analysis of ROC curves. Medical Decision Making. 2008 (to appear).

[15] van Houwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine. 1993;12:2272-2284.

[16] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: Multivariate approach and meta-regression. Statistics in Medicine. 2002;21:589-624.

CHAPTER 2

The Binomial Distribution of Meta-analysis was Preferred to Model Within-Study Variability

Abstract

Objective: When studies report proportions such as sensitivity or specificity, it is customary to meta-analyze them using the DerSimonian and Laird random effects model. This method approximates the within-study variability of the proportion by a normal distribution, which may lead to bias for several reasons. Alternatively an exact likelihood approach based on the binomial within-study distribution can be used. This method can easily be performed in standard statistical packages. We investigate the performance of the standard method and the alternative approach.
Study Design and Setting: We compare the two approaches through a simulation study, in terms of bias, means squared error and coverage probabilities. We varied the size of the overall sensitivity or specificity, the between-studies variance, the within-study sample sizes and the number of studies. The methods are illustrated using a published meta-analysis data set.
Results: The exact likelihood approach performs always better than the approximate approach and gives unbiased estimates. The coverage probability, in particular for the profile likelihood, is also reasonably acceptable. In contrast, the approximate approach gives huge bias with very poor coverage probability in many cases.
Conclusion: The exact likelihood approach is the method of preference and should be used whenever feasible.

2.1 Introduction

In this paper we consider meta-analysis of proportions. Very frequently occurring examples of proportions being meta-analyzed are sensitivity or specificity of a diagnostic test. Therefore this article is written from a diagnostic research perspective, though the results apply to meta-analysis of proportions in general, such as prevalences or incidences.

Meta-analytic methods for a diagnostic test depend on the type of data that is available from the different studies. In most medical articles, the commonly reported measures of diagnostic test accuracy are sensitivity and/or specificity. Alternatively, other measures such as diagnostic odds ratio, predictive values, area under the receiver operating characteristic (ROC) curve, are reported.

Statistical methods to pool the results of diagnostic test measures from different studies lay on different assumptions. For example it might be assumed that the observed differences between individual study results are only due to sampling variation, leading to what is called a fixed effect analysis. When an estimate of the sensitivity or specificity is reported in a single study, the simplest method to get a summary measure is to calculate the average sensitivity and/or specificity, possibly with weights depending on the within-study sample sizes or standard errors. However, this approach is usually inappropriate because it is likely that variability beyond chance can be attributed to between-study differences [1, 2]. Some of the between-study variability could be accounted for using explanatory variables in a regression analysis. But mostly not all heterogeneity can be explained and a random effects model is used in the statistical analysis, which allows between-studies heterogeneity [3, 4].

In the last decade many random effects methods have been developed to relax the fixed effect assumptions in meta-analysis [5, 6, 7, 8] of diagnostic tests [9, 10]. Some of these methods enable analyzing sensitivity and specificity jointly. However, in the medical literature numerous meta-analyses are published in which one is interested in meta-analyzing only sensitivity or specificity, and in this paper we concentrate on this situation. Then the standard way of analysis is with the DerSimonian and Laird [6] random effects model. It is not well known that this method can be heavily biased when it is applied to proportions, such as specificities or sensitivities, though some authors have mentioned this [5, 11, 12]. Chang *et al* [11] have proposed a method that repairs the bias. However, this article has been cited only once since the year 2001, showing that in practice this method is not used. It might be due to the difficulty to perform the method easily in standard statistical packages. The reason for the standard method being biased is that the binomial within-study likelihood of the sensitivity or specificity is approximated using a normal likelihood. It is well known that this approximation can be bad if the proportion is close to one or zero, and/or the sample size is small. So bias can be expected if this is the case in a meta-analysis. However, even if the normal approximation would be good enough for ordinary applications, bias could be introduced since the use of the normal approximation in meta-analysis ignores the correlation between the estimated proportion and its variance. We come back to this point in the next section. Nowadays standard statistical packages allow for fitting generalized linear mixed models (G*ized*LMM). This makes it very easy to use the exact binomial within-study distribution of the estimated sensitivity or specificity instead of a normal approximation of it. In this article we call the latter the approximate method and the former the exact method.

The purpose of this article is to compare the performance of the two modeling approaches, approximate and exact, through a simulation study. In Section 2.2 both methods are discussed. In Section 2.3 we describe the design of the simulation study and in Section 2.4 we present the results. In Section 2.5 we apply the methods on real meta-analysis data. We end with a discussion in Section 2.6. We used SAS software (version 9.1) to simulate the data and to estimate the parameters for the models discussed in Section 2.2.

2.2 Random Effects Model

In a situation where the interest is to meta-analyze sensitivities or specificities separately, the commonly used method is the DerSimonian and Laird [6] random effects model. In the remainder of this paper we will talk about meta-analyzing sensitivities, but all the results apply to specificities as well. In fact, the results apply to any meta-analysis where the target parameter is a proportion or probability and each study contributes a sample size and a number of "successes". Unlike a fixed effect model, a random effects model allows that sensitivities vary across studies beyond that expected from within-study sampling variability alone. More specifically, the true logit sensitivities, η_i, defined as $ln(sensitivity/(1-sensitivity))$, are assumed to follow a normal distribution:

$$\eta_i \sim N(\eta, \tau^2)$$

Here i denotes the number of study, η the true mean logit sensitivity. The parameter τ^2 is called the between-studies variance and it describes the variability between the true logit sensitivities of the different studies. The within-study sampling variability could be modeled by using the approximate normal likelihood or the exact binomial likelihood for the observed number of positive test results.

2.2.1 Approximate method

This is the standard method in practice. Different transformations of the observed proportion, such as the probit, $log(-log))$ or the $arcsine$ could be used and approximated by a normal distribution. In this paper we have chosen the logit transformation, since it is the predominant choice in practice. If m_i is the total number of subjects with the disease of study i, and x_i is the observed number of true positive test results in the group with the disease, then the observed logit sensitivity, $\hat{\eta}_i = ln(x_i/(m_i - x_i))$ is assumed to follow an approximate normal distribution with mean η_i, and within-study variance calculated from the observed data:

$$\hat{\eta}_i \sim N(\eta_i, \hat{\sigma}_i^2) \quad \text{with} \quad \hat{\sigma}_i^2 = \frac{1}{x_i} + \frac{1}{m_i - x_i}$$

If x_i or $m_i - x_i$ is zero the logit sensitivity and the within-study variance will be undefined. To avoid this problem 0.5 should be added to x_i and $m_i - x_i$ for all studies, including those with no zero [13, 14]. The effect of adding 0.5 may bias the results [14]. Further, usually there is a high correlation between $\hat{\eta}_i$ and $\hat{\sigma}_i^2$. The correlation is positive if η is positive and negative if η is negative, leading to a bias towards zero in the estimate of the overall logit-sensitivity, or a bias towards 0.5 in the estimate of the overall sensitivity. This is because both the mean and the variance of $\hat{\eta}_i$ are determined by the same parameter. The effect of this correlation in a random effects meta-analysis was discussed by several authors [5, 11, 12]. Though they suggested a correction to reduce the bias in the estimate of η due to this correlation, Chang *et al* [11] mentioned that the estimated between-studies variance remained still biased even if we let grow the number of studies included. Using the approximate normal likelihood for the within logit sensitivity, the model is a linear random effects model and the parameters can be estimated by standard likelihood procedures using a Linear Mixed Model (LMM) program, which is available in many standard statistical packages. For example, in SAS the procedure MIXED, in S-Plus/R the function lme can be used. As discussed by Turner *et al* [15] three methods, maximum likelihood (ML), restricted (residual) maximum likelihood (REML) and the method of moments proposed by DerSimonian and Laird, are available to estimate the random effects model. They differ mainly on the estimation of the between-studies variance, in which ML gives a downward bias. In the biomedical literature, usually the method of moments proposed by DerSimonian and Laird or the REML estimator is used for estimating the heterogeneity parameter [16]. The REML estimator is the iterative equivalent of the DerSimonian and Laird estimator and gives very similar results [6]. In this paper we used the REML method, which can be specified in the method option of the MIXED or lme procedures.

2.2.2 Exact method

Here we use the fact that the observed number of true positive test results x_i follows a binomial distribution:

$$x_i \sim binomial(\pi_i, m_i)$$

where $\pi_i = 1/(1+e^{-\eta_i})$ and m_i is the total number of subjects with the disease of study i. In this case there is no need to add 0.5 even if a zero count is encountered and there is no problem anymore of correlation between the observed values and their variance because there is no more need of calculating the $\hat{\eta}_i$ and its within-study variance, $\hat{\sigma}^2_\eta$ from the observed data and approximating by a normal distribution. Now the variance is inherent from the binomial distribution. In the approximate method, the distribution of x_i is basically approximated by a normal distribution with mean $m_i\pi$ and variance $m_i\pi(1-\pi)$, which is estimated by a normal distribution with mean $m_i\hat{\pi}$ and variance $m_i\hat{\pi}(1-\hat{\pi})$. Clearly the estimated mean and variance are correlated, and, since the variance estimate is treated as a fixed and known number, this correlation is not modeled, which causes bias in the approximate method.

Now the model is a G*ized*LMM and the parameters can be estimated by standard likelihood procedures. The practical disadvantage is that software is much more scarce and not yet available in all statistical packages. We used the NLMIXED procedure from the SAS package [17]. It is also possible to use the recently included GLIMMIX procedure in the SAS package, which is still experimental in SAS version 9.1. The GLIMMIX procedure allows more random effects but it has the disadvantage that it uses an approximation instead of the true log likelihood. In contrast, although the number of random effects that can be practically managed is limited, NLMIXED uses very accurate integrating techniques to calculate the true likelihood. Unlike the MIXED procedure, the NLMIXED procedure only implements ML. This is because the analog to the REML method in NLMIXED would involve a high dimensional integral over all of the fixed effects parameters and this integral is typically not available in closed form [17]. Hence we used the ML method for the exact approach.

The exact binomial likelihood approach as used here leads to a logistic regression model with a random intercept, and is therefore analogous to the 'individual patient data methods (IPD)' as used by Turner *et al* [15]. They consider meta-analysis of treatment effect log odds ratios and used the MLwiN software [18] to fit the model. In contrast to NLMIXED, this program uses an approximation of the likelihood instead of the true one. Turner *et al* [15] also suggested the use of bootstrapping techniques to estimate τ^2, but this approach is not implemented in NLMIXED and was not incorporated in our simulation study. In the appendix we have given the syntax needed to fit the models following the approximate and the exact methods.

2.3 Simulation Study

A simulation study was carried out to compare the performance of the two methods, approximate and exact, discussed in Section 2.2. We investigated the effect of the number of studies included in the meta-analysis, the mean within-study sample size, the between-study variability and the true median sensitivity. The data were simulated in two steps. First, the true logit sensitivity, η_i, was simulated from a normal distribution with a given mean logit sensitivity η and between-studies variance τ^2. Secondly, the within-study data were simulated from a binomial distribution with a probability $\pi_i = 1/(1 + e^{-\eta_i})$ and within-study sample size m_i. In practice the m_i's vary across studies included in the meta-analysis. In some meta-analyses the range of the size of studies is as big as 1500 or more (for example [19, 20]). To accommodate this variation, the m_i's were generated from a normal distribution and rounded to the nearest integer. Two different vales were considered for the mean m_i (standard deviation): 40 (30) and 500 (450). The minimum study size was set to be 10, that is, if the generated m_i was less than 10 then it was taken to be 10. Consequently 40 and 500 are no more the means for the simulated sample sizes but the medians and the realized standard deviation becomes a bit smaller than 30 and 450 respectively. We considered 12 different situations and for each situation we did the simulation assuming a different number of studies (10, 25, 50, 100) included in the meta-analysis at a time, i.e. we simulated 48 scenarios in total. All values assigned to the parameters in the different scenarios were based on real data sets from the medical literature (for example [21, 22, 23]). An overview of the simulated scenarios is given in Table 2.1.

Table 2.1: The different scenarios used in the simulation study. Each subset of 4 scenarios corresponds to 10, 25, 50 and 100 studies in the meta-analysis. The corresponding true median sensitivity for 0.41, 1.11 and 2.57 are 0.60, 0.75 and 0.93 respectively.

Scenario	η	τ^2	m	Standard deviation of m
1-4	0.41	0.3	40	30
5-8	0.41	0.3	500	450
9-12	0.41	1.0	40	30
13-16	0.41	1.0	500	450
17-20	1.11	0.3	40	30
21-24	1.11	0.3	500	450
25-28	1.11	1.0	40	30
29-32	1.11	1.0	500	450
33-36	2.57	0.3	40	30
37-40	2.57	0.3	500	450
41-44	2.57	1.0	40	30
45-48	2.57	1.0	500	450

Each scenario was replicated 1000 times and the simulated data sets were analyzed according to the approximate and exact approach using the SAS procedures MIXED and NLMIXED respectively. We concentrated on the estimation of the mean logit sensitivity η and between-study variance τ^2. The estimated results were compared using bias (difference between the mean estimate and the true value of the parameter), mean squared error (MSE) and coverage probability of the 95% confidence interval (the frequency in which the true value falls in the confidence interval). In meta-analysis, mostly Wald type confidence intervals are used. For η, the Wald type confidence interval is $\hat{\eta} \pm 1.96 * se(\hat{\eta})$. A disadvantage of this confidence interval might be that when the number of studies is small, the standard error of $\hat{\eta}$ is underestimated due to the fact that the uncertainty in the estimate of τ^2 is not accounted for. This problem may be solved using a profile likelihood based confidence interval [24], which is also discussed by several authors in the context of meta-analysis (for example, [8, 25, 26, 27, 28]). Turner *et al* [15] also discussed a bootstrapping technique to provide confidence intervals for η and τ^2. This method is directly available in the MLWin software [18]. Recently, Knapp *et al* [16] proposed a new approach for a confidence interval of the heterogeneity parameter. In this article we restricted to the Wald and profile likelihood approaches.

The profile log likelihood of η is defined as $pl(\eta) = max_{\tau^2} l(\eta, \tau^2)$ where $l(\eta, \tau^2)$ is the log likelihood for η and τ^2. The 95% profile likelihood CI for η is then given by all values that satisfy $pl(\eta) > pl(\hat{\eta}) - 1.92$ (1.92 is the 95% percentile of the χ^2_1 distribution (3.84) divided by two). The Wald-type CI for τ^2 was calculated through a logarithmic transformation as $\hat{\tau}^2 exp(\pm 1.96 SE(\hat{\tau}^2)/\hat{\tau}^2)$. In a similar way as for η, we also calculated a profile likelihood CI for τ^2. The profile likelihood for both approaches is based on ML only, because the likelihood ratio test statistics computed directly from REML for the fixed effect parameter may not be valid [29, 30]. This type of CI cannot be automatically done in NLMIXED and needs some extra programming. For example, to find the profile likelihood CI of η, first the model is fitted in NLMIXED, and the ML value while estimating both η and τ^2 is saved. The value of the profile likelihood for a given η-value is calculated by running NLMIXED keeping η fixed to this value. Then the two η-values with profile likelihood equal to $ML - 1.92$ were found iteratively using the bisection method, using a self-written macro.

2.4 Simulation Results

The results from the simulations are presented in Figure 2.1(a), 2.1(b), 2.2(a) and 2.2(b), and Table 2.2. Figure 2.1(a) and 2.1(b) shows the biases and MSEs for the mean logit sensitivity η. It can be seen from Figure 2.1(a) that the exact likelihood approach yields estimates of η that are quite unbiased regardless of the different scenarios; i.e. the expected value of the estimated η using the exact method is almost equal to the true value, and always closer to the true value than the approximate likelihood method. The bias in the approximate method varies considerably with the within-study sample size and true mean logit sensitivity. It increases dramatically when the within-study sample size is smaller and the true median sensitivity is larger. For example, when the median within-study sample size is 40 and the median sensitivity is 0.93, the estimated logit sensitivity is biased downward by about 35% or more regardless of the true between-studies variance and the number of studies included. Concerning the effect of between-studies variances, the larger the between-studies variance the more the approximate method underestimates logit sensitivity. However, the number of studies included in the meta-analysis does not make much difference on the bias.

The MSEs in Figure 2.1(b) are corrected for the number of studies included in the meta-analysis by N/100, i.e. the vertical axis in Figure 2.1(b) is MSE*N/100. Comparing the two approaches in terms of the MSEs, the approximate method tends to be worse (larger MSE) for the scenarios with small within-study sample size and medium median sensitivity, or large median sensitivity, in which case the difference in the MSEs between the two methods increases with the number of studies included in the meta-analysis. The constant MSE*N/100 in the figure indicate the fact that the MSE, before multiplying by N/100, decreases with the number of studies included in the meta-analysis regardless of the methods used and the different scenarios.

The coverage probabilities based on the Wald type confidence intervals and the profile likelihood confidence intervals are presented in Table 2.2. Regardless of the method used to construct the confidence intervals, the exact method performs better than the approximate almost always. When N=10, the coverage probabilities of the approximate method are particularly bad for larger sensitivity and small within-study sample size. The coverage probabilities of the exact method are quite reasonable except for the Wald type confidence interval. The intervals based on the profile likelihoods give a valuable improvement upon the Wald type intervals. When $N \geq 25$, not much difference is observed between the two methods of constructing the confidence interval: Wald and profile likelihood. In this case the profile likelihood intervals improve the coverage probability only slightly. In all scenarios the profile likelihood confidence interval of the exact method behaves very satisfactorily. Generally, the approximate likelihood method performed only slightly worse than the exact in the case of small sensitivity (0.60). For the scenarios with large sensitivity (0.93), the coverage percentages are dramatically bad. For some scenarios the coverage dropped down to even 0 percent. For the intermediate value of the sensitivity the approximate method is considerably worse than the exact method when the within-study sample size is small.

Table 2.2: Coverage probabilities for η and τ^2 using approximate and exact methods.

True parameter values				η				τ^2			
				Approximate		Exact		Approximate		Exact	
N[1]	η	τ^2	n[2]	Wald	PL[3]	Wald	PL	Wald	PL	Wald	PL
10	0.41	0.30	40	0.91	0.93	0.91	0.93	1.00	0.94	0.99	0.96
			500	0.91	0.92	0.89	0.92	0.93	0.92	0.90	0.92
		1.00	40	0.93	0.94	0.92	0.94	0.94	0.88	0.94	0.92
			500	0.91	0.93	0.90	0.93	0.93	0.93	0.92	0.94
	1.11	0.30	40	0.90	0.92	0.91	0.93	0.99	0.95	0.98	0.96
			500	0.89	0.91	0.88	0.91	0.92	0.91	0.89	0.91
		1.00	40	0.90	0.92	0.92	0.93	0.96	0.88	0.94	0.91
			500	0.92	0.93	0.91	0.93	0.93	0.92	0.91	0.94
	2.57	0.30	40	0.70	0.79	0.93	0.94	0.84	0.96	0.88	0.96
			500	0.90	0.91	0.90	0.93	0.95	0.92	0.92	0.92
		1.00	40	0.67	0.75	0.90	0.92	0.96	0.80	0.98	0.95
			500	0.90	0.92	0.90	0.92	0.93	0.91	0.90	0.92

[1] number of studies included in the meta-analysis

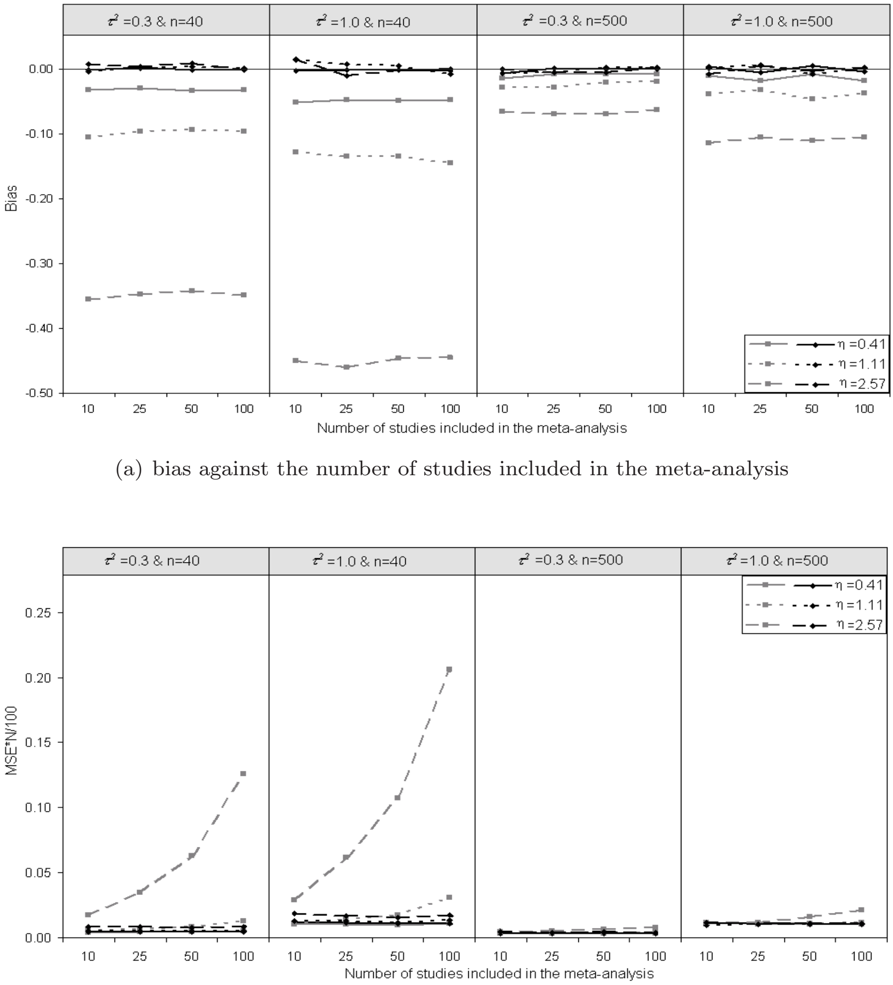
[2] median within-study sample size

[3] profile likelihood

25	0.41	0.30	40	0.92	0.93	0.93	0.93	0.97	0.89	0.96	0.93
			500	0.94	0.94	0.94	0.94	0.95	0.94	0.93	0.95
		1.00	40	0.94	0.95	0.95	0.96	0.92	0.89	0.95	0.95
			500	0.93	0.93	0.92	0.93	0.94	0.94	0.93	0.94
	1.11	0.30	40	0.87	0.90	0.94	0.95	0.99	0.87	0.98	0.93
			500	0.92	0.93	0.91	0.92	0.93	0.93	0.93	0.93
		1.00	40	0.88	0.90	0.92	0.93	0.87	0.84	0.92	0.92
			500	0.93	0.94	0.95	0.95	0.93	0.93	0.93	0.94
	2.57	0.30	40	0.34	0.43	0.93	0.94	0.93	0.82	0.97	0.96
			500	0.90	0.91	0.92	0.93	0.94	0.93	0.93	0.94
		1.00	40	0.39	0.46	0.94	0.94	0.80	0.64	0.96	0.94
			500	0.91	0.91	0.93	0.94	0.93	0.93	0.93	0.93
50	0.41	0.30	40	0.94	0.94	0.95	0.95	0.93	0.88	0.96	0.94
			500	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.94
		1.00	40	0.94	0.94	0.95	0.95	0.82	0.80	0.95	0.95
			500	0.95	0.95	0.94	0.94	0.93	0.93	0.94	0.94
	1.11	0.30	40	0.79	0.81	0.94	0.94	0.91	0.84	0.95	0.94
			500	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.95
		1.00	40	0.82	0.83	0.95	0.95	0.78	0.74	0.95	0.95
			500	0.94	0.95	0.94	0.95	0.94	0.93	0.95	0.96
	2.57	0.30	40	0.06	0.09	0.96	0.96	0.97	0.62	0.97	0.94
			500	0.87	0.87	0.93	0.93	0.93	0.93	0.94	0.95
		1.00	40	0.10	0.14	0.94	0.94	0.46	0.40	0.95	0.94
			500	0.88	0.88	0.95	0.95	0.92	0.92	0.94	0.94
100	0.41	0.30	40	0.90	0.91	0.94	0.95	0.84	0.81	0.95	0.94
			500	0.95	0.95	0.95	0.95	0.93	0.93	0.94	0.94
		1.00	40	0.92	0.92	0.94	0.94	0.71	0.68	0.95	0.95
			500	0.95	0.95	0.95	0.95	0.92	0.92	0.94	0.94
	1.11	0.30	40	0.66	0.67	0.95	0.95	0.80	0.74	0.95	0.94
			500	0.92	0.92	0.95	0.95	0.93	0.93	0.94	0.94
		1.00	40	0.70	0.71	0.94	0.94	0.58	0.55	0.95	0.95
			500	0.93	0.93	0.93	0.93	0.92	0.92	0.94	0.95
	2.57	0.30	40	0.00	0.00	0.95	0.95	0.64	0.36	0.97	0.95
			500	0.79	0.80	0.95	0.95	0.93	0.92	0.94	0.94
		1.00	40	0.00	0.01	0.95	0.95	0.15	0.13	0.94	0.94
			500	0.82	0.82	0.95	0.95	0.89	0.89	0.96	0.96

Similarly, the estimation of the between-studies variance is investigated using the bias, MSE and coverage probabilities. The simulation results are presented in Figure 2.2(a) and 2.2(b) and Table 2.2. Broadly speaking, the results are analogous to the estimates of the mean logit sensitivity. From Figure 2.2(a) it appears that the exact method always performs better than the approximate method in terms of bias, especially in the scenarios with intermediate and high value of the sensitivity. In scenarios with a larger number of studies (50-100) the exact method is practically unbiased, but when the number of studies included is small and the true between-study variance is large the exact method underestimates the between-study variance by about 10%. The pattern of the simulation results for the MSEs (Figure 2.2(b)) is also similar to that of the mean logit sensitivity.

The coverage probabilities are presented in Table 2.2. In some scenarios the Wald based coverage probabilities are greater than the nominal level. For $N = 10$,



(a) bias against the number of studies included in the meta-analysis

(b) MSE*N/100 against the number of studies included in the meta-analysis

Figure 2.1: Simulation results for η. The bias (Figure 2.1(a)) and MSE (Figure 2.1(b)) are given for the approximate likelihood method, gray lines, and the exact likelihood method, black lines. The true between-studies variance and median within-study sample size are given at the top of each plot.

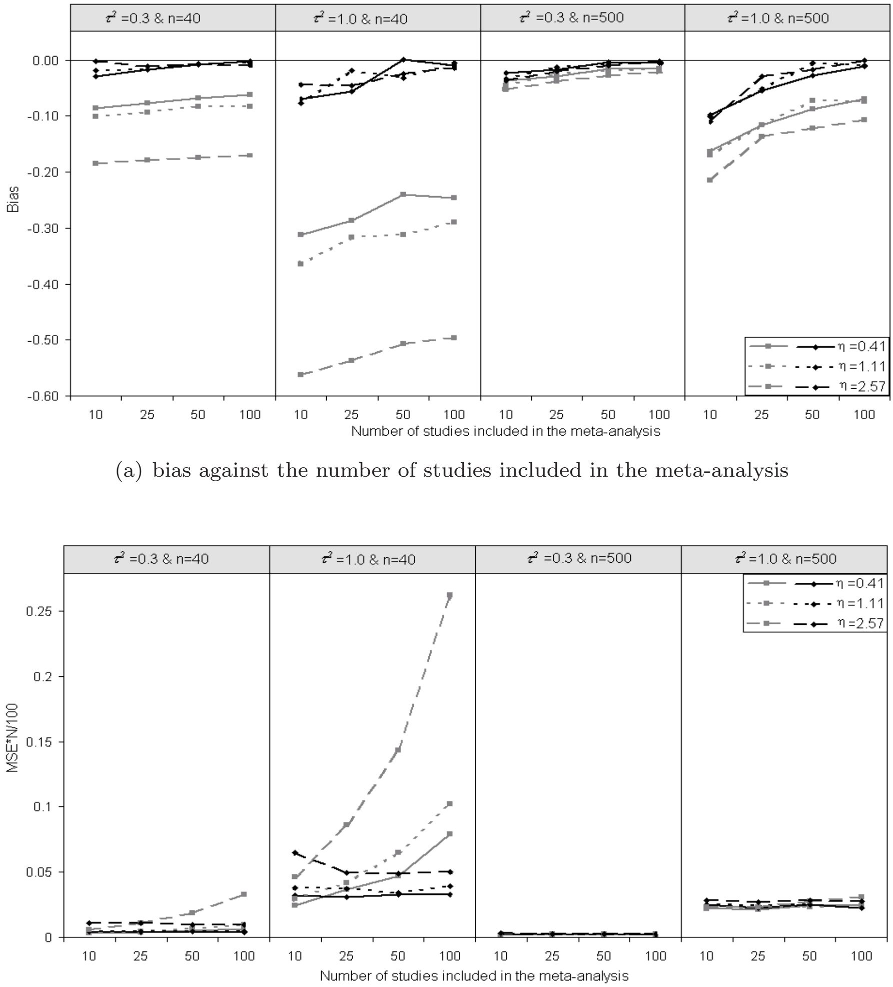
the difference between the exact and approximate method is not large. In most scenarios the profile likelihood based intervals have better coverage than the Wald based. Overall, the profile likelihood based interval of the exact method behaves the best. For $N \geq 25$, the Wald based confidence intervals of the approximate method behave worse than those of the exact method, in particular if the within-study sample size is small and the value of τ^2 is large. The profile likelihood method improves the coverage probabilities a little bit for the approximate method, but still they are very bad in many scenarios. For the exact method, the Wald based confidence intervals are already quite good, and they are improved by the profile likelihood method.

In summary, the approximate likelihood method gives a biased estimate for η and τ^2, even a considerably large bias when the true median within-study sample size is smaller and median sensitivity is larger. This result is in consistent with the findings of Sidik and Jonkman [31] who reported that τ^2 is underestimated when the approximate method using ML or REML estimation techniques is applied. Other authors [5, 32] also noted the downward bias in the estimation of the between-studies variance. The coverage probabilities for η and τ^2 of the approximate method are also far from the nominal value in the same region. On the contrary, the exact method gives unbiased estimates for η with a reasonable coverage probability, in particular for the profile likelihood method. It also gives unbiased estimates for τ^2 except when the true between-studies variance is large and the number of studies is small, in that case there is a slight downward bias. This might be due to the fact that the maximum likelihood estimate of the variance parameter is biased for small sample sizes even in the simple independently and identically distributed observations. The coverage probabilities, in particular the ones based on the profile likelihood, are reasonably acceptable.

2.5 Data Example

To illustrate the methods discussed in this article, we re-analyzed the data of a published meta-analysis [33]. Patwardhan *et al* [33] present data from fifteen studies to assess the operating characteristics of positron emission tomography (PET) by using fluorine 18 fluorodexyglucose (FDG). They performed a literature search in the MEDLINE, CINAHL, and HealthSTAR databases published between 1989 and 2003. Articles were selected if FDG PET was performed with a dedicated scanner and the resolution was specified, if standard criteria were used for the diagnosis of Alzheimer disease, if at least 12 human subjects with Alzheimer disease were enrolled in the study, if clinical diagnosis or histopathologic findings were used as the reference standard, and if sufficient data were provided to construct a 2x2 table. Out of the fifteen studies only nine of them allowed to construct a 2x2 table with reasonable certainty and they believed these studies were suitable for meta-analysis. The authors pooled sensitivity and specificity separately using a random effects model in Meta-Test (version 6.0) software. We also re-analyzed only the nine studies. The data and the SAS code are given in the Appendix.

Sensitivity and specificity were pooled separately using both the approximate and exact likelihood method. Some studies had zero counts and therefore we added 0.5 in each of the 2x2 tables to avoid undefined values for the approximate like-



(a) bias against the number of studies included in the meta-analysis

(b) MSE*N/100 against the number of studies included in the meta-analysis

Figure 2.2: Simulation results for τ^2. The bias (Figure 2.2(a)) and MSE (Figure 2.2(b)) are given for the approximate likelihood method, gray lines and the exact likelihood method, black lines. The true between-studies variance and median within-study sample size are given at the top of each plot.

Table 2.3: Parameter estimates (standard error) and Wald type confidence intervals based on the normal approximation and profile likelihood based confidence interval.

		95% Confidence Intervals	
Parameter	Estimate (s.e.)	Wald	Profile likelihood
Approximate likelihood method			
logit(sensitivity)	1.71(0.33)	[1.07, 2.34]	[1.12, 2.52]
Sensitivity	0.85(0.04)	[0.74, 0.91]	[0.76, 0.93]
$\tau^2_{logit(sensitivity)}$	0.37(0.35)	[0.06, 2.39]	[0.01, 2.08]
logit(specificity)	1.61(0.33)	[0.96, 2.26]	[1.03, 2.46]
Specificity	0.83(0.05)	[0.72, 0.91]	[0.74, 0.92]
$\tau^2_{logit(specificity)}$	0.29(0.36)	[0.03, 3.24]	[0.10, 2.13]
Exact likelihood method			
logit(sensitivity)	2.20(0.44)	[1.34, 3.06]	[1.40, 3.40]
Sensitivity	0.90(0.04)	[0.79, 0.96]	[0.80, 0.97]
$\tau^2_{logit(sensitivity)}$	0.97(0.80)	[0.19, 4.93]	[0.17, 5.28]
logit(specificity)	2.27(0.49)	[1.31, 3.23]	[1.35, 3.59]
Specificity	0.91(0.04)	[0.79, 0.96]	[0.79, 0.97]
$\tau^2_{logit(specificity)}$	1.23(1.01)	[0.24, 6.17]	[0.21, 6.59]

lihood method. The parameter estimates (standard error) and the corresponding 95% Wald type and profile likelihood based confidence intervals are tabulated in Table 2.3.

The magnitudes of the parameter estimates from the exact likelihood are larger than from the approximate likelihood method. Back transforming the estimates, the median sensitivities are 85.1% and 90.0%, and the median specificities are 84.0% and 90.6% for the approximate and exact likelihood methods, respectively. That is, the estimates from the approximate method are lower for the median sensitivity by 4.9% and for the median specificity by 6.6% compared to the exact method. The differences between the two methods for the estimates of the between-studies variances are considerable as well. The estimated between-studies variances for the logit sensitivity are 0.48 and 0.97 using the approximate and exact method, respectively. For the logit specificity the estimates are 0.42 and 1.23, respectively. Comparing the confidence intervals based on Wald and profile likelihood for logit sensitivity, η and logit specificity, ξ, the profile likelihood intervals are wider, which is expected since the profile likelihood takes into account the uncertainty in the between-studies variance. Also the confidence intervals from the exact method are wider compared to the approximate methods for the variance parameters. In summary, the parameter estimates from the practical data example follow the same pattern as the simulation study, i.e. the approximate method gives lower estimates compared to the exact method for a given parameter. Furthermore, the example shows that the differences between the approximate and exact method in practice are not negligible.

2.6 Discussion

In numerous medical articles sensitivities or specificities, or more generally proportions are analyzed, nowadays almost invariably with the DerSimonian and Laird [6] random effects model. This model uses a normal distribution for the logit transformed true probabilities. Alternatively, one could assume a beta distribution for the true probabilities. Then the model can be fitted in a statistical package such as EGRET. However this model is not used in practice, may be due to the fact that many statistical packages allow only a normal distribution for the random effects. We restricted in this article to the standard method of DerSimonian and Laird. Instead of the usual logit transformation of the observed data, other transformations such as the probit, $log(-log)$, arcsine could be used, and implemented in the same program. We do not expect that the results of this paper would change substantially if another transformation was used and approximated by a normal distribution. The reason is, that there will always be a correlation between the estimate and the within-study variance as they are determined by the same parameter, which, if not accounted for in the model, may lead to biased parameter estimates [11]. Hence we restricted to the standard logit transformation.

In this paper we compared the use of the approximate normal within-study likelihood, which is used in practice, with the alternative exact binomial likelihood. Calculation of the exact binomial likelihood involves an approximation of the integral. In NLMIXED the method of Gaussian quadrature is used, with the number of quadrature points to be specified by the user or automatically by SAS. The larger that number is chosen, the better the approximation, but at the cost of more computational time. For example, Carlin *et al* [34] have shown that for binary outcome longitudinal data a reasonably large number of quadrature points (that is 20) is required to ensure convergence on model parameter estimates. In our data example, to study the impact of the number of quadrature points we fitted the model for varying number of quadrature points. It turned out the estimates (standard error) of sensitivity and specificity did not change for a number of quadrature points greater than or equal to 10 and 15 respectively. We used 20 quadrature points for our simulation study.

Our simulations have shown that the approximate method yields biased estimates for the overall sensitivity or specificity as well as for the between-studies variance, the bias being especially considerable in cases with smaller within-study sample sizes, larger between-studies variance and larger values of the overall sensitivity, as is frequently reported in diagnostic tests. In these cases also the coverage percentages of the confidence intervals are far off the nominal values. Considering possible bias in analyzing log odds ratios is beyond the scope of this paper. However, we expect that in analyzing log odds ratios the bias might be less of a problem, since it might be at least partly cancel out. The bad performance of the approximate method is mainly caused by the fact that it does not adjust for the correlation between the estimate of the sensitivity or specificity and its standard error. Also the addition of 0.5 when there is a zero count adds to the bias [14, 35, 36]. Although it was mentioned in the literature that the standard random effects method could be biased when the parameter to be estimated is a proportion [5, 11, 12], it is still generally used. A possible explanation is that there were no practically feasible methods available that address this bias. However, the implementation of procedures for the generalized linear mixed model in standard packages has made it practically feasible nowadays to use the exact within-study distribution of the estimated sensitivity or specificity. To carry out the exact method, the sample size and the number of positive test results is needed. In practice, these quantities will always be available, though sometimes indirectly. For instance, if only the estimated proportion and its standard error are given, the sample size and number of positive results can be easily calculated. In this paper we have compared the exact method with the standard method through an extensive simulation study. We accounted for possibly important factors such as the number of studies included in the meta-analysis, the magnitude of the mean logit sensitivity, the between-studies variance and the within studies sample size. We have shown that in all scenarios studied the exact method outperformed the approximate method with respect to bias of the estimated mean logit sensitivity and coverage percentages of the corresponding confidence interval. The exact method yielded unbiased estimates of the logit sensitivity in all scenarios with reasonable coverage percentages of the Wald confidence interval, with the exception of the scenarios where the number of studies in the meta-analysis is small. Mostly the coverage probabilities were slightly lower than the nominal value. This could be due to the fact that the standard Wald method does not adjust for the between-studies variance being estimated. A profile likelihood based confidence interval, which allows for the uncertainty in the estimated between-studies variance, appeared to improve the coverage percentage to an acceptable level close enough to the nominal level. The Wald type confidence interval can be automatically done in SAS. For the profile likelihood method, some extra programming is needed. A SAS macro is available from the authors on request.

Concerning estimation of the between-studies variance we have shown that the approximate method yielded underestimates in all scenarios studied, with bias of considerable magnitude and coverage probabilities far from the nominal level in many cases. The exact method always outperformed the approximate method, although there was some bias left for cases where the number of studies in the meta-analysis was small and a coverage percentage of the Wald type confidence interval a bit less than 0.90 for some scenarios, which could be corrected using profile likelihood. In the spirit of REML estimation, a possible improvement might be to multiply the estimate by $k/(k-1)$, where k is the number of studies in the meta-analysis. Another possibility might be the use of a bootstrapping technique similar to Turner *et al* [15], which is directly available in the MLWin software [18].

We did some further simulations when the true parameters were simulated from a skewed distribution where most of the proportions are close to 1. We used Fernandez and Steel's [37] approach to introduce skewness into a unimodal normal distribution. In most cases the exact method outperformed the approximate method, with only slight bias ($\cong 0.10$) for the exact method left for scenarios with highly skewed true distribution and small within-study size. The coverage probability from the exact method was also better than the approximate method in most cases and close to the nominal level, especially when we used profile likelihood based confidence interval (results not presented and available from the authors on request).

Though we focused in this paper on sensitivities or specificities, the results apply more generally to meta-analyzing proportions, such as prevalences or incidences of a disease. The proportions can also be corrected for important covariates, using the same generalized linear mixed model programs.

In a situation when two end points, for example, sensitivity and specificity, are presented in a study and when there is a need to incorporate the correlation that might exist between these two measures, a bivariate meta-analysis approach can be used [8, 10, 25]. Reitsma *et al* [10] assume the within-study error distribution of sensitivity and specificity to follow a normal distribution. However an exact binomial likelihood can also be implemented in generalized linear mixed model programs [38], for example, SAS NLMIXED or S-Plus / R nlme. Although we did no simulations for the bivariate model, it is very likely that the results of the univariate meta-analysis carry over to the bivariate case.

In this article the models were fitted using classical likelihood methods. An alternative would be to use a Bayesian hierarchical modeling approach [34, 39], which can be carried out using the publicly available software Win BUGS [40].

Our overall conclusion from this paper is that in many cases the standard approximate method falls short and that the exact method should be used, preferably accompanied by profile likelihood based confidence intervals, whenever that is feasible in practice.

Appendix

In this appendix the SAS syntax is given to estimate the parameters using approximate and exact likelihood methods for the meta-analysis data used in Section 2.5. The data is given in the table below. The SAS code for the MIXED and NLMIXED procedure is given to pool sensitivity, but it can also be used to pool specificity.

Study	TP(xi)	FN	FP	TN	mi	logitsens	est
15	33	6	5	35	39	1.640	0.184
17	18	6	5	10	24	1.046	0.208
19	20	13	0	41	33	0.418	0.123
20	19	0	0	19	19	3.664	2.051
22	44	6	10	19	50	1.924	0.176
24	18	3	1	9	21	1.665	0.340
25	27	1	4	21	28	2.909	0.703
29	21	0	1	9	21	3.761	2.047
30	18	1	1	20	19	2.512	0.721

The meaning of the variables that are used in the table and SAS code below are the following:
study = a number given for a study
xi = number of patients with true positive test result
mi = within-study sample size in the diseased group
logitsens = observed logit sensitivity (=ln((xi+0.5)/(mi-xi+0.5))
est[4]= estimated within-study variance of logit sensitivity = 1/(xi+0.5)+1/(mi-xi+0.5))
0.5 is added in each of the two by two tables when we calculate logitsens and est to avoid undefined values.

```
/* Approximate likelihood method using SAS procedure MIXED*/
proc mixed data = d method = REML;
     class  study;
     model logitsens = / intercept Solution cl df=1000;  /*df=1000
is specified to get Wald type confidence interval instead of the t*/
        random intercept /  subject = study ;
        repeated /group = study;
/*dataest is the name of the data set that contains only the variable
 called est and 10 lines. The first value is a starting value for
 the between-studies variance. The next nine values are the estimated
 within-study variances (=1/(xi+0.5) + 1/(mi-xi+0.5)). eqcons is used
 to specify that the within-study variance are assumed to be known*/
        parms / parmsdata = dataest eqcons= 2 to 10;
run;
/* The Exact approach using the SAS procedure NLMIXED*/
proc nlmixed data=d df = 1000;
        parms mtlnsens=2.0   vtlnsens=0.8;  /*Initial values*/
        pi = 1/(1+exp(-tlnsens));
/*tlnsens=is the unknown true logit sensitivity*/
        model xi~binomial(mi,pi);
        random tlnsens ~ normal(mtlnsens , vtlnsens) subject=study;
run;
```

[4] This is the prescribed name by SAS of the variable that contains the variances.

References

[1] Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC *et al.* Guidelines for meta-analyses Evaluating Diagnostic Tests. Annals of Internal Medicine 1994; 120(8): 667-676.

[2] Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of hetrogenity in systematic reviews of diagnostic tests. Statistics in Medicine 2002; 21:1525-1537.

[3] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted. Statistics in Medicine 2002; 21: 1559-1573.

[4] Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 1998; 17: 841-856.

[5] Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effect regression model for meta-analysis. Statistics in Medicine 1995; 4: 395-411.

[6] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7(3): 177-188.

[7] Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in meta-analysis: an application. Statistics in Medicine 2003; 22:1335-1353.

[8] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine 2002; 21:589-624.

[9] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 2001; 20:2865-2884.

[10] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005; 58: 982-990.

[11] Chang BH, Waternaux C, Lipsitz S. Meta-analysis of binary data: which study variance estimate to use? Statistics in Medicine 2001; 20: 1947-1956.

[12] Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. Statistics in Medicine 1999; 14: 395-411.

[13] Cox DR. The Analysis of Binary Data. Methuen, London, 1970.

[14] 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. Statistics in Medicine 1993; 12: 1293-1316.

[15] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2000; 19: 3417-3432.

[16] Knapp G, Biggerstaff BJ, Hartung J. Assessing the Amount of Heterogeneity in Random-Effects Meta-Analysis. Biometrical Journal 2006; 48(2): 271-285.

[17] SAS Institute Inc 2004. SAS/STAT 9.1 User's Guide. Cary, NC: SAS Institute Inc.

[18] Rasbash J, Steele F, Browne W, Prosse B. A User's Guide to MLWin. *Version 2* London: Institute of Education; 2004.

[19] Mol BWJ, Lijmer JG, Ankum WM, van der Veen F, Bossuyt PMM. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. Human Reproduction 1998; 13(11): 3220-3227.

[20] Cruciani M, Nardi S, Malena M, Bosco O, Serpelloni G, Mengoli C. Systematic review of the accuracy of the ParaSightTM-F test in the diagnosis of Plasmodium falciparum malaria. Med Sci Monit 2004; 10(7): MT81-88.

[21] Bipat S, Glas AS, Slors FJM, Zwinderman AH, Bossuyt PMM, Stoker J. Rectal Cancer: Local Staging and Assessment of Lymph Node Involvment with Endoluminal US, CT, and MR Imaging-A Meta-Analysis. Radiology 2004; 232: 773-783.

[22] Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between Compliance with Methodological Standards of Diagnostic Research and Reported Test Accuracy: Meta-Analysis of Focused Assessment of US for Trauma. Radiology 2005; 236: 102-111.

[23] Cruciani M, Marcati P, Malena M, Bosco O, Serpelloni G, Mengoli C. Meta-analysis of diagnostic procedures for Pneumocystis carinii Pneumonia in HIV-1-infected patients. European Respiratory Journal 2002; 20: 982-989.

[24] Cox DR, Hinkley DV. Theoretical Statistics. London: Chapman and Hall; 1974.

[25] van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine 1993; 12:2273-2284.

[26] Brockwell SE, Gordon R. A comparison of statistical methods for meta-analysis. Statistics in Medicine 2001; 20:825-840.

[27] Sidik K. Simple heterogeneity variance estimation for meta-analysis. Applied Statistics 2005; 54:367-384.

[28] Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Statistics in Medicine 1996; 15:619-629.

[29] Roger JH, Kenward MG. Repeated measures using proc mixed instead of proc glm. Proceedings of the First Annual South-East SAS Users Group conference. SAS Institute: Cary NC, 1993; pp 199-208.

[30] Welham SJ, Thompson R. Likelihood Ratio Tests for Fixed Model Terms using Residual Maximum Likelihood. Journal of Royal Statistical Society, series B 1997; 59: 701-714.

[31] Sidik K, Jonkman J N. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine 2007; 26: 1964-1981.

[32] Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine 1999; 18: 2693-2708.

[33] Patwardhan MB, McCrory DC, Matchar DB, Samsa GP, Rutschmann OT. Alzheimer Disease: Operating Characterstics of PET A Meta-analysis. Radiology 2004; 231:73-80.

[34] Carlin JB, Wolfe R. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. Biostatistics 2001; 2: 397-416.

[35] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine. 2004; 23: 1351-75.

[36] Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Statistics in Medicine. 2007; 26:53-77.

[37] Fernandez C. Steel MFG. On Bayesian modeling of fat tails and skewness. Journal of American Statistical Association. 1998; 93:359-71.

[38] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007; 8(2): 239-251.

[39] Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approach to random-effects meta-analysis: A comparative study. Statistics in Medicine 1995; 14: 2685-2699.

[40] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 1.4.1. MRC Biostatistics unit: Cambridge 2004. Program available at http://www.mrc-bsu.cam.ac.uk/bugs

CHAPTER 3

Bivariate Random Effects Meta-analysis of ROC curves

Abstract

Meta-analysis of ROC-curve data is often done with fixed effects models, which suffer many shortcomings. Some random effects models have been proposed to execute a meta-analysis of ROC-curve data, but these models are not often used in practice. Straightforward modeling techniques for multivariate random effects meta-analysis of ROC-curve data are needed. The first aim of this paper is to present a practical method that addresses the drawbacks of the fixed effects SROC method of Littenberg and Moses. Sensitivities and specificities are analysed simultaneously using a bivariate random effects model. The second aim is to show that other summary ROC curves can also be derived from the bivariate model through different characterisations of the estimated bivariate normal distribution. Thereby we show that the bivariate random effects approach not only extends the SROC approach, but provides a unifying framework for other approaches as well.

The authors bring the statistical meta-analysis of ROC curve data back into a framework of relatively standard multivariate meta-analysis with random effects. The analyses can be easily carried out in standard statistical software. The syntax needed for the software package SAS (Proc NLMIXED) that is used throughout this paper is given in the appendix.

3.1 Introduction

For a thorough assessment of the effectiveness of a specific treatment it is common to execute a meta-analysis of randomised clinical trials reported in the literature. The same is done for the assessment of the characteristics of a diagnostic test to distinguish patients having a certain disease from patients not having that disease. Meta-analyses to assess the reliability, accuracy and impact of diagnostic tests are essential to guide optimal test selection and the appropriate interpretation of test results [1]. However, the designs of test accuracy evaluations differ from the designs of studies that evaluate the effectiveness of treatments, which means that different criteria are needed when assessing study quality and potential for bias. Additionally, often each evaluation of diagnostic tests reports a pair of related summary statistics (for example sensitivity and specificity) rather than a single statistic, requiring alternative statistical methods for pooling study results [1]. Receiver Operating Characteristic (ROC) curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds.

In the last 15 years, several methods for meta-analysis of diagnostic tests have become available [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. The proposed methods depend on the type of data that are available. Some [6, 8] are designed to be used when individual patient data of the studies are available. Others are applicable when each study provides an estimate of the area under the ROC curve [12]. Still others are applicable to the situation where per study only one estimated pair of sensitivity and specificity (corresponding to possibly different diagnostic thresholds) is available. In this article we focus on this last situation, which is by far the most common in practice. For this situation the aim of the meta-analysis is to estimate the overall ROC curve of the (continuous) diagnostic marker.

Probably the most well known and most commonly used method in practice is the Summary ROC (SROC) method proposed by Littenberg and Moses [2] and Moses *et al.* [3]. They plotted the difference versus the sum of the logit(true positive rate) and logit(false positive rate) from each study. Then they fitted three types of regression lines (robust, unweighted and weighted) to these points. Finally they transformed the line to ROC space.

Although frequently used, the SROC method has a number of serious shortcomings. The first aim of this article is to present an approach that extends the SROC method, addresses its drawbacks and is still easily carried out in practice using familiar statistical packages like SAS. The method follows the general multivariate approach as described in van Houwelingen *et al.* [13] and Arends *et al.* [14, 15]. The second aim of this paper is to show how other summary ROC curves (regression lines in the logit space) can be derived from the bivariate model, such as the logit(true positive rate) on logit (false positive rate), logit (false positive rate) on logit(true positive rate), the first principal component, or the curve corresponding to the method of Rutter and Gatsonis [9].

In Section 3.2 we introduce two data sets that will be used as examples. In Section 3.3 we give an overview of the SROC method and we briefly discuss its shortcomings. In Section 3.4 we briefly discuss other methods proposed in the literature. In Section 3.5 the new approach is presented. In Section 3.6 the methods are applied on the two example data sets and the results are presented. We use the SAS procedures Proc Mixed and Proc NLMixed for the analyses and give the syntax in the Appendix. Finally we end with a discussion in Section 3.7.

3.2 Data examples

To illustrate the methods discussed in this paper, we apply them to two meta-analysis data sets, one relatively small (29 studies) data set and one large data set (149 studies).

Example 1: FNAC of the Breast [16]

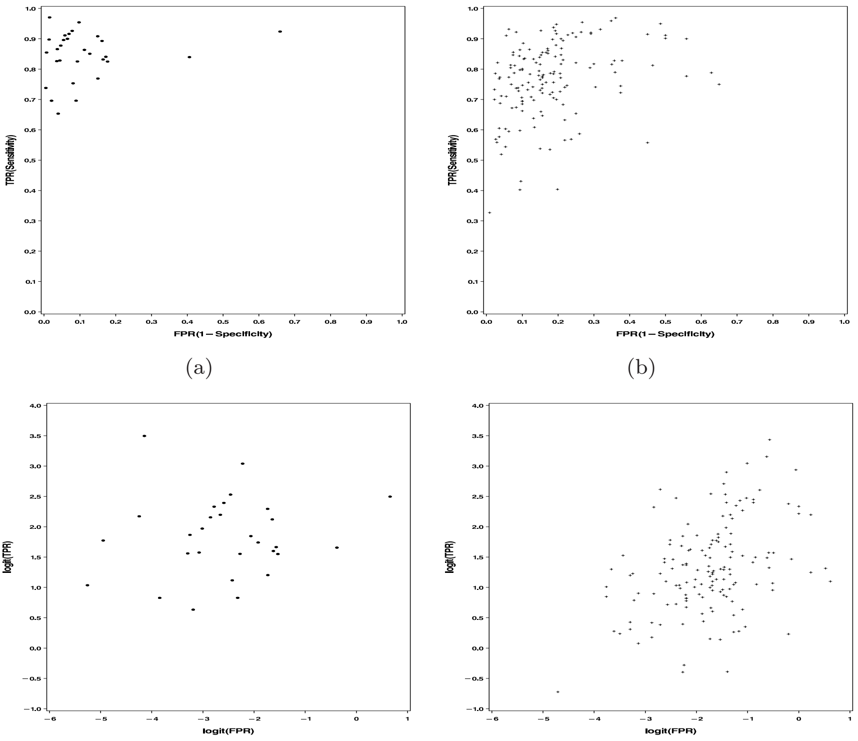
Giard and Hermans [16] present 29 studies evaluating the accuracy of fine-needle aspiration cytologic examination (FNAC) of the breast to assess presence or absence of breast cancer. FNAC provides a non-operative way of obtaining cells for establishment of the nature of a breast lump and therefore plays a pivotal role in the preoperative diagnostic process [16, 17, 18, 19]. The sensitivity and specificity of FNAC were determined for each study. Sensitivity was defined as the probability of a malignant or suspect test result in patients with cancer. Specificity was defined as the probability of absence of abnormal cells in the patients without cancer [16]. Table 3.1 shows the frequencies of the FNAC outcomes given the final diagnosis of benign or malignant breast disease.

Table 3.1: Example 1: data from clinical studies on patients with a breast mass who underwent a fine-needle aspiration cytological examination (FNAC). Patients are cross-classified according to their final diagnosis (benign or malignant breast disease) and their FNAC result.

	FNAC results for Patients with					
	benign disease			malignant disease		
Study	FP (Y_0)	TN	Total (n_0)	TP (Y_1)	FN	Total (n_1)
1	70	939	1009	979	89	1068
2	3	163	166	51	22	73
3	55	894	949	1569	152	1721
4	25	259	284	35	15	50
5	4	121	125	59	12	71
6	18	216	234	56	4	60
7	602	3117	3719	329	39	368
8	10	213	223	125	17	142
9	88	499	587	211	63	274
10	0	31	31	49	1	50
11	26	643	669	336	178	514
12	147	746	893	210	42	252
13	5	25	30	16	3	19
14	16	356	372	258	53	311
15	9	107	116	56	18	74
16	16	112	128	162	28	190
17	6	112	118	116	13	129
18	99	145	244	65	12	77
19	5	78	83	94	10	104
20	0	70	70	26	4	30
21	28	136	164	1318	249	1567
22	55	539	594	569	120	689
23	1	287	288	46	16	62
24	13	76	89	64	6	70
25	1	104	105	39	4	43
26	16	426	442	132	20	152
27	17	161	178	470	22	492
28	25	200	225	28	4	32
29	43	22	65	42	3	45

In Table 3.1, Y_1 is the number of patients with a malignant or suspect test result in the patients with cancer. The total number of patients with cancer is n_1. Y_0 is the number of patients with a malignant or suspect test result in the n_0 patients without cancer. The true positive rate TPR, or sensitivity, is estimated for a study by Y_1/n_1, and the false positive rate FPR, which is 1 minus the specificity, by Y_0/n_0. See Figure 3.1(a) for a plot of the estimated TPRs against the estimated FPRs and Figure 3.1(c) for the estimated TPRs and FPRs on the logit scale.

The estimated TPRs and FPRs vary considerably across studies. Also, the proportions of patients with benign or malignant disease according to the final diagnosis differed substantially. At the time of publication (1992), no reasonable methods to summarize diagnostic test data across several studies were available. In this paper we use the data to fit the standard fixed effects SROC model as well as the proposed random effects models.



(a)

(b)

(c)

(d)

Figure 3.1: Observed sensitivity against (1-specificity) of data reported across 29 studies that were originally meta-analyzed by Giard and Hermans [16] (left side of picture) and across 149 studies that were originally meta-analysed by Heijenbrok-Kal *et al.* [20] (right side of picture) on the original scale and on logit transformed scale.

Example 2: Imaging tests for coronary artery disease [20]

Heijenbrok-Kal [20] searched PubMed from January 1990 through May 2003 for meta-analytic studies on the diagnostic performance of imaging tests for coronary artery disease. In all meta-analyses included in her paper, angiography is the reference standard and the source numbers of true and false positives and true and false negatives are reported. Duplicate source studies are excluded. This resulted in a total of 246 patient series including 24,761 patients who underwent one of eight different imaging technologies for coronary artery disease. The coronary tests showed little difference in diagnostic performance.
To illustrate our approach we choose from the 246 source studies only those in which the performance of an exercise or stress echo was investigated. This resulted in 149 studies including 13,303 patients. In Figure 3.1(b) a plot is given of the estimated TPRs against the estimated FPRs and Figure 3.1(d) represents the estimated TPRs and FPRs on the logit scale.

3.3 The standard SROC method

The starting point of a meta-analysis of ROC curve data is a number of studies providing information on a continuous diagnostic marker or variable M. In the different studies possibly different thresholds for M are used to obtain a dichotomous diagnostic test. The data provided by each study are the number of patients with a positive test result (y_1) and the total number of patients (n_1) in the group with the disease, and the number of patients with a positive test result (y_0) and the total number of patients (n_0) in the group without the disease. The aim is to estimate the overall ROC curve of the diagnostic marker M based on the available data from the different studies. The standard method used in practice is the SROC method of Littenberg and Moses [2], which proceeds as follows. The underlying model assumes that there exists a transformation of the continuous diagnostic variable M such that the transformed test, X, follows a logistic distribution both in the population without the disease and in the population with the disease. In other words, it is assumed that the transformation that makes the distribution of M logistic in the non-diseased (which always exists) makes the distribution simultaneously logistic for those with the disease. We assume that the transformation is done such that large values of X correspond with the diseased population.

The cumulative distribution of X in the healthy and the diseased populations is given by

$$Pr(X < x|healthy) = \frac{e^x}{1+e^x} \quad \text{and} Pr(X < x|disease) = \frac{e^{-\alpha+\beta x}}{1+e^{-\alpha+\beta x}} \tag{3.1}$$

for some values of $\alpha \geq 0$ and $\beta > 0$. The difference between the mean value in the population with the disease and without the disease is α/β, and the ratio between the standard deviation of the diseased and the healthy population is $1/\beta$. Thus $0 < \beta < 1$ corresponds with a higher variance in the population with the disease and $\beta > 1$ with a smaller variance. Figure 3.2 gives a graphical illustration with the interpretation of α and β, where clearly $0 < \beta < 1$.

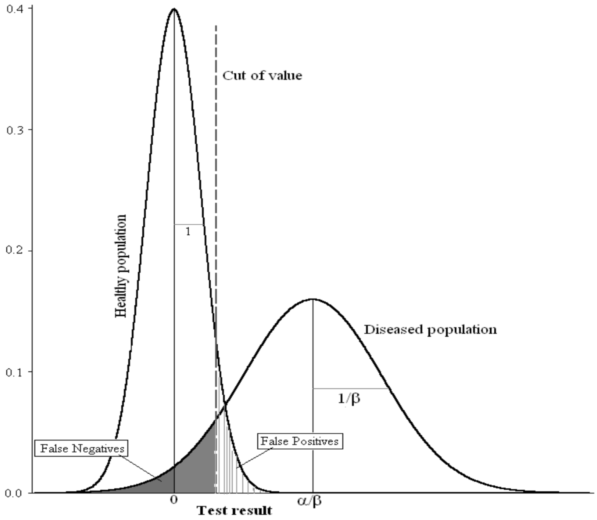


Figure 3.2: Graphical illustration with interpretation of α and β

If λ denotes the threshold X-value for the test being declared positive, then according to (3.1) the probability of a false positive result is $1 - e^{\lambda}/(1 + e^{\lambda})$ and hence $logit(FPR) = -\lambda$. Similarly we have $logit(TPR) = \alpha - \beta\lambda$. In the following we will use the notation:

$$\xi = logit(FPR) = -\lambda$$
$$\eta = logit(TPR) = \alpha - \beta\lambda$$

This implies the linear relationship

$$\eta = \alpha + \beta\xi \tag{3.2}$$

Following Rutter and Gatsonis [9], α can be called the accuracy parameter and β the scale or asymmetry parameter. If $\beta = 1$, the resulting ROC curve is symmetric (with respect to the minus 45 diagonal), otherwise it is asymmetric. In the SROC approach of Littenberg & Moses the relation (3.2) is written as

$$\eta - \xi = \alpha' + \beta'(\eta + \xi)$$

with $\alpha' = 2\alpha/(\beta + 1)$ (with $\alpha' \geq 0$) and $\beta' = (\beta - 1)/(\beta + 1)$ (with $-1 < \beta' < 1$). If D and S are the estimated values of $\eta - \xi$ and $\eta + \xi$ from a study (to avoid division by zero, 0.5 is added to all numbers in the 2x2 table of a study), then approximately

$$D = \alpha' + \beta' S \tag{3.3}$$

and the values of α' and β' are estimated by a simple weighted or unweighted linear regression. The weights are chosen proportional to the inverse variance of D. D is interpreted as the log odds ratio of a positive test result for diseased individuals relative to healthy individuals, and is often called the diagnostic odds ratio. Its estimated variance is

$$\frac{1}{y_0+0.5}+\frac{1}{n_0-y_0+0.5}+\frac{1}{y_1+0.5}+\frac{1}{n_1-y_1+0.5} \tag{3.4}$$

The summary ROC curve is obtained by transforming the estimate of (3.3) to the ROC space. A value of $\beta' \neq 0$ indicates that the curve is asymmetric. The advantage of the SROC method, which explains its popularity, is that it is very simple to understand and can be carried out in any statistical package. Despite this important advantage of simplicity, a number of critical comments can be made.

First of all, the SROC method is a fixed effects method, i.e. it assumes that the values of α and β do not vary across studies. Thus variation is due only to the threshold effect and within-study sampling variability. However, in many practical cases it is likely that there is between-studies variation beyond those sources. Study characteristics such as technical aspects of the diagnostic test, patient selection, study settings, experience of readers etc. are among the potential contributors to between-study variation in the estimates of diagnostic performance [9]. Modern meta-analytical methods take possible variation across studies into account by introducing random effects [13, 21, 22, 23, 24]. If there is between-study variation, a fixed effects model can give biased estimates and typically underestimates standard errors [21, 25].

Second, the independent variable S in the regression equation (3.3) is measured with measurement error, which should be taken into account. As a result, regression to the mean [26] and attenuation due to measurement errors [27] could seriously bias the slope of the regression line [13]. Thus, not taking into account the measurement error in S leads to bias in β' (in general towards zero) and α' and therefore also in β (in general towards one) and α [24].

Third, D and S are correlated within a study, positively or negatively depending on the study. In the standard fixed effects SROC model this correlation is ignored. Although probably the correlation is usually small in practice, the consequence of ignoring it is not obvious.

Fourth, it is reasonable that the different studies should be somehow weighed in the analysis, in particular if the studies vary substantially in size. If there is both between- and within-study variation then weighing by the inverse of within-study variances, as is done in the weighted SROC approach, will not be optimal.

Finally, to avoid undefined log odds, log odds ratios and their variances, quite arbitrarily 0.5 is added to the numbers in the fourfold tables of the trials. As Moses *et al.* [3] showed, the effect of this adjustment can be surprisingly large. Adding 0.5 to all cells tends to push an estimated ROC curve away from the desirable northwest corner of ROC space. The standard SROC method has to do this because it does not use the true binomial distribution of the number of positive test results within a group. It would be preferable to not use this artificial and arbitrary correction.

In Section 3.5 we present a method that does not have the described disadvantages of the SROC method and can still be carried out easily in standard statistical packages. In addition, we show how other summary ROC curves (regression lines in the logit space) can be derived from our method. But first we discuss in Section 3.4 some other methods proposed in the literature.

3.4 Other methods proposed in the literature

Kardaun and Kardaun [28] also assume model (3.1) and exploit the linear relationship $\eta_i = \alpha + \beta\xi_i$ where $i = 1, \ldots, k$ denotes the number of the study. Using straightforward approximate likelihood methods all $k+2$ parameters (including the ξ_i's) are estimated. The estimation method is called *approximate* likelihood, since, instead of the exact likelihood based on the true conditional distribution of the estimated ξ (called $\hat{\xi}$) and η (called $\hat{\eta}$) given ξ and η, an approximate likelihood based on the familiar normal approximations $\hat{\xi} \cong N(\xi, 1/y_0 + 1/(n_0 - y_0))$ and $\hat{\eta} \cong N(\eta, 1/y_1+1/(n_1-y_1))$ is used. The drawbacks of the method of Kardaun and Kardaun [28] are, first, that the number of estimated parameters is proportional to the number of trials, hence standard likelihood theory does not apply [29]. For instance, consistency of the estimates when the number of studies tends to infinity is not guaranteed. Second, their computer-intensive method based on profile likelihood is not very practical. Third, it is a fixed effects model, and fourth, the arbitrary value of 0.5 is added to avoid undefined log odds and their variances; thus the first and last drawbacks mentioned in the previous section for the SROC method still apply.

Recently Rutter and Gatsonis [9] proposed a hierarchical Bayesian regression approach, that does not have the drawbacks mentioned in the previous section for the SROC method. They assumed the following model. Let π_{i0} be the true FPR in the non-diseased and π_{i1} be the true TPR in the diseased populations. Then $Y_0 \sim Binomial(n_0, \pi_{i0})$ and $Y_1 \sim Binomial(n_1, \pi_{i1})$. Defining $\xi_i = logit(\pi_{i0})$ and $\eta_i = logit(\pi_{i1})$, the following relationship is assumed to hold between ξ_i and η_i :

$$\begin{aligned} \xi_i &= (\theta_i + \alpha_i X_0)e^{-\beta X_0} \\ \eta_i &= (\theta_i + \alpha_i X_1)e^{-\beta X_1} \end{aligned} \tag{3.5}$$

where X_0 and X_1 are chosen to be -1/2 and +1/2 respectively. This implies the linear relationship

$$\eta_i = \alpha_i e^{-\beta/2} + e^{-\beta}\xi_i \tag{3.6}$$

For equation (3.6), α is called the accuracy parameter, because it measures the difference between TPR and FPR, and β is called the scale parameter. With this parameterisation, if $\beta \neq 0$ the ROC curve is asymmetric.

The between-studies variation is modelled by assuming that α_i and θ_i are independent and normally distributed:

$$\begin{pmatrix} \alpha_i \\ \theta_i \end{pmatrix} = N\left(\begin{pmatrix} \bar{\alpha} \\ \bar{\theta} \end{pmatrix} \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\theta^2 \end{pmatrix} \right) \tag{3.7}$$

To compute a summary ROC curve, Rutter and Gatsonis [9] plug in the estimates for $\bar{\alpha}$ and β into the linear relation (3.6) and transform it into the ROC space. The method allows for between-study variation by modeling the accuracy parameter α with a random effect. A practical disadvantage is that Rutter and Gatsonis [9] compute the estimates in a Bayesian way using Markov Chain Monte Carlo (MCMC) simulation with the BUGS software, which is rather complicated. MCMC estimation requires programming, simulation, evaluation of convergence and model adequacy, and synthesis of simulation results. Implementation of MCMC simulation entails non-trivial analysis tasks including evaluation of convergence and the adequacy of prior distributions, and these tasks require statistical expertise. As the authors mention, this is a high price that has to be paid for the advantages of the hierarchical SROC model. Furthermore, Rutter and Gatsonis [9] use a relatively complicated parameterisation, which can make it difficult for the meta-analyst to fully understand what (s)he is doing. Macaskill [30] shows how the model of Rutter and Gatsonis can be fitted in a non-Bayesian way using the SAS NLMixed program for generalized linear mixed models. This makes the model of Rutter and Gatsonis model much more practical.

Recently a straightforward random effects extension of the method of Littenberg and Moses [2] has been used in some medical applications [31, 32, 33], using the STATA program Metareg [34]. This method is as follows. Let η_i and ξ_i again denote the true logit(TPR) and logit(FPR) for study i. Let $D_i = \eta_i - \xi_i$ be the true log odds ratio and $S_i = \eta_i + \xi_i$. The corresponding estimates are given by $\hat{\xi}_i$, $\hat{\eta}_i$, $\hat{D}_i$, and $\hat{S}_i$ respectively. Then the model is:

$$\hat{D}_i = \alpha_i + \beta \hat{S}_i + \varepsilon_i \tag{3.8}$$

with $\varepsilon_i \cong N(0, \frac{1}{y_0+0.5} + \frac{1}{n_0-y_0+0.5} + \frac{1}{y_1+0.5} + \frac{1}{n_1-y_1+0.5})$ and $\alpha_i \cong N(\bar{\alpha}, \sigma_\alpha^2)$.

In this model, all studies have a common slope β , but the intercepts vary randomly between-studies according to a normal distribution. The overall ROC line is $\eta = \bar{\alpha} + \beta\xi$, where the individual study lines vary randomly around this line with between-studies standard deviation σ_α. This is the standard random effects meta-regression model and there are many statistical packages in which this model can be fitted, such as SAS, STATA and R/S-Plus.

Measurement error of $\hat{D}_i$ is correctly accounted for, but the measurement error in $\hat{S}_i$ is still neglected. Another drawback for sparse data sets is that it is not possible to use the underlying binomial distributions for $\hat{D}_i$ and $\hat{S}_i$ instead of the normal approximations.

3.5 Alternative approach

In numerous medical articles sensitivities or specificities are meta-analysed separately by the standard random effects model of DerSimonian and Laird [21]. The method we propose is a direct extension of this approach. We analyse sensitivities and specificities simultaneously using a two-dimensional random effects model. We will show that the model implies a linear relationship between η and ξ, and can be seen as an extension of the SROC method of Littenberg and Moses [2]. In Section 3.5.1 we introduce our model. In Section 3.5.2 we discuss several types of summary ROC curves. In Section 3.5.3 we discuss the relation with the approach of Rutter and Gatsonis [9]. Throughout we follow a two-level hierarchical modelling approach, explicitly modelling the within and between-studies variability.

3.5.1 The bivariate model

The standard way of meta-analysing false positive rates of a diagnostic test in the medical literature is the DerSimonian and Laird [21] random effects model:

$$\xi_i \cong N(\bar{\xi}, \sigma_\xi^2) \quad \text{with} \quad \hat{\xi}_i \cong N(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0})$$

Here $\hat{\xi}_i$ and ξ_i are the observed and true logit(FPR)of study i, respectively. Note the well-known formula for the standard error of an estimated log odds. The parameter $\bar{\xi}$ describes the overall mean logit false positive rate and σ_ξ^2 describes the between-studies variance in true logit false positive rates. Similarly, true positive rates are analyzed using the model:

$$\eta_i \cong N(\bar{\eta}, \sigma_\eta^2) \quad \text{with} \quad \hat{\eta}_i \cong N(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1})$$

The straightforward generalisation is to assume a bivariate normal model for the the pair (ξ_i, η_i):

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \cong N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{pmatrix} \right) \tag{3.9}$$

Note that this model implies the standard univariate random effects meta-analysis model for the ξ_i and η_i separately, but now allows that ξ_i and η_i are correlated. This model fits in the framework of bivariate meta-analysis as originally introduced by van Houwelingen *et al.* [35]. Later on McIntosh [36] and Arends *et al.* [14] used this model to investigate the relationship between baseline risk and size of treatment effect in clinical trials meta-analysis. In van Houwelingen *et al.* [13] bivariate meta-analysis was generalized to multivariate meta-analysis and it was shown how standard General Linear Mixed Model programs can be used to fit these models. An example of a tri-variate meta-analysis is given by Arends *et al.* [15]. The most simple characterisation of the overall accuracy of the diagnostic test would be to take the estimated $\bar{\xi}$ and $\bar{\eta}$ and transform them to the ROC space. A more extensive description would be to characterise the bivariate normal distribution by a line and transform that line to the ROC space. Note that the bivariate normal distribution implies a linear association between ξ_i and η_i. However, as will be discussed in the next section, different lines might be employed, leading to different summary ROC curves. For example, the regression line of η_i on ξ_i could be used. Standard normal distribution theory tells that the regression line of η_i on ξ_i has intercept α and slope β given by

$$\alpha = \bar{\eta} - \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}\bar{\xi} \quad \text{and} \quad \beta = \frac{\sigma_{\xi\eta}}{\sigma_\xi^2} \tag{3.10}$$

The residual variance of the regression, given by $\sigma_{\eta|\xi}^2 = \sigma_\eta^2 - \frac{\sigma_{\xi\eta}^2}{\sigma_\xi^2}$, describes the variation in the true sensitivities between-studies that have the same specificity. In Section 3.5.2 we discuss some alternative summary ROC curves.

Similarly, as in the above univariate models for meta-analysing specificities and sensitivities separately, we model the within-study sampling variability using the fact that the estimated logit transformed FPR, $\hat{\xi}_i$, and TPR, $\hat{\eta}_i$, are independent and approximately normally distributed:

$$\hat{\xi}_i|\xi_i \cong N(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0}) \quad \text{and} \quad \hat{\eta}_i|\eta_i \cong N(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1}) \qquad (3.11)$$

If one or more of the denominators are close to zero, 0.5 should be added to the denominators, as in (3.4). The equations (3.9) and (3.11) together specify a general linear mixed model (GLMM), and the parameters can be estimated by (restricted) maximum likelihood using a GLMM program. Subsequently the intercept α and the slope β of a summary line can be calculated, using for instance (3.10) or one of the formulas given in the next subsection if another type of summary ROC curve is preferred. Standard errors of α and β can be calculated with the delta method. Many statistical packages provide a GLMM program. We used Proc Mixed from the SAS package. The syntax is given in the appendix. Proc Mixed does not give estimates and standard errors of user defined derived parameters, thus we have to calculate the estimates of α and β by hand, though the calculations are very simple. SAS users can avoid these hand calculations, since the model can also be fitted in Proc NLMixed. This program provides estimates and standard errors of user defined derived parameters. The syntax needed for Proc NLMixed is given in the appendix. Another possibility in Proc NLMixed is to reparameterise the model in such a way that one immediately gets the estimates and standard errors for the parameters of interest.

We call the GLMM approach the approximate likelihood approach, because an approximate (normal) model denoted by equation (3.11) is used for the within-study sampling variability. The practical advantage is that the model remains a GLMM, for which much software is available. The approximate likelihood approach works well for larger data sets [13]. As a rule of thumb, the requirement 'all denominators in equation (3.11) larger than or equal to 5' might be adopted, though this is probably too severe.

In our model the first drawback of the SROC method as mentioned in Section 3.3, that it is a fixed effects model, no longer applies. Also the model does not suffer from the second, third and fourth drawbacks. The problem of measurement error (the second drawback) is avoided by assuming a distribution for ξ_i. In general there are two ways of dealing with measurement error, the structural and the functional approach [13]. Our approach is in the spirit of the structural approach, similar to Arends *et al.* [14, 15] and van Houwelingen *et al.* [13], which has the important advantage that the parameters can be estimated by straightforward likelihood methods.

The third drawback does not apply, because $\hat{\xi}_i$ and $\hat{\eta}_i$ are independent within-studies. Even if we would formulate the model in terms of D and S, as is done in the standard SROC method, then there would be no problem since the correlation can be easily modelled in the GLMM. The fourth drawback does not apply since the likelihood method implicitly uses the "correct" weighting based on within- as well as between-study variation. The fifth drawback, i.e. arbitrary adding 0.5 to the numbers in the fourfold table to avoid undefined log odds, still applies since we assumed an approximate within-study model. If we want to address this drawback as well, the true distribution of $\hat{\xi}_i = Y_{0i}/n_{0i}$ and $\hat{\eta}_i = Y_{1i}/n_{1i}$ should be used. Given the true $FPR_i = (1 + exp(-\xi_i))^{-1}$ and $TPR_i = (1 + exp(-\eta_i))^{-1}$ of study i, the observed test positive numbers Y_{0i} in the healthy group and Y_{1i} in the diseased group follow binomial distributions:

$$Y_{0i}|n_{0i} \cong Binomial(n_{0i}, FPR_i); \quad Y_{1i}|n_{1i} \cong Binomial(n_{1i}, TPR_i) \tag{3.12}$$

The equations (3.9) and (3.12) together now specify a General*ized* Linear Mixed Model. This model has the advantage that the fifth drawback no longer applies, but a practical disadvantage is that software for *Gized*LMMs is not available in many packages. We again used Proc NLMixed of SAS. A syntax example is given in the appendix. We call this the *exact* likelihood approach, since the likelihood is based on the exact (i.e. binomial) within-study distribution of the data.

3.5.2 Choice of summary ROC curve

Above we have seen that a summary ROC curve can be obtained through a characterisation of the estimated bivariate normal distribution given by (3.9). One possibility is to take the regression line of η_i on ξ_i, as we did above. However, there are other possibilities as well. For example, we could take the regression line of ξ_i on η_i. We now discuss this and other possible choices.

1. The regression line of η_i on ξ_i:

   $$\eta = \bar{\eta} + \frac{\sigma_{\xi\eta}}{\sigma_{\xi}^2}(\xi - \bar{\xi}) \tag{3.13}$$

   This summary line estimates the mean logit transformed sensitivity given a specific value for the logit transformed 1- specificity. When transformed to the ROC space, the summary ROC curve estimates the median TPR given a specific value for the FPR.

2. The regression line of ξ_i on η_i :

   $$\xi = \bar{\xi} + \frac{\sigma_{\xi\eta}}{\sigma_{\eta}^2}(\eta - \bar{\eta})$$

   which is equivalent to:

   $$\eta = \bar{\eta} + \frac{\sigma_{\eta}^2}{\sigma_{\xi}\eta}(\xi - \bar{\xi}) \tag{3.14}$$

   There is no a priori reason to regress η_i on ξ_i instead of the other way around. Therefore an alternative summary line is obtained by regressing ξ_i on η_i. This summary line characterises the mean logit transformed 1-specificity given a specific value for the logit transformed sensitivity. When transformed to the ROC space, the summary ROC curve characterises the median FPR given a specific value for the TPR.

3. The regression line of D_i on S_i :
   Let $D_i = \eta_i - \xi_i$ and $S_i = \eta_i + \xi_i$, as in the classical SROC method. From

(3.9) it follows that the covariance of D and S is equal to $\sigma_\eta^2 - \sigma_\xi^2$ and the variance of S is equal to $\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta}$. The regression line therefore is

$$D = \bar{D} + \frac{\sigma_\eta^2 - \sigma_\xi^2}{\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta}}(S - \bar{S})$$

The popularity of this summary line is possibly explained by the fact that it has an appealing interpretation. Given S, which can be interpreted as a proxy for the positivity criterion of the diagnostic test, this regression line estimates D, which can be interpreted as the diagnostic log odds ratio.
In terms of η and ξ the regression line is

$$\eta = \bar{\eta} + \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}(\xi - \bar{\xi}) \tag{3.15}$$

This method is a kind of compromise between the vertical way of looking in the first method (median TPR given a specific value for the FPR) and the horizontal way of looking in the second method (median FPR given a specific value for the TPR), since its slope lies between the slopes in (3.13) and (3.14).

4. The Rutter and Gatsonis [9] summary ROC curve:
Their method leads to the summary line (see Section 3.5.3)

$$\eta = \bar{\eta} + \frac{\sigma_\eta}{\sigma_\xi}(\xi - \bar{\xi}) \tag{3.16}$$

This line can also be interpreted as a sort of compromise between the regression of η_i on ξ_i and that of ξ_i on η_i, since the slope is equal to the geometric mean of the slopes of the two regression lines in (3.13) and (3.14).

5. The major axis method:
The last possibility we mention is to characterize the bivariate normal distribution between ξ and η by the major axis that runs through the extreme points of the contour ellipses (defined by points having the same density) of the estimated bivariate distribution. This results in the summary line [37]

$$\eta = \bar{\eta} + \frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{(\sigma_\eta^2 - \sigma_\xi^2)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}(\xi - \bar{\xi}) \tag{3.17}$$

In fact, taking this line is analogous to summarizing a two dimensional distribution by its first principal component.
The summary ROC curves of methods 3-5 are symmetric in ξ and η; that is, if the labels of diseased and non-diseased test results and disease status are interchanged, the summary ROC curve does not change. For all of the mentioned summary lines, standard errors for the slope, intercept and for η at a given value for ξ can be calculated using the delta method. Confidence

intervals for the slope and intercept, and a confidence band for the summary line, are calculated using standard methods. A confidence band for the summary ROC curve is obtained by transforming the confidence band of the summary line. No extra programming or hand calculations are needed if a program like SAS Proc NLMIXED is used that allows user defined derived parameters.

3.5.3 Relationship with model of Rutter and Gatsonis

From (3.5) and (3.7) it follows that the model of Rutter & Gatsonis can be written as

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} (\bar{\theta} + X_0\bar{\alpha})e^{-X_0\beta} \\ (\bar{\theta} + X_1\bar{\alpha})e^{-X_1\beta} \end{pmatrix}, \Psi \right)$$

$$\text{where} \quad \Psi = \begin{pmatrix} (\sigma_\theta^2 + X_0^2\sigma_\alpha^2)e^{-2X_0\beta} & (\sigma_\theta^2 + X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta} \\ (\sigma_\theta^2 + X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta} & (\sigma_\theta^2 + X_1^2\sigma_\alpha^2)e^{-2X_1\beta} \end{pmatrix}$$

This specifies a bivariate normal distribution for (ξ_i, η_i), just as we do in (3.9). Note that the number of parameters is the same, too. Thus the two models are essentially the same, only the parameterisation is different. Rutter and Gatsonis [9] choose $X_0 = -1/2$ and $X_1 = 1/2$ and do not discuss other choices. One can check that their labeling leads to the summary line given by (3.16), with slope σ_η/σ_ξ. All other choices such that $X_0 = -X_1$ also lead to σ_η/σ_ξ. Alternative choices for X_0 and X_1 lead to other summary lines. For instance, the choice $X_0 = 0$ and $X_1 = 1$ leads to the η on ξ regression line given by (3.13). The choice $X_0 = 1$ and $X_1 = 0$ leads to the ξ on η regression given by (3.14). One can show that it is not possible to specify X_0 and X_1 such that it leads to the D on S regression line (3.15).

We conclude that our bivariate model is in principle identical to that of Rutter and Gatsonis [9]. A minor difference is the different parametrization. Another minor difference is that the slope in the Rutter and Gatsonis model is $e^{-\beta}$, and this it is restricted to be positive. We do not restrict the slope in our model, although in practice negative slope estimates will typically not occur. An important practical difference is that Rutter and Gatsonis follow a laborious Bayesian estimation approach, while our method can be carried out conveniently using standard statistical packages. Furthermore, we think our method is easier to understand, since it simply assumes a standard random effects model for the sensitivities and specificities simultaneously.

3.6 Results

3.6.1 The bivariate model

Example 1: FNAC of the Breast [16]

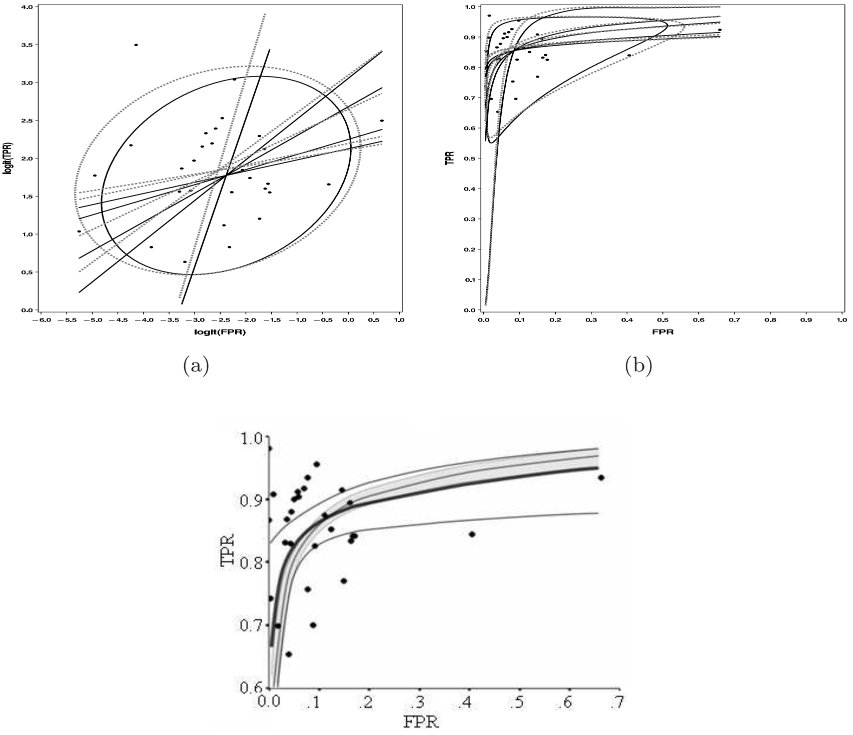
We fitted the bivariate model as described in Section 3.5.1 on the data of the 29 studies of the meta-analysis of Giard *et al.*[16]. The estimates of the means

Table 3.2: First data example: FNAC of the breast [16]. In the upper part estimates are given of the random intercept model (Section 3.5.1), using approximate as well as exact likelihood. In the lower part the parameter estimates are given for the five different choices of the summary ROC curves discussed in Section 3.5.2.

Parameter	Approximate likelihood		Exact likelihood	
	Estimate (se)		Estimate (se)	
$\bar{\eta}$	1.774 (0.114)		1.839 (0.119)	
$\bar{\xi}$	-2.384 (0.201)		-2.547 (0.225)	
σ_{η}^{2}	0.286 (0.093)		0.316 (0.104)	
σ_{ξ}^{2}	0.990 (0.313)		1.297 (0.411)	
$\sigma_{\xi\eta}$	0.146 (0.132)		0.141 (0.155)	
Type of SROC	Approximate likelihood		Exact likelihood	
	$\alpha(se)$	$\beta(se)$	$\alpha(se)$	$\beta(se)$
1. η on ξ	2.126 (0.32)	0.148 (0.13)	2.115 (0.32)	0.108 (0.12)
2. ξ on η	6.431 (3.95)	1.954 (1.65)	7.560 (6.13)	2.246 (2.39)
3. D on S	2.680 (0.37)	0.380 (0.15)	2.647 (0.37)	0.318 (0.14)
4. Rutter & Gats.	3.054 (0.31)	0.537 (0.12)	3.096 (0.32)	0.494 (0.11)
5. Major axis	2.249 (0.42)	0.199 (0.17)	2.196 (0.41)	0.141 (0.15)

and variances of η_i and ξ_i resulting from the approximate and exact likelihood approach are presented in the upper part of Table 3.2. Based on these estimates, the results for the five different choices of the summary ROC curve (Section 3.5.2) are presented in the lower part of Table 3.2. In Figure 3.3 the different ROC curves are depicted, in the logit-logit space as well as in the ROC space. Also the 95% coverage regions are given. These regions are based on the fitted bivariate distribution and estimate the area that contains approximately 95% of the true pairs of $(logit(FPR), logit(TPR))$ and (FPR, TPR) respectively.

From Table 3.2 and Figure 3.3 it is clear that the results of the exact and approximate approach are similar in this data example. The exact approach results in a somewhat more favourable average sensitivity and specificity. This was to be expected beforehand for two reasons. First, as mentioned in Section 3.3, adding 0.5 to the numbers in the fourfold table, as is done in the approximate approach, results in estimated mean sensitivity and specificity that are biased downwards, pushing the ROC curve away from the left upper corner. Second, as shown by Chang *et al.* [38], even if it is not needed to add 0.5, the estimates of the mean sensitivity and specificity are still somewhat biased towards 0.5. This is due to the fact that the approximate approach does not account for the correlation between the logit(TPR) and its variance, and between the logit(FPR) and its variance.
From Table 3.2 and Figure 3.3 it is clear that the difference among the different types of the summary ROC curve is substantial, especially for the first two choices $'\eta$ on ξ' and $'\xi$ on η'. As one can see on the basis of the formulas for the slopes given in Section 3.5.2, the first two types ($'\eta$ on ξ' and $'\xi$ on η') give a kind of lower and upper bound for the estimated summary ROC curves, and types 3 to 5 lie between these two curves. In fact, the slopes of choices 3 to 5 could be considered different kinds of 'weighted averages' of the slopes of methods 1 and 2. In this



(a)

(b)

(c)

Figure 3.3: SROC curves for the five different choices of the summary ROC curve, as a graphical illustration of Table 3.2. The curves are presented in logit-logit space (Figure 3.3(a)) as well as in the ROC space (Figure 3.3(b)). Also the 95% coverage regions are given as an ellipse in Figure 3.3(a) and a ’triangle’ in Figure 3.3(b). The solid lines present the results of the approximate likelihood, the grey dashed lines present the results of the exact likelihood. The lines in 3.3(a) and 3.3(b) represents for the ’ξ on η’, ’R&G’, ’D on S’, ’Major Axis’ and ’η on ξ’ in decreasing order of slope (from the right top to bottom). In Figure 3.3(c) the SROC curves of the random intercept model (solid lines) versus the fixed Littenberg & Moses model (grey dashed lines) are given together with their confidence intervals.

Table 3.3: Second data example: Imaging tests for coronary artery disease [20]. In the upper part estimates are given of the random intercept model (Section 3.5.1), using approximate as well as exact likelihood. In the lower part the parameter estimates are given for the five different choices of the summary ROC curves discussed in Section 3.5.2.

Parameter	Approximate likelihood		Exact likelihood	
	Estimate (se)		Estimate (se)	
$\bar{\eta}$	1.257 (0.057)		1.339 (0.061)	
$\bar{\xi}$	-1.560 (0.071)		-1.851 (0.085)	
σ^2_η	0.333 (0.057)		0.406 (0.066)	
σ^2_ξ	0.337 (0.074)		0.585 (0.117)	
$\sigma_{\xi\eta}$	0.182 (0.049)		0.272 (0.065)	
Type of SROC	Approximate likelihood		Exact likelihood	
	$\alpha(se)$	$\beta(se)$	$\alpha(se)$	$\beta(se)$
1. η on ξ	2.098 (0.21)	0.540 (0.13)	2.199 (0.18)	0.465 (0.10)
2. ξ on η	4.106 (0.69)	1.827 (0.45)	4.102 (0.58)	1.493 (0.31)
3. D on S	2.802 (0.26)	0.991 (0.17)	2.802 (0.23)	0.791 (0.12)
4. Rutter & Gats.	2.805 (0.21)	0.993 (0.13)	2.880 (0.19)	0.833 (0.10)
5. Major axis	2.796 (0.37)	0.987 (0.24)	2.677 (0.28)	0.723 (0.15)

example the curves for approaches 3, 4 and 5 lie closer to the regression of η on ξ, but in general that is not the case. Results depend on the variances of ξ_i and η_i, and the covariance between them. The more similar the variances of ξ_i and η_i are, the more similar will be the results of approaches 3-5.

For all summary ROC curves given in Figure 3.3(a) and 3.3(b) a 95% confidence band can be calculated. As an example, we have drawn in Figure 3.3(c) the $'D$ on S' summary ROC curve together with its 95% confidence band. From all 5 types of summary ROC curves, this $'D$ on S' summary ROC curve should be most comparable to the standard summary ROC curve from the Littenberg&Moses[2] (L&M) approach, which also estimates the regression of D on S. To compare the two, we have also drawn the L&M summary curve and its confidence band in Figure 3.3(c). The L&M summary ROC curve has a slope that is considerably steeper than our $'D$ on S' curve, leading to larger estimated sensitivities if the specificity is small, and smaller estimated sensitivities if the specificity is large. This is not a general pattern, as will be seen from the second data example. Furthermore, it is seen that the L&M approach grossly underestimates the variability in the data, leading to a much too narrow confidence band. This is due to the fact that the L&M approach is based on a fixed effects model, which erroneously assumes that there is no between-studies variability.

Example 2: Imaging tests for coronary artery disease [20]

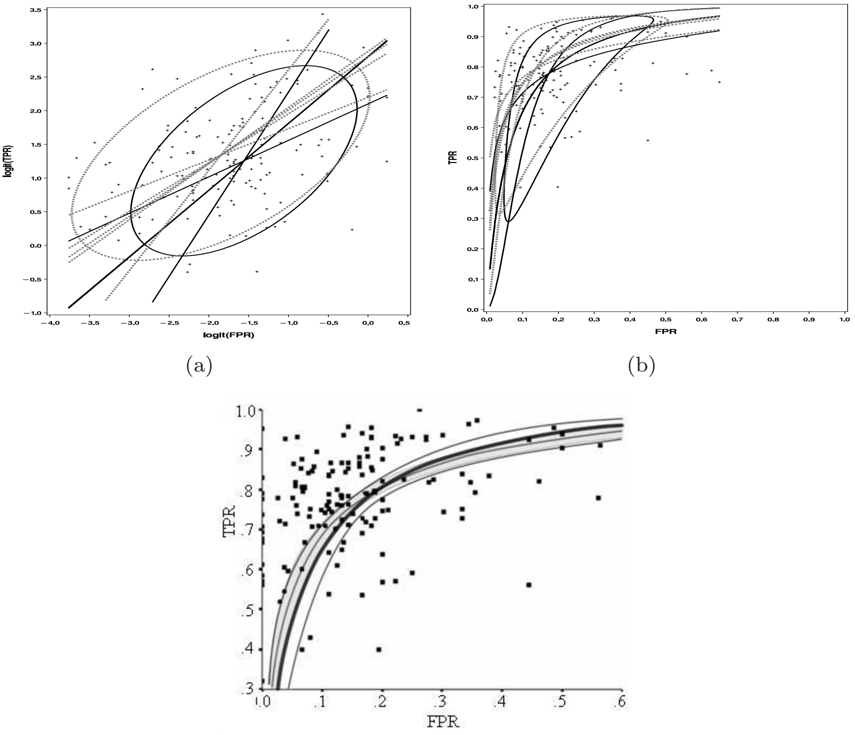
We fitted the bivariate model of Section 3.5.1 on the data of the 149 studies included in the meta-analysis of Heijenbrok-Kal *et al.* [20]. The estimates of the means and variances of η_i and ξ_i based on the approximate and exact likelihood approach are presented in the first part of Table 3.3. Based on these estimates, the results for the five different choices of the summary ROC curve (Section 3.5.2) are presented in the lower part of Table 3.3. In Figure 3.4(a) and 3.4(b) the different summary ROC curves are given for the exact and the approximate approach, in the logit-logit space as well as in the original ROC space.

In this example the results of the approximate and exact likelihood approach are also similar. In the approximate likelihood approach the variances of ξ and η are almost equal, which results in very little differences among methods 3 to 5. For the exact likelihood approach the difference between the two variances is somewhat larger, leading to somewhat larger differences between the methods 3 to 5. Notice that in Figure 3.4 considerably more than 5 percent of the studies fall outside the 95% coverage region. However, this is expected since the coverage ellipse describes the variation between the true pairs of sensitivity and specificity, while the points in the plot represent the estimates (observed) pairs of sensitivity and specificity. The observed points, of course, should show more variation due to within-study sampling variability. In Figure 3.4(c) we compare again our $'D$ on S' summary ROC with the standard L&M one. In contrast to the previous example, now the slope of the L&M ROC is smaller than that of our $'D$ on S' curve. Again it is clear that L&M method leads to smaller standard errors.

3.7 Discussion

Meta-analysis of diagnostic tests requires statistical techniques that analyse pairs of related summary statistics (e.g. sensitivity and specificity) rather than a single statistic. In the literature numerous meta-analyses are published in which one is interested in meta-analysing only sensitivities or only specificities. For these situations the standard method of analysis is the DerSimonian-Laird univariate random effects model. The method we propose in this article is a direct extension of that approach. We analyse sensitivities and specificities simultaneously using a two-dimensional random effects model. Our method could also be seen as an extension of the approach of Littenberg and Moses[2]. Their model implies a linear relationship between the logit transformed sensitivity and specificity, which can be transformed into ROC space to obtain a summary ROC curve. Despite its many drawbacks, their approach still seems to be the most popular method for meta-analysis of diagnostic accuracy data where pairs of sensitivity and specificity per study are available. This is probably due to the fact that the method is very easy to carry out in practice. In this article we have shown that the bivariate meta-analysis model addresses all its shortcomings.

The method of Rutter and Gatsonis[9] is also an appropriate alternative for the Littenberg and Moses method and recently it was pointed out how this method can be performed in a non-Bayesian way using standard statistical software[30]. We have shown that their model is essential the same as ours, only the parameterisation is different and it leads to a special type of summary ROC. In our opinion the way their model is less straightforward and more difficult to understand. The way we present the bivariate model fits into the standard framework of multivariate meta-analysis[13, 25, 35, 39] thereby bringing the meta-analysis of this kind of diagnostic back into mainstream meta-analysis methods, which can



(a)

(b)

(c)

Figure 3.4: Second data-example of Heijenbrok-Kal *et al.* ROC curves for the five different choices of the summary ROC curve, as a graphical illustration of Table 3.3. The curves are presented in logit-logit space (Figure 3.4(a)) as well as in the ROC space (Figure 3.4(b)). Also the 95% coverage regions are given as an ellipse in Figure 3.4(a) and a 'triangle' in Figure 3.4(b). The solid lines present the results of the approximate likelihood, the grey dashed lines present the results of the exact likelihood. The lines in 3.4(a) and 3.4(b) represents for the 'ξ on η', 'R&G', 'D on S', 'Major Axis' and 'η on ξ' in decreasing order of slope (from the right top to bottom). In Figure 3.4(c) the SROC curves of our random intercept model (solid lines) versus the fixed Littenberg & Moses model (grey dashed lines) are given together with their confidence intervals.

be fitted straightforwardly in standard statistical software. The SAS syntax we used for the examples is given in the appendix to make the method easily accessible to meta-analysists. The approach presented in this paper is straightforwardly extended with covariates. In a General Linear Mixed Model program like Proc Mixed both the mean logit(FPR) and mean logit(TPR) can be allowed to depend on covariates. If a Generalized Linear Mixed Model program such as SAS Proc NLMixed is used, there are many more possibilities, depending on how the model is parameterized.

Our contribution in this paper has been to compare different types of summary ROC curves by showing how they are all related to our bivariate model. We discussed 5 types of summary ROC curves, each of which has its own interpretation and properties. In the Littenberg and Moses approach, the choice is made explicitly as the regression of D on S. In the approach of Rutter & Gatsonis[9], the choice is implicitly made, and we pointed out that it is a kind of geometric mean between the regression line of logit(TPR) on logit(FPR) and the regression line of logit(FPR) on logit(TPR). Thus the two methods estimate different summary curves and the resulting curves are therefore in principle not the same. If one wants to describe the median sensitivity of studies with a fixed value of the specificity, one can choose the regression of η_i on ξ_i, but if one prefers to describe the median specificity with a fixed sensitivity, one can choose the regression of ξ_i on η_i. The other three summary ROC curves are compromises between these two options.

Notice that in our approach no assumptions about individual study curves have to be made. For example, the method does not require an underlying continuous diagnostic test, and hence it can also be applied to intrinsically dichotomous tests. Our bivariate random effects model simply leads to a description of the distribution of the pairs (ξ_i, η_i). The summary ROC curve is just a one-dimensional representation of this distribution and cannot be interpreted as a kind of average curve or a curve typical for the study specific ROC curves. It can have a shape that is very different from the study specific shapes. The different kinds of summary ROC curves discussed in Section 3.5.2 are still interpretable, even if the test is intrinsically dichotomous. E.g. the regression of η_i on ξ_i simply describes the median sensitivity of studies with a fixed value of the specificity.

We fitted our models with standard software based on straightforward likelihood methods. In our examples this approach worked well, although sometimes some convergence problems were met. In our two clinical data examples these convergence problems were addressed by specifying better starting values. However, we can imagine that, especially for small meta-analyses, this could be more of a problem. An alternative is to fit the models in a Bayesian way. This can be done using the free available software program WinBugs [40]. The Bayesian approach has the advantage of being more flexible; for instance, one can assume non-normal parametric distributions for logit(TPR) and logit(FPR). Also in applications with a relatively small number of studies, the Bayesian method might perform better, since the standard likelihood is based on large sample theory. A disadvantage is that it is more time consuming and is less easily done by non-statisticians.

The bivariate model we proposed in this paper can be fitted using approximate or exact likelihood. Using approximate likelihood has the advantage that a

General Linear Mixed Model program can be used, which is widely available. The exact likelihood method can only be used if one has an appropriate Gneral*ized* Linear Mixed Model program available. Unfortunately these programs are still rather scarce. Simulation studies are needed to compare the two approaches.

Appendix

In this appendix we provide the SAS syntax needed to reproduce the results given in Table 3.2 and 3.3. First we describe the data format that is needed. We have two records per study, one for the diseased and one for the healthy group, as in the following table.

study	group	n	npos	disease	healthy	y	est
.	.	.	.	.	.	.	0.20
.	.	.	.	.	.	.	0.10
.	.	.	.	.	.	.	0.20
1	1	1009	70	0	1	-2.58974	0.01525
1	2	1068	979	1	0	-2.58974	0.01219
2	3	166	3	0	1	-3.84405	0.29183
2	4	73	51	1	0	-3.84405	0.06386
3	5	949	25	0	1	-2.77988	0.01914
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The meaning of the variables is:
study = number of the study
group = unique identifier for the diseased and healthy group
n = number per group
npos = number with positive diagnostic test
disease = 0 for healthy group, = 1 for diseased group
healthy = 1 for healthy group, = 0 for diseased group
y = ln(npos/(n-npos))
est = 1/(npos+0.5) + 1/(n-npos+0.5)
The first three lines have only a non-missing value for the variable est. These three values serve as starting values for the variances and covariance of ξ_i and η_i. The following syntax produces the approximate likelihood results given in the upper part of Table 3.2.

```
proc mixed cl method=ml data=giardcol;  *call procedure; cl asks for confidence
                       intervals of covariance parameters;
class study d group;   *study, d and group are classification variables;
model y=disease healthy/noint s cl covb ddf=1000,1000;  *model with indicator
                       variables disease and healthy as explanatory variables
                       for log odds. Covb asks for covariance matrix of fixed
                       effects parameters;
random disease healthy/subject=study type=un s;  *indicators of diseased and
                       healthy group are random effects, possibly correlated
                       within-study and independent between-studies, covariance
                       matrix is unstructured. Print empirical Bayes estimates ’s’;
parms/parmsdata=giardcol eqcons=4 to 61;  *data file giardcol.sd2 contains the
                       variable ’est’ with starting values for the three
                       covariance parameters of the random effects together with
                       the 58 within  study-group variances. The latter are
                       assumed to be known and should be kept fixed;
repeated/group=group;  *each group in a study (diseased and healthy) has its own
                       within-study arm variance; within-study estimation errors
                       are independent;
run;
```

The different summary ROC curves have to be calculated by hand based on the output of the program. The same results can also be obtained with SAS Proc NLmixed. The advantage is that the parameters of the summary ROC curves can be specified as derived parameters. The syntax is as follows.

```
proc nlmixed data=giardcol; * call procedure;
parms meaneta=1.8 meanksi=-2.4 vareta=0.3 varksi=1 covksieta=0.15 ; *choose
                        starting values for means, variances and covariance;
model y~normal(eta*disease+ksi*healthy,est);   *log odds y are alternately
                        normally distributed around eta and ksi; disease
                        and healthy are indicator variables, 'est' contains
                        the within-study variances;
random ksi eta~normal([meanksi,meaneta],[varksi,covksieta,vareta]) subject=study;
                        * the shrunk parameters ksi and eta are normally
                        distributed around their common means, with between
                       study variances varksi and vareta and covariance covksieta;
estimate 'eta on ksi: beta' covksieta/varksi; *estimate slope of ksi on eta
                        regression line;
estimate 'eta on ksi: alpha' meaneta-covksieta/varksi*meanksi; *estimate
                        intercept of eta on ksi regression line;
estimate 'ksi on eta: beta' vareta/covksieta; *estimate slope of ksi on
                        eta regression line;
estimate 'ksi on eta: alpha' meaneta-vareta/covksieta*meanksi; *estimate
                        intercept of ksi on eta regression line;
estimate 'D on S: beta' (vareta+covksieta)/(varksi+covksieta); *estimate
                        slope of D on S regression line;
estimate 'D on S: alpha' meaneta-(vareta+covksieta)/(varksi+covksieta)*meanksi;
                        *estimate intercept of D on S regression line;
estimate 'R&G: beta' (vareta**0.5)/(varksi**0.5); *estimate slope of R&G
                        regression line;
estimate 'R&G: alpha' meaneta-(vareta**0.5)/(varksi**0.5)*meanksi; *estimate
                        intercept of R&G regression line;
estimate 'major axis: beta' (vareta-varksi+((vareta-varksi)**2+4*covksieta**2)
*0.5)/(2*covksieta);        *estimate slope of major axis regression line;
estimate 'major axis: alpha' meaneta-(vareta-varksi+((vareta-varksi)**2+
4*covksieta**2)**0.5)/(2*covksieta)*meanksi; * estimate intercept of
                            major axis regression line;
run;
```

The following syntax reproduces the right half (exact likelihood) of Table 3.2.

```
proc nlmixed data=giardcol; *call procedure;
parms meaneta=1.8 meanksi=-2.4 vareta=0.3 varksi=1 covksieta=0.15 ; *choose
                    starting values for means, variances and covariance;
pi = 1/(1+exp(-(eta*disease+ksi*healthy))); *calculating the 'true' TPR
                    and FPR (pi);
model npos~binomial(n,pi); *the positive numbers in both groups follow
                    binomial distributions;
random ksi eta ~ normal([meanksi,meaneta],[varksi,covksieta,vareta])
subject=study;     *the shrunk parameters ksi and eta are normally
                    distributed around their common means, with between
                    study variances varksi and vareta and covariance covksieta;
estimate 'eta on ksi: beta' covksieta/varksi; *estimate slope of eta
                    on ksi regression line;
estimate 'eta on ksi: alpha' meaneta-covksieta/varksi*meanksi; *estimate
                    intercept of eta on ksi regression line;
estimate 'ksi on eta: beta' vareta/covksieta; *estimate slope of ksi
                    on eta regression line;
estimate 'ksi on eta: alpha' meaneta-vareta/covksieta*meanksi; *estimate
```

```
                    intercept of ksi on eta regression line;
estimate ’D on S: beta’ (vareta+covksieta)/(varksi+covksieta); *estimate
                    slope of D on S regression line;
estimate ’D on S: alpha’ meaneta-(vareta+covksieta)/(varksi+covksieta)
*meanksi;         *estimate intercept of D on S regression line;
estimate ’R&G: beta’ (vareta**0.5)/(varksi**0.5); *estimate slope of
                    R&G regression line;
estimate ’R&G: alpha’ meaneta-(vareta**0.5)/(varksi**0.5)*meanksi;
                    *estimate intercept of R&G regression line;
estimate ’major axis: beta’ (vareta-varksi+((vareta-varksi)**2 +
4*covksieta**2)**0.5)/(2*covksieta); *estimate slope of major
                    axis regression line;
estimate ’major axis: alpha’ meaneta-(vareta-varksi+((vareta-varksi)**2
+ 4*covksieta**2)**0.5)/(2*covksieta)*meanksi;   *estimate intercept of
                    major axis regression line;
run;
```

References

[1] Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith DG, Altman DG, eds. Systematic reviews in health care: Meta-analysis in context. London: BMJ Publishing Group; 2001.

[2] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conflicting reports - a new metaanalytic method. Medical Decision Making 1993;13:313-321.

[3] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-1316.

[4] Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995;117:167-178.

[5] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48:119-130.

[6] Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of roc curves. Med Decis Making 1999;19:252-264.

[7] Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. Journal of Clinical Epidemiology 1999;52:943-951.

[8] Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000;20:430-439.

[9] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865-2884.

[10] Walter SD. Properties of the summary receiver operating characteristic (sroc) curve for diagnostic test data. Statistics in Medicine 2002;21:1237-1256.

[11] Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of tresholds. Biometrics 2003;59:936-946.

[12] McClish DK. Combining and comparing area estimates across studies or strata. Med Decis Making 1992;12:274-279.

[13] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: Multivariate approach and meta-regression. Stat Med 2002;21:589-624.

[14] Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: Three clinical meta-re-analyses. Statistics in Medicine 2000;19:3497-3518.

[15] Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in a meta-analysis:An application. Statistics in Medicine 2003;22:1335-1353.

[16] Giard RWM, Hermans J. The value of aspiration cytologic examination of the breast - a statistical review of the medical literature. Cancer 1992;69:2104-2110.

[17] Mushlin AI. Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. Ann Intern Med 1985;103:79-85.

[18] Committee. HaPP. The use of diagnostic tests for screening and evaluating breast lesions. Ann Intern Med 1985;103:147-151.

[19] Dixon JM, Clarke PJ, Crucioli V, Dehn TCB, Lee ECG, Greenal MJ. Reduction of the surgical excision rate in benign breast disease using fine needle aspiration cytology with immediate reporting. Br J Surg 1987;74:1014-1016.

[20] Heijenbrok MH. Assessment of diagnostic imaging technologies for cardiovascular disease. Rotterdam: Epidemiology and Biostatistics, Erasmus MC. 2004.

[21] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:177-188.

[22] Normand SL. Meta-analysis: Formulating, evaluating, combining, and reporting. Stat Med 1999;18:321-359.

[23] Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 1998;17:841-856.

[24] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 1994;309:1351-1355.

[25] Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. Statistics in Medicine 1998;17:2537-2550.

[26] Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). Statistics in Medicine 1994;13:293-296.

[27] Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London: Chapman & Hall; 1995.

[28] Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990;29:12-22.

[29] Rothman K, Greenland S. Modern epidemiology. Philadelphia: Lippincott Williams & Wilkins; 1998.

[30] Macaskill P. Empirical bayes estimates generated in a hierarchical summary roc analysis agrees closely with those of a full bayesian analysis. Journal of Clinical Epidemiology 2004;57:925-932.

[31] Visser K, Hunink MG. Peripheral arterial disease: Gadolinium-enhanced mr angiography versus color-guided duplex us–a meta-analysis. Radiology 2000;216:67-77.

[32] Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic resonance angiography compared with digital subtraction angiography in carotid artery stenosis: A systematic review. Stroke 2003;34:1324-1332.

[33] Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. Mr imaging of the menisci and cruciate ligaments: A systematic review. Radiology 2003;226:837-848.

[34] Harbord R, Steichen T. Metareg: Stata module to perform meta-analysis regression. 2004: Statistical Software Components S4446201.

[35] van Houwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine 1993;12:2272-2284.

[36] McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. Statistics in Medicine 1996;15:1713-1728.

[37] Johnson R, Wichern D. Applied multivariate statistical analysis. London: Prentice Hall; 2002.

[38] Chang B, Waternaux C, Lipsitz S. Meta-analysis of binary data: Which within study variance estimate to use? Statistics in Medicine 2001;20:1947-1956.

[39] Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. Psychological Methods 1996;1:227-235.

[40] Spiegelhalter DJ, Thomas A, Best NG, Lunn D. Winbugs user manual, version 1.4. Cambridge: MRC Biostatistics Unit; 2003.

CHAPTER 4

Meta-analysis of Diagnostic Studies: a Comparison of Random Intercept, Normal-Normal and Binomial-Normal Bivariate Summary ROC Approaches

Abstract

The summary ROC method of Littenberg and Moses (L&M) has many shortcomings and recently improvements upon it were proposed. In this paper we consider three recently introduced refinements of the L&M SROC that aim to meet its shortcomings. All the three methods assume a linear relationship between the difference and sum of the logit transformed true and false positive rates. The first one is based on a random intercept linear meta-regression model (RI). The other two are based on the bivariate normal random effects meta-analysis model but with a different within-study model of sensitivity and specificity: the approximate normal distribution (normal-normal model (NN)) or the binomial distribution (binomial-normal model (BN)). We evaluate the performances of the three methods in a simulation study. The parameters to be compared are the associated intercept, slope and residual variance, using bias, mean squared error and coverage probabilities. We varied the overall sensitivity and specificity, the between-studies variance, the within-study sample size and the number of studies. The methods are illustrated using published meta-analysis data.

In general the BN method performed better than the RI and NN method. It always gave unbiased estimates of the intercept and slope parameter. The coverage probabilities were also reasonably acceptable, unless the number of studies was very small. In contrast the RI and NN methods could produce large biases with poor coverage probabilities, especially when sample sizes of individual studies were small or when sensitivities or specificities were close to 1. Although this was rare in our simulations, the bivariate methods can suffer from non-convergence mostly due to the correlation being close to ±1.

4.1 Introduction

In meta-analysis of diagnostic test accuracy studies often summary ROC curves are presented. When test results are presented by one two-by-two table per study, the SROC method of Littenberg and Moses (L&M) and Moses *et al.* [1, 2] has been used for long as the standard method in practice to construct a summary ROC curve. Their approach is based on a simple regression of the difference of logit(sensitivity) and logit(1-specificity) on the sum of the two, and is easily carried out in any statistical package. However, as pointed out by several authors [3, 4, 5], it has serious shortcomings. First, it does not account for the within- and between-studies variability in an appropriate way. Second, it assumes that the predictor variable is measured error free. Third, it does not account for the within-study correlation between the response and the predictor variable. Fourth, it needs a correction factor when there is a zero cell count. Therefore more appropriate methods have been proposed, among them the hierarchical SROC method [5] and the SROC curves based on the random effects bivariate approach [3, 4]. In this article we consider three refinements of the L&M SROC method that overcome some or all of its shortcomings and compare their performances through a simulation study.

The first refinement is the random intercept meta-regression approach, which has been introduced recently in practice (see for example [6, 7, 8]). It appropriately deals with the within- and between-studies variability and therefore meets the first of the above mentioned shortcomings. It is an obvious improvement upon the L&M method and in this paper we will refer to it as the random intercept (RI) SROC approach. It can be easily carried out in practice using one of the widely available random coefficients regression programs. One of the potential disadvantages of the RI method is that it treats the predictor variable in the regression as error free, which might lead to bias in the slope of the regression line [9, 10]. No research has been done to investigate if this is serious in practice.

The other two refinements stem from the bivariate random effects model proposed for the joint analysis of pairs of sensitivities and specificities [3]. Arends *et al.* [4] discuss five different possible choices of SROC curves based on the bivariate random effects model. In this paper we consider the L&M type SROC. The difference between the two refinements is the way they handle the within-study distribution of sensitivity and specificity. This can be done in two ways, by an approximate normal distribution [3, 4] or by a binomial distribution [4, 11]. In this paper we will refer to these methods as the normal-normal bivariate random effects (NN) and the binomial-normal bivariate random effects (BN) SROC approach, respectively. Compared to the RI method, the NN and BN method are more complicated and less easily carried out in practice. The NN can be carried out in linear mixed model programs in which certain options are implemented. The BN model can be fitted in a generalized linear mixed model programs.

Recently, Hamza *et al.* [12] compared, in an extensive simulation study, the approximate normal with the binomial within-study models in the univariate case where either only sensitivities or only specificities were meta-analyzed, and concluded that the binomial distribution, in many cases, gives substantially less bias in the estimate of the overall sensitivity or specificity. Chu and Cole [13] and Riley *et al.* [14], in (limited) simulation studies, showed that these results also hold for the bivariate case. However, these results do not immediately carry over to SROC curves, because the biases in the estimates of the mean sensitivity and specificity could cancel out. Therefore the aim of this article is to compare by a simulation study the performances of the RI, NN and BN SROC approaches to estimate a SROC curve. We investigated the effect of, 1. the number of studies included in the meta-analysis, 2. the within-study sample size, 3. the between-studies variability and 4. the true median sensitivity and specificity.

In Section 4.2 we shortly review the L&M method and introduce the three different random effects models to be compared. In Section 4.3 we describe the design of the simulation study and in Section 4.4 the simulation results. In Section 4.5 we illustrate the methods using published meta-analysis data. We end with a discussion in Section 4.6.

4.2 Method

The methods in this paper assume the setting where each study presents per group, diseased or non-diseased, the number of test positive and test negative subjects. Alternatively a pair of sensitivity and specificity with standard errors and the number of diseased and non-diseased subjects might be given. Let x_1 denote the number of true positives, n_1 the total number of diseased subjects, x_0

the number of false positives and n_0 the total number of non-diseased subjects. Then the observed sensitivity and the observed specificity are given by x_1/n_1 and $(n_0 - x_0)/n_0$ respectively.

4.2.1 Standard L&M approach

The standard SROC method of Littenberg and Moses [1, 2] estimates the following simple regression model:

$$\hat{D}_i = \alpha + \beta\hat{S}_i + e_i \tag{4.1}$$

Here i denotes the number of the study. $\hat{D}_i$ and $\hat{S}_i$ are calculated as the difference and the sum of $logit(sensitivity) = ln(x_{1i}/(n_{1i} - x_{1i}))$ and $logit(1 - specificity) = ln(x_{0i}/(n_{0i} - x_{0i}))$, respectively, and e_i is the random residual error. D is the log of the diagnostic odds ratio as it is the logarithm of the odds of a positive test result among diseased individuals relative to the odds of positivity among healthy individuals. Therefore it is considered to be a measure of the discriminative power of the test in the i^{th} study. S can be considered as a measure of test positivity as it directly depends on the choice of the threshold. The parameter α is called the accuracy parameter, as it is the log diagnostic odds ratio at the point where $S = 0$ (where sensitivity is equal to specificity). The interpretation of β is less straightforward than α. It measures the asymmetry of the ROC curve. If $\beta = 0$ the curve is symmetric and the overall common odds ratio is given by e^{α}. For a detailed discussion of the interpretation of these parameters we refer to [15].

Model (4.1) is estimated using *weighted* or *unweighted* least squares. The *weighted* analysis assumes that $e_i \sim N(0, \phi se_i^2)$ where se_i is the estimated standard error of the log odds ratio of study i, $\hat{D}_i$, considered as known. ϕ is a dispersion parameter meant to capture between-studies heterogeneity. Note that this is less appropriate since it is multiplicative instead of additive. In pratice ϕ is often assumed to be one and then the method reduces to a fixed effects approach. The *unweighted* analysis assumes $e_i \sim N(0, \sigma^2)$, where σ^2 is estimated from the data. An obvious disadvantage is that it does not take into account the differences in sample size between-studies. In fact, the term σ^2 stands for a mix of within-study variation, different between-studies, and between-studies variation, and does not explicitly separate these two sources of variation. Thus both the *weighted* as the *unweighted* variant do not adequately model the within and between-studies variation, but were motivated by the fact that the model could be fit by ordinary linear regression programs. Arends *et al.* [4] and Rutter *et al.* [5] also discuss several other shortcomings of the SROC method. The RI approach discussed in Section 4.2.2 appropriately assumes $e_i \sim N(0, \sigma^2 + se_i^2)$. This is an obvious improvement upon both variants of the L&M method and hence we did not include the L&M method in the simulation study.

4.2.2 RI SROC approach

A straightforward improvement upon L&M's method is provided by the standard random effects meta-regression model as developed in the last decade, see for instance [16, 17]. This method has recently been applied in the medical literature (for example [7, 18, 19]). The model can be written as a random intercept model as follows.

$$\hat{D}_i = \alpha_i + \beta \hat{S}_i + e_i \quad \text{with} \quad \alpha_i \sim N(\bar{\alpha}, \sigma^2_\alpha) \tag{4.2}$$

The variance of σ^2_α quantifies the variability between-studies having the same threshold. The observed log diagnostic odds ratio, $\hat{D}_i$, is assumed to follow a normal distribution with mean $\alpha_i + \beta \hat{S}_i$ and within-study variance calculated from the data as $\hat{\sigma}^2_{Di} = 1/x_{1i} + 1/(n_{1i} - x_{1i}) + 1/x_{0i} + 1/(n_{0i} - x_{0i})$. The $\hat{\sigma}^2_{Di}$'s are assumed to be known, an assumption that might be less appropriate when the sample size is small. Often this within-study variance is correlated with the estimated parameter $\hat{D}_i$, which can lead to bias in the parameter estimates [16, 20, 21]. Since model (4.2) is a simple linear random effects model, the parameters can be estimated using a linear mixed model program.

Continuity Correction: When a zero count is encountered in one or more of the cells, $\hat{D}_i$ and its variance are undefined, and hence a continuity correction has to be added to the numbers in the fourfold table for the RI and NN methods. Sweeting *et al.* [22] discussed different (constant) correction factors. In this paper we added 0.5, the usual correction factor in practice, throughout all simulations, except for some selected scenarios where also 0.1 was considered to see the effect of it.

4.2.3 Bivariate Random Effects approach

Among others, Reitsma *et al.* [3] and Arends *et al.* [4] advocated the bivariate approach to meta-analyze sensitivity and specificity jointly, incorporating the correlation that might exist between these two measures. The model can be given as:

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix}, \begin{pmatrix} \sigma^2_\xi & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma^2_\eta \end{pmatrix} \right) \tag{4.3}$$

Here ξ_i is the true $logit(1 - specificity)$ and η_i is the true $logit(sensitivity)$ of studies i. σ^2_ξ and σ^2_η are the between-studies variances of ξ_i and η_i respectively and $\sigma_{\xi\eta}$ is the between-studies covariance between ξ_i and η_i. The mean $logit(1 - specificity)$ and $logit(sensitivity)$ over studies are given by $\bar{\xi}$ and $\bar{\eta}$. Note that, since sensitivity and specificity tend to be negatively correlated, ξ_i and η_i tend to be positively correlated. Without loss of generality, one can also model $logit(specificity)$ instead of $logit(1 - specificity)$ (for example see [3]).

The bivariate method leads to an estimate of the two-dimensional distribution of the underlying true sensitivities and specificities. A summary ROC curve can be considered just as the curve corresponding to a line that characterizes the estimated bivariate normal distribution. Arends *et al.* [4] discussed five choices for such a line, one of them being the regression line of $D_i = \eta_i - \xi_i$ on $S_i = \eta_i + \xi_i$. For this choice, the bivariate method can be seen as a direct extension of the L&M or RI SROC approach. Unlike these two methods, that assume S_i to be error free, the bivariate approach accounts for error in S_i. In this paper we consider the regression line of D on S and equivalently model (4.3) can be re-written as follows:

$$\begin{pmatrix} S_i \\ D_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{S} \\ \bar{D} \end{pmatrix}, \begin{pmatrix} \sigma_S^2 & \sigma_{SD} \\ \sigma_{SD} & \sigma_D^2 \end{pmatrix} \right) \tag{4.4}$$

The parameters for the regression line $D_i = \alpha_i + \beta S_i$ can be derived from (4.4) as

$$\beta = \frac{\sigma_{SD}}{\sigma_S^2} \quad , \quad \bar{\alpha} = \bar{D} - \frac{\sigma_{SD}}{\sigma_S^2}\bar{S} \quad \text{and} \quad \sigma_\alpha^2 = \sigma_D^2 - \frac{\sigma_{SD}^2}{\sigma_S^2}$$

The standard error of these parameters can be calculated using the Delta method. The model is completed by modeling the within-study variability. In this paper we considered two distributions to model the within-study variability, the normal and the binomial.

NN SROC approach

This method approximates the within-study distribution of the observed data by a bivariate normal distribution

$$\begin{pmatrix} \hat{S}_i \\ \hat{D}_i \end{pmatrix} \sim N\left(\begin{pmatrix} S_i \\ D_i \end{pmatrix}, \begin{pmatrix} \hat{\sigma}_S^2 & \hat{\sigma}_{SD} \\ \hat{\sigma}_{SD} & \hat{\sigma}_D^2 \end{pmatrix} \right) \tag{4.5}$$

with $\hat{\sigma}_S^2 = \hat{\sigma}_D^2 = \frac{1}{x_{0i}} + \frac{1}{n_{0i}-x_{0i}} + \frac{1}{x_{1i}} + \frac{1}{n_{1i}-x_{1i}}$ and $\hat{\sigma}_{SD} = \frac{1}{x_{1i}} + \frac{1}{n_{1i}-x_{1i}} - \frac{1}{x_{0i}} - \frac{1}{n_{0i}-x_{0i}}$. Note that the within-study correlation is accounted for in (4.5) and the covariance matrix is assumed to be known, as usually done in meta-analysis. Model (4.4 & 4.5) together specify a linear random effects model and the parameters can be estimated by likelihood methods in a linear mixed model program.

BN SROC approach

Chu and Cole [13] and Arends *et al.* [4] used the binomial distribution for the within-study variability. Now the within-study distribution of $\hat{D}_i$ and $\hat{S}_i$ is implicitly modeled by specifying:

$$x_{1i} \sim \mathit{binomial}(n_{1i}, \pi_{1i}) \quad \text{and} \quad x_{0i} \sim \mathit{binomial}(n_{0i}, \pi_{0i}) \tag{4.6}$$

with $\pi_{1i} = \frac{e^{0.5*(S_i+D_i)}}{1+e^{0.5*(S_i+D_i)}}$ and $\pi_{0i} = \frac{e^{0.5*(S_i-D_i)}}{1+e^{0.5*(S_i-D_i)}}$. Now the model is a generalized linear mixed model and a practical disadvantage is that software for these models is scarce and not yet available in many statistical packages. Apart from the parametrization, the model of Rutter and Gatsonis [5] is identical to the BN [4, 11]. However, as shown by Arends *et al.* [4], their method estimates another type of SROC curve than the one considered in this paper. Chu and Cole [13] and Hamza *et al.* [12] showed the advantage of using a binomial distribution over the normal distribution in estimating the (logit transformed) sensitivity and specificity, and their between-studies variances. They showed that approximating the within-study variability by a normal distribution leads to a downward bias in η (sensitivity) and ξ(specificity) and their between-studies variances in many cases. However, this does not necessarily mean the BN outperforms the other two methods on estimating the SROC curve. The reason is that still the estimate of $\bar{D} = \bar{\eta} - \bar{\xi}$ could be approximately unbiased, i.e. the biases may cancel out at least partially since they are in the same direction, and also for the β, which is the ratio of (co)variance parameters. If so, in practice the RI or NN method might be preferable due to their simplicity for fitting them.

4.3 Simulation Study

4.3.1 Data Simulation

First the true η_i and ξ_i were simulated from a bivariate normal distribution, with specified true values for $\bar{\eta}$, $\bar{\xi}$, σ^2_η and σ^2_ξ, and correlation between ξ and $\eta(\rho_{\xi\eta})$. The choices of the true values were based on real meta-analysis data sets from the medical literature (for example [23, 24, 25]). Most often the estimated sensitivities and specificities in the real data sets run from 0.50 to 0.98 and we assigned three pairs of values representing the situations where both sensitivity and specificity are large, where both are small and where one is small and the other large. The different true values in the simulations are given in Table 4.1.

Table 4.1: Different scenarios used in the simulation

$(\bar{\xi}, \bar{\eta})$	(MSP^a, MSE^b)	$(\sigma^2_\xi, \sigma^2_\eta)$	$\rho_{\xi\eta}$	$n_i(sd)$	N
(-2.20, 2.94)	(0.90, 0.95)	(0.5, 0.5)	0.2	40(30)	10
(-0.85, 2.94)	(0.70, 0.95)	(1.2, 1.2)	0.5	500(450)	25
(-0.85, 0.62)	(0.70, 0.62)	(0.5, 1.2)			50
					100

[a] Median specificity
[b] Median sensitivity

Secondly, the within-study number of positive test results for the diseased and non-diseased cases were simulated from binomial distributions with probabilities $\pi_{1i} = 1/(1 + e^{-\eta_i})$ and $\pi_{0i} = 1/(1 + e^{-\xi_i})$ and within-study sample sizes n_{1i} for the diseased and n_{0i} for the non-diseased subjects. In practice the n_i's vary across studies included in the meta-analysis. In some meta-analyses the range of the within-study size (both in diseased and non-diseased or treated and control group) is as big as 3000 or more, see for example references [18, 25, 26]. To represent this variation, the n_{0i}'s and n_{1i}'s were generated independently from a normal distribution and rounded to the nearest integer. As shown in Table 4.1, two different values, 40(30) and 500(450), were considered for the mean n_i's (standard deviation). If a sample size less than 10 was drawn, its value was set to 10. The latter makes 40 and 500 no more the means of the simulated sample sizes, but the medians, and the realized standard deviation becomes a bit smaller than 30 and 450 respectively. We considered every possible combination of the initial values given in Table 4.1, in total 144 scenarios. Besides, a large (0.9) and zero correlation ($\rho_{\xi\eta}$) was considered for some selected scenarios.

4.3.2 Estimation of the simulated data and Comparison of methods

The simulation was replicated 1000 times for each scenario. The RI and NN models were fitted using restricted maximum likelihood (REML) in SAS Proc MIXED, and the BN model was fitted using ordinary maximum likelihood (ML) in SAS Proc NLMIXED. If one of the three models did not converge for a given simulated data set, then this data set was dropped and the simulation was continued until we got 1000 data sets for which all the three models converged. We concentrated on the estimation of the mean log diagnostic odds ratio ($\bar{D} = \alpha + \beta\bar{S}$), slope ($\beta$) and residual variance ($\sigma^2_{D|S}$). Note that the regression line of D on S can be written as $D = \bar{D} + \beta(S - \bar{S})$. The three methods were compared using bias (difference between the mean estimate and the true parameter), mean squared error (MSE) and coverage probability of the 95% confidence interval (the frequency in which the true value falls in the interval). The coverage probabilities were calculated using Wald type confidence intervals based on the normal distribution. A disadvantage of this type of confidence interval might be that when the number of studies is small, the standard errors are underestimated due to the fact that the uncertainty in the estimate of the covariance matrix is not accounted for. Using a t-distribution instead of the standard normal might solve the problem to some extent. However, the t-distribution again is only approximate, and the appropriate number of degrees of freedom needs to be estimated from the data instead of using $k - 1$, where k is the number of studies. Furthermore, estimating the degrees of freedom is rather complex especially for the derived parameters. Another alternative would be to use the profile likelihood [27] which takes into account the uncertainty on the estimated covariance matrix. Here we restricted to the Wald type confidence interval. For example, the 95% confidence interval of β is $\hat{\beta} \pm 1.96se(\hat{\beta})$. The 95% confidence interval for the residual variance is constructed using the log transformation, i.e. the confidence interval is given by $\hat{\sigma}^2_\alpha exp(\pm 1.96se(\hat{\sigma}^2_\alpha)/\hat{\sigma}^2_\alpha)$.

4.4 Simulation Results

The simulation results for the possible combinations of parameter values from Table 4.1 were collected and investigated for the possible effects of the different factors on the results of the three different approaches discussed in Section 4.2. Generally speaking, the biases and coverage probabilities did not vary largely with the true between-studies variances and the correlation of sensitivity and specificity compared to the other factors such as the median within-study size and true median value of sensitivity and specificity. The results of the selected scenarios with correlation 0 and 0.9 also lead to similar results. Hence we only present the results for the 24 scenario's with $\sigma^2_\xi = 0.5$, $\sigma^2_\eta = 1.2$ and $\rho_{\xi\eta}$= 0.2. The results for the whole set of scenarios are available from the authors on request. The bias, coverage probability and MSE for the mean log diagnostic odds ratio, slope and residual variance are given in Tables 4.2, 4.3 and 4.4 respectively. Our conclusions also did not change when a correction factor of 0.1 instead of 0.5 was used for the RI and NN methods.

Table 4.2 shows the bias, coverage probability and MSE of the mean log diagnostic odds ratio from the three different modeling approaches. Most often the BN method gives quite unbiased estimates even when the number of studies was small, and reasonably acceptable coverage probabilities, except for N=10, in which case the Wald type confidence interval may not be appropriate. In almost all scenarios it is better than the RI and NN methods in terms of bias and coverage probability. It also usually gives the smallest MSEs, except for some scenarios where the MSE is slightly in favor of the RI and NN methods. The RI as well as the NN methods always underestimate the mean log diagnostic odds ratio, often by a substantial amount. The coverage probabilities from the RI and NN methods are mostly considerably lower than the nominal level, and sometimes go almost to zero. When the within-study sample size is small and in particular if also the median sensitivity or specificity is large, use of both RI and NN methods is more problematic in terms of bias and coverage probabilities. In most scenarios, the performance of the RI method is better compared to the NN method, although the difference is mostly only small. As expected, the bias and coverage probability from the RI and NN methods are better for larger within-studies sample size.

The bias, coverage probability and MSE for the slope parameter are tabulated in Table 4.3. Like for the mean log diagnostic odds ratio the BN method performs reasonably well in terms of bias and coverage probability regardless of the scenario. The coverage probabilities are not satisfactory when N=10. The RI and NN methods perform well only when the within-study size is large. The coverage probabilities for these scenarios are comparable with the corresponding results from the BN method. When the within-study size is small and the median sensitivity or specificity is large these methods, especially the NN method, are highly biased. For these scenarios the coverage probabilities from the RI and NN methods are far from the nominal level, and even worse when the number of studies is 100. Overall, the performances of both RI and NN methods in terms of bias and coverage probability are similar. Comparing the MSE from the three approaches, the RI method gives the smallest value most often. The NN method performs worst in terms of the MSE. As expected, the bias and coverage probability from the RI and NN methods are better for larger within-studies sample size.

The residual variance is the important parameter to characterize the amount of heterogeneity between-studies having the same value for the threshold S. The bias, coverage probabilities and MSE from the three methods are tabulated in Table 4.4. In all scenarios the residual variance is underestimated by all methods. Comparing the three methods, the BN largely performs the best both in terms of bias and coverage probability. However, some bias is still left when the number of studies is 25 or less. When N=10, the results from all the three methods are not satisfactory. The coverage of the BN method, although better than for the RI and NN methods, is not satisfactory for a small number of studies. It is smaller than the nominal level for the large within-study sample size, and larger than the nominal level (close to 100%) for the small within-study sample size. The BN method also gives the smallest MSE most often compared to the RI and NN methods. Bias and coverage probability of the RI and NN methods is bad in many scenarios, in particular if the within sample size is small. Overall, there is no important difference in performance between the RI and NN methods.

Table 4.2: Simulation results for the mean log diagnostic odds ratio, when $\rho_{\xi\eta} = 0.2$, $\sigma^2_\eta = 1.2$ and $\sigma^2_\xi = 0.5$. (In the table, N=Number of studies included in the meta-analysis and n=within-study number of subjects in the (non-)diseased group)

		True Values				Bias			Coverage Probability			MSE	
N	n	η	ξ	$\bar{D}$	RI	NN	BN	RI	NN	BN	RI	NN	BN
10	40	0.62	-0.85	1.5	-0.092	-0.142	-0.007	0.916	0.903	0.906	0.165	0.166	0.213
		2.94	-2.2	5.1	-0.577	-0.861	-0.055	0.774	0.436	0.933	0.491	0.867	0.295
			-0.85	0.9	-0.417	-0.668	-0.076	0.829	0.583	0.921	0.335	0.573	0.274
	500	0.62	-0.85	1.5	-0.025	-0.033	0.004	0.917	0.926	0.882	0.139	0.139	0.150
		2.94	-2.2	5.1	-0.146	-0.206	-0.007	0.897	0.890	0.880	0.168	0.188	0.171
			-0.85	0.9	-0.103	-0.162	-0.003	0.901	0.883	0.860	0.159	0.174	0.172
25	40	0.62	-0.85	1.5	-0.097	-0.151	-0.004	0.904	0.888	0.910	0.068	0.083	0.068
		2.94	-2.2	5.1	-0.588	-0.865	-0.025	0.410	0.066	0.933	0.412	0.807	0.108
			-0.85	0.9	-0.412	-0.657	-0.023	0.633	0.204	0.935	0.228	0.481	0.083
	500	0.62	-0.85	1.5	-0.022	-0.031	0.010	0.931	0.935	0.916	0.056	0.059	0.057
		2.94	-2.2	5.1	-0.161	-0.212	-0.011	0.861	0.845	0.913	0.084	0.104	0.061
			-0.85	0.9	-0.100	-0.155	0.004	0.919	0.887	0.925	0.065	0.084	0.058
50	40	0.62	-0.85	1.5	-0.098	-0.155	-0.002	0.898	0.839	0.933	0.038	0.053	0.033
		2.94	-2.2	5.1	-0.580	-0.854	-0.007	0.118	0.003	0.938	0.368	0.758	0.051
			-0.85	0.9	-0.411	-0.645	-0.001	0.375	0.029	0.939	0.198	0.442	0.040
	500	0.62	-0.85	1.5	-0.036	-0.049	-0.005	0.937	0.942	0.940	0.027	0.031	0.026
		2.94	-2.2	5.1	-0.150	-0.205	0.001	0.828	0.764	0.933	0.051	0.073	0.030
			-0.85	0.9	-0.118	-0.170	-0.011	0.878	0.822	0.944	0.039	0.057	0.027
100	40	0.62	-0.85	1.5	-0.096	-0.154	0.004	0.873	0.744	0.939	0.023	0.038	0.017
		2.94	-2.2	5.1	-0.593	-0.868	-0.011	0.005	0.000	0.942	0.367	0.767	0.023
			-0.85	0.9	-0.423	-0.655	-0.009	0.087	0.001	0.952	0.194	0.443	0.019
	500	0.62	-0.85	1.5	-0.041	-0.053	-0.008	0.923	0.921	0.929	0.015	0.017	0.013
		2.94	-2.2	5.1	-0.154	-0.210	-0.001	0.730	0.584	0.932	0.038	0.059	0.015
			-0.85	0.9	-0.109	-0.171	-0.003	0.846	0.709	0.949	0.024	0.043	0.013

Table 4.3: Simulation results for the slope (true value = 0.377), when $\rho_{\xi\eta} = 0.2$, $\sigma^2_\eta = 1.2$ and $\sigma^2_\xi = 0.5$

	True Values				Bias			Coverage Probability			MSE	
N	n	η	ξ	RI	NN	BN	RI	NN	BN	RI	NN	BN
10	40	0.62	-0.85	-0.087	-0.011	0.011	0.912	0.935	0.905	0.098	0.224	0.241
		2.94	-2.2	-0.138	-0.090	0.013	0.947	0.975	0.928	0.096	0.448	0.300
			-0.85	-0.098	-0.191	0.023	0.950	0.964	0.923	0.092	0.360	0.236
	500	0.62	-0.85	-0.023	0.003	0.012	0.912	0.888	0.878	0.085	0.106	0.111
		2.94	-2.2	-0.034	-0.017	0.017	0.911	0.904	0.883	0.093	0.127	0.143
			-0.85	-0.020	-0.020	0.012	0.904	0.895	0.860	0.096	0.146	0.177
25	40	0.62	-0.85	-0.071	-0.002	0.008	0.923	0.951	0.932	0.034	0.053	0.049
		2.94	-2.2	-0.146	-0.102	0.002	0.892	0.945	0.951	0.057	0.181	0.099
			-0.85	-0.112	-0.219	-0.001	0.891	0.895	0.936	0.046	0.167	0.080
	500	0.62	-0.85	-0.029	-0.009	-0.001	0.942	0.935	0.934	0.028	0.031	0.031
		2.94	-2.2	-0.035	-0.028	0.002	0.932	0.934	0.923	0.030	0.036	0.036
			-0.85	-0.021	-0.032	0.005	0.937	0.935	0.930	0.030	0.036	0.035
50	40	0.62	-0.85	-0.066	-0.001	0.010	0.907	0.954	0.940	0.018	0.021	0.021
		2.94	-2.2	-0.142	-0.099	-0.005	0.826	0.926	0.934	0.038	0.084	0.042
			-0.85	-0.131	-0.229	-0.015	0.816	0.776	0.955	0.033	0.096	0.030
	500	0.62	-0.85	-0.023	-0.005	0.002	0.934	0.948	0.942	0.014	0.014	0.014
		2.94	-2.2	-0.034	-0.031	0.000	0.928	0.947	0.947	0.015	0.017	0.016
			-0.85	-0.024	-0.036	0.003	0.946	0.935	0.937	0.014	0.017	0.016
100	40	0.62	-0.85	-0.073	-0.010	0.002	0.856	0.959	0.949	0.012	0.010	0.010
		2.94	-2.2	-0.141	-0.085	0.001	0.682	0.894	0.933	0.028	0.032	0.019
			-0.85	-0.128	-0.211	-0.008	0.704	0.648	0.942	0.024	0.063	0.014
	500	0.62	-0.85	-0.026	-0.008	-0.001	0.928	0.952	0.953	0.007	0.007	0.007
		2.94	-2.2	-0.032	-0.028	0.004	0.918	0.918	0.938	0.008	0.009	0.008
			-0.85	-0.026	-0.037	-0.001	0.950	0.939	0.956	0.007	0.008	0.007

Table 4.4: Simulation results for the residual variance (true value = 1.281), when $\rho_{\xi\eta} = 0.2$, $\sigma^2_\eta = 1.2$ and $\sigma^2_\xi = 0.5$

	True Values				Bias		Coverage Probability			MSE		
N	n	η	ξ	RI	NN	BN	RI	NN	BN	RI	NN	BN
10	40	0.62	-0.85	-0.069	-0.358	-0.236	0.993	1.000	0.992	0.477	0.433	0.473
		2.94	-2.2	-0.245	-0.524	-0.077	0.999	0.998	0.952	0.478	0.521	0.711
			-0.85	-0.268	-0.453	-0.176	0.999	1.000	0.983	0.431	0.439	0.528
	500	0.62	-0.85	-0.038	-0.193	-0.249	0.944	0.887	0.851	0.373	0.308	0.314
		2.94	-2.2	-0.125	-0.252	-0.305	0.949	0.912	0.866	0.393	0.365	0.375
			-0.85	-0.099	-0.228	-0.273	0.956	0.908	0.860	0.387	0.335	0.360
25	40	0.62	-0.85	-0.177	-0.331	-0.118	0.954	0.915	0.947	0.192	0.235	0.205
		2.94	-2.2	-0.493	-0.604	-0.094	0.999	1.000	0.983	0.394	0.484	0.337
			-0.85	-0.465	-0.515	-0.154	0.995	0.989	0.986	0.354	0.372	0.242
	500	0.62	-0.85	-0.048	-0.109	-0.100	0.932	0.912	0.909	0.130	0.125	0.126
		2.94	-2.2	-0.105	-0.149	-0.109	0.940	0.922	0.921	0.150	0.142	0.151
			-0.85	-0.082	-0.112	-0.078	0.944	0.928	0.931	0.129	0.120	0.126
50	40	0.62	-0.85	-0.184	-0.304	-0.058	0.915	0.822	0.940	0.114	0.158	0.107
		2.94	-2.2	-0.554	-0.609	-0.096	0.935	0.865	0.987	0.382	0.434	0.170
			-0.85	-0.477	-0.481	-0.082	0.833	0.770	0.950	0.305	0.300	0.139
	500	0.62	-0.85	-0.058	-0.090	-0.057	0.940	0.928	0.935	0.062	0.061	0.062
		2.94	-2.2	-0.099	-0.115	-0.049	0.928	0.915	0.938	0.078	0.077	0.076
			-0.85	-0.123	-0.121	-0.059	0.909	0.906	0.927	0.078	0.074	0.074
100	40	0.62	-0.85	-0.191	-0.297	-0.029	0.868	0.699	0.935	0.073	0.118	0.051
		2.94	-2.2	-0.541	-0.574	-0.033	0.494	0.371	0.956	0.334	0.364	0.087
			-0.85	-0.472	-0.456	-0.044	0.490	0.471	0.959	0.259	0.241	0.063
	500	0.62	-0.85	-0.051	-0.071	-0.026	0.939	0.924	0.946	0.034	0.035	0.034
		2.94	-2.2	-0.104	-0.110	-0.030	0.901	0.900	0.936	0.044	0.042	0.038
			-0.85	-0.112	-0.099	-0.030	0.897	0.902	0.939	0.041	0.037	0.033

Table 4.5: Parameter estimates (standard error) and 95% confidence interval for the MR imaging techniques data.

Parameter	Method		
	RI	NN	BN
$\bar{D}$	4.411(0.276)	4.129(0.248)	4.675(0.271)
	[3.870, 4.952]	[3.643, 4.615]	[4.143, 5.207]
β	-0.277(0.283)	-0.476(0.990)	-0.668(0.508)
	[-0.832, 0.279]	[-2.417, 1.464]	[-1.664, 0.328]
$\sigma^2_{D\|S}$	1.290(0.541)	1.032(0.478)	1.038(0.561)
	[0.567, 2.937]	[0.416, 2.559]	[0.359, 2.995]

4.5 Data Example

To illustrate the methods discussed in this article we re-analyzed the data of a published meta-analysis [19]. Oei *et al.* [19] present data from 29 studies on the diagnostic performance of magnetic resonance imaging (MRI) of the menisci and cruciate ligaments to assess the effect of study design characteristics and magnetic field strength performances. For illustration of the methods discussed in Section 4.2, we used the meta-analysis data presented for the medial meniscal tears, which includes 27 studies. The mean number of diseased and non-diseased patients are 66.4 and 58.9 respectively. It has zero counts and therefore 0.5 is added to each of the two by two table to avoid undefined values when the RI and NN methods are applied. The three random effects methods discussed in Section 4.2 are fitted and the parameter estimates (standard errors) and the 95% confidence intervals are tabulated in Table 4.5. The syntax needed to fit the BN method in SAS Procedure NLMIXED is given in the appendix, and for the RI and NN it is available from the authors on request.

As shown in Table 4.5, the results from the three methods differ substantially. In the simulation study we showed that the RI and NN methods underestimate the average log diagnostic odds ratio $\bar{D}$ regardless of the scenario used. The estimates in this example from the RI and NN methods are lower by 0.484 and 0.617 respectively compared to the BN method. The estimated slopes resulting from the RI and NN methods are also lower than the BN method by 0.391 and 0.192 respectively. Notice that the standard error of the slope of the RI method is remarkably low compared to the other two methods. This is in agreement with the results of the simulation where the RI method gave the smallest MSE. These conclusions are not changed if we analyze the data after removing the two outlying cases (with a within-study size of 32 and 69) in Figure 4.1. In the data example, some other sources of bias and variation may exist and should be assessed and addressed. This is out of the scope of this article. The estimates are transformed into the ROC space and the ROC curves for the three approaches with the 95% coverage regions from the bivariate approaches are given in Figure 4.1. The ROC curves cross each other due to the different estimates of the slope parameter. The differences in the ROC curves are not negligible, in particular when the specificity is greater than 0.90. For example the estimated sensitivity when specificity is

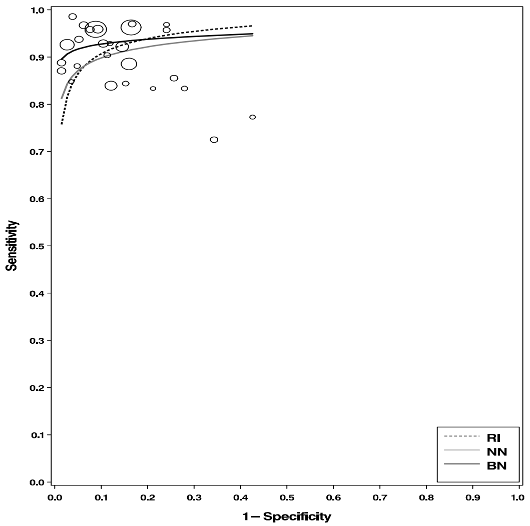


Figure 4.1: ROC curves from the three approaches for the MRI data. The circles are proportional to the sample size of individual studies.

0.95 from the RI, NN and BN methods respectively are 0.86, 0.87 and 0.92. The simulation study from Section 4.3 indicates that the BN method is likely the best. This is in agreement with Figure 4.1, where the BN method seems to fit well, especially in the area where 1-specificity is smaller than 0.10. There are also small differences in the estimated residual variance. In summary, the estimated SROC curves from the three different methods differ substantially and the patterns are in agreement with the results of the simulation study.

4.6 Discussion

When in a meta-analysis one pair of sensitivity and specificity is available per study, the SROC method of Littenberg and Moses [1, 2] has been used for more than a decade to estimate a summary ROC curve. This method has several shortcomings and recently more appropriate methods have been introduced. In this paper we compared the performance of three of these methods, the RI, NN and BN method, through an extended simulation study. We took into account possible important factors such as the number of studies included in the meta-analysis, the within-studies sample size, the order of magnitude of sensitivity and specificity, and the between-studies variances and covariance (correlation). We have shown that in almost all studied scenarios the BN method outperformed the RI and NN methods in terms of bias and coverage probability for all the parameters of the summary ROC curve. However, the differences were not always practically relevant, in particular when the within sample sizes were large. Unexpectedly, in many cases the simple RI method outperformed the more sophisticated NN method. Largely speaking, at least if the number of studies in the meta-analysis is larger ($N \geq 10$), we advise the medical researcher to use the BN method whenever feasible. If the BN method is not readily available to the medical researcher we advise to have a closer look at our simulation results to make a choice between the RI and NN method, which will be then in most cases the simple RI. In the small numbers of studies case ($N = 10$) the results of the comparisons were less straightforward. The BN was still the best in terms of bias, but was outperformed in a number of scenarios by the other methods, mostly the RI method, in terms of coverage probability or mean squared error. In particular for the slope parameter, the RI was better than the BN method in mean squared error. The explanation is that by taking into account the measurement error in S, the bivariate methods reduce the bias of the slope parameter on the expense of increasing the variability [9, 10]. The smaller standard error of the estimated slope of the RI method leads to a smaller MSE in spite of its larger bias. The Wald type confidence intervals that we used for the BN method did not always have satisfactory coverage probabilities when N=10 and hence caution should be taken on using this type of confidence intervals when N is small. The coverage probability could be improved by some kind of degrees of freedom correction or by using profile likelihood. When N=10, none of the three methods gave satisfactory results for the residual variance in terms of bias and coverage probability. In the spirit of REML estimation a possible improvement for the BN method that might be further investigated is to multiply the estimates by $N/(N-2)$. However no one does this in practice and it is beyond the scope of this paper to evaluate the performance of such correction factor.

All three methods can be carried in widely available statistical packages, also extended with covariates. The RI and NN methods can be fitted using a linear mixed model program, which is provided by many statistical packages now. We used the SAS procedure MIXED. Practically speaking, the RI method would be the most attractive, because it is most easy to understand and can be carried out also in meta-regression programs, which are widely available. The BN method has to be fitted in a generalized linear mixed model program, which are relatively scarce. We used SAS Proc NLMIXED. Other possibilities would have been for instance R/S-Plus nlme or the GLLAMM package of Stata. Numerically speaking, the BN approach is the most complicated. Maximizing the likelihood involves the numerical calculation of an integral. In NLMIXED the method of Gaussian quadrature is used to approximate the integral, with the number of quadrature points to be specified by the user or automatically by SAS. The larger the number is chosen, the better the approximation but at the cost of more computational time. In practice we would advise to choose it rather high, for instance 50 or larger. In our simulations we had to keep it rather low. To study the impact of the number of quadrature points we took a sample from our simulated data sets and fitted the BN method on these data sets with varying number of quadrature points. It turned out that estimates and standard errors hardly changed when the number of quadrature points was greater than 10. Hence we used 10 quadrature points throughout the simulation study. Besides fitting generalized linear mixed models can suffer from non-convergence problems, especially when the sample size is small. In our simulations the overall non-convergence rate was approximately 3.5%, and mostly the problem happened when $N = 10 (\approx 10\%)$. In general in case of non-convergence, there are many different options to try in SAS Proc NLMIXED [28]. For example, one can change the initial values by using a grid search specification to obtain a set of good feasible starting values or specify the minimum number of iterations or change the optimization technique. For a detailed discussion of different choices of quadrature options, optimization methods and convergence criteria in NLMIXED we refer to the extensive manual [28]. However, for the bivariate meta-analysis models the problem is mostly in the data itself. NLMIXED reported non-convergence mostly because of the covariance matrix not being positive definite, caused by the maximum likelihood estimator of the between-studies correlation being one. We do not have a clear-cut opinion about what is best to do in that situation, and some further research for this situation is needed. One possibility is to check that the program has correctly converged to a ML estimate with correlation 1, and just work with the results. Another possibility is to study the results of the scenarios of our simulation study that are most close to the application data set. If the NN and/or the RI perform well in these scenarios, one of these can be chosen. However, mostly the convergence problem is also shared by the NN method. Alternatively the BN model could be fitted following a Bayesian approach, for instance using the publicly available WinBUGS software [29]. For more discussion of the problems associated with estimating the between-studies correlation we also refer to Riley *et al.* [14].

Appendix (SAS syntax)

In this section we present the SAS syntax used to fit the BN methods discussed in Section 4.2.3. First, we read in the original data which include the first author, publication year, true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

```
data mrimage ;
    input author $ Pub_year study TP FP TN FN ;
    nd = TP + FN ;  *number of patients in the diseased group (nd);
    nnd = FP + TN ; *number of patients in the non-diseased group (nnd);
    datalines;
    Boere       1991    1   58      6   63  2
    Fischer     1991    2   270     33  170 10
        .       .       .   .       .   .   .
        .       .       .   .       .   .   .
        .       .       .   .       .   .   .
    Elvenes     2000    27  15      6   20  0
run;
data BN ;  * Data step to re-arrange the data for the BN  method;
    set mrimage;
/*Disease status indicators; dis= diseased, ndis=non-diseased*/
       dis= 1; ndis=0;  y = TP ; n = nd; output;
       dis= 0; ndis=1;  y = FP ; n = nnd; output;
run;
proc nlmixed data = BN df=1000 qpoints = 10 ;
   parms malpha = 2.3  beta = 0.2  mS=0.5  valpha=0.7  vS=0.7 ;
   Di = ai + beta*Si;
   theta = exp(0.5*(Si + Di*(dis - ndis)));
   pi = theta/ (1 + theta);
   model y ~ binomial(n, pi);
   random ai si~ normal([malpha, mS], [valpha,0,vS]) subject=study;
run;
```

References

[1] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conflicting reports - a new meta-analytic method. Medical Decision Making. 1993; 13:313-321.

[2] 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. Statistics in Medicine. 1993; 12:1293-1316.

[3] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58: 982-990.

[4] Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-kal MH, Hunink MGM, Stijnen T. Multivariate random effects meta-analysis of ROC curves. (submitted)

[5] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine. 2001; 20:2865-2884.

[6] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. Journal of Clinical Epidemiology. 1995; 48:119-130.

[7] Visser K, Hunink MG. Peripheral arterial disease: Gadolinium-enhanced MR angiography versus color-guided duplex US-a meta-analysis. Radiology. 2000; 216(1):67-77.

[8] Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic resonance angiography compared with digital subtraction angiography in carotid artery stenosis- a systematic review. Stroke. 2003; 34(5):1324-31.

[9] Carroll RJ, Ruppert D, Stefanski LA. Measurement error in Nonlinear Models. London: Chapman & Hall; 1995.

[10] Kendall MG, Stuart A. The Advanced Theory of Statistics. Volume II: Inference and Relationship. Griffin: London, 1973.

[11] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007; 8(2): 239-251.

[12] Hamza TH, van Houwelingen HC, Stijnen T. Random effects meta-analysis of proportions: The binomial distribution should be used to model the within-study variability. Journal of Clinical Epidemiology (In press).

[13] Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Journal of Clinical Epidemiology. 2006; 59(12) 1331-1333

[14] Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Medical Research Methodology. 2007; 7(3): 1-15.

[15] Walter SD. Properties of the summary receiver operating characterstics (SROC) curve for diagnostic test data. Statistics in Medicine. 1995; 14(4): 395-411.

[16] Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Statistics in Medicine. 1995; 14(4): 395-411.

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

[18] Cruciani M, Nardi S, Malena M, Bosco O, Serpelloni G, Mengoli C. Systematic review of the accuracy of the parasight TM-F test in the diagnosis of plasmodium falciparum malaria. Med Sci Monit. 2004; 10(7): MT81-MT88.

[19] Oei EHG, Nikken JJ, Verstijnen ACM, Ginai AZ, Hunink MGM. MR Imaging of the menisci and cruciate ligaments: a systematic review. Radiology. 2003; 226(3): 837-848.

[20] Chang BH, Waternaux C, Lipsitz S. Meta-analysis of binary data: which within study variance estimate to use? Statistics in Medicine. 2001; 20(13): 1947-1956.

[21] Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. Statistics in medicine. 1999; 18(6): 643-654.

[22] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine. 2004; 23(9): 1351-75.

[23] Bipat S, Glas AS, Slors FJM, Zwinderman AH, Bossuyt PMM, Stoker J. Rectal Cancer: Local Staging and Assessment of Lymph Node Involvement with Endoluminal US, CT, and MR Imaging-A Meta-Analysis. Radiology. 2004; 232(3):773-783.

[24] Bipat S, van Leeuwen MS, Comans EFI, *et al.* Colorectal Liver Metastases: CT, MR Imaging, and PET for Diagnosis-Meta-analysis. Radiology. 2005; 237(1):123-131.

[25] Ioannidis JPA, Salem D, Chew PW, Lau J. Accuracy and Clinical Effect of Out-of-Hospital Electrocardiography in the Diagnosis of Acute Cardiac Ischemia: A Meta-Analysis. Annals of Emergency Medicine. 2001; 37(5): 461-470

[26] Mol BWJ, Lijmer JG, Ankum WM, van der Veen F, Bossuyt PMM. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. Human Reproduction. 1998; 13(11): 3220-3227.

[27] Cox DR, Hinkley DV. Theoretical Statistics. London: Chapman and Hall; 1974.

[28] SAS Institute Inc 2004. SAS/STAT 9.1 User's Guide. Cary, NC: SAS Institute Inc.

[29] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 1.4.1. MRC Biostatistics unit: Cambridge 2004.

CHAPTER 5

Bivariate Random-Effects Meta-Regression of Diagnostic Tests: an Application

Abstract

Objective: The bivariate random effects (BRE) meta-analysis of diagnostic tests has been discussed and applied by many authors without adjusting for covariates. At the hand of a case-study we extended the BRE method to investigate the effect of covariates on sensitivity, specificity and, in particular, on diagnostic performance as characterized by a summary ROC curve.
Study design and setting: A two-by-two table is available for 149 studies on the diagnostic performance of three echo-tests (dobutamine-echo, dipyridamole-echo and exercise-echo test) for the diagnosis of coronary artery disease. We allow the overall specificity and sensitivity to depend on the covariates, and the covariance structure to depend on the type of tests. Drawing conclusion on the effect of covariates on sensitivity and specificity is straightforward but not on ROC curves.
Results: The overall sensitivity and specificity from the three tests are statistically significantly different. Exercise-echo is highly sensitive followed by dobutamine-echo. Dipyridamole-echo is highly specific over the whole range of publication years. The results on the diagnostic performance depend on the type of summary ROC curve. In all cases the summary ROC curves of the tests could be assumed ordered, having the same shape.
Conclusion: The BRE meta-regression is a convenient and appropriate tool to investigate the effect of covariates on sensitivity and specificity separately, and on the diagnostic performance of tests as measured by the choice of summary ROC curves.

5.1 Introduction

Meta-analysis of diagnostic tests aims to combine estimates of diagnostic test accuracy measures across related studies. Researchers who investigate the efficacy of diagnostic test accuracy often present sensitivity and specificity or the corresponding two-by-two table. Besides, study level covariates are often presented, for example publication year or subtype of the diagnostic tests. In practice such data are analyzed mostly using the summary receiver operating characteristics (SROC) method of Moses *et al.* [1, 2], which analyses sensitivities and specificities jointly, or using the random effects method of DerSimonian and Laird [3], which analyses sensitivities and specificities separately. The latter approach takes into account the heterogeneity across studies, but does not take into account the relation between sensitivity and specificity. The SROC method has a number of serious shortcomings, as pointed out for instance by Arends *et al.* [4] and Rutter *et al.* [5]. Recently more sophisticated methods, which meet these shortcomings, have been proposed in the literature [4, 5, 6].

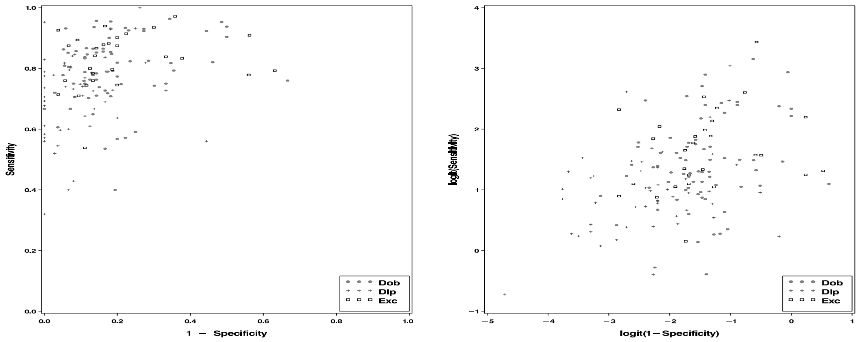
Reitsma *et al.* [6] and Arends *et al.* [4] advocate the bivariate random-effects method, which has many potential advantages over the SROC method and the univariate random effects methods. Harbord *et al.* [7] and Arends *et al.* [4] showed that the model of Rutter and Gatsonis [5] is in fact another parameterization of the bivariate model. These authors primarily discussed fitting the bivariate model without covariates, and only mentioned the possibility to extend the model by allowing sensitivity and specificity to depend on covariates. Adjusting for covariates may be important to correct for differences in the characteristics of studies [8] or to evaluate if diagnostic technologies depend on one or more characteristics of the studies involved. Although it is straightforward to extend the model with covariates and parameters that describe how sensitivity and specificity separately depend on covariates, the analysis is not trivial since in practice mostly the main question is how the relation between sensitivity and specificity, as expressed by an ROC curve, is influenced by covariates. In this paper we present a case study of a meta-analysis data set analyzed with the bivariate random effects model, where the focus is on the effect of covariates, in particular on how to investigate their effect on the summary ROC curve. The meta-analysis concerns 149 studies, each providing a two-by-two table of a non-invasive imaging test to detect coronary artery disease. There are three different types of tests and the interest is in differences in diagnostic performance between them, possibly adjusted for other covariates. The models can be fitted with standard statistical packages. We used the SAS package.

In Section 5.2 we describe the meta-analysis data, in Section 5.3 the bivariate model with covariates is discussed, in Section 5.4 we describe the analyses and the results, and Section 5.5 contains the discussion.

5.2 Data: Imaging tests for coronary artery disease

In this paper we used the meta-analysis data of Heijenbrok-Kal [9]. She searched PubMed from January 1990 through May 2003 for meta-analytic studies on the diagnostic performance of non-invasive imaging test for the diagnosis of coronary artery disease. Articles in the English language were included if they met the following criteria: the diagnostic performance of imaging tests for coronary artery disease was studied, coronary angiography was used as the reference standard, the absolute number of true positive, false negative, true negative and false positive results of the source studies were available or derivable from the meta-analysis and the meta-analytic study was published after 1990. Only data on imaging tests that were still in use were collected.

The data included study level covariates such as publication year, type of diagnostic test used, mean age, and percentage of men. The meta-analysis data set of Heijenbrok-Kal [9] consisted of a total of 246 patient series including 24,761 patients who underwent nine different types of imaging tests for coronary artery disease. In this paper we focus on the comparison of the diagnostic performance of the three echo tests, i.e. the dobutamine-echo, the dipyridamole-echo and the exercise-echo test, using a subset of 149 studies. The data are presented in Figure 5.1(a) and 5.1(b). In Figure 5.1(a) the observed sensitivity is plotted against one minus the observed specificity, and in Figure 5.1(b) the logit-transformed sensitivity is plotted against the logit-transformed 1-specificity.



(a)

(b)

Figure 5.1: 5.1(a) Sensitivity against (1-Specificity), and 5.1(b) logit(Sensitivity) against logit(1-Specificity)

5.3 The bivariate random effects meta-regression model

Reitsma *et al.* [6] and Arends *et al.* [4] discussed the bivariate approach to meta-analyzing sensitivity and specificity jointly, incorporating the correlation that might exist between these two measures. Their approach preserves the two-dimensional nature of the original data and they fitted models without covariates. After a brief introduction to the standard bivariate model we show in this section how it is extended with covariates.

5.3.1 The bivariate model without covariates

Let x_{1i} and n_{1i} be the number of subjects with a positive test result and the total number of subjects in the group with the disease of study i, respectively, and let x_{0i} and n_{0i} be the number of subjects with a positive test result and the total number of subjects in the group without the disease, respectively. Then the observed sensitivity is x_{1i}/n_{1i} and the observed specificity is $(n_{0i} - x_{0i})/n_{0i}$. Although other transformations such as the log(-log) or the probit might be used, it is common practice to transform the data with the logit transformation. We denote the observed logit(1-specificity) and logit(sensitivity) of study i by $\hat{\xi}_i$ and $\hat{\eta}_i$, respectively. The corresponding true study specific logit(1-specificity) and logit(sensitivity) are denoted by ξ_i and η_i, respectively.

Between-Studies model:

The standard bivariate model as discussed in [6] and [4] assumes a bivariate normal distribution for(ξ_i, η_i):

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{pmatrix} \right) \qquad (5.1)$$

Note that the model is the straightforward generalization of the well known univariate random effects meta-analysis model of DerSimonian and Laird [3]. Without loss of generality, one can also model $logit(specificity)$ instead of $logit(1 - specificity)$ (for example [6]). The within-study variability is modeled using an approximate normal distribution [6, 4] or the binomial distribution [4, 7, 10].

Approximate normal within-studies model:

Here the observed logit(1-specificity), $\hat{\xi}_i$ and the observed logit(sensitivity), $\hat{\eta}_i$, are assumed to be independent and to follow a normal distribution with mean ξ_i and η_i and variances that are calculated from the data. As usual, these variances are assumed to be known in the maximum likelihood parameter estimation procedure, an assumption that might be less appropriate when the study sample sizes are small.

$$\begin{aligned} \hat{\xi}_i &\sim N(\xi_i, \tfrac{1}{x_{0i}} + \tfrac{1}{n_{0i}-x_{0i}}) \\ \hat{\eta}_i &\sim N(\eta_i, \tfrac{1}{x_{1i}} + \tfrac{1}{n_{1i}-x_{1i}}) \end{aligned} \tag{5.2}$$

If for any study i one of the denominators in the variances is zero, a correction factor, usually 0.5, should be added to all denominators in all studies [2, 11] to avoid undefined values.

Binomial within-studies model:

Arends *et al.*, Harbord *et al.* and Chu and Cole [4, 7, 10] discuss the use of the binomial distribution for the observed numbers of positive test results. Then the approximate within-study distribution in (5.2) is replaced by the binomial distribution:

$$\begin{aligned} x_{0i} &\sim binomial(\tfrac{e^{\xi_i}}{1+e^{\xi_i}}, n_{0i}) \\ x_{1i} &\sim binomial(\tfrac{e^{\eta_i}}{1+e^{\eta_i}}, n_{1i}) \end{aligned} \tag{5.3}$$

Here $\frac{e^{\xi_i}}{1+e^{\xi_i}}$ and $\frac{e^{\eta_i}}{1+e^{\eta_i}}$ are the probabilities of false and true positive test results respectively, and ξ_i and η_i are the true logit transformed (1-specificity) and sensitivity from (5.1).

The bivariate models given by (5.1-5.3) can be fitted using general(ized) linear mixed model procedures in standard statistical packages, such as the SAS procedures (NL)MIXED or the R/S-Plus programs (n)lme. Hamza *et al.* [12], Hamza *et al.* [13] and Chu and Cole [10] compared in extensive simulation experiments the binomial and approximate normal within-study models, and showed that in general the performance of the binomial within-study model is much better. Hence, in this paper, we only report the results of the binomial within-study model.

Summary ROC curves:

In meta-analysis of this type of diagnostic test data it is customary to present summary ROC curves. A summary ROC curve can be obtained by characterizing the estimated bivariate normal distribution in (5.1) by a line. Transformation of this line from the logit scale to the probability scale then gives a summary ROC curve. Arends *et al.* [4] discussed five choices for such a line. The most straightforward choice may be the regression line of η on ξ, which is given by:

$$y = i + \gamma x \quad \text{with} \quad \gamma = \frac{\sigma_{\xi\eta}}{\sigma_\xi^2} \quad \text{and} \quad i = \bar{\eta} - \gamma\bar{\xi} \tag{5.4}$$

where y and x are logit(sensitivity) and logit(1-specificity). In the spirit of the popular method of Littenberg and Moses [1, 2] another choice might be the D on S regression line, where D and S are defined as $D = \eta - \xi$ and $S = \eta + \xi$ respectively. Then the summary line is given by

$$y = i + \gamma x \quad \text{with} \quad \gamma = \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}} \quad \text{and} \quad i = \bar{\eta} - \gamma\bar{\xi} \tag{5.5}$$

where y and x have similar definition as in (5.4).

This ROC curve is popular because of the interpretation of the variables and parameters involved. D is the diagnostic log odds ratio, and S is a measure for test positivity directly related to the threshold. The intercept i is the diagnostic odds ratio if sensitivity is equal to specificity, or the common diagnostic odds ratio if $\gamma = 0$. The slope γ is interpreted as an asymmetry parameter; if it is zero, the summary ROC curve is symmetric around the line on which sensitivity and specificity are equal.

In this paper we restrict the analyses to these two choices, but the analyses could be done in the same way for the other summary ROC curves as well.

5.3.2 Extension with covariates

In this section we discuss the extension with covariates. To study the effect of covariates on specificity and sensitivity, and also on ROC curves, the model can be extended by replacing $\bar{\xi}$ and $\bar{\eta}$ in (5.1) by a linear combination of study level covariates. For example, we can replace the $\bar{\xi}$ by $\alpha_0 + \sum_{r=1}^{p} \alpha_r Y_r$ and $\bar{\eta}$ by $\beta_0 + \sum_{r=1}^{q} \beta_r Z_r$, where the Y's and Z's are covariates measured at group (diseased or non-diseased) or at study level. Of course, some or all of the Y's might be identical to some or all of the Z's. In the present case-study we choose the covariates Y and Z to be identical. Without loss of generality we can always write the model as

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p \\ \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{pmatrix} \right) \tag{5.6}$$

Here the X_i's are measured on group level. If a covariate is not available for either specificity or sensitivity, or if one wants to use a covariate only for either specificity or sensitivity, the corresponding α_i or β_i is fixed to zero. Again the model can be fitted with a general(ized) linear mixed model program, and standard likelihood ratio or Wald tests can be used to test hypotheses for the regression coefficients.

Summary ROC curves now depend on the covariates too. For example, the η on ξ or the D on S summary line is obtained by replacing $\bar{\xi}$ and $\bar{\eta}$ in (5.4 or 5.5) by their linear predictors:

$$y = (\beta_0 - \gamma\alpha_0) + (\beta_1 - \gamma\alpha_1)X_1 + \ldots + (\beta_p - \gamma\alpha_p)X_p + \gamma x \tag{5.7}$$

with

$$\gamma = \frac{\sigma_{\xi\eta}}{\sigma_\xi^2} \quad \text{or} \quad \gamma = \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}$$

The covariance matrix in (5.6) can also be allowed to depend on the covariates. For instance, in our example we allow the covariance matrix Σ to depend on the three types of the diagnostic test ($k = 1, 2, 3$):

$$\Sigma_k = \begin{pmatrix} \sigma_{\xi k}^2 & \sigma_{\xi\eta k} \\ \sigma_{\xi\eta k} & \sigma_{\eta k}^2 \end{pmatrix}$$

Note that the slope, γ, in all the choices of regression lines is not any more constant across the tests, and it is therefore denoted by γ_k. If the covariance matrix does not depend on the type of test, the lines corresponding to the different covariate values are parallel, and the corresponding summary ROC curves do not cross. A higher intercept then means a uniformly higher sensitivity given the specificity and hence a higher area under the summary ROC curve. If the covariance matrix depends on k then the summary ROC curves corresponding to different test types possibly cross and a higher intercept does not necessarily correspond with a higher area under the ROC curve. Note that if the covariance matrix does not depend on k, the summary ROC curves do not cross regardless of the choice of type of summary ROC curve. This is also true for the other types of summary ROC curves not considered here. It is remarkable that if the covariance matrix depends on k, it is possible that the curves do not cross for one choice, but can cross for another choice. From (5.7) we see that a necessary and sufficient condition for the η on ξ regression lines being parallel is that

$$\sigma_{\xi\eta k} = \gamma\sigma_{\xi k}^2 \tag{5.8}$$

with γ not depending on k, while it is clear that this is not sufficient for the D on S regression lines being parallel. For this it follows from the second part of (5.7) that

$$\sigma_{\xi\eta k} = (\gamma\sigma_{\xi k}^2 - \sigma_{\eta k}^2)/(1 - \gamma) \tag{5.9}$$

Of course, it is much more easy to draw conclusions on the effect of covariates on the summary ROC curve if the curves corresponding to different covariate patterns do not cross. To test this null hypothesis, either (5.8) or (5.9) should be tested, depending on the choice of the type of summary ROC curve. In practice however, it is easier to test the hypothesis that the whole covariance matrix $\sum_k$ does not depend on k. If it depends on k, still the lines might be parallel and we need to test more specifically (5.8) or (5.9).

5.4 Analyses and results

In this section we analyse the data of our example. We start with investigating how the overall sensitivity and specificity are associated with type of diagnostic test and year of publication. Next we will investigate the effect of these covariates on the summary ROC curves.

5.4.1 Sensitivity and Specificity

In our data set we have three different diagnostic tests: dobutamine-echo, dipyridamole-echo and exercise-echo, which can be represented through two dummy variables Z_1 and Z_2. Here Z_1 is 1 if the test is the dipyridamole-echo, else Z_1 is 0; Z_2 is 1 if the test is exercise-echo, else Z_2 is 0. Furthermore we have the year of publication (Yr) as a continuous covariate. Let k =1 (dobutamine-echo), = 2 (dipyridamole-echo) and = 3 (exercise-echo) denote the different diagnostic tests, and let i be the number of the study within a diagnostic test group. We start with fitting the following saturated model.

$$\left(\begin{array}{c} \xi_{ik} \\ \eta_{ik} \end{array} \right) \sim N\left(\left(\begin{array}{c} \bar{\xi} \\ \bar{\eta} \end{array} \right), \left(\begin{array}{cc} \sigma^2_{\xi k} & \sigma_{\xi\eta k} \\ \sigma_{\xi\eta k} & \sigma^2_{\eta k} \end{array} \right) \right) \qquad (5.10)$$

with

$$\bar{\xi} = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 Yr_i + \alpha_4 Z_{1i} Yr_i + \alpha_5 Z_{2i} Yr_i$$
$$\bar{\eta} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Yr_i + \beta_4 Z_{1i} Yr_i + \beta_5 Z_{2i} Yr_i$$

Note that the model allows for possible interaction between publication year and diagnostic test and also allows the covariance matrix to be different for the different diagnostic tests. The model was fitted with SAS Proc NLMIXED. The syntax is given in the appendix. In the spirit of the general guidelines for mixed model building given by Verbeeke and Molenberghs [14] (chapter 9) we first try to simplify the covariance structure. By the likelihood ratio test we tested the null hypothesis that the covariance matrix does not depend on the diagnostic test k. The result was $\chi^2 = 2.6$ on df = 6, P = 0.857, so it is reasonable to assume that the covariance matrix is constant. Further reduction of the covariance structure does not make sense. Next in the model with constant covariance matrix the interactions between diagnostic test and publication year were tested. The result of the global test on all interaction terms is significant ($\chi^2 = 10.2$ on df = 4, P = 0.037). The interaction with respect to the sensitivity was not significant (LR χ^2 = 2.4 on df = 2, P = 0.301), while it was significant for the specificity (LR $\chi^2 = 7.3$ on df = 2, P = 0.026). In SAS Proc NLMIXED, also Wald tests can conveniently be carried out using the '*contrast*' statement. These tests resulted in very similar p-values. As a final model we fitted a model with only the interaction between test and publication year for the specificity. The estimated parameters are given in Table 5.1 and the figures for estimated sensitivity and specificity plotted against publication year are given in Figure 5.2(a) and 5.2(b).

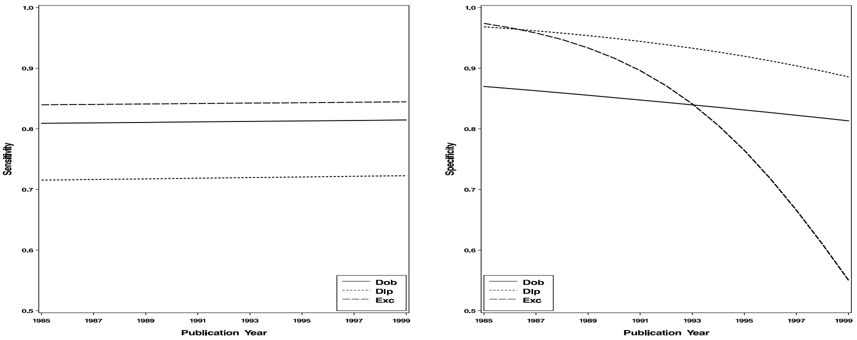
We can draw the following conclusions. With respect to the sensitivity, there are differences between the diagnostic tests (P < 0.0001, global Wald test). The exercise-echo has the highest sensitivity, followed by the dobutamine- and dipyridamole-echo. More specifically, there is no significant difference between the exercise-echo and the dobutamine-echo (95% CI on the logit scale: -0.09 to 0.52), but the difference between the dipyridamole-echo and the dobutamine-echo and between the exercise-echo and the dipyridamole-echo are very significant, with 95% confidence intervals (-0.78, -0.27) and (-1.05,-0.42), respectively. The association of year of publication with the sensitivity is small and far from statistically significant (p-value = 0.899). This is also illustrated in Figure (5.2(a)) where all the three tests

Table 5.1: Parameter estimates of logit specificity and sensitivity assuming there is no interaction term for sensitivity and assuming a constant covariance matrix.

Parameter	logit(1-Spec.)	Spec.[1]	logit(1-Sens.)	Sens.
Constant	-1.747(0.225)	0.834(0.014)	1.457(0.132)	0.813(0.012)
Dip.	-1.181(0.335)	0.925(0.010)	-0.522(0.129)	0.720(0.020)
Exc.	-0.651(0.381)	0.795(0.024)	0.212(0.153)	0.843(0.017)
Yr	0.031(0.040)		0.003(0.020)	
Dip*Yr	0.067(0.060)			
Exc*Yr	0.213(0.076)			
σ^2_ξ	0.322(0.078)			
$\sigma_{\xi\eta}$	0.162(0.049)			
σ^2_η	0.318(0.055)			

have almost constant sensitivity over the publication years.

With respect to the specificity, there is a statistically significant interaction between diagnostic test and year of publication (P= 0.021). Averaged over the three tests, the effect of publication year on specificity is very significantly negative ($P < 0.0001$), with 95% confidence interval (-0.185, -0.064). Averaged over the range of publication years, there are significant differences between the three tests (global Wald test $P < 0.0001$). More specifically, there is no significant difference between logit(1-specificity) of exercise- and dobutamine-echo (95% CI for difference -0.091 to 0.614). However, there is a significant difference between logit(1-specificity) of dipyridamole- and exercise-echo (95% CI for difference -1.564 to -0.746), and between dipyridamole- and dobutamine-echo (95% CI for difference -1.247 to -0.539). Equivalently for the average publication year, we can conclude that dipyridamole-echo is highly specific followed by dobutamine- and exercise-echo consecutively. Figure (5.2(b)) illustrates the changes in specificity of the three tests over the publication years.



(a) (b)

Figure 5.2: 5.2(a) Sensitivity against publication year, and 5.2(b) Specificity against publication year

5.4.2 Summary ROC curve

As shown in Section 5.4.1, the test on equality of the covariance matrices was far from significant and hence we can assume that they are equal. However, to show how one can proceed if the covariance matrices turn out to be different, we proceed the analyses assuming unequal covariance matrices.

The η on ξ type summary ROC curve:

To investigate the effect of the covariates on summary ROC curves, we start again with the model given by (5.10). The resulting η on ξ regression line is given by (5.7) with slope $\gamma_k = \sigma_{\xi\eta k}/\sigma^2_{\xi k}$ depending on the diagnostic test k. The effect of covariates on the summary ROC curves is much easier to judge and to describe if the slope parameter is the same for all diagnostic tests, i.e. if (5.8) holds. Therefore we fit the following model.

$$\begin{pmatrix} \xi_{ik} \\ \eta_{ik} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\xi k} & \gamma\sigma^2_{\xi k} \\ \gamma\sigma^2_{\xi k} & \sigma^2_{\eta k} \end{pmatrix} \right) \tag{5.11}$$

where $\bar{\xi}$ and $\bar{\eta}$ are as in (5.10). Note that in this model the ratio of the covariance between ξ and η and the variance of ξ, γ, is independent of k, hence the regression lines of the three diagnostic tests have the same slope. The likelihood ratio test comparing this model with (5.10) gives LR $\chi^2 = 0.4$ on df = 2, P = 0.82, so it is reasonable to adopt this model. The results are given in Table 5.2.

The parameters for the regression line of η on ξ are calculated by (5.7). The standard errors can be calculated using the delta-method, and hypotheses can be tested with Wald tests. In SAS Proc NLMIXED this can conveniently be done with the '*estimate*' and '*contrast*' statements. The estimated parameters of the regression line are given in Table 5.3.

Table 5.2: Parameter estimates of logit specificity and sensitivity for the regression line of η on ξ assuming a constant slope prameter.

<table>
<tr><th></th><th>logit(1-specificity) (se)</th><th colspan="2">logit(sensitivity) (se)</th></tr>
<tr><td>Intercept</td><td>-1.688(0.214)</td><td colspan="2">1.618(0.200)</td></tr>
<tr><td>Dipyridamole</td><td>-1.269(0.330)</td><td colspan="2">-0.789(0.237)</td></tr>
<tr><td>Exercise</td><td>-0.700(0.418)</td><td colspan="2">0.209(0.335)</td></tr>
<tr><td>Publication year</td><td>0.021(0.038)</td><td colspan="2">-0.029(0.036)</td></tr>
<tr><td>Dipyridamole*Pub year</td><td>0.088(0.059)</td><td colspan="2">0.060(0.044)</td></tr>
<tr><td>Exercise*Pub year</td><td>0.213(0.085)</td><td colspan="2">-0.009(0.069)</td></tr>
<tr><td colspan="4"><i>Covariance parameter estimates for the 3 tests assuming the slope is constant</i></td></tr>
<tr><td></td><td>Dobutamine</td><td>Dipyridamole</td><td>Exercise</td></tr>
<tr><td>σ^2_ξ</td><td>0.247(0.088)</td><td>0.289(0.162)</td><td>0.472(0.186)</td></tr>
<tr><td>σ^2_η</td><td>0.328(0.080)</td><td>0.256(0.081)</td><td>0.358(0.127)</td></tr>
<tr><td>Slope</td><td>0.526(0.158)</td><td></td><td></td></tr>
</table>

Table 5.3: Parameter estimates of the η on ξ regression line assuming a constant slope parameter with and without interaction between tests and year of publication.

Parameter	Based on model (5.11)		Based on model (5.12)	
	Estimate(se)	P-value	Estimate(se)	P-value
Intercept	2.506(0.341)	<.0001	2.377(0.327)	<.0001
Dipyridamole	-0.122(0.321)	0.705	-0.140(0.178)	0.431
Exercise	0.577(0.357)	0.109	0.115(0.167)	0.494
Publication year	-0.040(0.036)	0.278	-0.039(0.026)	0.136
Dipyridamole*Pub year	0.014(0.049)	0.772		
Exercise*Pub year	-0.120(0.078)	0.123		
Slope	0.526(0.158)	0.001	0.445(0.141)	0.002

The Wald test on the two interaction terms with publication year gave a p-value = 0.168, so the following final model was fitted.

$$\left(\begin{array}{c}\xi_{ik}\\ \eta_{ik}\end{array}\right) \sim N\left(\left(\begin{array}{c}\bar{\xi}\\ \bar{\eta}^*\end{array}\right), \left(\begin{array}{cc}\sigma^2_{\xi k} & \gamma\sigma^2_{\xi k}\\ \gamma\sigma^2_{\xi k} & \sigma^2_{\eta k}\end{array}\right)\right) \tag{5.12}$$

where $\bar{\xi}$ is as in (5.10) and $\bar{\eta}^*$ is replaced by

$$\bar{\eta}^* = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Yr_i + \gamma\alpha_4 Z_{1i} Yr_i + \gamma\alpha_5 Z_{2i} Yr_i$$

As an alternative to the aforementioned Wald test, one can also compare model (5.11) and (5.12) using the LR test. This gives a similar result (χ^2= 3.6 on df=2 and p-value=0.165). The summary ROC based estimate on this model is given in the right half of Table 5.3. The Wald test on $\beta_1 - \gamma\alpha_1 = \beta_2 - \gamma\alpha_2 = 0$ shows that there are no significant differences between the three diagnostic tests (p-value = 0.550). The estimated summary ROC curves are shown in Figure 5.3. The exercise echo gives the highest area under the ROC curve followed by dobutamine-echo, though the differences are not statistically significant. The publication year has no significant effect on the summary ROC curves.

The D on S type summary ROC curve:

The same analyses can be done if the D on S regression line is chosen to obtain a summary ROC curve. We start again with the model in (5.10) and compare with the model in (5.13) which assumes that the D on S regression lines corresponding to the three tests are parallel.

$$\left(\begin{array}{c}\xi_{ik}\\ \eta_{ik}\end{array}\right) \sim N\left(\left(\begin{array}{c}\bar{\xi}\\ \bar{\eta}\end{array}\right), \left(\begin{array}{cc}\sigma^2_{\xi k} & \psi\\ \psi & \sigma^2_{\eta k}\end{array}\right)\right) \tag{5.13}$$

with

$$\psi = \frac{\sigma^2_{\eta k} - \gamma\sigma^2_{\xi k}}{\gamma - 1}$$

where γ is the slope after rewriting the D on S regression in terms of η and ξ. This model can also be fitted in SAS Proc NLMIXED (the SAS syntax is available

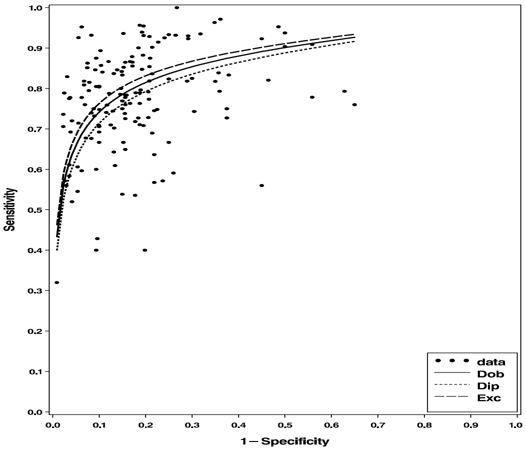


Figure 5.3: ROC curves based on the regression line of η on ξ, for the mean publication year

from the authors on request). Comparing this model with model (5.10), where the regression lines may be non-parallel using the likelihood ratio test results in a large p-value ($\chi^2 = 1.2$ on df $= 2$, P$= 0.549$). To test the interaction terms for the diagnostic tests with publication year, on the D on S regression line we fitted the following model:

$$\begin{pmatrix} \xi_{ik} \\ \eta_{ik} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta}^{\dagger} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\xi k} & \psi \\ \psi & \sigma^2_{\eta k} \end{pmatrix} \right) \tag{5.14}$$

where

$$\bar{\eta}^{\dagger} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Yr_i + \gamma\alpha_4 Z_{1i} Yr_i + \gamma\alpha_5 Z_{2i} Yr_i$$

Note that when the estimated $\gamma \approx 1$, there might be a convergence problem on fitting model (5.13) and (5.14) due to the $1/(\gamma - 1)$ in the covariance structure. In such cases, it is possible to reparametrize, for example by taking the bivariate normal distribution for the true S_i and D_i and directly fit the models in NLMIXED in a similar way as model (5.11) and (5.12). Transforming the regression parameters of the D on S regression into η on ξ can easily be done using the *estimate* statement in NLMIXED.

The likelihood ratio test comparing model (5.13) and (5.14) gives a p-value not far from significant result ($\chi^2 = 5.1$ on df $= 2$ and P=0.078). Hence we adopted model (5.13) as our final model. The results are shown in Table 5.4. From the table it is seen that the interaction between exercise and publication year is significant, implying that which diagnostic test is preferable depends on publication year. For example, exercise echo has the highest accuracy approximately until 1992 than dipyridamole echo and until 1994 than dobutamine echo. For the average publication year, there is no significant difference among the three tests(p-value=0.772), but there might be significant differences at different year of

Table 5.4: Parameter estimates of the D on S regression line assuming a constant slope parameter and with interaction between tests and year of publication. The estimates are given after rearranging terms in terms of $\eta(\xi)$

Parameter	Estimate(se)	P-value
Intercept	3.370(0.456)	<.0001
Dipyridamole	0.515(0.435)	0.237
Exercise	0.898(0.474)	0.059
Publication year	-0.049(0.047)	0.294
Dipyridamole*Pub year	-0.031(0.067)	0.643
Exercise*Pub year	-0.223(0.102)	0.029
Slope	1.039(0.217)	<.0001

publication. Probably clinically most relevant are the differences in the most recent years covered by the meta-analysis, for example in the year 1999. The overall test on differences between the three diagnostic tests are not far from significant in this given year (p-value = 0.084). Looking at the differences between the different pairs we can say that, for any given value of specificity, the logit sensitivity of the dobutamine test is 1.111(se = 0.543, p-value=0.041) higher than that of the exercise test, and the logit sensitivity of the dipyridamole is 1.345(se = 0.614, p-value=0.029) higher than that of the exercise test, while the dobutamine and dipyridamole are not statistically significantly different (estimated difference 0.234, se = 0.347, p-value=0.499). The resulting summary ROC curves for publication year 1999 are given in Figure 5.4.

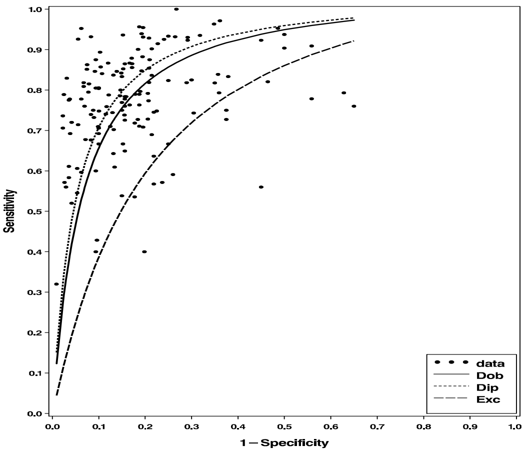


Figure 5.4: ROC curves based on the regression line of D on S, for publication year 1999

5.5 Discussion

Despite its shortcomings, the SROC method of Littenberg and Moses[1, 2] has been the standard method in practice for more than a decade to estimate a summary ROC curve in diagnostic test meta-analysis. Recently the bivariate random effects meta-analysis model [15] was advocated to be used for this type of data [4, 6] and we expect that it will soon replace the SROC method as the standard method in practice.

The bivariate approach has many advantages over the traditional SROC. It handles properly the within and between-study correlation that might exist between specificity and sensitivity or D and S, it takes into account the measurement error in S or ξ in the sprit of structural approach to measurement error [16, 17] and the possible heterogeneity across studies. By using the binomial distribution for the measurement error model, one can also avoid the bias due to addition of a correction factor in case of a zero count [4, 7, 12]. The bivariate approach is flexible to produce different summary outcome measures, such as sensitivity, specificity, diagnostic odds ratio, likelihood ratio and SROC curves. It can easily be fitted in many standard statistical packages. So far no examples have been published in the literature where the bivariate approach is extended with covariates for the meta-analysis of diagnostic tests. In this paper we apply the bivariate meta-regression approach to compare three different imaging tests for coronary artery disease [9], adjusted for other covariates. We used the bivariate meta-regression approach for two purposes. Firstly we wanted to compare sensitivity and specificity of the different diagnostic tests in a bivariate meta-regression model. Secondly we wanted to compare the diagnostic accuracies of the tests through the use of summary ROC curves. In principle the first goal can also be met by carrying out the standard univariate random effects regression approach on sensitivities and specificities separately. However, sensitivity and specificity are often negatively correlated within-studies, and ignoring this correlation would be inappropriate or suboptimal [4, 5, 6, 18]. The bivariate model takes into account both the within and between-study correlation and is more advantageous, over the univariate, as the within and between-study covariances are larger. If there are studies with missing sensitivity or specificity, and the missing data is not completely random, the univariate regressions might be biased, while the bivariate regression is still valid. Even if one is interested only in one of the outcomes, sensitivity or specificity, Riley *et al.* [19, 20] pointed out that the bivariate approach 'borrows strength' across outcomes and thus may increase the precision of the results.

The second aim of our analysis was to compare the diagnostic accuracy of the three diagnostic tests through the use of summary ROC curves. As pointed out by Arends *et al.* [4], there are several possible choices for a summary ROC curve. They mentioned 5 choices: the regression of η on ξ, the regression of ξ on η, the D on S regression, the major axis of the bivariate normal distribution and the one of Rutter and Gatsonis [5], which has slope σ_η/σ_ξ. All lines pass through the centre of the bivariate normal distribution (ξ, η), but have different slopes and of course different intercepts. The resulting summary ROC curves can be substantially different, and, if there is only one pair of specificity and sensitivity per study available, it is unidentifiable from the data which one resembles most the ROC curves of the individual studies. In this paper we have chosen to work with the η on ξ and the D on S summary ROC's, but the same analyses can be done for the other choices as well. The comparison of different diagnostic tests is greatly simplified if it is reasonable to assume that their covariance matrices are the same. Then summary ROC's are ordered and do not cross, independent of the choice of the type of summary ROC. In our application, the test on equality of the covariance matrices was far from significant, so we could have assumed that they are equal. Instead, just to show how one can proceed if the covariance matrices turn out to be different, we did the analyses not assuming equal covariance matrices. Then the choice of type of summary ROC really matters. It can happen for instance that the summary ROC's cross for one type, and do not cross for another type. In that case the results of the comparison of the diagnostic tests can be substantially different. However we think that usually in practice equality of the covariance matrices will be a reasonable assumption. One has to realise, however, that even then still the conclusions on differences in diagnostic performance as measured by the summary ROC curve may depend on the choice of the summary ROC curve type. For instance, it can happen that there are no differences for one choice, while there are non-negligible differences for another choice. Only if the effect of the covariates on $\bar{\xi}$ is zero the differences do not depend on the choice of the summary ROC curve.

In our case study we could assume that slope parameters were equal, thereby reducing the comparison to comparing the intercepts. However, if the slope parameters of the diagnostic tests are significantly different, the comparison is less straightforward. One possibility then is to focus on a specific value of the specificity, and then to compare the sensitivities or vice versa. Another possibility would be to compare the tests using the (partial)area under the ROC curve (see references, [21, 22] for example on how to calculate area under the curve and their standard errors).

The bivariate approach using maximum likelihood estimation can give convergence problems when the number of studies is small or when there is a large correlation between the two outcome measures. The problem can be more pronounced for the bivariate meta-regression, in particular when the covariance is allowed to depend on covariates. In this application we were not bothered by convergence problems, as long as the starting values were chosen appropriately. In SAS there are different options to avoid the non-convergence to some extent [23]. For example, when we use the procedure NLMIXED, we can change the initial values by using a grid search specification to obtain a set of good feasible starting values, or change the optimization technique. An alternative would be to use a Bayesian hierarchical modeling approach which can be carried out using the publicly available software WINBUGS [24].

Appendix (SAS syntax)

In this section we present the SAS syntax used to fit the bivariate random effects method with covariates. We used the NLMIXED procedure and the meta-analysis data described in Section 5.2. The data set is re-arranged in a vertical order and is given below.

TEST	Yr	Study	TP	FN	TN	FP	dob	dip	ex	y	n
echo-dip	3	101	18	8	4	0	0	1	0	18[a]	26[a]
echo-dip	3	101	18	8	4	0	0	1	0	0[b]	4[b]
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
echo-dob	1	124	44	7	18	1	1	0	0	44	51
echo-dob	1	124	44	7	18	1	1	0	0	1	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
echo-ex	7	275	151	43	22	28	0	0	1	151	194
echo-ex	7	275	151	43	22	28	0	0	1	28	50

The variables are defined as:
Yr = publication year - 1990
Study = a number given for a study
dob = dobutamine-echo; dip = dipyridamole-echo; ex = exercise-echo
y = number of patients with [a]true positive (TP) [b]false positive (FP) results
n = total number of patients in the group [a]with the disease (TP + FN) and [b]without the disease (FP + TN)

SAS Syntax for Model 5.10

A bivariate model with covariates assuming the covariance matrix depends on the type of diagnostic test used, and also includes the interaction between publication year and diagnostic tests used. The parameter names used in the syntax stand for: a0 & bo = constant, a1 & b1 = Dipyridamole-echo, a2 & b2 = Exercise-echo, a3 & b3 = Publication year, a4 & b4 = Dipyridamole-echo*Publication year, a5 & b5 = Exercise-echo*Publication year. The a's belong to logit(1-specificity) and the b's to logit(sensitivity).

```
proc nlmixed data = exbiv qpoints=20  miniter=60; /*call for procedure*/
parms a0=-1.57   a1=-0.95 a2=0.09  a3=0.04  a4=0   a5=0
     b0=1.47  b1=-0.53 b2=0.22  b3=0.04  b4=0   b5=0
        sx0=0.3  sx1=0.4  sx2=0.7  c0=0.1   c1=0.2 c2=0.2
        se0=0.3  se1=0.3  se2=0.4; /*starting values*/
/*txi is the true logit(1-specificity) and teta is the true logit(sensitivit
    theta = exp(txi*ndis + teta*dis );
    pi =  theta / (1 + theta);   /*calculate pi from true xi's and eta's*/
/*y is the false positives and true  positives in one column*/
model y ~ binomial(n, pi) ;
    random txi teta ~ normal ([a0+a1*dip+a2*ex+a3*Yr+a4*Yr*dip+a5*Yr*ex,
                              b0+b1*dip+b2*ex+b3*Yr+b4*Yr*dip+b5*Yr*ex],
/*the covariance matrix which depends  on the type of diagnostic tests*/
                     [sx0*dob+sx1*dip+sx2*ex, c0*dob+c1*dip+c2*ex,
                      se0*dob+se1*dip+se2*ex]) subject=study ;
/*contrast statements to test the interaction terms*/
    contrast 'no interaction pubyr and test wrt spec' a4,a5;
    contrast 'no interaction pubyr and test wrt sens' b4,b5;
    contrast 'no interaction pubyr and test' a4, a5, b4,b5;
run;
```

References

[1] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conflicting reports - a new meta-analytic method. Medical Decision Making. 1993; 13:313-321.

[2] 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. Statistics in Medicine. 1993; 12:1293-1316.

[3] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7:177-188.

[4] Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-kal, Hunink MGM, Stijnen T. Multivariate random effects meta-analysis of ROC curves. (In press)

[5] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 2001; 20:2865-2884.

[6] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005; 58: 982-990.

[7] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2006; 1(1):1-21

[8] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002; 21: 1559-1573.

[9] Heijenbrok-Kal MH, Hunink MGM, Fleischmann KE. Stress echocardiography, stress SPECT and electron-beam computed tomography for the assessment of coronary artery disease: a meta-analysis of diagnostic performance. Am Heart J. (In press).

[10] Chu H, Cole SR. Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the Editor). Journal of Clinical Epidemiology. 2006; 59: 1331-1331.

[11] Cox DR. The Analysis of Binary Data, Methuen, London, 1970.

[12] Hamza TH, van Houwelingen HC, Stijnen T. Random effects meta-analysis of proportions: The binomial distribution should be used to model the within-study variability. Journal of Clinical Epidemiology. 2008; 61(1): 41-51.

[13] Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal and binomial-normal bivariate Summary ROC approaches. Medical Decision Making (to appear).

[14] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New-York: Springer-Verlag, 2000.

[15] van Houwelingen JC, Zwinderman AH, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine 1993; 12:2273-84.

[16] Carroll RJ, Ruppert D, Stefanski LA. Measurement error in Nonlinear Models. London: Chapman & Hall; 1995.

[17] Kendall MG, Stuart A. The Advanced Theory of Statistics. Volume II: Inference and Relationship. Griffin: London, 1973.

[18] Walter SD, Jadad AR. Meta-analysis of screening data: a survey of the literature. Statistics in Medicine 1999; 18(24): 3409-24.

[19] Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Statistics in Medicine. 2007; 26: 78-97.

[20] Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Medical Research Methodology. 2007; 7(3): 1-15.

[21] Liu H, Li G, Cumberland WG, Wu T. Testing Statistical Significant of the Area under a Receiving Operating Characteristics Curve for Repeated Measures Design with Bootstrapping. Journal of Data Science. 2005; 3:257-278.

[22] Molodianovitch K, Faraggi D, Reiser B. Comparing the Areas Under Two Correlated ROC curves: Parametric and Non-Parametric Approaches. Biometrical Journal. 2006; 48:745-757.

[23] SAS Institute Inc 2004. SAS/STAT(r) 9.1 User's Guide. Cary, NC: SAS Institute Inc.

[24] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 2.0. MRC Biostatistics unit: Cambridge 2004. Program available at http://www.mrc-bsu.cam.ac.uk/bugs.

CHAPTER 6

Multivariate Random Effects Meta-analysis of Diagnostic Tests with Multiple Thresholds

Abstract

Bivariate random effects meta-analysis of diagnostic tests is becoming a well established approach when studies present one two-by-two table or one pair of sensitivity and specificity. When studies present multiple thresholds for test positivity, usually meta-analysts reduce the data to a two-by-two table or take one threshold value at a time and apply the well developed meta-analytic approaches. However, this approach does not fully exploit the data. In this paper we generalize the bivariate random effects approach to the situation where test results are presented with equal number of k thresholds for test positivity, resulting in a 2 by $(k+1)$ table per study. The model can be fitted with standard likelihood procedures in statistical packages such as SAS (Proc NLMIXED).We follow a multivariate random effects approach; i.e., we assume that each study estimates a study specific ROC curve that can be viewed as randomly sampled from the population of all ROC curves of such studies. In contrast to the bivariate case, where nothing can be said about the shape of study specific ROC curves without additional untestable assumptions, the multivariate model can be used to describe study specific ROC curves. The models are easily extended with study level covariates. The method is illustrated using published meta-analysis data.

6.1 Introduction

Meta-analysis of diagnostic accuracy studies depends on the type of data that is available from different studies. The most frequently reported measures of diagnostic test accuracy are sensitivity and specificity or a two by two table, i.e. with a single threshold value. Meta-analytic methodologies for such kind of data has been developed to summarize sensitivity and specificity separately or jointly in a fixed or random effects context, for example [1, 2, 3, 4, 5, 6]. In recent years the bivariate random effects meta-analysis of diagnostic tests has become a well established approach, which can easily be fitted in many statistical packages [1, 2]. The bivariate approach has many advantages over separate random effects meta-analysis of sensitivity and specificity and the traditional summary receiver operating characteristics (SROC) method of Littenberg and Moses [1, 2, 4]. Besides it is flexible to derive different outcome measures, such as overall sensitivity and/or specificity, diagnostic odds ratio and SROC curves, from the estimated parameters.

In this article we consider the situation where diagnostic test results are presented using more than one threshold, i.e. outcome is reported in more than two categories. One straightforward approach often followed in practice is to dichotomize the test results into two categories and apply the well developed bivariate methods separately for each of the thresholds. When data is presented for many thresholds, a ROC can be calculated per study, and meta-analytic methods have been developed to derive a SROC from them [7, 8, 9]. Poon [10] discusses a latent normal distribution model for analysing ordinal responses with applications in meta-analysis. This model also can be applied to multiple threshold diagnostic meta-analysis data. Specifically for diagnostic accuracy studies, Dukic *et al.* [11] discussed both ordinal regression and hierarchical approaches based on latent variable modeling.

The above approaches are no direct extensions of the nowadays popular bivariate meta-analysis approach for the one threshold case. The aim of this article is to generalize this approach to the situation where test results are presented using more than one threshold or in more than two categories. Not necessarily in all studies the same number of categories is presented, however, in this article we restrict to the case where the number of categories is equal across studies. Our approach can easily be implemented in standard statistical packages. In Section 6.2 we briefly revised the bivariate random effects approach, in Section 6.3 we introduce the multivariate approach to meta-analyse studies that report test results with more than one thresholds, in Section 6.4 we illustrate the methods using a published meta-analysis data and Section 6.5 contains the discussion.

6.2 The Bivariate Random Effects (BREM) Approach

For the situation where each study presents one pair of sensitivity and specificity with corresponding standard errors, the bivariate meta-analysis approach [12] has become a well established method [1, 2, 13]. The approach preserves the two-dimensional nature of the original data taking into account the between-studies correlation of sensitivity and specificity. It can be seen as an improvement on the method of Littenberg and Moses [4], which has been the standard method to construct a SROC for more than a decade.

In this section first we will introduce the bivariate random effects model (BREM) in its standard form. Subsequently we will derive another form of the model, which starts from a model for study specific ROCs and has a different parametrization. This formulation of the model is the natural one to generalize to the case where we have two or more pairs of specificity and sensitivity per study. This formulation also sheds more light on the interpretation of SROCs, which is problematic in the case where only one pair of sensitivity and specificity is available. This issue seems to have been overlooked in the literature.

For study i, denote $\xi_i = logit(1 - specificity_i)$ and $\eta_i = logit(sensitivity_i)$. Let x_{1i} be the number of true positives, n_{1i} the total number of diseased subjects, x_{0i} the number of false positives and n_{0i} the total number of non-diseased subjects. Then the observed sensitivity and specificity for a given study i are x_{1i}/n_{1i} and $(n_{0i} - x_{0i})/n_{0i}$ respectively. Note that sensitivity and specificity tend to be negatively correlated because of explicit or implicit differences in the thresholds. Therefore ξ_i and η_i will tend to be positively correlated. The between-studies model [1] is given by:

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix} \begin{pmatrix} \sigma_\xi^2 & \sigma_\xi\eta \\ \sigma_\xi\eta & \sigma_\eta^2 \end{pmatrix} \right) \tag{6.1}$$

The bivariate distribution of true logit transformed sensitivities and 1-specificities can be characterized by different lines. Back transforming such a line by taking the inverse logits gives a SROC. Since there are several reasonable choices for lines characterizing a bivariate normal distribution, several types of SROCs are possible. For example, a straightforward choice would be the regression of η on ξ.

Table 6.1: Different choices of summary lines as resulting from the BREM: $y = \alpha + \beta x$ where $y = logit(sensitivity)$ and $x = logit(1 - specificity)$. R&G denotes the SROC as resulting from the method of Rutter and Gatsonis [6]

Parameter	Type of regression line				
	η on ξ	ξ on η	D on S	R & G	Major Axis
β	$\frac{\sigma_{\xi\eta}}{\sigma_\xi^2}$	$\frac{\sigma_\eta^2}{\sigma_{\xi\eta}}$	$\frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}$	$\frac{\sigma_\eta}{\sigma_\xi}$	$\frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{(\sigma_\eta^2 - \sigma_\xi^2)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}$
α	$\bar{\eta} - \beta\bar{\xi}$				

However, since the roles of ξ and η are interchangeable, the regression line of ξ on η is an evenly reasonable choice. The method of Littenberg and Moses chooses the regression line of $D = \eta - \xi$ on $S = \eta + \xi$. Table 6.1 gives an overview of 5 different choices as distinguished by Arends *et al.* [2].

The different SROCs can be vastly different in applications, see for instance Arends *et al.* [2] and the examples in Section 6.4. The BREM approach as introduced by Reitsma *et al.* [1] and discussed by Arends *et al.* [2] does not assume anything about study specific curves. The method simply leads to an estimated underlying bivariate distribution of the true sensitivities and specificities as reported by the different studies included in the meta-analysis. This means that the chosen SROC does not necessarily correspond with the true curves of the studies. The true study specific curves might have a substantially different shape, and the SROC cannot be interpreted as a kind of average or overall ROC representative for the ROCs of the different studies. There might even be no study specific curves at all, in case the diagnostic test cannot be thought of as a continuous test. However, this does not mean that the analysis does not make sense in this case, since the existence of study specific ROC curves is not assumed by the method. In the remainder of this section we introduce a new formulation of the BREM, which starts with the study specific ROCs. This will make clear under which extra assumption the BREM describes the distribution of study specific ROCs and the calculated SROC can be considered to be a real overall SROC.

Suppose that in the (η, ξ) space the study specific ROC curves are straight lines with a common slope β. The lines of the different studies then only differ in level, characterised by the intercept α_i for study i:

$$\eta_i = \alpha_i + \beta\xi_i \tag{6.2}$$

We assume that the α_i's are normally distributed with mean $\bar{\alpha}$ and variance σ_α^2. The observations consist of an estimate $(\hat{\xi}_i, \hat{\eta}_i)$ of one pair (ξ_i, η_i) per study. To be able to estimate the parameters, we have to assume a model that describes how these pairs arise across studies. A straightforward assumption is that the ξ_i values are drawn from a normal distribution with mean $\bar{\xi}$ and variance σ_ξ^2. This leads to the following marginal model for (ξ_i, η_i):

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\xi} \\ \alpha + \beta\bar{\eta} \end{pmatrix} \begin{pmatrix} \sigma_\xi^2 & \sigma_{\alpha\xi} + \beta\sigma_\xi^2 \\ \sigma_{\alpha\xi} + \beta\sigma_\xi^2 & \sigma_\alpha^2 + \beta^2\sigma_\xi^2 + 2\beta\sigma_{\alpha\xi} \end{pmatrix} \right) \tag{6.3}$$

This model is just the same as (6.1), only with a different parametrization. However, the number of parameters is one more, which means that one of them is unidentifiable. To make the model identifiable, we need a further assumption on how the ξ_i's in the different studies are selected. For instance we could assume that $\sigma_{\alpha\xi}$ is zero. This means that the individual investigators, in selecting their ξ_i value, are not lead by the level of their line. However it is perfectly conceivable that an investigator who happens to have a ROC that is relatively low, tends to choose a relatively high value for his ξ_i if a high sensitivity is preferred, or just a relatively low value of ξ_i if high specificity is preferred.

If we assume that the correlation between α_i and ξ_i is zero, it is easy to see that β is given by the slope of the regression line of η on ξ. In this case the η on ξ type SROC is the true SROC in the sense that it really can be interpreted as such. In the (ξ, η) space it is just the average line over the population of studies, in the ROC space it can be interpreted as a kind of median ROC.

Another assumption could be that the correlation between η and α is zero. This means that we assume $\sigma_{\alpha\xi} = -\sigma_\alpha^2/\beta$, and it is easily seen that $\beta = -\sigma_\alpha^2/\sigma_{\alpha\xi}$, the slope of the regression of ξ on η. Thus under this assumption the ξ on η type SROC is the real one.

More general, we could assume that some linear combination $a\xi + b\eta$ of ξ and η is not correlated to α, for some value of a and b. We have already seen that if $a = 1$ and $b = 0$, the η on ξ type SROC is the correct one. If $a = 0$ and $b = 1$, then the ξ on η type is the correct one. If we assume $a = b = 1$, then one can check easily that β is equal to the slope of the regression of $D = \eta - \xi$ on $S = \eta + \xi$, and the Littenberg & Moses type SROC is the correct one. It is also easy to check that the assumption $a = \beta$ and $b = 1$ leads to the Rutter & Gatsonis type SROC.

We conclude that in the situation where we have only one pair of sensitivity and specificity per study a calculated SROC can only be interpreted as a real overall ROC under an untestable assumption. This changes as soon as more pairs of sensitivity and specificity are available per study.

The within-study variability can be modeled using an approximate normal distribution [2, 1] or a binomial distribution [2, 13, 14]. Hamza *et al.* [15] compared in extensive simulation experiments the binomial and approximate normal within-study models, and showed that in general the performance of the binomial within-study model is much better. Chu and Cole [14] also showed similar results using a selected number of simulations. Therefore in this paper we restrict to the binomial within-study model. For the approximate approach we refer to [1, 2].

The within-studies model is based on the binomial distribution of the number of false positive (x_{0i}) and true positive (x_{1i}) test results. More specifically we assume:

$$\begin{aligned} x_{0i} &\sim binomial\left(\tfrac{e^{\xi_i}}{1+e^{\xi_i}}, n_{0i}\right) \\ x_{1i} &\sim binomial\left(\tfrac{e^{\alpha_i+\beta\xi_i}}{1+e^{\alpha_i+\beta\xi_i}}, n_{1i}\right) \end{aligned} \tag{6.4}$$

The ξ_i and $\eta_i = \alpha_i + \beta\xi_i$ are the true logit transformed (1-specificity) and sensitivity from (6.2). The bivariate model given by (6.2-6.4) can be fitted using generalized linear mixed model procedures in standard statistical packages, such as the SAS procedure NLMIXED, STATA *gllamm* or the R/S-Plus program *nlme*.

6.3 Multivariate Random Effects Meta-analysis (MREM)

In this section we consider studies where a single test is administered and the results are reported using $J-1$ thresholds or, equivalently, with J ordered categories. Let the number of non-diseased and diseased patients with test result in category j from the i^{th} study be given by x_{0ij} and x_{1ij}, respectively. The total number of non-diseased and diseased patients for study i is denoted by $n_{0i} = \sum_j x_{0ij}$ and $n_{1i} = \sum_j x_{1ij}$, respectively.

6.3.1 The Model

Let the true logit transformed $1 - specificity$ and *sensitivity* for a given threshold j be denoted by ξ_{ij} and η_{ij} respectively, where ξ_{ij}'s and η_{ij}'s are ordered in the j index. We assume a hierarchical model that is a direct generalization of model 6.2-6.3. In contrast to the one threshold case (bivariate approach), when we have more than one threshold (multiple points per study), the SROC curve is identifiable. The between and within-studies models are given as follows:

Between-studies model:

1. Model for the relation between ξ_{ij} and η_{ij}:
   Within a study we assume a linear relation with common slope β and study specific intercept α_i.

$$\eta_{ij} = \alpha_i + \beta\xi_{ij} \quad \text{with} \quad \alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2) \tag{6.5}$$

2. Model for the ξ_{ij}'s:

$$\xi_{ij} = \bar{\xi}_j + \Delta_i + \delta_{ij} \tag{6.6}$$

   Here $\bar{\xi}_j$ is the mean ξ_{ij} over studies, Δ_i represents the study specific systematic deviation of the ξ_{ij}'s from the overall means $\bar{\xi}_j$, δ_{ij} represents the random residual deviation. The Δ_i's can be assumed to follow some parametric or non-parametric distribution. In this article we assume a normal distribution given by $\Delta_i \sim N(0, \sigma_\Delta^2)$. The δ_{ij}'s are assumed to be independent and follow a normal distribution, $\delta_{ij} \sim N(0, \sigma_\delta^2)$. Furthermore, the δ_{ij}'s are assumed to be independent of the Δ_i and α_i. The covariance between α_i and Δ_i is denoted by $\sigma_{\alpha\Delta}$. A negative $\sigma_{\alpha\Delta}$ for instance would mean that in studies with a relatively small α_i the ξ_{ij} tend to be chosen relatively high. The above assumptions (in 6.5 & 6.6) lead to the following marginal between-studies model:

$$\begin{pmatrix} \alpha_i \\ \xi_{i1} \\ \vdots \\ \vdots \\ \xi_{i,J-1} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\alpha} \\ \bar{\xi}_1 \\ \vdots \\ \vdots \\ \bar{\xi}_{J-1} \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \sigma_{\alpha\Delta} & \cdots & \cdots & \sigma_{\alpha\Delta} \\ \sigma_{\alpha\Delta} & \sigma^2_\Delta + \sigma^2_\delta & \sigma^2_\Delta & \cdots & \sigma^2_\Delta \\ \vdots & \sigma^2_\Delta & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \sigma^2_\Delta \\ \sigma_{\alpha\Delta} & \sigma^2_\Delta & \cdots & \sigma^2_\Delta & \sigma^2_\Delta + \sigma^2_\delta \end{pmatrix}\right) \tag{6.7}$$

Note that the covariance structure for the ξ_{ij}'s is of compound symmetry. However, another choice such as Toeplitz, auto-regressive or unstructured might be chosen as well.

Within-study model:

Given $(\alpha_i, \beta, \xi_{i1}, \ldots, \xi_{i,J-1})$, the observed number of subjects in the non-diseased $(\mathbf{x}_{0i1}, \ldots, \mathbf{x}_{0iJ})$ and diseased $(\mathbf{x}_{1i1}, \ldots, \mathbf{x}_{1iJ})$ groups have independent multinomial distributions with parameters $(\pi_{0i1}, \ldots, \pi_{0iJ})$ and $(\pi_{1i1}, \ldots, \pi_{1iJ})$, where

$$\pi_{0ij} = \begin{cases} \frac{exp\{\xi_{ij}\}}{1+exp\{\xi_{ij}\}} & , \quad \text{for} \quad j = 1 \\ \frac{exp\{\xi_{ij}\}}{1+exp\{\xi_{ij}\}} - \frac{exp\{\xi_{i,j-1}\}}{1+exp\{\xi_{i,j-1}\}} & , \quad \text{for} \quad j = 2, \ldots, J-1 \\ 1 - \frac{exp\{\xi_{i,j-1}\}}{1+exp\{\xi_{i,j-1}\}} & , \quad \text{for} \quad j = J \end{cases} \tag{6.8}$$

$$\pi_{1ij} = \begin{cases} \frac{exp\{\eta_{ij}\}}{1+exp\{\eta_{ij}\}} & , \quad \text{for} \quad j = 1 \\ \frac{exp\{\eta_{ij}\}}{1+exp\{\eta_{ij}\}} - \frac{exp\{\eta_{i,j-1}\}}{1+exp\{\eta_{i,j-1}\}} & , \quad \text{for} \quad j = 2, \ldots, J-1 \\ 1 - \frac{exp\{\eta_{i,j-1}\}}{1+exp\{\eta_{i,j-1}\}} & , \quad \text{for} \quad j = J \end{cases} \tag{6.9}$$

The probability density function (pdf) given the π_{0ij}'s and π_{1ij}'s of the observations of the i^{th} study is given by:

$$f(x_{0ij}, x_{1ij} | n_{0i}, n_{1i}, \pi_{0ij}, \pi_{1ij}) = \frac{n_{0i}! n_{1i}!}{\prod_{j=1}^{J} x_{0ij}! x_{1ij}!} \prod_{j=1}^{J} \pi_{0ij}^{x_{0ij}} \pi_{1ij}^{x_{1ij}} \tag{6.10}$$

Inference on the parameters is obtained through the standard likelihood method based on the marginal density for the data, which is calculated by integrating out the random effects $\mathbf{B} = (\alpha, \xi_1, \ldots, \xi_{J-1})'$. Then the contribution of the i^{th} study to the likelihood is

$$L(\theta | x_{0ij}, x_{1ij}) = \int f(x_{0ij}, x_{1ij} | n_{0i}, n_{1i}, \pi_{0ij}, \pi_{1ij}) g(\mathbf{B}) d\mathbf{B} \tag{6.11}$$

As seen from 6.7, we assumed a multivariate normal distribution for the random effects. However, the density $g(\mathbf{B})$ can also be assumed to belong to some other parametric family of distributions [16].

Fitting the model:

1. We fitted the model using Proc NLMIXED of SAS. This procedure does not support directly the multinomial distribution. However, the procedure allows a user specified log-likelihood function. This is easily done for the multinomial distribution. In the appendix the syntax for an example is given. The NLMIXED procedure calculates the likelihood function by numerical integration, using adaptive Gaussian quadrature.

2. NLMIXED allows user specified likelihoods, but many other programs do not. Usually the binomial distribution is supported, therefore we also mention another possibility to fit the model in Generalized Mixed Model programs. The trick is to write the multinomial pdf as a sequence of conditional univariate pdf's, i.e. the pdf of $(x_{Di1}, \ldots, x_{DiJ})$ is expressed as:

   $$p(x_{Di1})p(x_{Di2}|x_{Di1})\ldots p(x_{Di,J-1}|x_{Di1}, \ldots, x_{Di,J-2})$$

   where D is the disease status, 0 or 1. These conditional distributions are all binomial [17] and given by:

   $$\begin{aligned} x_{0ij}|x_{0i1}, \ldots, x_{0i,j-1} &\sim \textit{binomial}\left(n_{0i} - \sum_{r=1}^{j-1} x_{0ir}, \frac{\pi_{0ij}}{1 - \sum_{r=1}^{j-1} \pi_{0ir}}\right) \\ x_{1ij}|x_{1i1}, \ldots, x_{1i,j-1} &\sim \textit{binomial}\left(n_{1i} - \sum_{r=1}^{j-1} x_{1ir}, \frac{\pi_{1ij}}{1 - \sum_{r=1}^{j-1} \pi_{1ir}}\right) \end{aligned} \tag{6.12}$$

   where π_{0ij} and π_{1ij} are calculated as in (6.8 & 6.9) with $j = 1, \ldots, J-1$.

6.4 Data Examples

To illustrate the methods discussed in this article, we apply them to two published meta-analysis data-sets. One is relatively large (29 studies) with three test result categories (2 thresholds) and one more test result categories for those who do not have satisfactory specimen for diagnosis (see Section 6.4.1 for the detail). The second data is small (10 studies) with five test result categories (4 thresholds). Here our objective is to fit the models discussed in 6.2 and 6.3, and to derive the SROC curves.

6.4.1 Example 1: Fine-needle aspiration cytologic examination

Giard and Hermans [18] present 29 studies evaluating the accuracy of fine-needle aspiration cytologic examination (FNAC) of the breast to assess the presence of breast cancer. FNAC provides a non-operative way of obtaining cells for the establishment of the nature of a breast lump and therefore plays a pivotal role in

Table 6.2: Two-by-four contingency table for study i for relating the FNAC outcome to the final diagnosis of breast lesion (The FNAC data is given in the Appendix).

FNAC outcome	Malignant	Suspect	Benign	Unsatisfactory	Total
Final diagnosis					
Malignant	x_{1i1}	x_{1i2}	x_{1i3}	x_{1i4}	n_{1i}
Benign	x_{0i1}	x_{0i2}	x_{0i3}	x_{0i4}	n_{0i}

the preoperative diagnostic process [18, 19]. The selected FNAC results were classified in the following four cytologic categories: definitely malignant, suspect for malignancy, benign, and unsatisfactory specimen for diagnosis (acellular aspiration)(Table 6.2).

The authors [18] determined the sensitivity and specificity of FNAC for each study by reducing the two-by-four table into a two-by-two table. They classified malignant and suspect test results as test result positive, and benign and unsatisfactory as test result negative. Here, following this classification, we applied first the BREM introduced in Section 6.2. The estimated means(standard error) were $\bar{\xi} = -2.547(0.226)$ and $\bar{\eta} = 1.839(0.119)$, and the covariance parameters were estimated as $\sigma^2_{\xi} = 1.306(0.416)$, $\sigma_{\xi\eta} = 0.139(0.156)$ and $\sigma^2_{\eta} = 0.317(0.105)$. From these estimates, the 5 different types of SROCs were calculated and depicted in Figure 6.1. The corresponding intercepts and slopes, and area under the curve (AUC) are given in Table 6.3. Notice that there are relatively large differences between these curves. As argued in Section 6.2, from the BREM the right SROC is not identifiable. The different curves correspond with assuming correlation of α_i and ξ_{ij} ($\rho_{\alpha\xi}$) 0.000, -0.976, -0.400, -0.620 and -0.085 respectively for types 1 to 5 mentioned in Table 6.1. Note that when the absolute correlation increases the test accuracy as expressed by the AUC increases. However the choice of the correlation remains questionable and one should be careful because it is not identifiable from the data without extra assumptions.

In the FNAC data set the last category is for those who do not have satisfactory specimen for diagnosis. Following the authors, we merged this group with the benign group, which resulted in a two by three table. Using this classification we fitted the MREM method discussed in Section 6.3. The estimated mean ξ_j's were $\bar{\xi}_1 = -7.084(0.408)$ and $\bar{\xi}_2 = -2.548(0.260)$, and the estimated variances and covariances were $\sigma^2_{\alpha} = 0.363(0.117)$, $\sigma_{\alpha\Delta} = -0.045(0.143)$, $\sigma_{\Delta} = -0.042(0.443)$ and $\sigma^2_{\delta} = 1.841(0.615)$. The test for the significance of the covariance between α_i and ξ_i in the MREM approach is not significant (likelihood ratio $\chi^2_1 \cong 1.00$, p-value $\cong 0.317$), and hence there is no indication for the choice of the ξ_i's to depend on the level of individual curves. The estimates for the intercept and slope, and AUC are given in Table 6.3. Based on these estimates and the estimated $\bar{\xi}_j$'s, the estimates of the $\bar{\eta}_j$'s are simply calculated using the formula $\bar{\eta}_j = \alpha + \beta\bar{\xi}_j$. The SROC from the MREM is depicted in Figure 6.1. Comparing the estimated SROC curves from the two approaches, one can see that the BREM approach underestimates or overestimates the SROC curves depending on the choice of the type of SROC. This can be seen clearly from the AUC of the SROC curves in

Table 6.3: Parameter estimates (standard errors) and AUC of the SROC curves from the BREM and MREM approaches for the FNAC data

Type of SROC	α	β	AUC
BREM			
η on ξ	2.110(0.321)	0.107(0.118)	0.882
ξ on η	7.636(6.307)	2.276(2.463)	0.955
D on S	2.643(0.371)	0.316(0.137)	0.918
Rutter and Gatsonis	3.094(0.319)	0.493(0.112)	0.935
Major axis	2.191(0.406)	0.138(0.153)	0.889
MREM	2.368(0.135)	0.224(0.016)	0.902

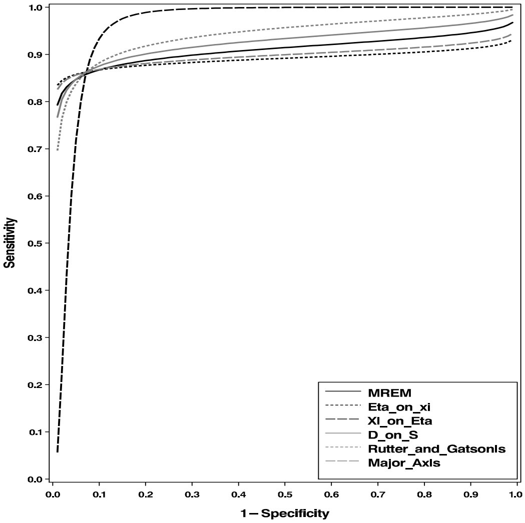


Figure 6.1: SROC curves from the five choices of BREM approach and MREM approach for the FNAC data set

Table 6.3. Of course, if we had chosen another cut point for the BREM, we would have ended up with a different SROC estimate for each of the choices.

In contrast to the BREM, the MREM provides estimates of the study specific ROCs. The program that we used, NLMIXED from SAS, gives the empirical Bayes estimates of the study specific random intercept α_i, which enables to draw study specific SROC curves. We give the study specific ROCs from the MREM approach in Figure 6.2. Of course, the BREM can also provide study specific curves, but only if an untestable assumption on the correlation between α_i and ξ_i is made.

6.4.2 Example 2: CAGE in screening for alcoholism

The CAGE questionnaire is a combination of four questions (resulting in a score from of 0 to 4) that can be used for the screening of patients for alcoholism. Aert-

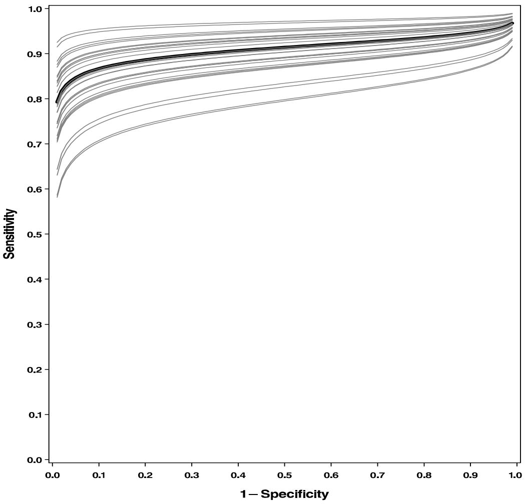


Figure 6.2: SROC curve (black line) with 29 study specific curves (gray lines) from the MREM approach for the FNAC data set

geerts *et al.* [20] performed a diagnostic meta-analysis of all published studies to evaluate the diagnostic value of the CAGE questionnaire. In total they present 10 studies published between January 1974 to December 2001, of which 5 were carried out in primary care populations and 5 in non-primary. In this data example we also include the study level covariate whether or not the patients are from primary care. If study is carried out in a primary care population then z_1 is assigned 1 else z_1 is 0.

In most cases a CAGE score of $\geq$ 2 is considered to indicate an alcohol problem. For the illustration of the BREM method discussed in Section 6.2 we therefore use the threshold of $\geq$ 2 as test positive. Now the mean structure in (6.1) or (6.3) is replaced by $\bar{\xi} = a_0 + a_1 z_1$ and $\bar{\eta} = b_0 + b_1 z_1$. The summary lines in the logit-logit space are then $y = \alpha + \gamma z_1 + \beta xb$ with $\alpha = b_0 - a_0\beta$, $\gamma = (b_1 - a_1\beta)z_1$ and β is given by the different choices given in Table 6.1. The estimates(standard error) were $a_0 = -2.135(0.390)$, $a_1 = -0.160(0.547)$, $b_0 = 0.982(0.393)$ and $b_1 = -0.084(0.552)$. The covariance parameters were estimated as $\sigma^2_\xi = 0.647(0.344)$, $\sigma_{\xi\eta} = 0.543(0.302)$ and $\sigma^2_\eta = 0.671(0.363)$. The resulting estimates of the 5 lines are given in Table 6.4 and the corresponding SROCs are depicted in Figure 6.3. Unlike the FNAC data example, the differences among the SROC curves are small.

The two-by-five tables from the CAGE meta-analysis were also analyzed using the MREM approach discussed in Section 6.3. Here the between-studies model in (6.5) can be rewritten as $\eta_{ij} = \alpha_i + \beta\xi_{ij} + \gamma z_1$ to adjust for the covariate z_1. In fact the ξ_{ij} can also be adjusted for study level covariates, but we did not do that in this example. The estimated mean ξ_{ij}'s were -5.171(0.297), -3.689(0.257), -2.221(0.240) and -1.054(0.233). The estimated variances and co-

Table 6.4: Parameter estimates (standard errors) and AUC (for non-primary care patients (non-PC), and for primary care patients (PC)) of the SROC curves from the BREM and MREM approaches for the CAGE data-set

Type of SROC	α	β	γ	AUC $non-PC$	AUC PC
BREM					
η on ξ	2.775(0.712)	0.840(0.316)	0.050(0.389)	0.886	0.890
ξ on η	3.618(0.827)	1.235(0.364)	0.113(0.479)	0.902	0.908
D on S	3.160(0.710)	1.020(0.314)	0.079(0.419)	0.895	0.900
R & G	3.156(0.657)	1.019(0.286)	0.079(0.417)	0.895	0.900
Major Axis	3.165(0.776)	1.023(0.347)	0.079(0.420)	0.895	0.900
MREM	2.537(0.312)	0.795(0.047)	0.207(0.382)	0.849	0.888

variances were $\sigma_\alpha^2 = 0.392(0.211)$, $\sigma_{\alpha\Delta} = -0.217(0.178)$, $\sigma_\Delta = 0.463(0.226)$ and $\sigma_\delta^2 = 0.036(0.022)$. The test for the correlation between the random intercept, α_i and ξ_i is not significant ($\chi^2 = 2.1$, p-value=0.147). Therefore there is no indication for the choice of the ξ_i's to depend on the level of individual curves. The estimated SROC parameters are given in Table 6.4.

As shown in Figure 6.3 and the AUCs from Table 6.4 the bivariate approach seems to over estimate the SROC curve, for any of the 5 choices of the type of the SROC. Again this would possibly be changed if we choose another cut-off point for positivity on the screening test for alcoholism.

6.5 Discussion

The summary ROC curve has been introduced as a way to assess the diagnostic accuracy of a diagnostic test in a meta-analysis [4, 5, 21, 7, 22]. For the most frequent situation, when one point per study is presented, the medical (and statistical) articles seem to have overlooked the problems inherent to SROCs based on studies with only one point. Although recent developments in the area have shown that the bivariate random effects meta-analysis approach has important advantages over the standard SROC approach of Littenberg and Moses [4, 1, 2], the problem of identifiability and therefore interpretability of the resulting SROC remains. When studies present more than one point per study, commonly the test results are reduced to two categories and meta-analysed using a well established approach such as the BREM, which is a suboptimal approach. In our data examples we illustrated this by considering a single cut-off value and applying the BREM approach. The results from the two data examples showed that differences between the estimated SROC curves based on the BREM approach can be large, as in the first example, or relatively small, as in the second example. The sizes of the differences depend on the values of the three covariance parameters. The η on ξ and ξ on η curves are always most extreme in the sense that the other three lie between them. Therefore a necessary and sufficient condition for the 5 different curves to be equal is the correlation being one, which is not very probable in practical situations. Equality of the variances of ξ_i and η_i is a sufficient condition

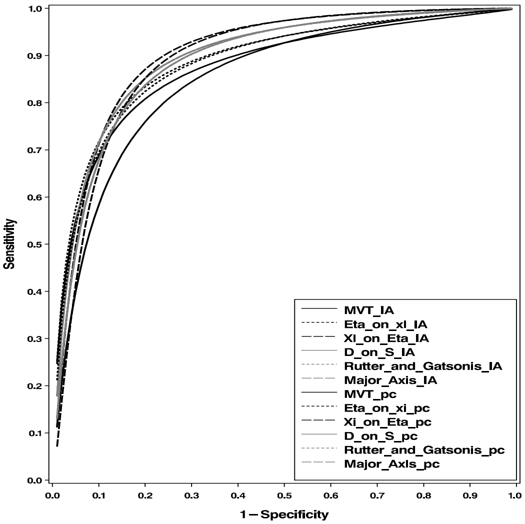


Figure 6.3: SROC curves from the five choices of BREM approach and MREM approach for the CAGE data set

for equality of the SROCs from the three intermediate approaches (D on S, Rutter and Gatsonis, and major axis). In the first example the variances differ a factor 4 and the correlation is relatively small (0.22). In the second example the variances are almost equal and the correlation is relatively large (0.82). This explains why the differences are large in the first and small in the second example.

In this article we generalized the BREM approach for one threshold to the situation where more than one point per study is available and the number of thresholds is equal across studies. In our opinion, the MREM approach is easily comprehensible and has several advantages. First, data of the full 2 by k table is used without losing any information by dichotomizing the test results. Second, different outcome measures can be derived from the fitted model, such as SROC curves and overall sensitivity and/or specificity for any choice of the threshold. Third, in contrast to the BREM approach, the summary ROC and the study specific curves are identifiable. Fourth, the model is symmetric in the ξ_{ij}'s and η_{ij}'s. Interchanging their role leads to the same model. Fifth, it is straightforward to include study level covariates. They can be added directly to the intercept and slope of the SROC, and also to the threshold values. Sixth, the MREM can be fitted in standard statistical packages without extra programming. In equation (6.7), we specified compound symmetry for the covariance structure of the ξ_{ij}'s. However, one can also choose another, possibly richer structure and simplify it using the likelihood ratio test.

A disadvantage in practice is that fitting the MREM can be complicated by convergence problems. We used NLMIXED from SAS, and in all the practical examples that we tried we could reach convergence. However, we noticed that the program is very sensitive to having good starting values. In our experience, good starting values are provided by first fitting the BREMs according to all possible cut-off values. In NLMIXED there are different options to overcome non-convergence. For a detailed discussion on what to do in case of non-convergence, we refer to the SAS/STAT manual [23]. Alternatively, a Bayesian method can be used to fit the MREM, for instance in the publicly available WinBUGS software [24].

Related work was done by Dukic and Gatsonis [11]. They used ordinal regression and a hierarchical approach based on latent variable modeling. To our knowledge their approach is rarely used in practice, probably due to the complexity inherent in fitting the models. The difference between Dukic and Gatsonis model and the MREM is mainly in the modeling of the ξ_{ij}'s. They treated them all as fixed parameters, while we modeled them using the standard multivariate meta-analysis model [12]. The motivation for this is to reduce the number of parameters and to correct for the measurement errors in the $\hat{\xi}_{ij}$'s. Similarly, Poon [10] has done a related work for the meta-analysis of ordinal clinical trial data. They followed a latent normal distribution modeling approach. Similar Dukic *et al.*, Poon put a constraint on the latent continuous variable of the control group; i.e. it is assumed to be fixed.

In this paper we focused on random intercept approach when test results are presented with equal number of threshold. However the assumption that the study specific ROC curves lines are parallel around the SROC curve can be relaxed by allowing the fixed slope parameter to be random. In our examples we were not able to get convergence using the NLMIXED procedure. A Bayesian approach might be used, but this is beyond the scope of this article. Secondly, in practice test results can possibly be presented with different number of threshold. While the MREM method analyse a 2 by k table with equal number of threshold, it should also be generalized to the situation where there are unequal number of thresholds. Besides that, to see how the MREM approach performs for small number of studies and small within-study sample size a simulation study could be considered.

Appendix

SAS syntax to fit the MREM model by writing the likelihood in the SAS procedure NLMIXED

```
proc nlmixed data = cage  df=1000    miniter=15  qpoints=5;
    parms   ma=2.6  b=0.8 mxi1=-5.2  mxi2=-3.7  mxi3=-2.2  mxi4=-1.0
        va=0.4    cavdi=-0.2  vdi=0.5  vdij=0.02  ;
            eta1 = a + b*xi1 ;
            eta2 = a + b*xi2 ;
            eta3 = a + b*xi3 ;
            eta4 = a + b*xi4 ;

            p01 = 1/(1+exp(-(xi1)));
            p02 = 1/(1+exp(-(xi2))) -  1/(1+exp(-(xi1))) ;
            p03 = 1/(1+exp(-(xi3))) -  1/(1+exp(-(xi2))) ;
            p04 = 1/(1+exp(-(xi4))) -  1/(1+exp(-(xi3))) ;
            p05 = 1 - 1/(1+exp(-(xi4))) ;
            p11 = 1/(1+exp(-(eta1)));
            p12 = 1/(1+exp(-(eta2))) - 1/(1+exp(-(eta1))) ;
            p13 = 1/(1+exp(-(eta3))) - 1/(1+exp(-(eta2))) ;
            p14 = 1/(1+exp(-(eta4))) - 1/(1+exp(-(eta3))) ;
            p15 = 1 - 1/(1+exp(-(eta4))) ;
if (p01^=0 and p02^=0 and p03^=0 and p04^=0 and p05^=0
and p11^=0 and p12^=0 and p13^=0 and p14^=0 and p15^=0) then
ll = n01*log(p01)+ n02*log(p02)+n03*log(p03)+n04*log(p04)+n05*log(p05)+
     n11*log(p11)+ n12*log(p12)+n13*log(p13)+n14*log(p14)+n15*log(p15) ;
else ll = -1**100 ;

model n11 ~ general(ll) ;
random a xi1 xi2 xi3 xi4 ~  normal([ma, mxi1, mxi2, mxi3, mxi4],
        [va,
         cavdi, vdi + vdij,
         cavdi, vdi,        vdi + vdij,
         cavdi, vdi,        vdi,        vdi + vdij,
         cavdi, vdi,        vdi,        vdi ,       vdi + vdij ])
subject = study   out = cagerand ;
run ;
```

Data from clinical studies on patients with a breast mass who underwent a fine-needle aspiration cytologic examination (FNAC) (where M=Malignant, S=Suspect, B=Benign and U=Unsatisfactory).

		FNAC result for patients with							
		malignant disease				benign disease			
Author	year	M	S	B	U	M	S	B	U
Linsk	1972	823	156	56	33	1	69	805	134
Furnival	1975	51	0	5	17	2	1	121	42
Zajdela	1975	1526	43	63	89	3	52	846	48
Wilson	1978	19	16	9	6	2	23	164	95

Thomas	1978	49	10	4	8	0	4	92	29
Duguid	1979	50	6	2	2	0	18	181	35
Kline	1979	240	89	35	4	0	602	2810	307
Gardecki	1980	109	16	6	11	0	10	146	67
Strawbridge	1981	141	70	24	39	3	85	326	173
Shabot	1982	46	3	0	1	0	0	29	2
Azzarelli	1983	262	74	65	113	3	23	381	262
Bell	1983	119	91	27	15	0	147	615	131
Norton	1984	8	8	1	2	0	5	9	16
Dixon	1984	222	36	24	29	0	16	275	81
Aretz	1984	26	30	14	4	0	9	93	14
Ulanow	1984	137	25	19	9	1	15	100	12
Wanebo	1984	93	23	1	12	0	6	102	10
Wollenberg	1985	52	13	11	1	0	99	132	13
Somers	1985	81	13	5	5	0	5	37	41
Lannin	1986	23	3	2	2	0	0	63	7
Eisenberg	1986	1050	268	72	177	0	28	68	68
Barrows	1986	481	88	48	72	2	53	338	201
Watson	1987	37	9	13	3	1	0	200	87
Hammond	1987	59	5	4	2	1	12	61	15
Dundas	1988	18	21	2	2	0	1	72	32
Smith	1988	110	22	8	12	0	16	307	119
Palombini	1988	446	24	15	7	0	17	151	10
Langmuir	1989	13	15	1	3	0	25	167	33
Wilkinson	1989	29	13	3	0	0	43	21	1

References

[1] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58: 982-990.

[2] Arends LR, Hamza TH, Van Houwelingen JC, Heijenbrok-kal, Hunink MGM, Stijnen Th. Multivariate random effects meta-analysis of ROC curves. Medical Decision Making. 2007 (to appear).

[3] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7(3): 177-188.

[4] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conflicting reports - a new meta-analytic method. Medical Decision Making. 1993; 13:313-321.

[5] 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. Statistics in Medicine. 1993; 12:1293-1316.

[6] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine. 2001; 20:2865-2884.

[7] Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidlines for Meta-analyses Evaluating Diagnostic Tests. Annals of Internal Medicine. 1994; 120:667-676.

[8] Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Medical Decision Making. 1988; 8:204-15.

[9] Kester ADM, Buntinx F. Meta-analysis of ROC curves. Medical Decision Making. 2000; 20:430-439.

[10] Poon WY. A latent normal distribution model for analysing ordinal responses with applications in meta-analysis. Statistics in Medicine. 2004; 23: 2155-2172.

[11] Dukic V, Gatsonis C. Meta-analyis of Diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003; 59: 936-946.

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

[13] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2006; 1(1):1-21

[14] Chu H, Cole SR. Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the Editor). Journal of Clinical Epidemiology. 2006; 59: 1331-1331.

[15] Hamza TH, van Houwelingen HC, Stijnen T. Random effects meta-analysis of proportions: The binomial distribution should be used to model the within-study variability. Journal of Clinical Epidemiology. 2007 (Available online).

[16] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New-York: Springer-Verlag, 2005.

[17] Agresti A. Categorical Data Analysis. Wiley & Sons, Inc., New Jersey, 2002.

[18] Giard RWM, Hermans J. The value of Aspiration Cytologic Examination of the Breast. A Statistical Review of the Medical Literature. Cancer. 1992; 2104-2110.

[19] Mushlin AI. Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. Annals Internal Medicine. 1985;103:79-85

[20] Aergeerts B, Buntinx F, Kester A. The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. Journal of Clinical Epidemiology. 2004; 57: 30-39.

[21] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. Journal of Clinical Epidemiology. 1993; 48(1): 119-130.

[22] Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. Medical Decision Making. 1993; 13: 253-256.

[23] SAS Institute Inc 2004. SAS/STAT(r) 9.1 User's Guide. Cary, NC: SAS Institute Inc.

[24] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 2.0. MRC Biostatistics unit: Cambridge 2004. Program available at http://www.mrc-bsu.cam.ac.uk/bugs

CHAPTER 7

Discussion

During the last decade advanced statistical methodology for meta-analysis of clinical trials has been introduced and discussed in the medical and statistical literature (for example, [1, 2]). In diagnostic studies, despite the introduction of some advanced methods [3, 4, 5], traditional fixed effect methods (for example [6, 7]) still have been used predominantly. That could be due to the fact that the new methods are complex and not readily implemented in standard software. In contrast, the volume of published diagnostic studies has expanded and the interest in meta-analysis of these studies has been rapidly growing. The aim of this thesis is to contribute to the statistical methodology for meta-analysis of diagnostic accuracy studies and to present new methods that address the drawbacks of the traditionally used methods in the medical literature.

The general(ized) mixed model as a framework for meta-analysis

In our opinion, an important contribution of this thesis is that we have presented methods that fit into the framework of the general or generalized mixed model. For instance, the normal-normal model for separate meta-analysis of sensitivities or specificities is a very simple special case of the general linear mixed model. Specifying a binomial within-study distribution instead of assuming a normal distribution makes it a generalized linear mixed model. Within the general(ized) linear mixed model, the bivariate approach for jointly analyzing pairs of sensitivities and specificities is a natural and straightforward extension of the univariate models. In our opinion, the use of the general(ized) linear mixed model as the statistical framework brings the meta-analysis of diagnostic data from a bundle of ad hoc methods back into the mainstream of statistical methods. The advent of powerful general(ized) mixed model software has made the use of these new methods practically feasible. All the approaches discussed in this thesis have the big advantage of being carried out in a widely available statistical packages, such as SAS, R/S-Plus and STATA. In this thesis we have illustrated our methods using the SAS software.

Binomial versus approximate normal within-study distribution

Traditionally in meta-analysis the within-study distribution of summary statistics is approximated by a normal distribution. For binary outcome, an alternative is to specify a binomial within-study likelihood. In practice this almost never done. In this thesis we advocate the use of the binomial within-study distribution. It removes the correlation between the estimated proportion and its standard error, which often leads to bias in the traditional approach. As shown in our simulation studies, the binomial-normal model often works much better than the approximate normal-normal model. In the case that a normal within-study approximation is employed and the numbers of events are relatively small, there is a need to add a more or less arbitrary correction to estimates and/or standard errors. There is quite a lot of literature on how to deal with small or zero numbers of events in meta-analysis and how to modify the formula for the standard errors, see for instance [8]. However, no ad hoc corrections are needed anymore when the binomial distribution is employed. A practical but temporary disadvantage of the binomial-normal method is that at the moment software for the generalized linear mixed model is still less widely available than for the general linear mixed model.

Small number of studies

As usual, the random effects parameters were assumed to follow a normal distribution in all our models. For the univariate and bivariate approaches, the impact of misspecification of the normality assumption was studied and it turned out the binomial-normal models are fairly robust if the number of studies is large. When the number of studies is small, bias is left in the estimated parameters. We think that relaxing the stringent normality assumption might eliminate some or all of the bias. For the between-studies covariance parameters, even if the underlying true distribution is normal, the maximum likelihood method underestimates the parameters if the number of studies available is small. We expect that a correction factor inline with the restricted maximum likelihood method may overcome this shortcoming.

Another problem in case of a small number of studies is that the construction of Wald type confidence intervals for the maximum likelihood estimates is often not appropriate. One of the reasons could be the fact that the uncertainty in the between-studies covariance parameters is not taken into account. A possible solution might be to use profile likelihood based confidence intervals [9], which is also discussed by several authors in the context of meta-analysis (for example [1, 10, 11]). The simulation study for the meta-analysis of proportions (chapter 2) revealed a large improvement of the coverage probability by using profile likelihood instead of Wald type confidence interval. We believe this result can also hold for the bivariate and multivariate approaches. Where the profile likelihood based confidence is hard to implement in practice, other alternatives [12, 13] such as bootstrapping technique, could be implemented.

Interpretation of summary ROC curves

The statistical methodology for diagnostic studies depends on the available data from each study. In most cases diagnostic test accuracy findings are presented in a two-by-two table or, equivalently, by estimates of sensitivity and specificity with their corresponding standard errors. In this case traditionally the method of Littenberg and Moses [6] and Moses *et al.* [7] has been used to estimate a summary ROC curve. However, this method has many drawbacks. In our opinion, the bivariate approach presented in chapter 3 is a much more principled and elegant method avoiding these drawbacks. Based on this model, several summary ROC curves can be defined. We presented five different reasonable choices, two being based on the two regression lines of the estimated bivariate normal distribution of $logit(sensitivity)$ and $logit(1 - specificity)$, one based on the regression of the difference between $logit(sensitivity)$ and $logit(1 - specificity)$ on the sum of the two (i.e. the SROC of Littenberg & Moses), the Rutter and Gatsonis [3] summary ROC curve and the one based on the major axis of the estimated bivariate normal distribution. In our opinion, a summary ROC curve is nothing more than a curve based on a characterization of the underlying bivariate normal distribution by a line. Our examples showed that the different choices can lead to substantially different summary ROC curves. Thus a summary ROC curve cannot be interpreted as a kind of average of the study specific ROC curves and its shape does not necessarily look like a typical study specific ROC curve. This seems to be completely overlooked in the statistical and medical literature. The ”true” summary ROC curve is unidentifiable. Only under an untestable assumption on how the reported pair of sensitivity and specificity is selected in the individual studies, the calculated summary ROC curve can be interpreted as an estimate of the true one.

When studies present more than one point per study, commonly the test results are reduced to two categories and meta-analyzed using a well established approach such as the bivariate. However, this is suboptimal. In this thesis we proposed an alternative multivariate approach that fits the available data directly. As discussed in chapter 6, the multivariate approach has several advantages, one of them being that the data of the full 2-by-k table is used without losing any information by dichotomizing the test results. More importantly, in contrast to the bivariate approach, the summary ROC curve and the study specific curves are identifiable. Moreover, it is straightforward to include study level covariates, both to the intercept and slope of the summary ROC curve. In our opinion, using as many points as are available per study should be strongly encouraged.

Further research

Some possible issues needing further investigation could be noted. First, as we discussed in chapter 3, when the diagnostic outcome measure of interest is a SROC curve, several different SROC curves can be derived based on the bivariate random effects approach. However, each of these choices is based on an untestable assumption. Therefore further research is needed on what is the most reasonable choice in a certain situation. Second, the multivariate approach (chapter 6) was limited to an equal number of thresholds across studies. In practice however, the number of thresholds often differs between-studies. Therefore there is a need to generalize our method to this more complicated situation. Third, more research could be done on methods that relax the normality assumption of the random effects parameters. For example, methods based on the functional approach in the sprit of Carroll *et al.* [14] , which assumes no distribution, could be developed, or methods in the spirit of the penalized Gaussian mixture distribution model [15] as implemented in the context of meta-analysis of clinical trials [16]. Fourth, in this thesis hardly any attention is paid to checking goodness-of-fit of our models. More research on goodness-of-fit methods would be welcome. Fifth, in this thesis we assumed a single diagnostic test was investigated per study. In practice, there are also studies in which multiple tests are compared. Then the results on different diagnostic tests are correlated within-studies, and methods are needed that account for this dependence. Sixth, throughout the thesis standard likelihood methods were implemented to fit our models. However, all models could be fitted using a Bayesian approach as well, for instance using the publicly available software package WinBUGS [17]. Research on comparing these approaches would be welcome, especially for cases where the number of studies included in the meta-analysis is small. Finally, studies of diagnostic accuracy are subject to different sources of bias and variation. The variation can be due to sampling error, clinical and methodological diversity. We believe that most of the sources of variation are accounted for by the random effects in the methods proposed in this thesis. However, the different sources of biases, such as verification bias, error in the reference, spectrum bias and publication bias, are not considered. These biases may hinder the validity of the statistical analysis and question the applicability of results. Therefore the consequences of the different sources of biases in the pooled estimates should be investigated.

References

[1] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: Multivariate approach and meta-regression. Stat Med 2002;21:589-624.

[2] Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in a meta-analysis:An application. Statistics in Medicine 2003;22:1335-1353.

[3] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 2001; 20:2865-2884.

[4] Dukic V, Gatsonis C. Meta-analyis of Diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003; 59: 936-946.

[5] Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med. 1990;29:12-22

[6] Littenberg B, Moses LE. Estimating diagnostic-accuracy from multiple conicting reports - a new meta-analytic method. Medical Decision Making. 1993; 13:313-321.

[7] 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. Statistics in Medicine. 1993; 12:1293-1316.

[8] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine. 2004; 23(9): 1351-75.

[9] Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall; 1974.

[10] van Houwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine 1993;12:2272-2284.

[11] Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Statistics in Medicine 1996; 15:619-29.

[12] Knapp G, Biggerstaff BJ, Hartung J. Assessing the amount of heterogeneity in random-effects meta-analysis. Biometrical Journal 2006;48: 271-285.

[13] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Statistic in Medicine 2000;19:3417-3432.

[14] Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London: Chapman & Hall; 1995.

[15] Ghidey W, Lesaffre E, Eilers P. Smooth random effects distribution in a linear mixed model. Biometrics 2004; 60:945-53.

[16] Ghidey W, Lesaffre E, Stijnen T. Semi-parametric modelling of the distribution of the baseline risk in meta-analysis. Statistics in Medicine 2007; 17893888.

[17] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 1.4.1. MRC Biostatistics unit: Cambridge 2004.

Summary

Samenvatting

Acknowledgements

About the author

Summary

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies. Statistical methodologies for meta-analysis of diagnostic studies depend on the available data from each study; most commonly a two-by-two table or sensitivity and specificity with the total number of subjects in the diseased and non-diseased groups are available. Some other types of data, ordinal or continuous data, also reported. In this thesis we consider studies that present a two-by-two table (chapter 2 - chapter 5) or a two-by-k table (chapter 6). Despite the availability of few advanced methods, often diagnostic studies are meta-analyzed using fixed effects approach. In this thesis we introduce random effects approaches for meta-analyses of diagnostic studies, and the inference is based on straightforward likelihood technique. All the approaches discussed have the advantage that they can be fitted in standard statistical packages.

Chapter 1 is a general introduction of the thesis. We introduce general terms and the traditionally used statistical methodologies for meta-analysis of diagnostic accuracy studies. Besides, the aims and outlines of the thesis are stated.

When the interest of meta-analysis is to pool proportions, such as sensitivity, specificity, incidence or prevalence, the normal-normal random effects approach has been used to take into account the heterogeneity across studies. This approach has shortcomings and an alternative exact likelihood (binomial-normal) random effects approach is discussed in chapter 2. Using extensive simulations, we showed that the binomial-normal random effects approach gives unbiased estimates with reasonably acceptable coverage probability compared to the normal-normal random effects approach. When the number of studies included is small, we also showed that profile likelihood based confidence intervals are superior over Wald type confidence intervals. Besides, the simulation shows the binomial-normal approach is fairly robust against misspecification of the normality assumption if the number of studies included is relatively large.

Most often, sensitivity and specificity are negatively correlated because of explicit and implicit differences in the threshold. Separate pooling of sensitivity and specificity is not appropriate unless the correlation between these parameters is zero. The bivariate random effects method discussed in chapter 3 incorporates the correlation that might exist between sensitivity and specificity. Two possible assumptions for the within-study distribution were discussed: approximate normal and binomial. The bivariate approach is flexible to derive different choices of outcome measures, such as sensitivity, specificity and summary ROC curves. Under an additional assumption the model gives individual study specific ROC curves as well. One important point that needs considerable attention is that different possible SROC curves can be derived from the bivariate approach based on different untestable assumptions. We discussed five possible choices; two being based on the two regression lines of the estimated bivariate normal distribution of logit(sensitivity) and logit(1-specificity), one based on the regression of the difference between logit(sensitivity) and logit(1- specificity) on the sum of the two (i.e.

the SROC of Littenberg & Moses), the Rutter and Gatsonis summary ROC curve and the one based on the major axis of the estimated bivariate normal distribution. Huge differences among the SROC curves of the different choices were observed, especially when there is a large difference among the estimated between-studies covariance and variances.

Despite all the shortcomings the SROC method of Littenberg and Moses has been used as a standard method for the last ten or fifteen years. A possible improvement of their method is assuming the test accuracy parameter to be study specific (random). Another alternative could be to use the SROC curve derived from the normal-normal or binomial-normal bivariate random effects approach. We discussed the pros and cons of these methods in chapter 4. The three methods were compared through an extensive simulation study. The simulation study revealed that the binomial-normal SROC method performs better than the other two methods in terms of bias and coverage probability. When the average size of studies is large, much difference was not found between the methods, though the performance of the binomial-normal bivariate approach was better. Comparing the normal-normal and the univariate approaches, the latter was preferred in most cases and we recommend a close watch of our simulation results if the binomial-normal bivariate approach is not readily implemented in practice.

The primary studies can also report study level covariates in addition to the two-by-two table. In such cases investigating the effects of covariates on sensitivity, specificity and the summary ROC curves through meta-regression may be of interest. In chapter 5 we discussed the bivariate random effects meta-regression approach with a case study at hand. We allowed both mean and covariance structure to depend on three types of diagnostic tests to be compared. Summarizing sensitivity and specificity, and comparing the significance difference of different tests in terms of these outcome measures is straightforward. However, comparing SROC curves of the different diagnostic tests is not trivial especially when the slope depends on the type of test used. We discussed different ways to compare the test results when there is slope-test dependence.

In chapter 6 we discussed statistical methods for the case where test results are presented with an equal number of multiple thresholds. The usual method is to dichotomize the data and apply the standard methods, which is suboptimal. We discussed a multivariate random effects approach that uses the full account of the 2-by-k table. One important characteristic of this approach is that the SROC curve is identifiable. Besides, the model can be fitted in standard statistical packages and study level covariates can be added in a straight forward manner.

In Chapter 7 we discussed the main findings of the thesis, limitations and recommendations for further study.

Samenvatting

Meta-analyse kan in het algemeen worden gedefinieerd als een kwantitatieve samenvatting en samenvoeging van resultaten van de resultaten van gerelateerde maar onderling onafhankelijke studies. Statistische methoden voor meta-analyse van diagnostische studies zijn afhankelijk van de beschikbare data van elke studie. Meestal zijn in diagnostische studies de sensitiviteit en specificiteit beschikbaar samen met het totaal aantal zieke en gezonde personen in de studie. Dit wordt ook wel weergegeven in zogenaamde 2-bij-2 kruistabellen. Andere type data, zoals ordinale of continue data, worden ook wel eens gerapporteerd. In dit proefschrift beschouwen we studies die een 2-bij-2 kruistabel (hoodstuk 2 t/m hoofdstuk 5) of een 2-bij-k kruistabel (hoofdstuk 6) weergeven. Ondanks de beschikbaarheid van een paar geavanceerde methoden, wordt een meta-analyse van diagnostische studies meestal gedaan met het zg. vaste-effecten model. In dit proefschrift introduceren we random-effecten modellen voor meta-analyses van diagnostische studies en zijn de gevolgtrekkingen gebaseerd op likelihood-schattingsmethoden. Alle modellen die worden besproken hebben als voordeel dat ze kunnen worden gebruikt in standaard statistische paketten.

Hoofdstuk 1 bevat een algemene introductie van het proefschrift. We introduceren hier de meest gebruikte meta-analyse termen en beschrijven de gebruikelijke statistische methoden voor meta-analyse van diagnostische studies. Bovendien worden de doelen en hoofdlijnen van het proefschrift weergegeven.

Als men in een meta-analyse voornamelijk is genteresseerd in het samenvoegen van proporties, zoals sensitiviteit, specificiteit, incidentie of prevalentie, dan wordt het normaal-normaal random-effecten model gebruikt om rekening te houden met heterogeniteit tussen studies. Dit model heeft allerlei nadelen en een alternatief random-effecten model dat is gebaseerd op exacte likelihood (binomiaal-normaal) wordt besproken in Hoofdstuk 2. Op grond van een uitgebreide simulatie-studie laten we zien dat het binomiaal-normaal random effecten model zuivere schattingen geeft met redelijk acceptabele coverage kansen vergeleken met het normaal-normaal random effecten model.

Als het aantal studies in de meta-analyse klein is, tonen we aan dat betrouwbaarheidsintervallen gebaseerd op profile likelihood superieur zijn aan betrouwbaarheidsintervallen gebaseerd op de Wald toets. Bovendien laat de simulatiestudie zien dat als het aantal studies in de meta-analyse relatief groot is, het binomiale-normale model redelijk robuust is tegen misspecificatie van de normaliteitsassumptie.

Sensitiviteit en specificiteit zijn meestal negatief gecorreleerd vanwege expliciete en impliciete verschillen in de gebruikte dremelwaarde. Het afzonderlijk poolen van sensitiviteit en specificiteit is niet geschikt, tenzij de correlatie tussen deze twee parameters nul is. Het bivariate random effecten model dat in Hoofdstuk 3 wordt besproken, houdt rekening met de correlatie die er zou kunnen zijn tussen de sensitiviteit en specificiteit. Twee mogelijke assumpties voor de binnen-studie verdeling worden besproken: de benaderende normale en de binomiale assumptie. Het bivariate model is flexibel zodat men kan kiezen uit verschillende uitkomstmaten zoals sensitiviteit, specificiteit en samenvattende ROC curven. Onder een extra aanname kan het model ook individuele studie-specifieke ROC curven geven. Een belangrijk punt dat extra aandacht verdient, is dat verschillende SROC curven kunnen worden afgeleid van de bivariate benadering, die allemaal zijn gebaseerd op verschillende niet-testbare assumpties. We bespreken vijf mogelijke keuzes, twee zijn gebaseerd op de twee regressielijnen van de geschatte bivariate normale verdeling van logit(sensitiviteit) en logit(1-specificiteit), n gebaseerd op de regressie van het verschil tussen logit(sensitiviteit) en logit(1-specificiteit) op de som van de twee (SROC model van Littenberg & Moses), de Rutter & Gatsonis samenvattende ROC curve en de lijn die gebaseerd is op de hoofd-as van de geschatte bivariate normale verdeling. We zien grote verschillen tussen de SROC curven van deze verschillende keuzes, vooral als er een groot verschil is tussen de geschatte tussen-studie covarianties en varianties.

Ondanks alle nadelen is de SROC methode van Littenberg en Moses al tien of vijftien jaar de standaard methode. Een mogelijke verbetering van hun methode is om de parameter die de accuratesse van de test weergeeft als studie-specifiek (of random) te veronderstellen. Een ander alternatief kan zijn om de SROC curve te gebruiken die volgt uit het normaal-normaal of binomiaal-normaal bivariaat random effecten model. We bespreken de voordelen en nadelen van deze methoden in Hoofdstuk 4. De drie methoden worden onderling vergeleken door middel van een uitgebreide simulatie studie. De simulatie studie laat zien dat de binomiaal-normaal SROC methode het beter doet dan de andere twee methoden op grond van hun bias en coverage kansen. Als de gemiddelde grootte van de studies in de meta-analyse groot is, wordt er niet veel verschil tussen de methoden gevonden, ook al is de binomiaal-normaal methode beter. Bij het vergelijken van de normaal-normaal en de univariate modellen bleek de laatste in de meeste gevallen de voorkeur te hebben en bevelen we aan om goed te kijken naar de resultaten van onze simulatie studie als het binomiale-normale bivariate model niet redelijk eenvoudig te implementeren is in de praktijk.

De individuele studies in de meta-analyse kunnen naast de standaard 2-bij-2 kruistabel ook covariaten op studie-niveau rapporteren. In zulke gevallen kan het interessant zijn om door middel van meta-regressie te onderzoeken wat voor effecten deze covariaten hebben op sensitiviteit, specificiteit en de samenvattende ROC curves. In Hoofdstuk 5 bespreken we het bivariate random effecten meta-regressie model aan de hand van een praktisch voorbeeld. We laten in ons model toe dat zowel de gemiddelde als de covariantie structuur af mogen hangen van drie typen van diagnostische tests die worden vergeleken. Het samenvatten van sensitiviteit en specificiteit, en het vergelijken van de verschillen in significantie van de verschillende tests in termen van deze uitkomstmaten is eenvoudig. Het is echter niet triviaal om de SROC curven van de verschillende diagnostische tests te vergelijken. Dit is vooral moeilijk als de helling afhankelijk is van de type test die wordt gebruikt. We bespreken verschillende manieren om de testresultaten te vergelijken als er een afhankelijkheid is tussen de helling en de test.

In Hoofdstuk 6 bespreken we statistische methoden voor de situatie waarin testresultaten worden gepresenteerd voor meerdere, maar wel een gelijk aantal drempelwaarden. De gebruikelijke methode is om de data te dichotomiseren en standaard methoden te gebruiken, maar dit is suboptimaal. We stellen voor om een multivariaat random effecten methode te gebruiken dat de volledige informatie van de 2-bij-k kruistabellen benut. Een belangrijk kenmerk van deze methode is dat de SROC curve identificeerbaar is. Bovendien kan het model worden gefit in standaard statistische paketten en kunnen op een eenvoudige manier covariaten op studieniveau worden toegevoegd.

In Hoofdstuk 7 bespreken we de belangrijkste bevingingen van dit proefschrift, evenals de beperkingen er van en doen we aanbevelingen voor vervolgonderzoek.

Acknowledgements

Glory to the Almighty God and Virgin Mary for giving me the endurance to complete this study. Many persons contributed to the success of this thesis in some way, in which I would like to take this opportunity to acknowledge them.

First and foremost I would like to express my deep and sincere gratitude to my promoter Prof. Dr. Theo Stijnen, his incessant effort has led to the realization of this thesis. Theo thanks for the scientific and personal guidance through out the course of my study, and also for the enthusiasm and inspiration, which was always there when I needed it. I would like also to thank my co-promoter Prof. Dr. Hans C. van Houwelingen for his constructive comments and continual discussion throughout this work. My sincere appreciation goes to Dr. Lidia Arends for making the Dutch version of the summary and also for the collaboration.

Thanks to many colleagues at the Department of Epidemiology and Biostatistics, Erasmus Medical Center with whom I shared many events: Bettina Hansen, Maria Ridder, Marek Molas, Sten Willemsen, Wim Hop, Paul Mulder, Prof. Dr. Emanuel Lesaffre, Dymph Wijnen, Prof. Dr. Myriam Hunink, Ylian Liem, Ineke van den Berg, Prof. Joke Bosch, Jolanda Veldhuis, Tien Nguyen, Majanka Heijnbrok-Kal, Jan-Jap Visser and Nano Suwarno. I also thank Simon Bignell.

I owe my most sincere gratitude to Dr. Wendimagegn Ghidey & Birkenesh Alemeneh whom I bothered much about my personal life and their kind help for all the troubles I came across. Doc, Birekey, thank you for your lovely and persistent care through out my stay in The Netherlands, and for the scientific discussion we have had. I also thank Fetene Bekele & Miheret Anteneh for their continual encouragement and moral support. My warm thanks are due to my Shimeles Gebeyehu & Selamawit Tsigea for their relentless motivation, advice and continual praying.

I am thankful for the spiritual, moral and social support received from all members of Ethiopian Orthodox Church in The Netherlands & Belgium, Rotterdam Sunday School, Delft *Gubae* and MK-Holland at different stage of my stay in The Netherlands. My specific gratitude, however, goes to M.S. Aba TesfaMariam, L.T. Qesis Mulugeta, M.M. Qesis Mezgebu, Qesis Daniel, Dn. Yosef, Engedashet (Abeba), Gashe Tadesse(Tadesu), Yafet, Dawit, W/o Masho, Genet, Haymanot, Aserate, Mulugeta(Fetelework), Daniel(Turey), Tewolde(Selam), Adehanit, Gashe Hailu, Mimi and Alemu(Woineshet).

Thanks also for many friends who share most of my burdens during the long journey regardless of the geographical distance, including, Dr. Fenta Mandefro, Dr. Zenebe Gebreegziabher, Dr. Alemayehu Wassie, Ayalew Getachew (Saba Chaneyalew), Tesema Ayalew, Samueal Kebede (Selamawit), Fikere Assefa, Mathewos Solomon, Hailu Shiferaw (Dr. Selamawit), Kinfe Alemu, Tariku Chala, Yonas Tadese, Ayalew Tadele, Kassahun Yimer, Berhanu Tadesse, Takele Mitiku, Sahle Miheret, Tesfaye Mekonen, Yewoineshet and Atetegeb Meazah.

Last but not list, I am thankful for my families and relatives: Serkalem Abebe, Hussien Hamza, Hawa Legese, Ahmed Muhey, Aleganesh Abebe, Dr. Alebel Aragea, Andualem Damtew, Tezera Dessie, Hannisa Berhe, Hareg Mekonen, Tsehayenesh Mengesha, Merema Muheye, Assefa Nurey, Turey, Alemayehu, Engdawork, Simeret, Firehiwot, Kiber, Wubeshet, Tesfay, Zeru, Emebet, Woyenetu, Lili and my dearest Tsigereda Wosen for their love, patience, encouragement, support and advices.

Taye
(***May - 28 - 2008***)

About the author

Taye Hussien Hamza was born on 26 September 1976 in Worailu, Ethiopia. He obtained his Bachelor degree in Mathematics (Minor Physics) (distinction) from Debub University, Ethiopia in 2000. The same year he joined Arbaminch University, Ethiopia as graduate assistant. In 2002 he joined Hasselt University, Diepenbeek, Belgium where he obtained his MSc degree in Applied Statistics (2003) and MSc degree in Biostatistics (2004). He did summer researches in the Institute for Materials Research, Diepenbeek, Belgium (2003) and UCB Pharma, Braine-l'Alleud, Belgium (2004).

In December 2004 he started the project described in this thesis at the Department of Epidemiology and Biostatistics, Erasmus University Medical Center. During this period he assists in teaching Biostatistics courses and collaborates with medical researchers in the Medical Center. The content of this thesis has been presented in different international conferences, for example, International Society for Clinical Biostatistics annual conference in Geneva (2006), Greece (2007), International Conference on Sources of Heterogeneity in Meta-analysis of Randomized Clinical Trials, Bremen, Germany (2007), First IBS Channel Network Conference, Rolduc, The Netherlands (2007), First Conference of the Central European Network Statistics and Life Sciences: Perspectives and Challenges, Munich, Germany (2008).

ISBN 978-90-9023002-3