POPULATION-AVERAGED MODELS FOR DIAGNOSTIC ACCURACY STUDIES AND META-ANALYSIS

James Murray Powers

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Dr. John S. Preisser
Dr. Haitao Chu
Dr. Jason Fine
Dr. Paul Stewart
Dr. Charles Poole

Abstract

JAMES MURRAY POWERS: Population-averaged models for diagnostic accuracy studies and meta-analysis
(Under the direction of Dr. John S. Preisser and Dr. Haitao Chu)

Modern medical decision making often involves one or more diagnostic tools (such as laboratory tests and/or radiographic images) that must be evaluated for their discriminatory ability to detect presence (or absence) of current health state. The first paper of this dissertation extends regression model diagnostics to the Receiver Operating Characteristic (ROC) curve generalized linear model (ROC-GLM) in the setting of individual-level data from a single study through application of generalized estimating equations (GEE) within a correlated binary data framework (Alonzo and Pepe, 2002). Motivated by the need for model diagnostics for the ROC-GLM model (Krzanowski and Hand, 2009), GEE cluster-deletion diagnostics (Preisser and Qaqish, 1996) are applied in an example data set to identify cases that have undue influence on the model parameters describing the ROC curve. In addition, deletion diagnostics are applied in an earlier stage in the estimation of the ROC-GLM, when a linear model is chosen to represent the relationship between the test measurement and covariates in the control subjects. The second paper presents a new model for diagnostic test accuracy meta-analysis. The common analysis framework for the meta-analysis of diagnostic studies is the generalized linear mixed model, in particular, the bivariate logistic-normal random effects model. Considering that such cluster-specific models are most appropriately used if the model for a given cluster (i.e. study) is of interest, a population-average (PA) model may be appropriate in diagnostic test meta-analysis settings where mean estimates of sensitivity and specificity are desired. A PA model for correlated binomial outcomes is estimated with GEE in the meta-analysis of two data sets. It is compared to an indirect method of estimation of PA parameters based on transformations of bivariate random effects model parameters. The third paper presents an analysis guide for a new SAS macro, PAMETA (Population-averaged meta-analysis), for fitting population-averaged (PA) diagnostic accuracy models with GEE as described in the second paper. The impact of covariates, influential clusters and observations is investigated in the analysis of two example data sets.

To my family for always believing in me.

Acknowledgments

The road traveled during this program has been a long one. Through the years there have been some people who have supported me unconditionally during this time. First, John Preisser and I have been working together since the beginning of my program, early on as my GRA advisor and eventually my dissertation co-advisor. John has always been supportive of my program progress and dissertation, provided excellent advice when I sought guidance, delivered constructive criticism when needed and always found time to help me with anything I needed.

To my best friend and partner for life, Kimberly, thanks for your patience during the particularly difficult times during this process.

I would also like to thank Quintiles, where I have worked for many years of this doctoral program. Quintiles provided financial support to me (and hundreds of other working students) over the years. In addition, Quintiles allowed for flexible work schedules when needed in order to pursue academic interests.

There are far too many people to list here who have helped me in various ways over the years. I provide here a final general acknowledgement to each person whom I have had the pleasure of working with has provided me with a key insight or skill that has allowed me to continue through this process.

Table of Contents

List of Tables x

List of Figures xii

1 Literature Review 1

1.1 Introduction to Diagnostic Test Accuracy 1

1.2 Introduction to ROC Curves 2

1.2.1 Notation and Properties 4

1.2.2 ROC Estimation Methods 5

1.3 Covariate Adjustment of ROC Curves 8

1.3.1 Indirect Regression Methods 9

1.3.2 Direct Regression Methods 11

1.3.3 ROC Regression Model Diagnostics 16

1.4 Meta-analysis of Diagnostic Tests 18

1.4.1 The Summary ROC curve 19

1.4.2 The Hierarchical sROC 20

1.4.3 Bivariate Random Effects Models 21

1.4.4 Generalized Linear Mixed Models 22

1.5 Motivating Examples 23

2 Identifying Influential Cases with the ROC-GLM 26

2.1 Introduction to Diagnostic Test Accuracy 26

2.2 Methods 29

2.2.1 Overview of ROC-GLM 29

2.2.2 GEE Estimation of the ROC-GLM 33

2.2.3 Cluster-deletion Diagnostics 36

2.3 Analysis of the Neonatal Audiology Data set 38

2.3.1 Reference distribution model step of ROC-GLM 39

2.3.2 Covariate Adjustment model step of ROC-GLM 41

2.3.3 Influence on Covariate-adjusted AUC 42

2.4 Discussion 43

3 A Semi-parametric PA Approach to Diagnostic Test Meta-Analysis 52

3.1 Introduction 52

3.2 A PA Model for Diagnostic Test Meta-analysis 55

3.2.1 GLMM Model Definition 55

3.2.2 PA Model Definitions and Estimation Procedures 56

3.2.3 Relationship of PA and SS model parameters 58

3.3 Data Examples and Analysis 61

3.3.1 Example 1: Catheter Segment Culture Data 61

3.3.2 Example 2: Simulated Correlated Binomial Data 63

3.4 Simulation Study 64

3.4.1 Simulation design 65

3.4.2 Simulation Study Results 66

3.5 Discussion 68

4 Implementation guide for PA diagnostic accuracy meta-analysis . . 76

4.1 Introduction to Diagnostic Test Accuracy 76

4.2 PA Model Definitions and Estimation Procedures 78

4.3 %PAMETA Macro overview and Implementation 81

4.3.1 Macro inputs and description 81

4.3.2 Output overview 83

4.4 Illustrative Data Sets and Analysis 83

4.4.1 Data set 1: Blood stream catheter infection data 83

4.4.2 Data set 2: Lymph node metastases data 89

4.5 Conclusions 92

5 Conclusion 101

Appendix I: Supplemental Tables to Chapter 3 103

Appendix II: Diag104.sas macro 108

Appendix III: Determination of ROC curve from PA model 115

Bibliography 117

List of Tables

2.1 Results of linear model for DPOAE in control subjects only, estimated with GEE to account for clustering of ears within subjects 49

2.2 Results of ROC-GLM model, case subjects standardized to controls, all data 50

2.3 Change in AUC with removal of influential subjects 51

3.1 Comparison of PA Logit and Probit models with heterogeneous scale for clusters for the catheter segment culture data , as well as the GLMM-converted analogs. All standard errors reported are empirical except for PA parameters ϕ_0, ϕ_1 and ρ_g which are estimated via a 2-stage bootstrap approach. 71

3.2 Results of GLMM model for Logit and Probit links for the catheter segment culture data (from Chu et. al (2010)) 72

3.3 Simulated Data: Comparison of Logit and Probit link functions for the PA method heterogeneous scale for clusters. All standard errors reported are empirical except for PA parameters ϕ_0, ϕ_1 and ρ_g which are estimated via a 2-stage bootstrap approach. 73

3.4 Simulated Data: Results of GLMM model for Logit and Probit links . . 74

3.5 Percent coverage of nominal 95% confidence intervals based upon empirical standard errors after fitting PA model as well as SS-model marginal converted results 75

4.1 Data from a Meta-Analysis of Studies on Semi-Quantitative (Type=1) or Quantitative (Type=2) Catheter Segment Culture for Diagnosis of Intravascular Device-Related Bloodstream Infection. (Source: Chu et al. (2010) 95

4.2 Comparison of Full and Reduced models for the catheter segment culture data (all data), for PA model with logit link (bias-corrected standard errors) 96

4.3 Studies 18 and 20 removed: Comparison of Full and Reduced models for the catheter segment culture data (all data),for PA model with logit link (bias-corrected standard errors) 97

4.4 Data from a Meta-Analysis of Studies on lymph node metastases. (Source: Klerkx et al. (2010) 98

4.5 Comparison of Full and Reduced models for the lymph node metastases data (all data), for PA model with logit link (bias-corrected standard errors) 99

4.6 Study 28 removed: Comparison of Full and Reduced models for the lymph node metastases data (all data), for PA model with logit link (bias-corrected standard errors) 100

List of Figures

2.1 Observation deletion diagnostics for Control Reference Distribution linear model portion of ROC-GLM. 45

2.2 Cluster deletion diagnostics for Control Reference Distribution linear model portion of ROC-GLM. DFBETAs are standardized by use of empirical standard errors. 46

2.3 Covariate adjusted ROC curves for Age (age=50 top line, age=40 middle line, age=30 lower line) for both full model (solid curves) and model without subjects 20409 and 11289 (dashed lines) 47

2.4 Cluster deletion diagnostics for ROC regression portion of ROC-GLM. 48

3.1 GLMM prediction region and PA model mean estimate with 95% confidence region. Upper left: Chu et al. 2010 data, logit link; Upper right: Chu et al. 2010 data, probit link; Lower left: Expanded and balanced data, logit link; Lower right: Expanded and balanced data, probit link. The confidence region is the smaller area contained within the prediction region. 70

4.1 Deletion diagnostics for catheter segment culture data. Upper left panel: Cluster (Study) level Cook's D; upper right: Observation (sensitivity and 1- specificity within a cluster) Cook's D; lower left: DFBETAS for 1-specificity; and, lower right: DFBETAS for sensitivity. 93

4.2 Deletion diagnostics for lymph node metastases data. Upper left panel: Cluster (Study) level Cook's D; upper right: Observation (sensitivity and 1- specificity within a cluster) Cook's D; lower left: DFBETAS for 1-specificity; and, lower right: DFBETAS for sensitivity. 94

Chapter 1

Literature Review

1.1 Introduction to Diagnostic Test Accuracy

Modern medical decision making often involves one or more diagnostic tools (such as laboratory tests and/or radiographic images). These diagnostic tools are developed using the most current technology available, and are often welcomed into medical practice with the hope of improving the care for patients. A diagnostic tool must be evaluated for it's discriminatory ability to detect presence (or absence) of current health state. The basic properties of quantitative evaluation of diagnostic tools were set forth over half a century ago in the field of signal detection. The quantitative properties of a decision tool involve assessments of how well the tool discriminates between states.

Using the notation of Pepe (2003), the variable for true disease status is defined as $D = 1$ for a diseased subject and $D = 0$ for a non-diseased subject. The notation of $\bar{D}$ for non-diseased and D for diseased subjects is also frequently used when displaying equations for the regression models. The variable Y is the result of the diagnostic test: $Y = 1$ indicates positive disease status, while $Y = 0$ denotes negative disease status.The measures of accuracy displayed next include the disease-specific classification probabilities false positive fraction (FPF) and true positive fraction

(TPF). TPF is also referred to as sensitivity, while $1 - FPF$ is also known as specificity. The first measures of accuracy of interest are those of quantifying the misclassification probabilities for each disease group. The ideal test would have no false positives or false negatives, since these are considered errors. The true and false positive fractions are defined as:

$$FPF = P[Y = 1|D = 0]$$
$$TPF = P[Y = 1|D = 1]$$

These quantities address the question: to what degree does the test reflect the true disease state? The ideal test has $FPF = 0$ and $TPF = 1$, while a completely uninformative test has $TPF = FPF$. The FPF and TPF can be considered either probabilities or fractions. However, these are often called false positive and true positive 'rates', which they are not (Pepe, 2003) because the numerator and denominator are in the same scale. There is a large body of literature concerning the analysis of binary tests, most of which is based upon the theory of 2x2 tables. Specific topics for binary tests including methods for a single test, multiple tests and regression models are summarized in Pepe (2003).

1.2 Introduction to ROC Curves

While it is natural to think of diagnostic tests in terms of a dichotomous outcome, in practice many tests are created on a continuous or ordinal scale and then possibly simplified into a dichotomous outcome (such as a pregnancy test). The previous section introduced measures of diagnostic accuracy and possible analysis models, all assuming the test of interest was dichotomous. Diagnostic tests that are measured on a continuous or ordinal scale are now examined.The consideration must be made that since there are more than just two possible outcomes of the test, there is now more than just one 2x2 table to consider. For each result of a given test, there is an associated set of accuracy measures. The Receiver Operating Characteristic (ROC) curve is a device that describes the range of tradeoffs between failing to detect disease and falsely identifying disease with the test (Pepe, 2003).

The development of the ROC curve can be traced back to the early 1950s where it was developed for signal detection and radar applications (Metz, 1986). In the 1950s, the first application of ROC to a medical test was completed when researchers attempted to quantify the ability of a Pap smear analyzer to discriminate between malignant and benign tissue samples (Zweig and Campbell, 1993). In the 1960s, ROC plots began to surface in psychology and psychophysics studies (Metz, 1986). Lusted (1960) provided the first paper on using "logical analysis" in radiology by presenting decision making tradeoffs with an ROC curve. The statistical development of ROC analysis can be traced initially to Patton (1978) who gives the first decidedly statistical summary presenting the probability theory in the context of a 2x2 decision analysis. Dorfman and Alf (1968) presented a maximum likelihood method that was used in early binormal ROC curve analysis, but this was not presented specifically as an ROC-specific method at the time.

Significant development of statistical methodology for ROC analysis came in the 1980's: for example, Metz (1986) was one of a number of articles published in the image evaluation area. In addition papers such as DeLong et al. (1988) offered the background theory for the empirical ROC curve and associated area under the curve (AUC). There was also a generalization to families of models including methods for applying generalized linear models (Tosteson and Begg, 1988). The 1990s continued the refinement of the binormal theory (Hanley, 1996; Metz et al., 1998) and a detailed exposition of the empirical theory (Hsieh and Turnbull, 1996).

1.2.1 Notation and Properties

Pepe (2003) summarized the attributes of ROC curves for evaluating diagnostic tests as providing a complete description of test performance, facilitating comparing and combining information across studies of the same test, guiding the choice of threshold value and providing a mechanism for relevant comparisons between different non-binary tests. Since the ROC curve transforms all results to the TPF and FPF scale, comparisons of different tests can be examined for the same disease, regardless of the units or scale of measurement.

The ROC curve can be viewed as a function that describes the distance between distributions. While the focus is typically on diagnostic tests, it is possible to use an ROC curve as an exploratory curve any time interest lies in the difference between the distribution of two groups. Brumback et al. (2006) provided an interpretation for the ROC curve when used to describe the differences between two treatment groups in a clinical trial, for example. By using a threshold , it is possible to transform a continuous test result into a dichotomous outcome. Assuming a test, Y , is positive if $Y \geq c$ and negative if $Y < c$ then the following represent the responses of controls and cases, respectively,

$$Y_{\bar{D}j}, j = 1, \ldots, n_{\bar{D}}$$

$$Y_{Di}, i = 1, \ldots, n_D$$

It is assumed that $Y_{\bar{D}i}$ and Y_{Di} are randomly selected from the population of test results associated with the diseased and non-diseased states (Pepe, 2003).

The definition of TPF and FPF may then be augmented:

$$TPF(c) = P[Y \geq c|D = 1]$$
$$FPF(c) = P[Y \geq c|D = 0]$$

The ROC curve is then defined as the entire set of TPF and FPF pairs after dichotomizing Y with different values of c :

$$ROC(\cdot) = [(FPF(c),\, TPF(c)), c \in (-\infty, \infty)]. \qquad (1.1)$$

When $c = \infty$, then $\lim_{c\to\infty} TPF(c) = 0$ and $\lim_{c\to\infty} FPF(c) = 0$, while at the opposite end of the interval we have $c = -\infty$, then $\lim_{c\to-\infty} TPF(c) = 1$ and $\lim_{c\to-\infty} FPF(c) = 1$. It is also possible to write the ROC curve as

$$ROC(\cdot) = [(t,\, ROC(t)), t \in (0, 1)] \qquad (1.2)$$

where $t = FPF(c)$ and $ROC(t) = TPF(c) = TPF(FPF^{-1}(c))$.
The ROC curve is a monotone increasing function mapping two [0,1] intervals. The uninformative test has an ROC curve that has unit slope through the unit square. In this case the distributions of test results for the diseased and non-diseased subjects are identical. On the other end of the spectrum the perfect test has an ROC curve that traces the left and upper limits of the unit square since $TPF(c) = 1$ and $FPF(c) = 0$.

1.2.2 ROC Estimation Methods

Empirical Method

DeLong et al. (1988) formally defined the basic properties of nonparametric empirical curves. These curves have a relationship to the Mann-Whitney U statistic (via the area under the curve) and are not smooth (resembling a Kaplan-Meier type of shape). Zweig and Campbell (1993) argued that continuous diagnostic tests should employ the purely nonparametric method since parametric methods were developed for ratings data. Hsieh and Turnbull (1996) later defined the asymptotic properties of the empirical ROC curve. The empirical ROC curve is a function only of the ranks of the data because it only depends on the relative orderings of the test results and their diseased status. For each possible cut-point c, the empirical estimates of TPF and FPF are, respectively:

$$\widehat{TPF}(c) = \sum_{i=1}^{n_D} I[Y_{Di} \geq c]/n_D \tag{1.3}$$

$$\widehat{FPF}(c) = \sum_{i=1}^{n_{\overline{D}}} I[Y_{\overline{D}j} \geq c]/n_{\overline{D}} \tag{1.4}$$

The empirical ROC curve is a plot of $\widehat{TPF}(c)$ versus $\widehat{FPF}(c)$ for all $c \in (-\infty, \infty)$ and denoted by $\widehat{ROC}_e(t)$. This is considered a discrete function because $\widehat{FPF}(c)$ can only take on values in increments of $1/\,n_{\overline{D}}$. Joining these points on a graph gives a step function with vertical jumps of $1/\,n_D$ corresponding to subjects from diseased subjects, while horizontal jumps of $1/\,n_{\overline{D}}$ are made from subjects in the non-diseased group. Ties within each group result in larger vertical or horizontal jumps, while ties in test results between diseased and non-diseased subjects result in diagonal jumps. A confidence band for the ROC curve was presented in Hsieh and Turnbull (1996). The topic of nonparametric confidence bands for the ROC curve is identified as an area requiring more statistical research (Pepe, 2003).

The empirical area under the curve is the Mann-Whitney U-statistic:

$$\widehat{AUC}_e = \sum_{j=1}^{n_{\overline{D}}} \sum_{i=1}^{n_D} \left(I[Y_{Di} > Y_{\overline{D}j}] + \frac{1}{2} I[Y_{Di} = Y_{\overline{D}j}] \right) /n_D\, n_{\overline{D}}. \tag{1.5}$$

When there are no ties between diseased and non-diseased observations the above expression simplifies to:

$$\widehat{AUC}_e = \sum_{j=1}^{n\overline{D}} \sum_{i=1}^{n_D} \left(I[Y_{Di} > Y_{\overline{D}j}] \right) / n_D \, n_{\overline{D}}. \tag{1.6}$$

Hanley and McNeil (1982) presented results for the asymptotic variance when observations are independent. DeLong et al. (1988) discussed an alternative representation of the asymptotic variance. The variability of the AUC is often calculated using the bootstrap, especially when the data are clustered (Pepe, 2003). In the case of clustered data, such as when a subject contributes multiple test data, bootstrap resampling is performed at the cluster level.

Examples of other nonparametric methods include Zou et al. (1997) and Zhou et al. (2002) who presented studies in kernel density smoothing. They both use kernels and bandwidth selection procedures, however they arrive at the smooth curve in a slightly different way. Their work provides some interesting theoretical results to help determine the theoretical basis and justification for smoothing in ROC curve analysis.

Parametric Method

With its foundation in Gaussian distribution theory, the binormal curve has become a common analysis tool for ROC curves, most commonly in the radiology imaging evaluation area. Metz (1986) and Metz et al. (1998) are just two of dozens of articles written by Charles Metz and colleagues. Despite being motivated based on Gaussian-distributed test results, later it will become evident that this condition may be relaxed. Given $Y_D \sim N(\mu_D, \sigma_D^2)$ and $Y_{D*} \sim N(\mu_{\overline{D}}, \sigma_{\overline{D}}^2)$ then the ROC curve is defined as

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)) \tag{1.7}$$

where $a = \frac{\mu_D - \mu_{\overline{D}}}{\sigma_D}$,$b = \frac{\sigma_{\overline{D}}}{\sigma_D}$ and Φ is the standard normal cumulative distribution function. Using the convention that larger test results are more indicative of disease $a > 0$, since $\mu_D > \mu_{\overline{D}.}$ The binormal method produces smooth curves which are aesthetically pleasing. Also, the binormal model is appealing for ordinal predictors as is often found in radiology studies for example (Metz et al., 1998).

1.3 Covariate Adjustment of ROC Curves

When considering binary tests it was observed that regression models could be fit for FPF and TPF separately, as well as predictive values and DLRs. Methods have been developed to fit models to continuous data, which are analogues to those fit to FPF and TPF. Covariate effects are evaluated on the non-disease reference distribution and the ROC curve which quantifies the discriminatory capacity of the test (Pepe, 2003).

Evaluation of covariate effects on the non-disease reference distribution F_X determines which factors affect the false positive fractions when a test threshold is fixed. Stated another way, it is determined whether thresholds should be defined differently for sub-populations with different covariate values in order to keep FPF constant across these subgroups. The methods for this are straightforward, making use of regression quantiles. When covariate effects are modeled on the ROC curve itself, the issue of interest is whether or not the covariates affect the ability of the test to discriminate disease from non-disease independent of threshold (Pepe, 2003).

Inference about the accuracy of a given test may be biased if covariate effects are neglected (Pepe, 2003). The classic case of confounding occurs when test results are related to covariates and the distributions of the covariates are different for both the diseased and non-diseased populations (Pepe, 2003). However, it is also possible to have bias when the distributions are the same in both populations. There are two cases to consider: the covariate affects only the test result, or the covariate affects the ROC curve but not the test results. The radiology literature has much discussion on attenuation of the ROC curve by ignoring covariates on the distribution of the test results (Rutter and Gatsonis, 2001). In the case of radiology studies, the 'reader-specific' ROC curve attenuates the overall ROC curve due to differing usage of the rating scale for a given image.

When deciding whether to present the pooled or covariate-specific curves, consideration should be given to what use the test result will have. If the test result and given threshold will be used for a given covariate level (such as age group), then the covariate-specific curve is of practical importance. However, if the test were to be used across all age groups then the pooled curve is more relevant. For the second case, where the covariate does not affect the test results of the non-diseased population but does affect the ROC curve, the pooled ROC curve can be thought of as a weighted average of the covariate-specific ROC curves. In this case the covariate level ROC curves are of interest. Pepe (2003) observed that in data analysis situations it may be useful to present both the pooled and covariate-specific curves.

1.3.1 Indirect Regression Methods

The first method for evaluating covariate effects on ROC curves was proposed by Tosteson and Begg (1988) and would later be followed up by Toledano and Gatsonis (1995). Although these two papers considered primarily ordinal data, the concepts apply more generally (Pepe, 1998).

For continuous data the approach of this method is to model F_X for both cases and controls, and then calculate the covariate-specific ROC curve for covariate values of interest. Tosteson and Begg (1988) achieved this by employing a location-scale ordinal regression model. Test results that follow a location-scale family yield parameters that describe the covariate effects of test results. In this case the parameters quantify covariate effects on the ROC curve. Note that the discrete ROC function framework is adopted for this model as opposed to the latent variable posture. The reason for this is that standard statistical packages do not handle the estimation properly due to the dependence of the scale parameter on disease and status and covariates.

It is possible to fit a location-scale model without specifying F_X, which is then a semi-parametric alternative to the previous methods listed above. In this case, quasi-likelihood may be used for estimation of the parameters. The induced curve ROC estimate does require an estimator for F_X. A proposed estimate is found in Pepe (1998), which estimates F_X with the empirical distribution of the standardized residuals which is very similar to a semi-parametric regression quantile estimator.

Location-scale models that incorporate random effects are often fit to acknowledge the correlations between test results. In the context of diagnostic accuracy evaluation, fitting these random effects models is no different than other applications (Pepe, 2003). Random effects models can also provide insight into test result variability. In the radiology setting multiple readers of a set of images present a level of correlation that may be important to quantify using random effects. The random effects formulation of the location-scale models presented previously would be of interest when there are a large number of readers and inference is to be generalized to the entire population of readers. Gatsonis (1995) followed by Ishwaran and Gatsonis (2000) presented advanced discussion of this topic. Etzioni et al. (1999) proposed a random effects model for longitudinal data regarding PSA testing. Finally, both Gonen and Heller (2010) and Devlin et al. (2010) have proposed models that are considered to be Lehmann Family models.

In summary, the indirect models assume a functional form of the distribution of the test results. This can impose unnecessary restrictions on the modeling process.

This prompted a new avenue of research into methods with less assumptions about the distributional form of the test results.

1.3.2 Direct Regression Methods

The first direct regression method was proposed by Pepe (1997) using ordinal data and applying the GEE for estimation. The individual test results are transformed to indicators that are then used for modeling. Using the notation of Pepe (2003), the variable for true disease status is defined as $D_i = 1$ for a diseased subject and $D_i = 0$ for a non-diseased subject. The notation of $\bar{D}$ for non-diseased and D for diseased subjects is used when displaying equations for the regression models. The variable Y_i is the diagnostic test result for subject i. Let $\{Y_{\bar{D}_j},\, j = 1,\, ...\, ,\, n_{\bar{D}}\}$ and $\{Y_{D_i},\, i = 1,\, ...\, ,\, n_D\}$ represent the ordinal or continuous responses of controls and cases, respectively, with larger values being more indicative of disease. It is assumed that Y_{D_i} and $Y_{\bar{D}_i}$ are randomly selected from the population of test results associated with the diseased and non-diseased states (Pepe, 2003). Next, we define a covariate vector, X, which contains the covariates that affect the test result distribution in control subjects, as well as those that affect the discrimination between cases and controls. Finally, we define a set of t discrete points $f = f_1,\, ...\, ,\, f_t$, on the x-axis of the ROC curve, chosen from the interval $(0, 1)$, over which the model will be fit. For these points, define

$$U_{it} = I[Y_i \geq F^{-1}_{D,\mathrm{X}j}(t)] - g(\alpha(t), \beta X) \tag{1.8}$$

Pepe (2000) then introduced the Receiver Operating Characteristic Generalized Linear Model (ROC-GLM), relaxing assumptions on the distribution of test results and improving on estimation using binary regression. This method uses ranks of the test results to create binary indicators as the response variable in the GLM:

$$U_{ij} = I(Y_{Di} \geq Y\bar{D}j) \tag{1.9}$$

There are two components to the ROC-GLM regression model. The first is the vector of covariate values X and the second is the specification of the ROC curve as a function of f. If $h_0(\cdot)$ and $g_0(\cdot)$ are monotone increasing (or decreasing) functions on $(0, 1)$ then

$$g(ROC_X(f)) = h_0(f) + \beta X \tag{1.10}$$

is an ROC-GLM regression model (Pepe, 2000).

Further work on the ROC-GLM occurred in Alonzo and Pepe (2002) and Pepe (2000). The concept of estimating a reference distribution for the control subjects and then standardizing case test results to these as “percentile” values are the basis of creating the model. We summarize, and then expand upon, the following 3 general steps required to perform a covariate adjustment of ROC curves using the ROC-GLM (Janes et al., 2009):

1. Estimate $PV_{DX_i} = F_X(Y_{DX_i})$, the percentile values of the test results for cases, where F_X is the distribution of test results in controls as a function of the covariates.

2. Estimate the cdf of the percentile values as a function of the covariates.

3. Specify the adjustment of the ROC curve as a function of the covariates. We then employ GEE for binary data to estimate the model parameters (covered in Section 2.2).

First, an estimate of F_X, the distribution of test results in the control group, is required. Essentially, we begin the process of standardizing the test results by finding the baseline relationship among the controls. A simple linear model could be specified (Janes et al., 2009) such that

$$Y_{\bar{D}_i} = \psi_0 + \psi_1^{'} X_i + \epsilon_i. \tag{1.11}$$

where ϵ_i are *i.i.d.* as $N(0, \sigma^2)$.
We observe that this is the first opportunity to apply ordinary linear model deletion diagnostics (such as Cook's D for simple linear models) in the estimation steps of the ROC-GLM. Given that the model in (2.1) is crucial to the remaining steps, it is proposed that deletion diagnostics be applied at this step to assess the control distribution model. We present the deletion diagnostics in a following section.
Having settled on a linear model in the previous step, and having assumed Gaussian errors for this linear model, then the percentile values for the cases are defined as

$$\widehat{PV}_{DX_i} = \Phi\left((Y_{DX_i} - \hat{\psi}_0 - \hat{\psi}_1^{'} X_i)/\hat{\sigma}\right). \tag{1.12}$$

If Gaussian errors and/or a linear relationship are too restrictive for a given application, there are other alternatives proposed. For example, Heagerty and Pepe (1999) propose an empirical estimation of the error distribution using the residuals of the linear model. Further, instead of assuming a linear relationship of the test result in the controls, one could use a stratified approach (Janes et al., 2009).

At this stage we have now standardized the test results for the cases as a function of the controls by the above step. Next, we must estimate the cdf of the percentile values.

In the second step, we make use of the fact that an ROC curve is essentially the cdf of the percentile values calculated above (Pepe, 2003). Defining the $h(f)$ as the cdf (recall that f are the chosen set of values on the x-axis of the ROC curve), we can write:

$$h_X(f) = ROC_X(f) = P(1 - PV_{DX} \leq f) = g(\beta_0 + \beta_1 g^{-1}(f)) \tag{1.13}$$

where $g(\cdot)$ gives a parametric form of the ROC curve; $g = \Phi$ is the standard normal c.d.f. and $g(\cdot) = exp(\cdot)/[1 + exp(\cdot)]$ is the logistic function giving binormal and bilogistic ROC curves respectively.

The result after this second step is an ROC curve that is not yet adjusted for covariates that discriminate between the cases and controls. However we do now have an ROC curve that is inherently adjusted for how covariates affect the test distribution results. This is quite important as Pepe (2003) demonstrates that ”pooled” or unadjusted ROC curves are biased.

The final step in the model specification is to create the inputs for a regression model using the newly created percentile value cdf, and the covariates that are assumed to affect the discriminatory capacity of the test. In other words, covariates that affect the intercept and/or slope of the ROC curve. Recall that we have defined T discrete points on the x-axis of the ROC curve over which to fit the model. We also define $U_{it} = I_{1-PV_{DX_i} \leq f_t}, t = 1, ..., T$, as the set of cumulative binary indicators which determine whether or not the percentile values are less than each choice of f. For example, if we chose $t = 10$ values of f then each subject would have a vector of 10 binary indicators for each percentile value within a cluster. Next, we define the covariates $X_D g^{-1}(f)$ as those that will enter the model as ones that affect discrimination. The complete model combining steps 2 and 3 is:

$$ROC_{X,X_D}(f) = g(\beta_0 + \beta_1 g^{-1}(f) + \beta_2^{'} X_D + \beta_3^{'} X_D g^{-1}(f)) \tag{1.14}$$

We may think of this final step as defining a model that has as its output a“baseline” ROC curve (from step 2 and in equation 2.3) and some additional model parameters that specify covariate-adjustments of that baseline curve. The previous steps also allow for flexibility in defining which covariates are important for adjusting the control test results distribution and those which affect the discrimination between cases and controls. At this point it is important to note that Pepe (2003) advocate using bootstrap standard errors for the estimates $\hat{\beta}$ from the fitted model. The reason is that since we do not have true independence between responses and covariates there could be bias in the standard errors. In the case of a covariate that affects both the test result distribution and the discriminatory capacity, this covariate would essentially influence both the responses U_{it} and the covariates in X.

Pepe (2003) suggested using the independence working covariance matrix for fitting the model. Any method for estimating the reference distribution (regression quantiles) may be used, though the empirical method is most robust (Pepe, 2003). The choice of f is important since this will determine the interval over which the model is to hold. The number of points in f, (denoted earlier as f_t) should be finite so that standard statistical software can handle the estimation. There is currently no method designed to choose the values in the domain that give optimally efficient results (Pepe, 2003). Alonzo and Pepe (2002) found relatively good efficiency for small values of f_t. Pepe (2003) suggests that in practice it is possible to estimate parameters with increasing the number of points in f_t, stopping when the decreases in standard errors become small.

Estimation of the ROC-GLM model proceeds following the generalized estimating equations (GEE) procedure (see Chapter 2 for details). Recall from above that $h_X(f)$ defines the basis for the ROC curve (having standardized cases to the control reference distribution. Typically the choice for link function f is the probit function which is the binormal model. One may choose the bi-logistic or any other basis. Alternatively, a semi-parametric formulation of the ROC-GLM is also available where

$h_X(f)$ is not formally parameterized (Cai and Pepe, 2002). Other research in this area include Cai and Moskowitz (2004) who proposed a profile MLE of ROC with binormal basis function (a special case of ROC-GLM with no covariates), as well as a pseudo-MLE (where covariates can be included). Pepe and Cai (2004) developed an extension of Cai and Pepe (2002) where semi-parametric ROC-GLM can be viewed as a transformation model of the placement values. In practice, assuming a probit or logistic basis function for the ROC curve is a reasonable assumption which eliminates the need for computation of methods such as Pepe and Cai (2004).

1.3.3 ROC Regression Model Diagnostics

Cai and Zheng (2007) introduced model checking diagnostics for the ROC-GLM. The asymptotic distributions are derived for cumulative residual-based model diagnostics for ROC regression models. The proposed method is an extension of model diagnostic procedures for traditional GLM models originally presented in Lin et al. (2002). The ROC-GLM extension of three model checks (adequacy of ROC-GLM model, link function and interaction of covariate effects with FPF) is based upon the semi-parametric ROC-GLM presented in Cai and Pepe (2002). Given the task of simultaneously evaluating the test result distributions as well as the relationship between them requires these important extensions. One practical application of this could be investigating the linearity of time in a longitudinal study. It is possible to test whether time enters the model linearly and adjust the coefficients by perhaps adding a quadratic term to the model.

It is natural to ask whether data from a single case (subject) has a large influence relative to other cases on the estimates in the marginal mean model. For the h-th element of β, interest is often in $(\hat{\beta}_h - \hat{\beta}_{h[i]})$, the difference in the parameter estimate with and without the i-th case included in the data. Preisser and Qaqish (1996) introduced computationally quick approximations for both observation- and cluster-deletion diagnostics for GEE. However, only the latter, which we call case-deletion, are relevant for this application because the observation-level diagnostics have no real interpretation in the ROC-GLM. Recall that the $U_{it}, t = 1, \ldots, n_i$, are a set of binary placement value indicators constructed for the i-th case in the course of applying the estimation method; they don't have any inherent meaning as individual data values.

Following the formulae of Hammill and Preisser (2006), the influence of the i-th case as given by the $p \times 1$ vector $(\hat{\beta}_1 - \hat{\beta}_{1[i]}, \ldots, \hat{\beta}_p - \hat{\beta}_{p[i]})$ can be approximated any further iterations following convergence of the GEE iteratively weighted least squares algorithm by

$$DFBETAC_i = M^{-1} D_i' V_i^{-1} (I - H_i)^{-1} r_i$$

where $H_i = D_i M^{-1} D_i' V_i^{-1}$ is the cluster leverage matrix. Note that $DFBETAC_i$ is a measure of the influence that each cluster has on the estimate of each parameter element of β. Further, there is a close relationship of the set of $DFBETAC_i, i = 1, \ldots, K$ with the bias-corrected variance estimator

$$V_{bc}(\widehat{\beta}) = \sum_{i=1}^{K} (DFBETAC_i)(DFBETAC_i)'$$

Standardization of $DFBETAC_i$ is achieved by dividing each of its elements by the standard error of its respective parameter estimate, usually based on the full data. Finally, a measure of the influence of the i-th cluster on the overall model fit can be estimated by Cook's D:

$$DCLS_i = (DFBETAC_i)' [\mathrm{var}(\hat{\beta})]^{-1} (DFBETAC_i)/p$$

where var($\hat{\beta}$) is estimated by either the empirical (as in Ziegler et al. (1998) and Preisser et al. (2012)) or bias-corrected variance estimators defined above. Additional details are provided in Chapters 2 and 4.

1.4 Meta-analysis of Diagnostic Tests

Evidence-based decisions in health care are becoming increasingly utilized. From pharmaceutical development programs to medical treatment regimens in practice, the heightened awareness of methods to analyze data in support of health care decisions requires quantitative methods for summarizing the evidence. Meta-analysis, decision analysis and cost-effectiveness analysis are the cornerstones of evidence-based medicine (Petitti, 2000). The meta-analysis of diagnostic tests is of particular interest in certain screening programs for certain diseases such as cancer. Cervical cancer screening in women and prostate cancer screening in men are both examples of heath screening programs that have a great deal of diagnostic test accuracy studies to draw from for meta-analysis.

Meta-analysis of clinical trials may be employed using various methods that attempt to find the mean effect, however for diagnostic studies the typical summary data points are two dimensional . These measures tend to be positively correlated since studies tend to vary in how test positivity is defined (Pepe, 2003). In the paragraphs below the evolution of the statistical methods for diagnostic test accuracy meta-analysis are presented. Pepe (2003) lists three benefits of meta-analysis for diagnostic tests: awareness within the research community of previous studies, explanation of discrepancies between individual study results and identification of common mistakes in study design thereby providing guidance for design of future studies. For the interested reader, two excellent books reviewing the broad spectrum of general meta-analysis considerations and statistical methods include Hedges and

Olkin (1985) and Petitti (2000).

1.4.1 The Summary ROC curve

Moses et al. (1993) propose a summary ROC curve for the set of values of TPF and FPF, which we denote as (TPF_k, FPF_k) for each of k studies summarized in the meta-analysis. The ROC-like curve, called sROC, is a curve that goes through the scatter plot of each TPF and FPF pair. In contrast to standard ROC analysis the resultant curve need not yield a monotonic curve (Walter, 2002).The regression equation proposed is $D = a + bS$ where $D = log(TPF/1 - TPF) - log(FPF/1 - FPF)$ which is equivalent to the diagnostic log-odds ratio, which conveys the test's accuracy from discriminating cases from non-cases, and $S = log(TPF/1 - TPF) + log(FPF/1 - FPF)$ which is an interpretation of the diagnostic threshold with high values corresponding to liberal inclusion criteria for cases. The regression equation is then fit with ordinary least squares assuming that D is approximately normally distributed for a given value of S. Weighted analysis may be employed (i.e. weighted least squares) to account for the heterogeneity of studies which is achieved through the sample variance of D. Pepe (2003) notes however that inaccurate studies with large sample sizes may then skew the results even further than just a regular unweighted analysis. If b is equivalent or nearly 0 then the overall $log(OR)$ may be used to summarize the studies since $a = log(OR)$. Conversely if $b \neq 0$ then the studies are heterogeneous with respect to OR. van Houwelingen et al. (2002) note that one simple refinement of the Moses et al. (1993) specification is to make the intercept a random effect. $D_i = \alpha_i + \beta S_i + e_i$ with $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$. Overall, this procedure converts the (TPF_k, FPF_k) to a diagnostic odds ratio, thereby removing the inherent bivariate properties of the diagnostic test result. For this reason, methods that preserve the bivariate properties are a more intuitive way to analyze these data. The trade-off then becomes the complexity of analysis methods (Pepe, 2003).

1.4.2 The Hierarchical sROC

Rutter and Gatsonis (2001) note three weaknesses of the Moses et al. (1993) method.

1. Both D and S are derived from the same set of random variables (TPF_k, FPF_k) thereby inducing dependence between the two.

2. Since S is measured with error this may introduce bias into the estimate of regression coefficients, and

3. The decision and potential differences between weighted and unweighted least squares for parameter estimation

Pepe (2003) also notes that the assumption that the true values (TPF_k, FPF_k) lie on the sROC if the true values were known is not appropriate because that would assume the only difference between studies is the threshold for test positivity, which is generally not the only source of variation. This method is based on the location-scale parametric formulation of the individual ROC curve presented previously. Here the model is extended to allow variation in the parameters that define an individual ROC curve to accommodate the meta-analysis setting. The assumed form of the ROC curve from the k^{th} study is $logit(TPF_k) = logit(FPF_k + \mu_k)\sigma_k^{-1} = (\theta_k + \mu_k)e^{-b_k}$. The (TPF_k, FPF_k) pairs from the k^{th} study are assumed to have a binomial distribution. Rutter and Gatsonis (2001) assume that θ_k and μ_k are independent where $\mu_k \sim N(M, \sigma_M^2)$ and $\theta_k \sim N(\Theta, \sigma_\theta^2)$ and also that b_k is a constant b across all studies. All of the parameters M, Θ, σ_θ^2, σ_M^2 and b may be estimated using maximum likelihood or Bayesian methods, as outlined in Rutter and Gatsonis (2001).

Pepe (2003) notes the following important attributes of this binomial regression framework:

1. accomodates between-study variability that can be modeled with covariates or that may be considered to be random
2. fitting procedures have a sound theoretical basis in maximum likelihood or Bayesian methodology

The drawbacks of this method seem to be in the complexity of the estimating algorithms with freely available software. Assessing model fit can also be difficult with these methods (Pepe, 2003).

1.4.3 Bivariate Random Effects Models

The hierarchical sROC approach of Rutter and Gatsonis (2001) has been criticized for being complex and requiring sophisticated statistical knowledge and programming skills (Reitsma et al., 2005). As a result the simpler Bivariate Random Effects Model has been presented as a more intuitive, easy-to-use model. As a great deal of the literature in this area comes from the applied medical and diagnostic statistics journals, it is no surprise that this method has been preferred since 2005.

Let $n_{i11}, n_{i00}, n_{i01}$ and n_{i10} represent the number of true positives, true negatives, false positives and false negatives (see Table 1), and n_{i1+} and n_{i0+} be the number of diseased and non-diseased subjects in the ith study from a meta-analysis, where studies are indexed as $i = 1, \ldots, K$.. The bivariate random effects model is specified by conditioning on the number of diseased and non-diseased in each study. Assume n_{i01} and n_{i11} are binomially distributed as $\text{Bin}(n_{i0+}, 1 - Sp_i)$ and $\text{Bin}(n_{i1+}, Se_i)$ conditionally on Sp_i and Se_i which are the specificity and sensitivity parameters for the ith diagnostic study, respectively. The expected sensitivity for a chosen specificity is given by

$logit(Se) = \mu_0 + \rho\sigma_\mu/\sigma_\nu[logit(Sp) - \nu_0] = (\mu_0 - \rho\nu_0\sigma_\mu/\sigma_\nu) + \rho\sigma_\mu/\sigma_\nu[logit(Sp)]$. Let $\theta = (\mu_0, \nu_0, \rho, \sigma_\mu, \sigma_\nu)$ be the parameters of interest from a bivariate random effects meta-analysis model and $\hat{\theta}$ be the MLE of θ with estimated variance covariance $\hat{\Sigma}$.

After the original publication of this method by Reitsma et al. (2005) a number of follow-up papers have sought improvements and refinements to this method. Arends et al. (2008) discuss 5 different choices for bivariate random effects models transformation of the sensitivities and specificities, noting that the within-study distribution of sensitivity and specificity can be handled in one of two ways: the normal-normal (approximate normal distribution) or the binomial-normal (binomial distribution). Riley et al. (2007) and Riley et al. (2008) investigate more closely the estimation of the between-study correlations to aid practitioners in understanding heterogeneity in the bivariate random effects model. The hierarchical sROC and BVRE models are similar under certain asumptions (Chu and Guo, 2010). A first attempt at unifying the underlying methods theoretically was proposed by Harbord et al. (2007). Chu and Guo (2009) then offered a correction and clarification of the notation of the two methods.

1.4.4 Generalized Linear Mixed Models

Chu and Guo (2010) note that previously only logit transformations were used in the bivariate random effects model. A natural extension of this is to consider other link functions such as the probit and complementary log-log. The resulting generalization is the generalized linear mixed model for diagnostic accuracy meta-analysis. The differentiation between the BVRE models and the current model is the specification that $g(Se_i) = \mu_i$ and $g(1 - Sp_i) = \nu_i$ where the random effects $(\mu_i, \nu_i)^T$ are bivariate normally distributed with mean μ and covariance matrix Σ.

Here, $g()$ is a montone link function (for example the logit link). Chu and Guo (2010) also note that any transformation of the sensitivity and specificity may be used.

The GLMM is defined as follows. Following the notation of Chu and Guo (2010), assume n_{i01} and n_{i11} are binomially distributed as $\text{Bin}(n_{i0+}, 1 - Sp_i)$ and $\text{Bin}(n_{i1+}, Se_i)$ conditionally on Sp_i and Se_i which are the specificity and sensitivity parameters for the ith diagnostic study, respectively. Next, define

$$g(1 - Sp_i) = \beta_0 + \nu_i \tag{1.15}$$

and

$$g(Se_i) = \beta_1 + \mu_i \tag{1.16}$$

where the random effects are assumed to be distributed as $(\nu_i, \mu_i)' \sim N(0, D)$, where

$$D = \begin{pmatrix} \sigma_0^2 & \rho_m\sigma_0\sigma_1 \\ \rho_m\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

Estimation of the parameters $\theta_m = (\beta_0, \beta_1, \rho_m, \sigma_0, \sigma_1)'$ using MLE methodology is performed using numerical procedures such as Gaussian quadrature (as found in SAS NLMIXED, for example).

1.5 Motivating Examples

Three data sets are used as analysis examples for the various methods found in Chapters 2-4. The first data set is analyzed in Chapter 2 as an example of a diagnostic accuracy study where the ROC-GLM approach (a PA model) is employed along with model diagnostics. The original publication for the example data are found in Norton et al. (2000). The sample comprises of 2742 infants and 5058 ears upon which three diagnostic screening tests (DPOAE, TEOAE and ABR) were performed. The gold standard reference test applied is an audiometric behavioral response test. The study was conducted at 6 different clinical centers. The above example data set is one that has been used extensively to demonstrate analysis methods for covariate-adjusted ROC curves. For example, Janes et al. (2009) use the data extensively to demonstrate various analysis options.

The next set of analyses are related to diagnostic accuracy meta-analysis. The first example data set for this topic is a meta-analysis of 33 diagnostic accuracy studies previously analyzed in Chu et al. (2010). The 33 studies studied semi-quantitative (19 studies) or quantitative (14 studies) catheter segment culture for the diagnosis of intravascular device-related blood stream infection. Chu et al. (2010) report that since there is no statistically significant difference between the semi-quantitative and quantitative methods, the data are combined together without including this potential covariate in any model. For demonstration purposes we investigate the covariate for type of catheter segment culture method (semi-quantitative or quantitative). The mean number (std. dev.) of diseased and non-diseased persons per study was 20 (19.8) and 237 (240.5) respectively. The gold standard was final diagnosis of blood-stream infection. The data are presented in Chapters 3 and 4.

The second example data set is a meta-analysis of 32 diagnostic accuracy studies previously analyzed in Klerkx et al. (2010). The diagnostic accuracy of gadolinium-enhanced MRI in detecting lymph node metastases using histopathologic test as the reference gold standard. The mean number (std. dev.) of diseased and non-diseased persons per study was 15(18.5) and 28 (30.4) respectively. Covariates for partial verification bias (PVB, 8 studies) and study design (case control, 6 studies or cohort, 26 studies) are available. The data are presented in Chapter 4.

The common theme throughout both analysis situations (single study and meta-analysis) is the fact that all methods employed are based on a population-average approach, specifically the implementation of GEE as the estimation engine. Further, deletion diagnostics are presented in both cases as a method of evaluating influential observations (both in the single study ROC-GLM setting and the meta-analysis setting).

Chapter 2

Identifying Influential Cases with the ROC-GLM

2.1 Introduction to Diagnostic Test Accuracy

Modern medical decision making often involves one or more diagnostic tools (such as laboratory tests and/or radiographic images). Tests are designed to discriminate between different states of health or medical conditions, e.g. cancer and no cancer. Diagnostic markers with improved accuracy or decreased cost are also being sought for established diseases. Screening biomarkers have the potential to detect disease at an early stage, when it is most treatable. Pre-screening markers are being investigated for their use in identifying subjects at high risk of the disease, who should be targeted for screening or disease-preventative interventions. Prognostic markers can be used, for example, to predict which patients will respond to treatment. In all of these settings, the primary question is how well the biomarker distinguishes between the two groups of individuals, the “cases” and the “controls”.

Receiver operating characteristic (ROC) curves are a well-accepted measure of accuracy for tests that yield ordinal or continuous results. Based on the notion of using a threshold to classify subjects as positive or negative, an ROC curve is a plot of the the true positive fraction (TPF) versus the false positive fraction (FPF) for all possible cutpoints. The TPF, also called the sensitivity, is the proportion of diseased subjects correctly detected by the test. On the other hand, FPF or (1-specificity) is defined as the proportion of non-diseased subjects erroneously deemed positive by the test. Thus, the ROC curve describes the whole range of possible operating characteristics for the test and hence its inherent capacity for distinguishing between diseased and non-diseased states.

The use of a regression framework to account for covariates in a diagnostic accuracy study was first proposed by Tosteson and Begg (1988) and would later be expanded upon by Toledano and Gatsonis (1995). Tosteson and Begg (1988) used regression models for the test outcome and inferred covariate effects on the corresponding ROC curves. Although these two papers considered primarily ordinal data, the concepts apply more generally Pepe (1998). The method of Pepe (1997) that directly models the ROC curve is a good practical choice for an ROC regression model because of its ease of interpretation. The model estimation approach was refined by Alonzo and Pepe (2002), and presented as the receiver operating characteristic generalized linear model (ROC-GLM), to allow for ease of fitting through application of generalized estimating equations (GEE) within a correlated binary data framework. The binary indicators are constructed from a diseased subject's test result according to whether it exceeds various specified quantiles of the distribution of test results from non-diseased subjects with the same covariates. It is this modeling approach that will be used for the remainder of this paper. For a complete review of alternative methods the interested reader is referred to Pepe (2003).

With the exception of outlier detection in univariate random effects meta-analysis Baker and Jackson (2008); Gumedze and Jackson (2011), there has been limited work on influence statistics for medical diagnostic studies. Krzanowski and Hand (2009) list model diagnostics as an area of future research for the ROC-GLM model. There currently exist diagnostics that focus on evaluating systematic departures from the ROC-GLM. In particular, Cai and Zheng (2007). present a global test for the ROC-GLM, a test for the link function and a test for the interaction between the basis function and covariates. However these are not designed to address the same questions as deletion diagnostics. The deletion diagnostics proposed by Preisser and Qaqish (1996) provide a sensitivity analysis tool for parameters in a GLM for clustered data for detection of isolated departures from the GLM assumptions. Since the ROC-GLM is a specialized GLM model reformulated to address diagnostic accuracy, the GEE cluster-deletion diagnostics may be applied to identify cases that have undue influence on the model parameters describing the ROC curve. As will be described in this article, the process of creating the final ROC-GLM requires three general steps: first, a reference distribution is created using only the controls; second, the cases (or diseased) observations are standardized to the control reference distribution; and finally,the standardized case observations are used to model the ROC curve. The opportunity to apply deletion diagnostics exists in steps one and three. To our knowledge, deletion diagnostics have not been presented alongside the ROC-GLM in any previous article.

In section 2, the ROC-GLM will be reviewed, followed by a description of the cluster-deletion diagnostics applied in the ROC-GLM context. In section 3, an example will be presented using data from the DPOAE data set Norton et al. (2000). In the example, children are measured for diagnostic accuracy of hearing tests against a gold standard, in either one or both ears. Finally, in section 4, the results of the analysis will be discussed followed by conclusionary comments in section 5.

2.2 Methods

2.2.1 Overview of ROC-GLM

The following notation is presented for the paragraphs that follow. The variable for true disease status is defined as $D_i = 1$ for a diseased subject and $D_i = 0$ for a non-diseased subject. Later the notation of $\bar{D}$ for non-diseased and D for diseased subjects will be used when displaying equations for the regression models. The variable Y_i is the diagnostic test result for subject i. Let $\{Y_{\bar{D}_j}, j = 1, \dots, n_{\bar{D}}\}$ and $\{Y_{D_i}, i = 1, \dots, n_D\}$ represent the ordinal or continuous responses of controls and cases, respectively, with larger values being more indicative of disease. It is assumed that Y_{D_i} and $Y_{\bar{D}_j}$ are randomly selected from the population of test results associated with the diseased and non-diseased states Pepe (2003). Next, a vector, X, is defined which contains the covariates that affect the test result distribution in control subjects, as well as those covariates, X_D, that affect the discrimination between cases and controls. Finally, a set of T discrete points $f = f_1, \dots, f_T$ is defined, on the x-axis of the ROC curve, chosen from the interval $(0, 1)$, over which the model will be fit. The choice of f is important since this will determine the interval over which the model is to hold. The number of points T should be assigned so that standard statistical software can handle the estimation. There is currently no method designed to choose the values in the domain that give optimally efficient results Pepe (2003). Alonzo and Pepe (2002) found relatively good efficiency for small values of f_t. Pepe (2003) suggests that in practice it is possible to estimate parameters with increasing the number of points in f_t, stopping when the decreases in standard errors become small.

The ROC-GLM is a regression model that provides covariate adjustment to a diagnostic accuracy analysis ROC curve. The following 3 general steps are required to perform a covariate adjustment of ROC curves using regression techniques Janes et al. (2009):

1. Estimate $PV_{DX_i} = F_X(Y_{DX_i})$, the percentile values of the test results for cases, where F_X is the distribution of test results in controls as a function of the covariates.

2. Estimate the cdf of the percentile values as a function of the covariates.

3. Specify the adjustment of the ROC curve as a function of the covariates. We then employ GEE for binary data to estimate the model parameters (covered in Section 2.2).

First, an estimate of F_X, the distribution of test results in the control group, is required. Essentially, the process of standardizing the test results begins by finding the baseline relationship among the controls. Different assumptions may be employed at this stage, the two most common being stratification and simple linear models Pepe (2003). For example, and the method used here for demonstration, a simple linear model could be specified Janes et al. (2009) such that the test measures in control subjects follow a linear relationship:

$$Y_{\bar{D}_i} = \psi_0 + \psi_1^{'} X_i + \epsilon_i. \tag{2.1}$$

where ϵ_i are *i.i.d.* as $N(0, \sigma^2)$. We recall from above that this particular model specification is not required for the ROC-GLM to hold, rather it is one option that is possible. In any case, this first step provides the first opportunity to apply ordinary linear model deletion diagnostics (such as Cook's D for simple linear models) in the estimation steps of the ROC-GLM. Given that the model in (2.1) is crucial to the remaining steps, it is proposed that deletion diagnostics be applied at this step to assess the control distribution model. The deletion diagnostics are presented in a following section.

Having settled on a linear model in the previous step, and having assumed Gaussian errors for this linear model, then the percentile values for the cases are defined as

$$\widehat{PV}_{DX_i} = \Phi\left((Y_{DX_i} - \hat{\psi}_0 - \hat{\psi}_1' X_i)/\hat{\sigma}\right). \tag{2.2}$$

If Gaussian errors and/or a linear relationship are too restrictive for a given application, there are other alternatives proposed. For example, Heagerty and Pepe (1999) propose an empirical estimation of the error distribution using the residuals of the linear model. Further, instead of assuming a linear relationship of the test result in the controls, one could use a stratified approach Janes et al. (2009). At this stage standardized test results for the cases as a function of the controls are completed by the above step. Next, the cdf of the percentile values is estimated.

In the second step, we make use of the fact that an ROC curve is essentially the cdf of the percentile values calculated above Pepe (2003). Defining the $h(f)$ as the cdf (recall that f are the chosen set of values on the x-axis of the ROC curve), it is possible to write:

$$h_X(f) = ROC_X(f) = P(1 - PV_{DX} \leq f) = g(\beta_0 + \beta_1 g^{-1}(f)) \tag{2.3}$$

where $g(\cdot)$ gives a parametric form of the ROC curve; $g = \Phi$ is the standard normal c.d.f. and $g(\cdot) = exp(\cdot)/[1 + exp(\cdot)]$ is the logistic function giving binormal and bilogistic ROC curves respectively. The result after this second step is an ROC curve that is not yet adjusted for covariates that discriminate between the cases and controls. However, the ROC curve is now inherently adjusted for covariates that affect the test distribution results. This is quite important as Pepe (2003) demonstrates that “pooled” or unadjusted ROC curves are biased.

The final step in the model specification is to create the inputs for a regression model using the newly created percentile value cdf, and the covariates that are assumed to affect the discriminatory capacity of the test. In other words, covariates that affect the intercept and/or slope of the ROC curve. Recall that T discrete points on the x-axis of the ROC curve over which to fit the model have been defined. Also defined are $U_{it} = I_{1-PV_{DX_i} \leq f_t}, t = 1, ..., T$, the set of cumulative binary indicators which determine whether or not the percentile values are less than each choice of f. For example, if $T = 10$ values of f are chosen, then each subject would have a vector of 10 binary indicators for each percentile value. Next, we define the covariates $X_D g^{-1}(f)$ as those that will enter the model as ones that affect discrimination. The complete model combining steps 2 and 3 is:

$$ROC_{X,X_D}(f) = g(\beta_0 + \beta_1 g^{-1}(f) + \beta_2^{'} X_D + \beta_3^{'} X_D g^{-1}(f)) \tag{2.4}$$

The link function $g^{-1}(\cdot)$ is often chosen to be the Probit link in the model above. The classical ROC curve typically employs the binormal basis function which is inherently a probit function. The binormal framework as an estimation method has its roots in works by Dorfman and Alf (1968); Metz (1986); Metz et al. (1998) among others as applied mainly to radiology imaging evaluation analysis. In the binormal framework, the distributions of both case and control observations are assumed to have a Gaussian distribution. That assumption is relaxed with the semi-parametric ROC-GLM. In this case the ”binormal” assumption refers only to the form of the ROC curve through its estimation via the GEE machinery. In the ROC-GLM model, the advantage of using the probit basis function is seen when interpreting the model parameters for covariates. A positive coefficient in this model is interpreted as the covariate adding diagnostic accuracy benefit to the model with higher values of the covariate, while a negative coefficient means lower values offer diagnostic accuracy benefit to the model. Other link functions are possible assuming the binomial variance function: the log link and logit link are possible and offer slightly different interpretations to the parameters of the regression coefficients. Pepe (2003) discusses all three links with examples and suggests that the probit model be used for its intuitive interpretation qualities.

This final step of the ROC-GLM process may be thought of as defining a model that has as its output a "baseline" ROC curve (from step 2 and in equation 2.3) and some additional model parameters that specify covariate-adjustments of that baseline curve. The previous steps also allow for flexibility in defining which covariates are important for adjusting the control test results distribution and those which affect the discrimination between cases and controls.

2.2.2 GEE Estimation of the ROC-GLM

Let $t = 1, 2, ..., T$ observations from $i = 1, 2, ..., K$ clusters where U_{it} is the response measure for the t-th observation in the ith cluster and x_{it} is a p x 1 vector of covariates. The mean $\mu_{it} = E(U_{it}|x_{it})$ is related to the covariates through the linear predictor η_{it} by $\mu_{it}(\beta) = g(\eta_{it})$. The variance of the response is $\text{var}(y_{ij}) = \phi v(\mu_{it})$ where $v(\cdot)$ is the variance function and ϕ is the scale parameter; since y_{it} is binary, $v(\mu_{it}) = \mu_{it}(1 - \mu_{it})$ and $\phi = 1$. The working covariance matrix for cluster i is $V_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$ where $A_i = \text{Diag}[v(\mu_{i1}), ..., v(\mu_{iT})]$, $R_i = R_i(\alpha)$ is the working correlation matrix depending on the nuisance parameter α and assumed not to vary by cluster; independence working correlation is advocated by Pepe (2003).

The linear predictor, η_i, includes the basis function and covariates for adjustment where, following Alonzo and Pepe (2002),

$$\eta_{it} = \beta_0 + \beta_1 g^{-1}(f_t) + \beta_2' X_{D_i} + \beta_3' X_{D_i} g^{-1}(f_t) \tag{2.5}$$

where $\beta = (\beta_0, \beta_1, \beta_2', \beta_3')'$. The GEE estimates are determined by iteratively solving

$$\sum_{i=1}^{K} D_i' V_i^{-1} r_i = 0 \tag{2.6}$$

where $r_i = y_i - \mu_i$, $D_i = \partial\mu_i/\partial\beta = (\partial\mu_i/\partial\eta_i)X_i^*$ and $X_i^* = \left(X_{i1}', ..., X_{it}', ..., X_{iT}'\right)'$, where $X_{it} = \left(1, g^{-1}(f(t)), X_{D_i}', X_{D_i}' g^{-1}(f(t))\right)$ and $g(f) = (g(f_1), ..., g(f_T))$. Under the marginal mean model for the binary indicators in equation (2.5) and working independence, the matrix components in estimating equations (2.6) become

$$D_i' V_i^{-1} r_i = \sum_{t \in T} X_{it}' \frac{\partial g(\eta_{it})}{\partial \eta_{it}} v_{it}^{-1} (y_{it} - \mu_{it}(\beta)) \tag{2.7}$$

Further, the variance function for the binary indicators are $v(\mu_i) = g(\eta_i)[1 - g(\eta_i)]$. Additionally, under a working independence correlation structure (i.e., $R_i = I_{n_i}$, where I_r is an $r \times r$ identity matrix),

$$V_i^{-1} = \left(A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}\right)^{-1} = A^{-1} = \text{Diag}\{g(\eta_i)^{-1}[1 - g(\eta_i)]^{-1}\}. \tag{2.8}$$

The empirical (sandwich) estimator of the covariance matrix of $\widehat{\beta}$ is given by

$$V_{emp}(\widehat{\beta}) = M^{-1} \left(\sum_{i=1}^{K} D_i' V_i^{-1} r_i r_i' V_i^{-1} D_i \right) M^{-1} \tag{2.9}$$

where $M = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)$ and $r_i = (r_{i1}, \ldots, r_{iT})'$ with $r_{it} = (U_{it} - \mu_{it})/\sqrt{v_{it}}$. The sandwich estimator is robust to mis-specification of the working correlation matrix in the sense that it is a consistent estimator as long as the marginal mean is specified correctly. When there are a small to moderate number of subjects (clusters), say between 15 and 50, a bias-corrected covariance estimator Mancl and DeRouen (2001), defined below, has been shown to have good finite sample properties (see also Lu et al. (2007)).

A summary statistic often expected with ROC analysis is the area under the curve (AUC). In this case we present covariate-adjusted AUC following the computational formulae presented in Janes et al. (2009). The adjusted AUC, denoted AAUC is defined as the mean of the case standardized placement values:

$$\widehat{AAUC} = \sum_{i=1}^{n_D} \widehat{PV}_{DZ_i}/n_D. \tag{2.10}$$

The AAUC is one measure which can be assessed for influential subjects in the sense that changes could potentially be more interpretable than using only the regression parameter diagnostics.

At this point it is important to note that Pepe (2003) advocate using bootstrap standard errors for the estimates $\hat{\beta}$ from of the fitted model, to account for uncertainty associated with estimating equation 2.1 in stage 1 for later use in stage 3. The corresponding theoretical results are quite complicated and not practically useful for inference. In practice, bootstrap replicates are created according to the design of the study. For clustered data, begin by sampling with replacement entire clusters (in the data analysis examples this would consist of $2T$ binary indicators from both ears of the child). Then proceed by fitting the complete ROC-GLM process according to the steps outlined above. The parameter estimates from each of r replicates are collected to construct estimates of the bootstrap standard error. The reason bootstrapping is necessary is because there does not exist true independence between the first and second stage models (i.e., the model for the outcomes in step 1 and for the ROC curve in step 3), there could be bias in the standard errors. In the case of a covariate that affects both the test result distribution (first stage) and the discriminatory capacity (second stage), this covariate would essentially influence both the responses U_{it} and the covariates in X.

2.2.3 Cluster-deletion Diagnostics

It is natural to ask whether data from a single cluster (subject) has a large influence relative to other clusters on the estimates in the marginal mean model given in equation (2.5). For the h-th element of β, interest is often in $(\hat{\beta}_h - \hat{\beta}_{h[i]})$, the difference in the parameter estimate with and without the i-th cluster included in the data. In the case of the ROC-GLM this will provide us with a measure of sensitivity of the parameters given deletion of a given cluster. Specifically, in the second stage binary regression model is fit to obtain estimates of the ROC parameters and associated covariates, insight is gained into how a given subject (cluster) influences the model. This manifests itself as a change in parameter estimates. However as noted above, and given the relationship between the parameters and the AUC, there is potential for impact there as well.

Preisser and Qaqish (1996) introduced computationally quick approximations for both observation- and cluster-deletion diagnostics for GEE. However, only the latter, called case-deletion, are relevant for this application because the observation-level diagnostics have no real interpretation in the second stage portion ROC-GLM where the actual ROC model is fit. Generally, only observation deletion diagnostics are used in the first stage of the model where a reference distribution is fit for the independent control subject observations. However in the case of clustered data such as those that arise with a diagnostic test result for both ears in the DPOAE data, cluster deletion diagnostics are also useful.

Recall that the $U_{it}, t = 1, \ldots, T$, are a set of binary placement value indicators constructed for the i-th case in the course of applying the estimation method; they don't have any inherent meaning as individual data values. Following the formulae of Hammill and Preisser (2006), the influence of the i-th cluster as given by the $p \times 1$ vector $(\hat{\beta}_1 - \hat{\beta}_{1[i]}, \ldots, \hat{\beta}_p - \hat{\beta}_{p[i]})$, where "$[i]$" denotes the i-th cluster excluded, can be approximated by

$$DFBETAC_i = M^{-1} D_i' V_i^{-1} (I - H_i)^{-1} r_i. \tag{2.11}$$

Note that $DFBETAC_i$ is a measure of the influence that each cluster has on the estimate of each parameter element of β. Next, we observe that the bias-corrected variance is related to 2.11 in the following way:

$$V_{bias-corr}(\hat{\beta}) = \sum_{i=1}^{k} (DFBETAC_i)(DFBETAC_i^{'}). \tag{2.12}$$

Standardization of $DFBETAC_i$ is achieved by dividing each of its elements by the standard error of its respective parameter estimate, usually based on the full data. Finally, a measure of the influence of the i-th cluster on the overall model fit can be estimated by Cook's D:

$$DCLS_i = (DFBETAC_i)' [\mathrm{var}(\hat{\beta})]^{-1} (DFBETAC_i)/p \tag{2.13}$$

where $\mathrm{var}(\hat{\beta})$ is estimated by either the empirical (as in Ziegler et al. (1998) and Preisser et al. (2012) or bias-corrected variance estimators defined above. Vens and Ziegler (2012) propose cut-off values for $DCLS_i$, $\chi^2_p(1 - \alpha)$ when cluster sizes are equal.

The goal of the cluster-deletion diagnostics in the ROC-GLM is to provide additional model diagnostics which are noted as one area of future work Pepe (2003);

Krzanowski and Hand (2009). As is demonstrated in the analysis below, there are some interesting aspects of what it means to be influential in the ROC-GLM, and that this complicated by the fact there are really two stages to the model as outlined previously.

2.3 Analysis of the Neonatal Audiology Data set

The original publication for the example data are found in Norton et al. (2000). The sample comprises of 2742 infants and 5058 ears upon which three diagnostic screening tests (DPOAE, TEOAE and ABR) were performed. The gold standard reference test applied is an audiometric behavioral response test. The study was conducted at 6 different clinical centers. The above example data set is one that has been used extensively to demonstrate analysis methods for covariate-adjusted ROC curves. For example, Janes et al. (2009) use the data extensively to demonstrate various analysis options.

Previous analysis of this data via the ROC-GLM was performed in Janes et al. (2009). The reference distribution adjustment model included a simple linear model for the test result with age (months) and gender (1=female, 0=male) as adjustment factors (covariates). Despite there being 2316 of 2724 subjects (85%) who had both ears tested, the model did not account for multiple (and possibly correlated) measures on a given subject. The ROC-GLM covariate adjustment was fit using a GEE with probit link with subject as the cluster. The clustered data, in this case subjects and ears tested are possibly correlated. We note here that the clustered nature of the data do not provide any additional complications with respect to how the model is created. As we will see in the sections that follow, clustering simply means that vectors of placement values for a given subject will be correlated, and further when creating bootstrapped estimates of parameter standard errors, we employ sampling with replacement with subjects selected as an entire cluster (in this case T or $2T$ placement values depending on whether diagnostic test data is available for one or both ears.) The model however is flexible to handle longitudinal data were it the case here.

In the sections that follow we present the ROC-GLM and covariate adjustment process focusing on identification of influential observations and clusters within the two adjustment models within the larger ROC-GLM analysis. To demonstrate concepts we focus on only 1 of the 3 screening tests (DPOAE), with higher scores being more indicative of hearing loss. Further, to maintain continuity with previous analyses and to highlight benefit of deletion diagnostics analysis, we preserve the model covariates in both portions of the ROC-GLM of Janes et al. (2009), highlighting places where the analyst must make important decisions about covariates for adjustment.

2.3.1 Reference distribution model step of ROC-GLM

As noted in section 2.1, the first important part of the process of creating a reference distribution is to estimate a plausible model. We use the control subjects to estimate a reference distribution and then standardize the cases to that reference. Therefore this section deals with the first two steps outlined in 2.1 (whereas the next section will deal only with influence in the ROC regression model).

We assume a linear relationship between the test results for DPOAE with the covariates age and gender. Previous analyses have assumed that ears within a given subject are not correlated. For this analysis we use a GEE to account for the fact that there is correlation potentially within measurements for a given subject. It should be noted that when comparing the results of this step to those in Janes et al. (2009) there is not much difference in the significance of the regression parameters or their estimates. This is likely due to the fact that there are a large amount of clusters

(2688). However we do depart from the Janes et al. (2009) reference model slightly by presenting the GEE to demonstrate the utility of both the cluster and observation level diagnostics. We define the reference distribution model for control subject i and observation j, $j = 1, 2$ as

$$Y_{ij} = \psi_0 + \psi_1 * AGE_i + \psi_2 * GENDER_i + \epsilon_{ij} \tag{2.14}$$

where ϵ_{ij} are correlated within subject (as opposed to *i.i.d.* in the previous analysis by Janes et al. (2009)). Note that the important result from this step is to ultimately obtain percentile values by which the case subjects can be standardized against.

The results of the model are presented in Table 2.1 (left columns). Clearly age is a significant covariate to adjust for in this analysis, while gender is not significant. Figures 1 and 2 display the Cook's D statistics for both observations (ears) and clusters (subjects) as well as the standardized DFBETAs for age (observation and cluster level). Given there are a maximum of only 2 individual observations within a cluster, the cluster level diagnostics are probably enough to determine influence. Figure 2 shows that subjects 20409 and 11289 influential in terms of the overall model (cluster Cook's D) as well as the regression parameter for age. Further inspection of the data for these subjects show that generally speaking subjects that are older tender to have much lower DPOAE scores. For example, the average test result for DPOAE amongst all infants greater than 45 months is -10.2. Subject 20409 is a male subject 52 months old at the time of gold standard measurement with a baseline DPOAE score of +17.1. The gold standard determination was no hearing loss yet the screening test would be considered possibly indicative of hearing loss. Subject 11289 is a female subject with a DPOAE +37.6 with a non-diseased reference status assigned at 46.4 months. For this sample the influential subjects, when removed, result in the model parameter estimates displayed in Table 2.1 (right columns). We notice that the intercept parameter estimate shifts toward zero by 0.608. The large sample size mitigates the impact of deleting these two subjects. Incidentally, and though not explored in Janes et al. (2009), the reduced model (including all subjects or with the 2 removed) includes only age, and is highly significant. At this point, the percentile values are calculated and cases (hearing loss ears) are standardized as described in section 2.1.

2.3.2 Covariate Adjustment model step of ROC-GLM

In the 3rd step of section 2.1, the ROC regression model is fit. The details of the model specification and estimation are described in section 2.2. At this stage we now have a dataset consisting of up to 20 binary indicators (10 per ear for each ear, noting some subjects have only 1 ear measured) of the 128 case subjects in the sample. Each of the indicators describes the placement of the standardized test results against the control reference distribution. In this way, only the cluster deletion diagnostics have an interpretation. Table 2.2 presents the results of the ROC-GLM model. The model fit includes a regression adjustment for the covariate age. The regression coefficient for age, β_3, is slightly non-significant at the 0.05 significance level. Figure 3 displays the covariate-adjusted ROC curves for age, for 30, 40 and 50 month old subjects. Despite the separation of curves, we conclude as do Janes et al. (2009) that there is no statistically significant impact of age on discrimination between cases and controls using the DPOAE screening test.

The deletion diagnostics for clusters are displayed in Figure 4. The bootstrapped covariance matrix is used for calculation of the deletion diagnostics. On the Cook's D plot (Figure 2.4, upper left) subjects 30276 and 50558 is highlighted as influential on the overall model fit. For the standardized cluster DFBETA for age we notice that these two subjects are at the opposite ends of the scale. For subject 30276, the value of the standardized DFBETA for age is -0.914. The unstandardized version provides a direct interpretation on the parameter estimate for age which was estimated as -0.01. Therefore removal of this subject changes the parameter estimate for age by approximately +0.01. The 95% CI now changes to be slightly significant from insignificant.

As in the first stage model, some interesting insights are available for consideration. Figure 2.3 (dashed lines) shows how the effect of removing subject 30276 affects the final ROC curve analysis. Once removed the 3 age-adjusted ROC curves are more spread apart. Recalling that if removed, the age term becomes significant, we conclude that subject 30276 is a highly influential subject with respect to the age-adjusted ROC curves. When the data that contributed to the model are considered we notice that the subject is a female subject, 43.1 months old at time of gold standard assessment and has measurements on both ears. The results of the DPOAE test are -16.1 and -19.2 respectively for each ear. For the 128 case subjects (those with hearing loss), on average, the older a subject is, the more positive the DPOAE. Given that subject 30276 has two ears both at the lower end of the range of data for diseased subjects, and recalling that standardization to the control distribution would have this subject's measurements be more typical of a control, it is not surprising to see the model flag this subject as influential.

2.3.3 Influence on Covariate-adjusted AUC

We have explored in the previous two sections the specific subjects and potential explanations for influence. Here, influence is quantified as a single summary statistic as is often done in ROC analyses. As presented earlier, we employ the AAUC as it is an adjusted version of the regular AUC for analyses without covariates. Table 2.3 displays the AAUC for 4 data scenarios: the full data, influential control subjects removed (2), influential case subjects removed (2), and all 4 influential subjects removed. A trend towards increasing AUC is observed, and smaller standard errors of the estimates. Overall it is concluded that, as a sensitivity analysis, these potentially influential subjects do not affect the AAUC significantly.

2.4 Discussion

In this article, deletion diagnostics as presented in Preisser and Qaqish (1996) were extended to the covariate adjustment setting in ROC curve analysis. Models for the control subject test result distribution, which forms the basis for evaluating standardization of case subject results, were examined more closely than has been acknowledged in previous papers such as Janes et al. (2009). By applying the deletion diagnostics at this first stage, the sensitivity to influential observations and clusters may be assessed. For the ROC-GLM portion, where ROC curves are adjusted for covariates (in this case age) a method to assess influential subjects (clusters) on the model parameters and overall fit is also presented. In addition, use of the AAUC may provide a simpler summary statistic upon which we hypothesis testing may be performed to simplify interpretation.

Norton et al. (2000) acknowledge the fact that the gold standard may in fact be in error in some cases due to the nature of the test and the subjects (infant children). The concept of an imperfect gold standard is covered in both Pepe (2003) and Krzanowski and Hand (2009). Given the deletion diagnostics have highlighted influential subjects in both the controls and cases that have profiles of screening test measurements that would suggest the opposite disease status, the deletion diagnostics may in fact be useful as a tool to identify impact of an imperfect gold standard. Examples of the individual subjects presented in the Results sections for each portion of the model suggest and possibly identify candidates where misdiagnosis via gold standard may have occurred. This is suggested as future work, where the connection between verification bias and imperfect gold standard measures, could be quantified potentially, or at a minimum provide guidance in this area.

Future work could also evaluate the influence of individual observations and/or clusters in relation to the ratio of the number of control subjects to the number of case subjects. In the present analysis there were approximately 18x the number of control subjects as case subjects. As this ratio decreases the hypothesis of interest would be whether the impact of a highly influential control subject would have downstream effects into the regression adjustment model for the cases. Also further investigation into imperfect gold standards and merging concepts from other literature in this area would add a potential tool for assessment of gold standards via the ROC-GLM.

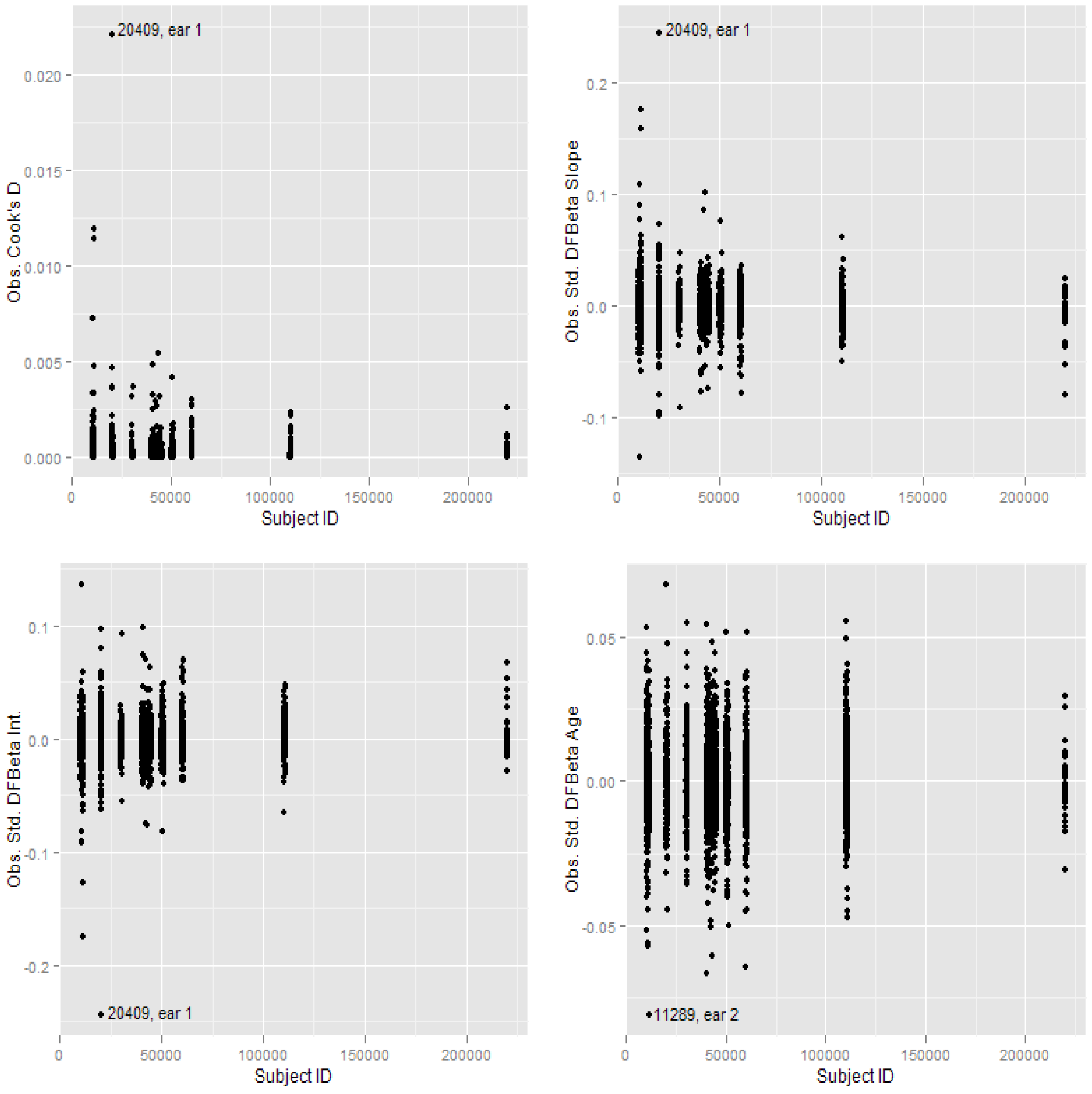

Figure 2.1: Observation deletion diagnostics for Control Reference Distribution linear model portion of ROC-GLM.

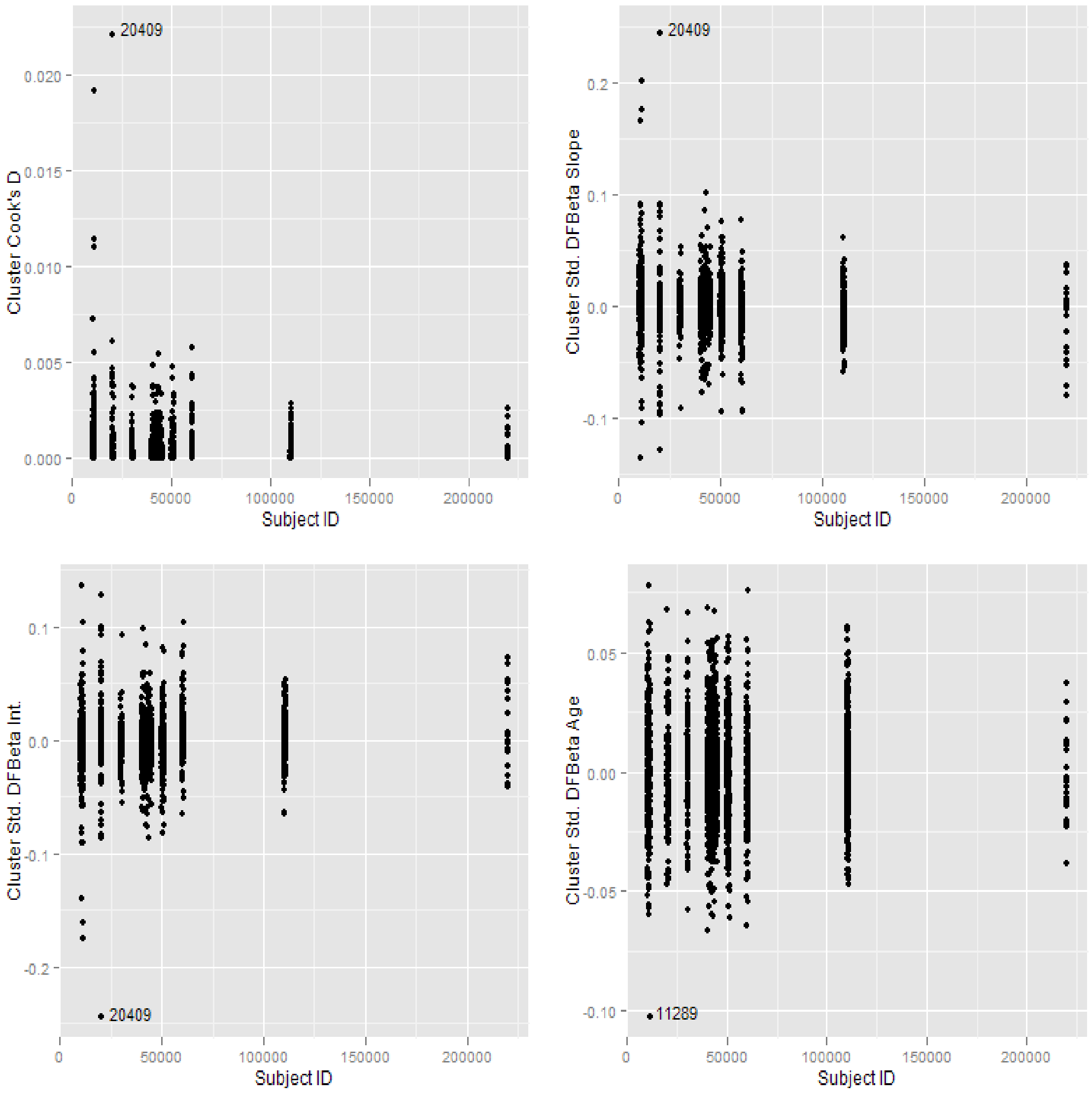

Figure 2.2: Cluster deletion diagnostics for Control Reference Distribution linear model portion of ROC-GLM. DFBETAs are standardized by use of empirical standard errors.

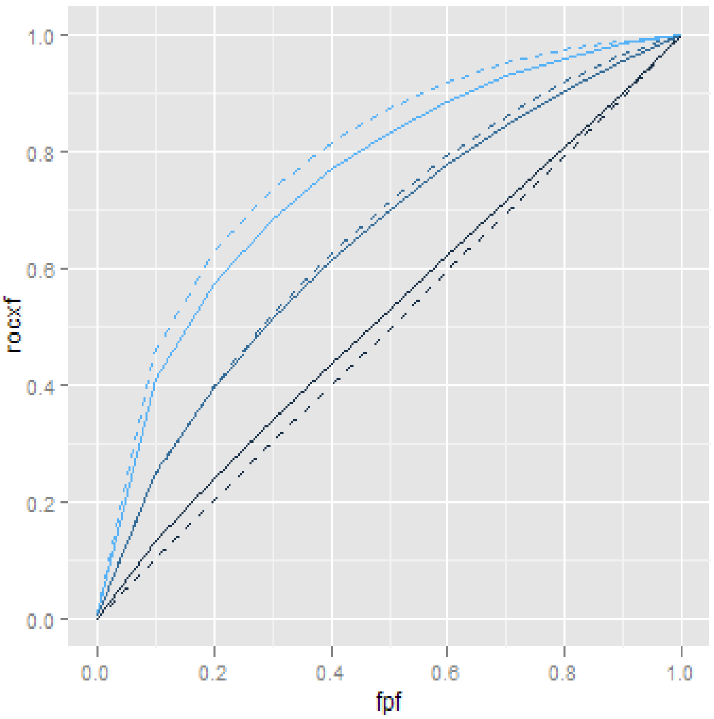

Figure 2.3: Covariate adjusted ROC curves for Age (age=50 top line, age=40 middle line, age=30 lower line) for both full model (solid curves) and model without subjects 20409 and 11289 (dashed lines)

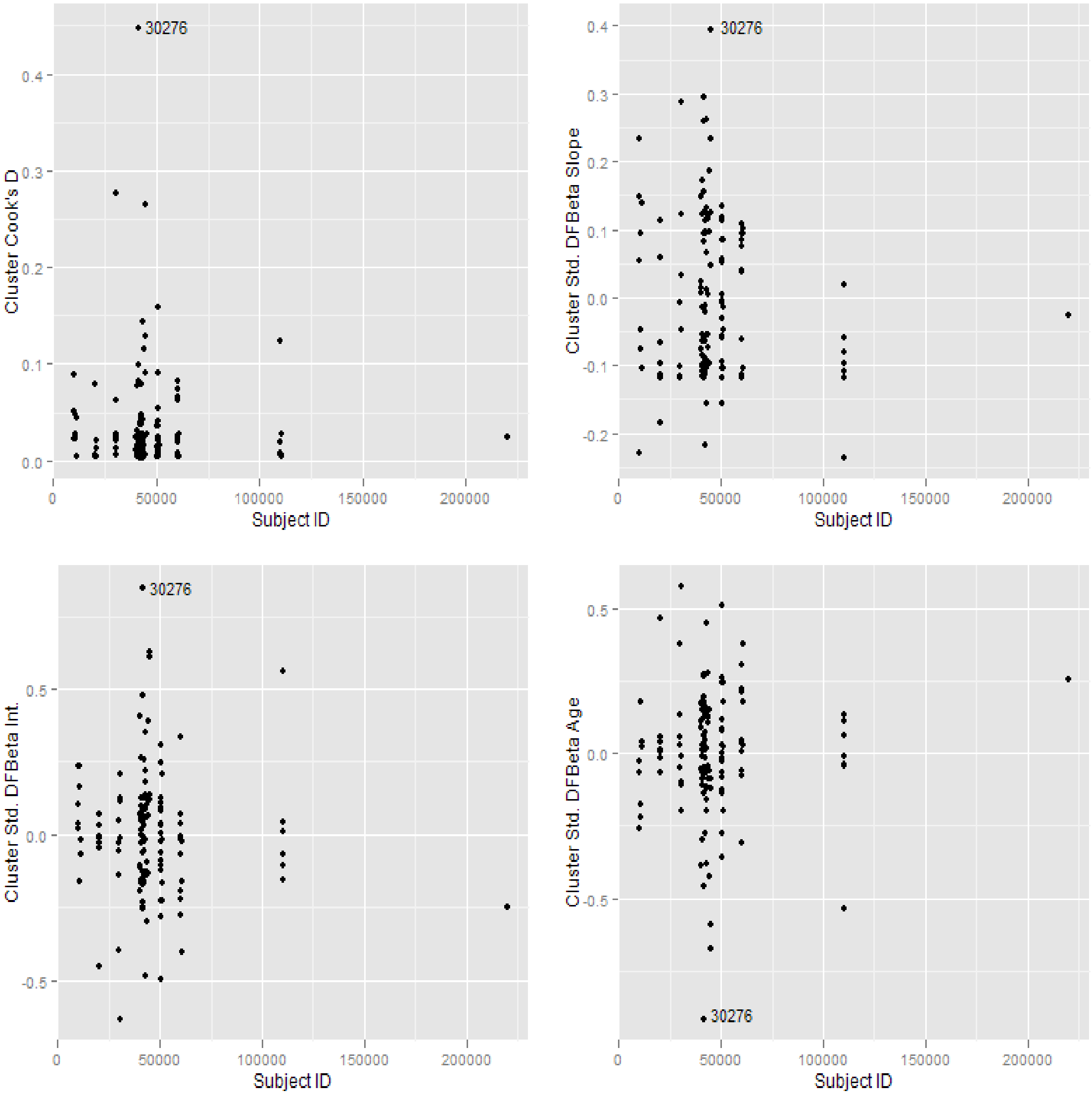

Figure 2.4: Cluster deletion diagnostics for ROC regression portion of ROC-GLM.

Table 2.1: Results of linear model for DPOAE in control subjects only, estimated with GEE to account for clustering of ears within subjects

	Full Data			20409 and 11289 removed		
Parameter	Est.	SE-emp	p-value	Est.	SE-emp	p-value
$\psi_{Int.}$	-1.676	1.551	0.280	-1.068	1.498	0.476
ψ_{Age}	-0.197	0.039	0.001	-0.214	0.038	0.001
ψ_{Gender}	0.294	0.263	0.265	0.308	0.262	0.239

Table 2.2: Results of ROC-GLM model, case subjects standardized to controls, all data

Parameter	Estimate	Bootstrap SE	Bias-corrected 95% CI
$\beta_{ROC Int.}$	-1.270	1.140	(-3.510, 0.884)
$\beta_{ROC Slope}$	-0.937	0.077	(-0.796,-1.110)
β_{Age}	0.045	0.030	(-0.010, 0.104)
Adjusted AUC	0.629	0.027	(0.573, 0.682)

Table 2.3: Change in AUC with removal of influential subjects

Scenario	AUC	Bootstrap SE	Bias-corrected 95% CI
All data	0.629	0.0267	(0.573,0.682)
Ctrl subj 20409 11289 removed	0.630	0.0247	(0.585,0.678)
Case subj 30276 50558 removed	0.642	0.0244	(0.596,0.693)
All 4 removed	0.643	0.0237	(0.592,0.688)

Chapter 3

A Semi-parametric PA Approach to Diagnostic Test Meta-Analysis

3.1 Introduction

Modern medical decision making often involves one or more diagnostic tools (such as laboratory tests and/or radiographic images). These diagnostic tools are developed using the most current technology available, and are often welcomed into medical practice with the hope of improving the care for patients. A diagnostic tool must be evaluated for it's discriminatory ability to detect presence (or absence) of current health state.

The meta-analysis of diagnostic tests is of particular interest in certain screening programs for certain diseases such as cancer. Pepe (2003) lists three benefits of meta-analysis for diagnostic tests: awareness within the research community of previous studies, explanation of discrepancies between individual study results and identification of common mistakes in study design thereby providing guidance for design of future studies.

The typical summary data points for studies chosen for a diagnostic accuracy meta-analysis are two dimensional: sensitivity and specificity. These measures tend to be negatively correlated since studies tend to vary in how test positivity is defined (Pepe, 2003).

Random effects models are intuitive in meta-analysis because the between-study heterogeneity is modeled explicitly. For diagnostic accuracy meta-analysis the bivariate random effects (BVRE) model (Reitsma et al., 2005) was the first model to propose an alternative to the computationally intensive and assumption-laden sROC model of Rutter and Gatsonis (2001). Chu and Guo (2010) also observe that previously only logit transformations were used in the bivariate random effects model. A natural extension of this is to consider other link functions such as the probit and complementary log-log. Further, the model formulation is expressed as a generalized linear mixed model (GLMM). Other key concepts regarding this approach include extensions to account for prevalence of disease (Leeflang et al., 2009; Chu and Guo, 2009; Ma et al., 2012) and sparse data models (Chu and Cole, 2006). All of the aforementioned random effects models are classified as subject-specific (SS) models.

Considering that SS models are most appropriately used if the model for a given cluster (i.e. study) is of interest, a population-average (PA) model may be appropriate in the current setting of diagnostic test meta-analysis. Although both of these approaches have their merits, the choice between them really depends on the research question being investigated. For covariates that do not vary within a cluster-like intervention condition, PA models are often recommended because of their regression parameter interpretation (Zeger et al., 1988). In the PA model, the regression parameter describes the average change in response across subsets of the population defined by the covariate. For cluster-specific models, the interpretation of the regression parameter is specific to a given cluster. In the case of meta-analysis, a random intercept model will provide median estimates of sensitivity and specificity, whereas a population-averaged model provides mean estimates. PA models have been previously recommended for diagnostic accuracy test results in the single study setting. For example, Wang et al. (2006) present a weighted least squares approach to compare predictive values of diagnostic tests; Martus et al. (2004) and Leisenring et al. (1997) present marginal regression models fit using GEE for diagnostic tests.

To our knowledge there has been no previous study of PA models for diagnostic test meta-analysis. We present a PA model that is estimated using estimating equations (GEE) procedures. In section 2, we define an overdispersed bivariate binomial model and compare it to the GLMM approach of Chu and Guo (2010). This includes a conversion of the SS parameter estimates to their PA equivalents using methodology presented in Zeger et al. (1988). Section 3 presents an example data set as well as results for both the GEE and GLMM approach for both the logit and probit links. Simulation studies are performed in Section 4 to investigate the finite sample performance of estimators of mean sensitivity and specificity based on the two approaches. Section 5 provides discussion of the analysis, as well as summary comments.

3.2 A PA Model for Diagnostic Test Meta-analysis

We present both a cluster-specific GLMM and a population-average model for the aim of estimating mean sensitivity and specificity across studies. We define the common notation as follows. Let $n_{i11}, n_{i00}, n_{i01}$ and n_{i10} represent the number of true positives, true negatives, false positives and false negatives, and n_{i1+} and n_{i0+} be the number of diseased and non-diseased subjects in the ith study from a meta-analysis, where studies are indexed as $i = 1, \ldots, K$. In the PA model, the marginal means are defined as $\mu^*_{i1} = E(n_{i11})/n_{i1+}$, which is the probability of a true positive, or sensitivity, and $\mu^*_{i0} = E(n_{i01})/n_{i0+}$, which is the probability of a false positive, or one minus specificity.

3.2.1 GLMM Model Definition

The GLMM is defined as follows. Following the notation of Chu and Guo (2010), assume n_{i01} and n_{i11} are binomially distributed as $\text{Bin}(n_{i0+}, 1 - Sp_i)$ and $\text{Bin}(n_{i1+}, Se_i)$ conditionally on Sp_i and Se_i which are the specificity and sensitivity parameters for the ith diagnostic study, respectively. Next, define

$$g(1 - Sp_i) = \beta_0 + \nu_i \tag{3.1}$$

and

$$g(Se_i) = \beta_1 + \mu_i \tag{3.2}$$

where the random effects are assumed to be distributed as $(\nu_i, \mu_i)' \sim N(0, D)$, where

$$D = \begin{pmatrix} \sigma_0^2 & \rho_m \sigma_0 \sigma_1 \\ \rho_m \sigma_0 \sigma_1 & \sigma_1^2 \end{pmatrix}$$

Estimation of the parameters $\theta_m = (\beta_0, \beta_1, \rho_m, \sigma_0, \sigma_1)'$ are based on the marginal likelihood function

$$m(\theta_m) = \prod_{i=1}^{k} \int \int p(n_{i0+}|\nu_i, \beta_0) p(n_{i1+}|\mu_i, \beta_1) f(\nu_i, \mu_i|\rho_m, \sigma_0, \sigma_1) d\nu_i d\mu_i \tag{3.3}$$

The negative log-likelihood function $f(\theta_m) = -\log\, m(\theta_m)$ (also known as the objective function) is minimized numerically in order to estimate θ_m, and the inverse Hessian matrix at the estimates provides an approximate variance-covariance matrix for the estimate of θ_m. Approximation is performed using numerical procedures such as Gaussian quadrature (SAS Institute, 2008).

The empirical (also known as “robust” or “sandwich”) standard errors are defined as

$$\mathrm{V}(\hat{\theta}_m) = [H(\hat{\theta}_m)]^{-1} \left(\sum_{i=1}^{k} g_i(\hat{\theta}_m) g_i(\hat{\theta}_m)' \right) [H(\hat{\theta}_m)]^{-1} \tag{3.4}$$

where H is the second derivative matrix of f and g_i is the first derivative of the contribution to f by the ith subject (SAS Institute, 2008). The empirical estimator of the variance-covariance matrix is robust to the mis-specification of the model under certain regularity conditions, in the sense that it consistently estimates the true variance even if the covariance structure is misspecified but is biased in small samples.

3.2.2 PA Model Definitions and Estimation Procedures

While study-level covariates could be incorporated into a PA model, we consider the mean model without covariates.

$$g(\mu^*_{ij}) = \beta^*_j,\ j = 0, 1 \tag{3.5}$$

where $g()$ is a monotone link function. For example the logit link is defined by, $g(\mu^*_{ij}) = \log[\mu^*_{ij}/1 - \mu^*_{ij}]$, and the probit link is defined by $g(\mu^*_{ij}) = \Phi^{-1}(\mu^*_{ij})$ where Φ

is the cdf of the standard normal distribution. For the model without covariates, all choices of monotone link functions give the same model. The variance is defined by

$$\mathrm{var}(n_{ij1}) = n_{ij+}h_{ij}\phi_j, \ j = 0, 1 \tag{3.6}$$

where $h_{ij} = \mu^*_{ij}(1 - \mu^*_{ij}), \ j = 0, 1$, are the variance functions, and $\phi_j \ (j = 0, 1)$ are the overdispersion parameters subject to natural restrictions $\phi_j \in (0, n_{ij+})$, for all i and j. In other words, our model constrains the overdispersion parameters to be constant across clusters. Additionally, define a common correlation, $\rho_g = \mathrm{corr}(n_{i01}, n_{i11})$.

A generalized estimating equations procedure is used to estimate the model parameters, $\theta_g = (\beta^*_0, \beta^*_1, \rho_g, \phi_0, \phi_1)^{'}$. Let $Y_i = (n_{i01}/n_{i0+}, n_{i11}/n_{i1+})$, and its expectation as $\mu^*_i = (\mu^*_{i0}, \mu^*_{i1})$. Given current estimates of ϕ_0, ϕ_1 and ρ_g, $\beta^* = (\beta^*_0, \beta^*_1)^{'}$ are estimated by solving

$$\sum_{i=1}^{K} D'_i(\beta^*)V_i^{-1}(\phi_0, \phi_1, \rho_g)(Y_i - \mu^*_i) = 0 \tag{3.7}$$

where $D'_i(\beta^*) = \partial\mu^*_i/\partial\beta^*$, $V_i(\phi_0, \phi_1, \rho_g) = A_i^{1/2}R_iA_i^{1/2}$, and
$A_i = \mathrm{Diag}\{\mathrm{var}(n_{i01})/n^2_{i0+}, \mathrm{var}(n_{i11})/n^2_{i1+}\}$. Finally, let $R_i = \rho_gJ_2 + (1 - \rho_g)I_2$, where I_r is the $r \times r$ identity matrix and $J_r = 1_r1^{'}_r$, where 1_r is a column of 1's.

The iteratively reweighted least squares GEE algorithm alternates between estimation of β^* in (3.7) and (ϕ_0, ϕ_1, ρ_g), the latter estimated by the method of moments (Liang and Zeger, 1986). The variance of $\hat{\beta^*}$ can be estimated by the model-based variance estimator

$$V_{model}(\hat{\beta^*}) = \Big(\sum_{i=1}^{K} D^{'}_iV_i^{-1}D_i\Big)^{-1} = M^{-1} \tag{3.8}$$

or the empirical variance estimator

$$V_{emp}(\hat{\beta}^*) = M^{-1}\Big(\sum_{i=1}^{K} D_i' V_i^{-1}(Y_i - \mu_i^*)(Y_i - \mu_i^*)' V_i^{-1} D_i\Big) M^{-1} \qquad (3.9)$$

For the three method of moment estimators (ϕ_0, ϕ_1, ρ_g) we use a 2-stage bootstrap for standard error estimates. These are typically not provided by software however in the context of comparing the PA and SS methods it provides some level ground for comparison. In the first stage of the bootstrap procedure we sample with replacement clusters equal to the original meta-analysis size. In the second stage we perform a stratified re-sampling procedure to produce randomly selected subjects within each of the studies. In this way we maintain both the original sample size of studies in the meta-analysis and the individual size of each study, all while creating 1000 bootstrap replicates. Finally the GEE estimation procedures are performed and the standard deviation of each of the bootstrap replicates of ϕ_0, ϕ_1 and ρ_g, respectively, constitutes bootstrap estimates of their standard errors.

3.2.3 Relationship of PA and SS model parameters

The comparison of the subject-specific and population-average methods require that their respective parameters be placed in a common context, that of the marginal model parameters. Under Gaussian random effects in the GLMM, the PA model in (3.5) and (3.6) with common marginal correlation ρ_g can be deduced (Zeger et al., 1988). In other words, θ_m has simple relationships (exact or approximate) to θ_g.

Computation of marginal mean parameters from SS model

The marginal mean simplifies or is easily approximated for the standard link functions. For the probit link the marginal parameters are obtained from the SS model via the following equation.

$$\beta_j^* = \frac{\beta_j}{\sqrt{\sigma_j^2 + 1}} \; j = 0, 1 \tag{3.10}$$

For the logit link, where the relationship is approximate, we have

$$\beta_j^* \approx \frac{\beta_j}{\sqrt{c^2\sigma_j^2 + 1}} \; j = 0, 1 \tag{3.11}$$

where $c = 16\sqrt{3}/(15\pi)$ (Zeger et al., 1988). The mean sensitivity and mean specificity may be estimated by 1-$g^{-1}(\hat{\beta}_0^*)$ and $g^{-1}(\hat{\beta}_1^*)$, respectively, where $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are calculated by plugging in estimates for the elements of θ_m into equations (3.10) and (3.11). Variance estimates, and therefore asymptotic confidence intervals, for them can be constructed via application of the delta method.

Computation of marginal covariance parameters from SS model

Unfortunately, no simple formulae exist for cov(Y_i), except for the identity link. However, an approximation via equation (3.4) of Zeger et al. (1988) is

$$\text{var}\Big(\frac{n_{i01}}{n_{i0+}}, \frac{n_{i11}}{n_{i1+}}\Big) \approx \begin{pmatrix} L_{i0}^2\sigma_0^2 & L_{i0}L_{i1}\sigma_{01} \\ L_{i0}L_{i1}\sigma_{01} & L_{i1}^2\sigma_1^2 \end{pmatrix} + \begin{pmatrix} \frac{\mu_{i0}^*(1-\mu_{i0}^*)}{n_{i0+}} & 0 \\ 0 & \frac{\mu_{i1}^*(1-\mu_{i1}^*)}{n_{i1+}} \end{pmatrix}$$

where $L_{ij} = \text{Diag}\{\partial g^{-1}(\beta_j)/\partial\beta_j\}$, $j = 0, 1$, $\sigma_{01} = \rho_m\sigma_0\sigma_1$, and $\mu_{ij}^* = g^{-1}(\beta_j^*)$ where β_j^* is given in equation (3.10) or (3.11). Equivalently, noting that for the model without covariates $\mu_j^* = \mu_{ij}^*$ and $L_j = L_{ij}$,

$$\text{var}(n_{ij1}/n_{ij+}) \approx L_j^2\sigma_j^2 + \frac{\mu_j^*(1-\mu_j^*)}{n_{ij+}}, \; j = 0, 1. \tag{3.12}$$

and

$$\mathrm{cov}\Big(\frac{n_{i01}}{n_{i0+}}, \frac{n_{i11}}{n_{i1+}}\Big) \approx L_0 L_1 \sigma_0 \sigma_1 \rho_m. \tag{3.13}$$

Under the probit link, $L_j = \phi(\beta_j)\phi(-\beta_j)$, where ϕ is the pdf standard normal function and $\mu_j^* = \Phi(\beta_j^*)$, $j = 0, 1$. Under the logit link, $L_j = \mathrm{logit}^{-1}(\beta_j) * \mathrm{logit}^{-1}(-\beta_j)$ and $\mu_j^* = \mathrm{logit}^{-1}(\beta_j^*)$, $j = 0, 1$.

Equation (3.12) implies $\mathrm{var}(n_{ij1}) \approx n_{ij+}h_{ij}\phi_{ij},\ j = 0, 1$, where

$$\phi_{ij} = \frac{n_{ij+}L_j^2\sigma_j^2}{\mu_j^*(1-\mu_j^*)} + 1,\ j = 0, 1.$$

Under an equal number of diseased $(n_{1+} = n_{i1+})$ and non-diseased $(n_{0+} = n_{i0+})$ cases across studies, the GLMM model above gives approximately constant variance and overdispersion,

$$\phi_j = \frac{n_{j+}L_j^2\sigma_j^2}{\mu_j^*(1-\mu_j^*)} + 1,\ j = 0, 1. \tag{3.14}$$

From equations Equation (3.12) and Equation (3.13), the GLMM implies that the marginal correlation for the i-th study is

$$\rho_{ig} = \frac{L_0 L_1 \sigma_0 \sigma_1 \rho_m}{\sqrt{[L_0^2\sigma_0^2 + \frac{\mu_0^*(1-\mu_0^*)}{n_{i0+}}][L_1^2\sigma_1^2 + \frac{\mu_1^*(1-\mu_1^*)}{n_{i1+}}]}} \tag{3.15}$$

Under an equal number of diseased $(n_{1+} = n_{i1+})$ and non-diseased $(n_{0+} = n_{i0+})$ cases across studies, the GLMM model above gives approximately constant correlation parameter,

$$\rho_g = \frac{L_0 L_1 \sigma_0 \sigma_1 \rho_m}{\sqrt{[L_0^2\sigma_0^2 + \frac{\mu_0^*(1-\mu_0^*)}{n_{0+}}][L_1^2\sigma_1^2 + \frac{\mu_1^*(1-\mu_1^*)}{n_{1+}}]}} \tag{3.16}$$

Thus, under the special case of equal number of diseased and non-diseased cases across studies, there is an approximate one-to-one transformation of the GLMM parameters θ_m to the marginal parameters θ_g. Thus, one way to obtain inference on the marginal model parameters is to fit a GLMM, transform $\hat{\theta}_m$ to $\hat{\theta}_g$, via a compound matrix function, i.e., $\hat{\theta}_g = F(\hat{\theta}_m)$, and use the multivariate delta method to obtain the asymptotic covariance matrix of $\hat{\theta}_g$ from the asymptotic covariance matrix of $\hat{\theta}_m$.

3.3 Data Examples and Analysis

Two data sets are presented for analysis to demonstrate the concepts presented in Section 2.

3.3.1 Example 1: Catheter Segment Culture Data

The first example data set is a meta-analysis of 33 diagnostic accuracy studies previously analyzed in Chu and Guo (2010). The 33 studies studied semi-quantitative (19 studies) or quantitative (14 studies) catheter segment culture for the diagnosis of intravascular device-related blood stream infection. The mean number (std. dev.) of diseased and non-diseased persons per study was 20 (19.8) and 237 (240.5) respectively. Chu and Guo (2010) report that since there is no statistically significant difference between the semi-quantitative and quantitative methods, the data are combined together without including this potential covariate in any model. The gold standard was final diagnosis of blood-stream infection. The data are presented in the Appendix, Table S1.

A PA model was fit to these bivariate binomial data with a cluster representing the bivariate binomials pairs. The generalized estimating equation in this case provides the estimation procedure. The logit link and an exchangeable working correlation were employed for these data. The probit link is then explored in a comparative fashion.

Two different perspectives were taken with regards to the scale parameter. As outlined in Section 2 is possible to have the scale parameter vary for sensitivity and specificity separately. Models for both constant scale within a cluster (Appendix Table S3) and varying scale within cluster (Table 3.1) are presented.

With respect to within-study variation, the scale parameter in the PA model reflects strong over-dispersion in the given sample of data. The ability to allow this to vary within each cluster seems to be important given the very different scale parameter results in the Table 3.1. The data are very skewed in terms of the study-level bivariate binomial proportions.

Table 3.1 displays the results of the PA model fit with a logit and probit link to the catheter segment culture data. Estimates of the parameters on the logit and probit scale as well as their standard errors are displayed along with estimates of the mean sensitivity and specificity (data scale). Note that the standard errors for the PA model parameters ϕ_0, ϕ_1 and ρ_g are calculated using a stratified bootstrap resampling procedure. Finally, a confidence region for this summary is displayed in Figure 2.1.

The use of GEE does not allow for a prediction region like that of GLMM, however a confidence region for the summary point is available (not available in GLMM) and displayed in Figure 3.1 (upper panels) for the GEE fit corresponding to Table 3.1 (logit and probit links with heterogeneous scale).

An elliptical confidence region was calculated using the method of Douglas (1993). This approximation for the confidence region ellipse has been used in other diagnostic accuracy meta-analysis studies, for example Chu and Guo (2010).

As we have fit a PA model to these data we have provided an estimate for what we may expect from the average study, assuming that our sample accurately reflects a good sample of similar studies from the population. This is fundamentally a different perspective than the SS model which would then seek to provide prediction for a future study, given study-specific estimates from a GLMM. Table 3.2 displays the results of the GLMM as fit by for example Chu and Guo (2010).

The population-average estimates are not directly available from the GLMM, however using the formulae provided in section 2.3 we can provide estimates. Table 3.1 displays the estimates of the GLMM parameters converted to their marginal counterparts, and compare these with the PA estimates for both the logit and probit links; given the unbalanced data $\bar{n}_{j+}, j = 0, 1$ was substituted into equations (3.14) and (3.16) for n_{j+}. We immediately notice that the estimates do not match up exactly, however they are closer for the probit link which is expected as the relationship is not an approximation as it is for the logit link. The overdispersion and extreme skew in the data are the cause for this, which is investigated further in the simulation study.

3.3.2 Example 2: Simulated Correlated Binomial Data

The second example is a single simulated data set, using similar methodology to that will be used for the simulation study. The study was simulated to have fixed cluster sizes of (25,175) with mean 1-specificity = 0.7 and mean sensitivity = 0.75, with an assumed correlation between the binomial proportions = 0.36. Bivariate overdispersed binomial data will be randomly generated using an algorithm for generation of correlated binary data based on the method of Emrich and Piedmonte (1991). The results of a single set of the simulated data are presented in the Appendix, Table S2.

For a single cluster for the above assumptions, as well as in each of the scenarios described in the simulation study in section 4, we generate correlated binary variates $Y = (Y_1^{(0)}, \ldots, Y_{n_0}^{(0)}, Y_{n_0+1}^{(1)}, \ldots, Y_n^{(1)})'$, distinguishing (with superscripts) the n observations as belonging to diseased and non-diseased groups with n_1 and $n_0 = n - n_1$ observations, respectively. Indexing observations with k, we define a model with means $\mathrm{E}(Y_k^{(0)}) = P(Y_k^{(0)} = 1) = \mu_0$ and $\mathrm{E}(Y_k^{(1)}) = P(Y_k^{(1)} = 1) = \mu_2$, and correlation structure with $\mathrm{corr}(Y_k^{(0)}, Y_{k'}^{(0)}) = \alpha_0$, $\mathrm{corr}(Y_k^{(1)}, Y_{k'}^{(1)}) = \alpha_1$, and $\mathrm{corr}(Y_k^{(0)}, Y_{k'}^{(1)}) = \alpha_2$, for $k \neq k'$. This model generates bivariate, overdispersed binomials taking $T_0 = Y_1^{(0)} + \ldots + Y_{n_0}^{(0)}$ and $T_1 = Y_{n_1+1}^{(1)} + \ldots + Y_n^{(1)}$. For a diagnostic testing study, T_0 and T_1 are the number of positive test results in groups without and with disease, respectively. It follows that $\mathrm{E}(T_0) = n_0\mu_0$, $\mathrm{E}(T_1) = n_1\mu_1$, $\mathrm{var}(T_0) = n_0\mu_0(1-\mu_0)\phi_0$, $\mathrm{var}(T_1) = n_1\mu_1(1-\mu_1)\phi_1$, and $\rho = \mathrm{corr}(T_0, T_1) = \alpha_2\sqrt{(n_0 n_1)/(\phi_0\phi_1)}$. Hall (2001) discusses that to generate (T_0, T_1) with correlation ρ, binomial parameters μ_0 and μ_1, and overdispersion ϕ_0 and ϕ_1 such that $\phi_d \in (0, n_d)$ (the natural range), respectively, take $\alpha_d = (\phi_d - 1)/(n_d - 1)$ for $d = 1, 2$ and $\alpha_2 = \rho\sqrt{\phi_0\phi_1/(n_0 n_1)}$.

The analysis of the simulated data proceeds in a similar fashion as Example 1. Tables 3.3 and 3.4 present analogous results as those found in Tables 3.1 and 3.2 above.

3.4 Simulation Study

Given the impact of skewed data and unbalanced clusters, the primary goal of the simulation study is to provide some insight as to how to perform a meta-analysis with data that is similar to example 1 where interest is in estimating the population averaged sensitivity and specificity. The simulation study will compare the estimation performance of the direct PA model method with estimation by GEE to the GLMM-converted PA parameter estimates for both the logit and probit link functions. Bivariate overdispersed binomial data will be randomly generated using an algorithm for generation of correlated binary data based on the method of Emrich and Piedmonte (1991), as described above. For each of sensitivity and specificity, bias, monte carlo standard errors, average of standard error estimates, and coverage of 95% confidence intervals will be evaluated.

3.4.1 Simulation design

The simulation study presented below has two main objectives: first, to demonstrate the properties of the PA and SS model estimators under fixed and varying cluster sizes; and, second, to investigate the relationship of the true correlated binomial proportions under different scenarios at values above 0.7 (which are typically observed in these types of meta-analysis, and tend to lead to skewed profiles of proportions in actual meta-analyses). In general, our hypotheses are that: first, percent relative bias and confidence interval coverage will be slightly better in the PA model than the marginal SS model equivalents, especially with smaller cluster sizes; and, second, that performance of scenarios with true mean sensitivity and specificity closer to 1 perform less well.

The design of the simulation study is as follows, where each item is a design factor, defining a total of 24 unique scenarios (each replicated 1000 times):

1. Cluster (study) sizes $K = 15$, $K = 25$ and $K = 50$

2. Cluster size: fixed across clusters or varies across clusters

   (a) Fixed option for (n_0, n_1): (175,25)

   (b) Varying within study: sample (n_0, n_1) from $MVN\left((\mu_{n_0}, \mu_{n_1})', \Sigma\right)$

      i. $\mu_{n0} = 175, \mu_{n1} = 25, \sigma_{n0} = 20, \sigma_{n1} = 5, \rho = 0.2$

3. Mean sensitivity and 1-specificity, generated as correlated binomial proportions

   (a) $\mu_0 = 0.7$ or 0.9

   (b) $\mu_1 = 0.75$ or 0.95

(c) The correlations are defined as $\alpha_0 = \alpha_1 = 0.05, \alpha_2 = 0.025$ which is equivalent to $\rho = \alpha_2 \sqrt{\frac{n_0 n_1}{[\alpha_0(n_0-1)+1][\alpha_1(n_1-1)+1]}}$ which equals 0.36 when $(n_0, n_1) = (175, 25)$.

For each scenario, the data set generated was analyzed using both the PA and SS models, as described in section 3.2. Both the logit and probit links were employed leading to a total of 4 analyses for each of the 24 simulation scenarios described above. Measures of performance include model convergence, percent relative bias and percent coverage of 95% confidence intervals. All simulations were run entirely using SAS software (SAS/IML for data generation, SAS/SQL for simulation results summary, SAS/IML macro Diag104.sas for PA models and SAS PROC NLMIXED for SS models).

3.4.2 Simulation Study Results

Convergence rates exceeded 99.8 in the PA model scenarios and 96.5 in the SS model scenarios (see Table S5 in the Appendix). Tables S4 and 3.5, summarize the findings of the simulation study with respect to percent relative bias and confidence interval coverage of simulated parameters.

For model convergence, we note that all PA models converged. While the SS models had very high convergence rates as well, a few trends are noticeable. For the 8 scenarios that were generated using mean 1-specificity and sensitivity $(0.9, 0.95)$, none converged 100% of the time. For cluster sizes of $K = 50$ we notice convergence was generally slightly better than for $K = 25$ and $K = 15$. Further, within a given cluster size we notice slightly better convergence of fixed cluster size over varying cluster size scenarios. This observation may suggest that for meta-analyses where the diagnostic accuracy is very high, the PA model may provide more stable convergence properties.

Percent relative bias (PRB) is generally small across the board (less than 1.6%

across the board). However there are some trends to be noted. First, we expect the PRB to be similar for the PA models between logit and probit since with no covariates the models are identical. The most noticeable trend is the difference in PRB for fixed and varying cluster sizes, most notably for the highest values of sensitivity (0.95). Even comparing within fixed or varying cluster sizes we notice that the groups of scenarios in the smaller cluster sizes ($K = 15$ and $K = 25$) have slightly higher PRB than those with larger cluster sizes ($K = 50$). Finally, within a given scenario we notice that generally speaking the PA estimated parameters from most scenarios have slightly smaller PRB than the SS model marginal estimates, though in general the magnitude of the PRB is same (most noticeably for higher sensitivity). From this we conclude that while not drastically lower the PA model does have good PRB qualities and even more so as we approach larger numbers of clusters with more uniform cluster sizes.

The summaries for coverage of 95% confidence intervals are all based on the empirical standard errors which are robust to model mis-specification. We would expect coverage to be slightly worse for K=15, then increasing for $K = 25$ and $K = 50$ and this is slightly evident. Coverage trends a bit lower for unequal cluster sizes, and within the unequal cluster sizes slightly lower coverage at the extremes (0.9,0.95) that drops just slightly below 90%.

In summary, the PA model performs well under the scenarios generated, noting that generally the PA and SS models trend together within a scenario. The scenarios with mean values $(0.9, 0.95)$ provide some insight into how models may have slightly lower convergence rates, slightly higher PRB and lower CI coverage. This observation is noteworthy as many meta-analyses of this type have a number of individual studies that fit this profile. The question of whether this simulation study takes into account within-cluster sample sizes that are very small (as is often observed in meta-analyses in practice) is addressed indirectly. For the varying cluster sample size scenarios denominators for 1-specificity that varied around $n_0 = 25$ generated some small sample sizes upon closer inspection. However given these small sample sizes are mixed in with larger ones due to random generation, perhaps their effect is not as clear.

3.5 Discussion

The binomial regression framework of the GLMM is the most prominent in the literature due to its ease of interpretation and flexibility. Also given that it is a random effects model, it has intuitive appeal because random effects approaches explicitly model heterogeneity between studies that is inherent in a meta-analysis. The inclusion of covariates and abundance of available software has allowed analysts easy access to this method. A limitation of the GLMM approach is that it does not easily provide population-averaged inference for sensitivity and specificity. In contrast, the PA approach proposed in this article directly estimates the average sensitivity and specificity. It can easily be extended to accommodate covariates, though that was not demonstrated in this article.

The PA method presented is very easy to implement with standard software for the assumption of a constant scale parameter. To account for heterogeneous scale, which is viewed as an essential part of the proposed approach, the Diag104.sas macro was used; this beta-test program will become the next version of the SAS macro for GEE originally introduced by Hammill and Preisser (2006). We know of no other software currently available to handle this. Otherwise GEE procedures in SAS such as PROC GENMOD may be used, though they cannot handle heterogeneous scale parameter estimation.

The arguments were based on the ideal setting of balanced clusters sizes with constant numbers of diseased and non-diseased subjects, respectively. The method's appropriateness for unbalanced clusters was investigated in the simulation study and revealed that the the PA model performs well in comparison to the SS model, in some cases slightly better.

The estimation of average sensitivity and specificity in the semi-parametric PA model approach does not lead directly to a summary ROC curve. The literature in this field seems to require a summary ROC curve however it is arguable whether this is useful for the practicing researcher. Future work in this area could begin by inverting the conversion equations used for the SS parameters to PA analogs we can then enter PA estimates and output SS converted estimates which would then allow for estimation of an ROC curve.

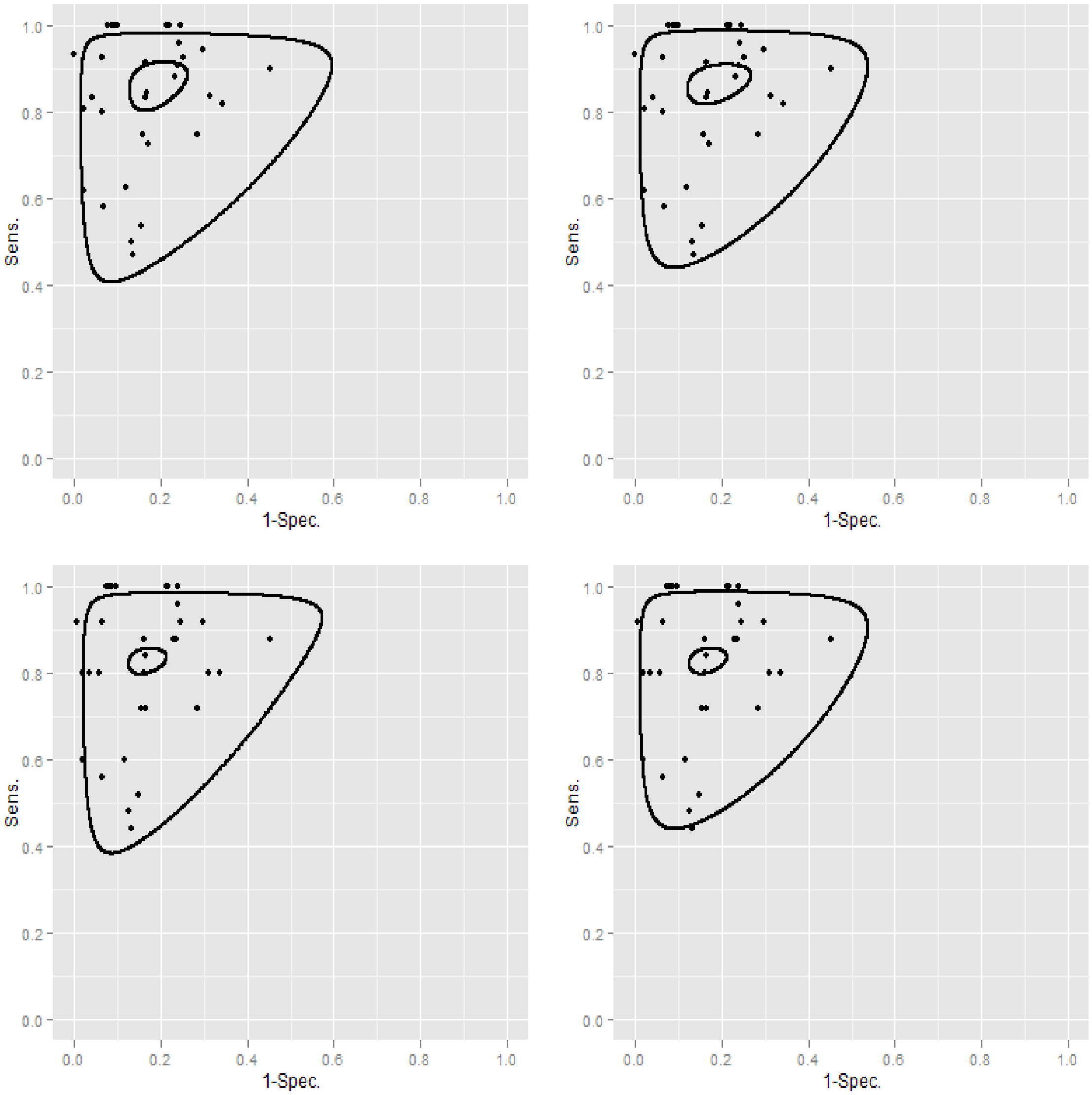

Figure 3.1: GLMM prediction region and PA model mean estimate with 95% confidence region. Upper left: Chu et al. 2010 data, logit link; Upper right: Chu et al. 2010 data, probit link; Lower left: Expanded and balanced data, logit link; Lower right: Expanded and balanced data, probit link. The confidence region is the smaller area contained within the prediction region.

Table 3.1: Comparison of PA Logit and Probit models with heterogeneous scale for clusters for the catheter segment culture data , as well as the GLMM-converted analogs. All standard errors reported are empirical except for PA parameters ϕ_0, ϕ_1 and ρ_g which are estimated via a 2-stage bootstrap approach.

Parameter	Logit	Logit GLMM conv.	Probit	Probit GLMM conv.
β_0^*	-1.901 (0.203)	-1.677 (0.127)	-1.127 (0.090)	-0.994 (0.072)
β_1^*	1.480 (0.183)	1.626 (0.184)	0.895 (0.113)	0.961 (0.104)
ϕ_0	20.35 (7.249)	18.33 (7.625)	20.35 (7.249)	19.74 (8.136)
ϕ_1	2.794 (1.005)	2.759 (0.719)	2.794 (1.005)	2.75 (0.746)
ρ_g	0.330 (0.200)	0.161 (0.150)	0.330 (0.200)	0.155 (0.151)
Mean Specificity	0.870 (0.023)	0.842 (0.017)	0.870(0.019)	0.840 (0.018)
Mean Sensitivity	0.815 (0.028)	0.836 (0.025)	0.815(0.030)	0.832 (0.026)

Table 3.2: Results of GLMM model for Logit and Probit links for the catheter segment culture data (from Chu et. al (2010))

Parameter	Logit	Probit
β_0	-1.909 (0.169)	-1.104 (0.088)
β_1	1.829 (0.222)	1.069 (0.120)
σ_0	0.925 (0.132)	0.483 (0.068)
σ_1	0.876 (0.213)	0.489 (0.117)
ρ_m	0.208 (0.190)	0.197 (0.192)
Median Specificity	0.871 (0.019)	0.864 (0.019)
Median Sensitivity	0.862 (0.026)	0.857 (0.027)

Table 3.3: Simulated Data: Comparison of Logit and Probit link functions for the PA method heterogeneous scale for clusters. All standard errors reported are empirical except for PA parameters ϕ_0, ϕ_1 and ρ_g which are estimated via a 2-stage bootstrap approach.

Parameter	Logit	Logit GLMM conv.	Probit	Probit GLMM conv.
β_0^*	-0.851 (0.131)	-0.864 (0.133)	-0.527 (0.079)	-0.527 (0.079)
β_1^*	0.934 (0.083)	0.950 (0.085)	0.580 (0.050)	0.580 (0.050)
ϕ_0	6.638 (2.287)	7.300 (2.504)	6.638 (2.287)	7.307 (2.514)
ϕ_1	2.457 (0.561)	2.358 (0.614)	2.457 (0.561)	2.315 (0.595)
ρ_g	0.212 (0.234)	0.160 (0.183)	0.212 (0.234)	0.160 (0.181)
Mean Specificity	0.703 (0.028)	0.703 (0.028)	0.703 (0.028)	0.703 (0.028)
Mean Sensitivity	0.721 (0.017)	0.721 (0.017)	0.721 (0.017)	0.721 (0.017)

Table 3.4: Simulated Data: Results of GLMM model for Logit and Probit links

Parameter	Logit	Probit
β_0	-0.904 (0.142)	-0.553 (0.089)
β_1	0.979 (0.094)	0.599 (0.055)
σ_0	0.527 (0.090)	0.315 (0.070)
σ_1	0.423 (0.143)	0.249 (0.046)
ρ_m	0.200 (0.259)	0.200 (0.258)
Median Specificity	0.704 (0.029)	0.705 (0.029)
Median Sensitivity	0.730 (0.021)	0.730 (0.021)

Table 3.5: Percent coverage of nominal 95% confidence intervals based upon empirical standard errors after fitting PA model as well as SS-model marginal converted results

				Logit				Probit			
Clus.				β_0^*		β_1^*		β_0^*		β_1^*	
Size	Type	(n_0, n_1)	(μ_0, μ_1)	PA	SS	PA	SS	PA	SS	PA	SS
K=15	Fix.	(175,25)	(0.7,0.75)	93	92	92	92	93	93	91	91
			(0.7,0.95)	91	90	91	93	90	87	91	92
			(0.9,0.75)	93	91	91	91	93	92	91	90
			(0.9,0.95)	91	93	93	93	91	90	93	93
	Uneq.	MVN(1)	(0.7,0.75)	88	90	92	91	88	89	92	91
			(0.7,0.95)	92	92	93	92	92	92	93	92
			(0.9,0.75)	89	90	90	89	89	89	90	89
			(0.9,0.95)	92	91	90	87	92	91	90	86
K=25	Fix.	(175,25)	(0.7,0.75)	94	94	92	93	94	93	93	92
			(0.7,0.95)	92	93	95	94	92	91	95	95
			(0.9,0.75)	94	92	91	90	94	94	91	90
			(0.9,0.95)	92	93	92	94	91	90	93	93
	Uneq.	MVN(1)	(0.7,0.75)	89	91	93	92	89	90	93	93
			(0.7,0.95)	93	94	94	94	93	94	94	94
			(0.9,0.75)	93	93	93	90	93	94	93	90
			(0.9,0.95)	92	93	91	86	92	91	90	87
K=50	Fix.	(175,25)	(0.7,0.75)	94	92	93	94	97	93	94	97
			(0.7,0.95)	94	94	94	93	93	93	92	93
			(0.9,0.75)	95	93	93	92	94	92	93	93
			(0.9,0.95)	94	93	94	95	94	93	94	93
	Uneq.	MVN(1)	(0.7,0.75)	83	88	95	95	84	85	95	95
			(0.7,0.95)	94	94	95	95	94	94	95	95
			(0.9,0.75)	92	93	94	92	92	92	94	93
			(0.9,0.95)	95	95	91	86	95	94	90	86

Chapter 4

Implementation guide for PA diagnostic accuracy meta-analysis

4.1 Introduction to Diagnostic Test Accuracy

Modern medical decision making often involves one or more diagnostic tools (such as laboratory tests and/or radiographic images). These diagnostic tools are developed using the most current technology available, and are often welcomed into medical practice with the hope of improving the care for patients. A diagnostic tool must be evaluated for it's discriminatory ability to detect presence (or absence) of current health state.

The meta-analysis of diagnostic tests is of particular interest in certain screening programs for certain diseases such as cancer. Pepe (2003) lists three benefits of meta-analysis for diagnostic tests: awareness within the research community of previous studies, explanation of discrepancies between individual study results and identification of common mistakes in study design thereby providing guidance for design of future studies.

The typical summary data points for studies chosen for a diagnostic accuracy meta-analysis are two dimensional: sensitivity and specificity. These measures tend to be negatively correlated since studies tend to vary in how test positivity is defined (Pepe, 2003). We may think of these type of data as correlated binomial outcomes within a cluster (study), with population-averaged (PA) models a possible choice for estimation given their ability to handle correlated outcomes (via generalized estimating equations (GEE) for example). In the PA model, the regression parameter describes the average change in response across subsets of the population defined by the covariate. For cluster-specific models, the interpretation of the regression parameter is specific to a given cluster. In the case of meta-analysis, a random intercept model will provide median estimates of sensitivity and specificity, whereas a population-averaged model provides mean estimates. PA models have been previously recommended for diagnostic accuracy test results in the single study setting. For example, Wang et al. (2006) present a weighted least squares approach to compare predictive values of diagnostic tests; Martus et al. (2004) and Leisenring et al. (1997) present marginal regression models fit using GEE for diagnostic tests.

In this paper we describe in detail the application of a PA model for diagnostic accuracy meta-analysis with covariates including investigation of influential clusters (studies) and observations within each cluster. To our knowledge there is no such analyst guide of this type. In section 2 we briefly describe the PA model definition and estimation procedures; section 3 describes the analysis macro %PAMETA; section 4 presents two data sets along with macro inputs and results for each respectively; finally, in section 5 we offer conclusions and commentary for future research directions.

4.2 PA Model Definitions and Estimation Procedures

Let $n_{i11}, n_{i00}, n_{i01}$ and n_{i10} represent the number of true positives, true negatives, false positives and false negatives, and n_{i1+} and n_{i0+} be the number of diseased and non-diseased subjects in the ith study from a meta-analysis, where studies are indexed as $i = 1, \ldots, K$. In the PA model, the marginal means are defined as $\mu_{i1} = E(n_{i11})/n_{i1+}$, which is the probability of a true positive, or sensitivity, and $\mu_{i0} = E(n_{i01})/n_{i0+}$, which is the probability of a false positive, or one minus specificity. We consider the mean model where study-level covariates could be incorporated if desired.

$$g(\mu_{ij}) = \beta_{0j} + x_{ij}^{'}\beta_j, \ j = 0, 1 \tag{4.1}$$

where $g()$ is a monotone link function. For example the logit link is defined by, $g(\mu_{ij}) = \log[\mu_{ij}/1 - \mu_{ij}]$, and the probit link is defined by $g(\mu_{ij}) = \Phi^{-1}(\mu_{ij})$ where Φ is the cdf of the standard normal distribution. For the model without covariates, all choices of monotone link functions give the same model. The variance is defined by

$$\mathrm{var}(n_{ij1}) = n_{ij+}h_{ij}\phi_j, \ j = 0, 1 \tag{4.2}$$

where $h_{ij} = \mu_{ij}(1 - \mu_{ij}), \ j = 0, 1$, are the variance functions, and ϕ_j $(j = 0, 1)$ are the overdispersion parameters subject to natural restrictions $\phi_j \in (0, n_{ij+})$, for all i and j. In other words, our model constrains the overdispersion parameters to be constant across clusters. Additionally, define a common correlation, $\rho_g = \mathrm{corr}(n_{i01}, n_{i11})$.

A generalized estimating equations procedure is used to estimate the model parameters, $\theta_g = (\beta_0, \beta_1, \rho_g, \phi_0, \phi_1)^{'}$. Let $Y_i = (n_{i01}/n_{i0+}, n_{i11}/n_{i1+})$, and its expectation as $\mu_i = (\mu_{i0}, \mu_{i1})$. Given current estimates of ϕ_0, ϕ_1 and ρ_g, $\beta = (\beta_{00}, \beta_{01}, \beta_0, \beta_1)^{'}$ are estimated by solving

$$\sum_{i=1}^{K} D_i'(\beta) V_i^{-1}(\phi_0, \phi_1, \rho_g)(Y_i - \mu_i) = 0 \tag{4.3}$$

where $D_i'(\beta) = \partial \mu_i / \partial \beta$, $V_i(\phi_0, \phi_1, \rho_g) = A_i^{1/2} R_i A_i^{1/2}$, and $A_i = \mathrm{Diag}\{\mathrm{var}(n_{i01})/n_{i0+}^2, \mathrm{var}(n_{i11})/n_{i1+}^2\}$. Finally, let $R_i = \rho_g J_2 + (1 - \rho_g) I_2$, where I_r is the $r \times r$ identity matrix and $J_r = 1_r 1_r^{'}$, where 1_r is a column of 1's.

The iteratively reweighted least squares GEE algorithm alternates between estimation of β in (4.3) and (ϕ_0, ϕ_1, ρ_g), the latter estimated by the method of moments (Liang and Zeger, 1986). The variance of $\hat{\beta}$ can be estimated by the model-based variance estimator

$$V_{model}(\hat{\beta}) = \Big(\sum_{i=1}^{K} D_i^{'} V_i^{-1} D_i \Big)^{-1} = M^{-1} \tag{4.4}$$

the empirical variance estimator

$$V_{emp}(\hat{\beta}) = M^{-1} \Big(\sum_{i=1}^{K} D_i^{'} V_i^{-1} r_i r_i^{'} V_i^{-1} D_i \Big) M^{-1} \tag{4.5}$$

or the bias-corrected variance estimator (Mancl and DeRouen, 2001)

$$V_{bias-corr}(\hat{\beta}) = M^{-1} \Big(\sum_{i=1}^{K} D_i^{'} V_i^{-1} (I - H_i)^{-1} r_i r_i^{'} (I - H_i)^{-1} V_i^{-1} D_i \Big) M^{-1} \tag{4.6}$$

where $r_i = (Y_i - \mu_i)$ and $H_i = D_i M^{-1} D_i^{'} V_i^{-1}$ is the cluster leverage matrix.

It is natural to ask whether data from a single cluster (study) has a large influence relative to other clusters on the estimates in the marginal mean model given in equation (4.1). For the h-th element of β, interest is often in $(\hat{\beta}_h - \hat{\beta}_{h[i]})$, the difference in the parameter estimate with and without the i-th cluster included in the data. Preisser and Qaqish (1996) introduced computationally quick approximations for both observation- and cluster-deletion diagnostics for GEE.

Following the formulae of Hammill and Preisser (2006), the influence of the i-th cluster as given by the $p \times 1$ vector $(\hat{\beta}_1 - \hat{\beta}_{1[i]}, \ldots, \hat{\beta}_p - \hat{\beta}_{p[i]})$, where "[$i$]" denotes the i-th cluster excluded, can be approximated by

$$DFBETAC_i = M^{-1} D_i' V_i^{-1} (I - H_i)^{-1} r_i. \tag{4.7}$$

Note that $DFBETAC_i$ is a measure of the influence that each cluster has on the estimate of each parameter element of β. We observe that (4.6) can be written as

$$V_{bias-corr}(\hat{\beta}) = \sum_{i=1}^{k} (DFBETAC_i)(DFBETAC_i'). \tag{4.8}$$

Standardization of $DFBETAC_i$ is achieved by dividing each of its elements by the standard error of its respective parameter estimate, usually based on the full data. Finally, a measure of the influence of the i-th cluster on the overall model fit can be estimated by Cook's D:

$$DCLS_i = (DFBETAC_i)'[\mathrm{var}(\hat{\beta})]^{-1}(DFBETAC_i)/p \tag{4.9}$$

where $\mathrm{var}(\hat{\beta})$ is estimated by either the empirical (as in Ziegler et al. (1998); Preisser et al. (2012) or bias-corrected variance estimators defined above. Vens and Ziegler (2012) propose cut-off values for $DCLS_i$, $\chi^2_p(1 - \alpha)$ when cluster sizes are equal.

4.3 %PAMETA Macro overview and Implementation

4.3.1 Macro inputs and description

The following SAS/ SAS-IML macro performs a population-averaged model for diagnostic accuracy studies. The macro centers around application of a second SAS/IML macro used for fitting PA models using GEE machinery (Diag104.sas, %GEE). %GEE was developed separately by the authors for use more generally when GEE methods are required. The complete documentation of %GEE is referred to separately in Appendix 1. We cannot use the SAS procedure PROC GENMOD to fit a GEE in this current context because it does not allow for heterogeneous estimates of the scale parameter for ϕ (i.e. a separate measure of overdispersion for the scale parameters for each of sensitivity and specificity respectively, which is automatically implemented here). The same applies for any software package that cannot handle heterogeneous overdispersion parameter estimation.

The following lists the options for the macro PAMETA (population-averaged meta-analysis) along with some brief explanation and detail around each option:

```
\%macro PAMETA (
    filenet= example: X:\\Mydirectory\,
         input the preferred file directory where the source
         data resides, the macro Diag104.sas resides and where
         all output written out from SAS will be stored.
         WARNING: do not forget the last backslash!

    sourcedat= example:mydata.sas,the name of the SAS data file.
        The format of the file must be as listed
        in the below example for one study. Two records per study
        are required with a study and observation level identifier.
        Additional columns contain covariables specifying the design
        matrix of choice (not shown).
        WARNING: do not forget the .sas extension.

                     STUDYID    Y     N  SESP
```

```
                   1       13    25   1
                   1      123   150   2

descstats= example:1,
    if =1 requests a basic descriptive study as part of output

yvar= example:Y,
    the numerator of the binomial proportion for
    sensitivity and specificity

nvar = example:N,
    the denominator of the binomial proportion for
    sensitivity and specificity

covars=example:ONESP SE,
    Here we must specify a design matrix through identification
    of covariates. Examples are given in the next section.
    An intercept is not automatically included.

studyid=example:ID,
    a unique identifier for study within the meta-analysis.

obsid=example:SESP,
    a unique identifier for observation
    (sensitivity, specificity) within study within
    the meta-analysis.

link=example:3,
    a numeric value for choice of link function,
    typically 3 (logit) or 5 (probit).

corr=example:1,
    a numeric value for choice of correlation structure
    typically 1 (independence) 3 (exchangeable)

outobsdiag=example:obsdiag,
    an output data set name for the observation level
    deletion diagnostics

outclusdiag=example:clusdiag,
    an output data set name for the cluster level
    deletion diagnostics
outcsv = example: 1,
    if =1 exports all outxxx datasets requested above
    in .csv format
```

4.3.2 Output overview

The macro returns the model output from the GEE routine as well as data scale conversions (i.e. on the scale of sensitivity and specificity). These results are displayed in the output window (or .lst file). Additionally one may utilize the output datasets described above (in both .sas and .csv formats for convenience).
The following is a brief summary of the macro output:

1. Estimates of model parameters (on both link scale and data scale), including 3 types of standard errors (output dataset: paramest)

2. Cluster (study) and observation-level deletion diagnostics for detection of potential influential studies and specific observations within those studies. The user may then use any graphics software to produce visuals. (output datasets: outobsdiag and outclusdiag)

3. The required data elements to plot a confidence region (explicit implementation is provided in Appendix 2)

4.4 Illustrative Data Sets and Analysis

4.4.1 Data set 1: Blood stream catheter infection data

Data and Descriptive Study

The first example data set is a meta-analysis of 33 diagnostic accuracy studies previously analyzed in Chu and Guo (2010). The 33 studies studied semi-quantitative

(19 studies) or quantitative (14 studies) catheter segment culture for the diagnosis of intravascular device-related blood stream infection. Chu and Guo (2010) report that since there is no statistically significant difference between the semi-quantitative and quantitative methods, the data are combined together without including this potential covariate in any model. For demonstration purposes we investigate the covariate for type of catheter segment culture method (semi-quantitative or quantitative). The mean number (std. dev.) of diseased and non-diseased persons per study was 20 (19.8) and 237 (240.5) respectively. The gold standard was final diagnosis of blood-stream infection. The data are presented in Table 4.1.

Model and macro inputs

We defined three models of interest *a priori*: a full model, reduced model and no covariates model. The full model is as follows

$$g(\mu_{ij}) = \beta_{0j} + x_{ij}\beta_j,\ j = 0, 1 \tag{4.10}$$

where $x_{i0} = I(j = 0)x_i$, $x_{i1} = I(j = 1)x_i$, and $x_i = 1$ if study type is semi-quantitative and $x_i = -1$ if it is quantitative. The input data set has the following structure:

```
StudyID  Y   N   SESP  TYPE ONESP  SE  ONESPTYP1  SETYP1
   1     36  85    1     1    1     0     1          0
   1      1  18    2     1    0     1     0          1
   21     4  60    1     2    1     0    -1          0
   21     2   8    2     2    0     1     0         -1
```

We interpret the full model parameters as follows: β_{00} is the “average” 1-specificity on the link scale (logit in this case); $\beta_{00} + \beta_0$ is the 1-specificity estimate for the semi-quantitative method on the link scale; and, $\beta_{00} - \beta_0$ is the 1-specificity estimate for the quantitative method on the link scale. Similarly for β_{01} and β_1 we obtain estimates for sensitivity (overall and by type). For the reduced model we assume a common study type effect for 1-specificity and sensitivity where $\beta_0 = \beta_1 = \beta$. Finally, all estimates are then available on the data scale by applying the inverse link (inverse logit in the current example).

As an example of the macro syntax inputs we assume a logit link model and include the covariates ONESPTYP1 for x_{i0} and SETYP1 for x_{i1} (for type of catheter segment culture analysis method). For the reduced model we include only x_i instead of x_{i0} and x_{i1}. For the analysis without covariates we would simply remove variables with the TYP1 suffix. The user must define which design matrix coding scheme (reference cell, effect etc.) is deemed most appropriate for the particular meta-analysis.

```
\%PAMETA(filenet = D:\AnalysisPaper3\ChuData\,
sourcedat = chu2010
yvar = ny,
nvar = nn,
covars = ONESP SE ONESPTYP1 SETYP1,
studyid = ID,
obsid = SESP,
link = 3,
corr = 4,
outobsdiag = chuobsdiag,
outclusdiag = chuclusdiag,
outcsv=0);
```

Results

For the first set of analyses, all data from the catheter segment culture data set were analyzed (see Table 4.1). Three models were fit: first, a model with separate Type effects for 1-specificity and sensitivity; second, a reduced model where a common Type effect is assumed; and, finally, a model with no covariates (for comparison purposes). In the macro input in the previous section only the “covars” option was changed at each stage (i.e. all models were fit with logit link and working correlation as exchangeable). The results of model estimation are presented in Table 4.2 and provides a view into the effects of the covariate Type on the parameter estimates for 1-specificity and sensitivity.

Considering the first column of Table 4.2, which is the model where we define separate effects of Type on each of sensitivity and specificity, we observe estimates of specificity are slightly different and the sensitivity estimates are essentially the same across types. A single df Wald Test was performed on the two regression parameters for Type to test whether model reduction to a common effect is appropriate. Define $H_0 : C\beta = 0$, where $C = [0\,0\,1\; -1]$ and $\beta = (\beta_{00}, \beta_{01}, \beta_0, \beta_1)^T$. The observed Wald statistic is 0.51 with associated p-value=0.777. We conclude that a common effect of Type may be appropriate in this case. The middle column of Table 4.2 displays this reduced model, where the regression parameter estimate reported is for the common effect of Type across sensitivity and specificity. The z-score (based upon the bias-corrected standard error) is 0.674 with an associated p-value of 0.714. As was noted in the pre-amble to the Chu and Guo (2010) paper, the effect of type of catheter segment culture quantification is not significant, and the two groups of studies may be included together for analysis. Finally, the right hand column displays the results of a model with no covariate effect for Type.

An additional observation of interest pertains to the reduced model of Table 2. Here we notice that the parameter estimate for the common covariate effect of Type is actually quite close to that of the full model which is not generally expected. In practice we expect the common effect parameter estimate to be more close to the mid-point between the two. However, there is an explanation of this: when a binomial outcome has a larger denominator it has a smaller variance. When this variance is inverted (which is essentially done when one inverts $A_i^{-1/2}$ in equation (3), and which is precisely done when one assumes R_i is the independent working correlation matrix), the resultant weight is larger than the weight of a binomial outcome with a larger variance. This is the reason that the estimate of a common effect for Type in the reduced model is closer to the estimate of TYP1 for ONESPTYP1 from the full model than the estimate of Type for SETYP1.

Next deletion diagnostics are considered to assess whether certain studies or observations within studies (i.e., values of 1-specificity and sensitivity within a cluster) are influential in the model. Using the model without covariates, we output the cluster and observation level deletion diagnostics (see Figure 4.1). There are two studies that seem to stand out as potential influential clusters: 18 and 20. Upon closer evaluation we observe that the change in ONESP (lower left panel), with studies 18 and 20 deleted is substantial. The Cook's D statistic for the cluster (upper left panel) and the observation (i.e. sensitivity or 1-specificity within a given study; upper right panel) both highlight this finding as well as the DFBETAS. Thus, one working hypothesis could be that re-fitting the model without studies 18 and 20 could lead to different results, and perhaps some insight into the slightly elevated SEs for the parameter estimates. When the source data for studies 18 and 20 are viewed in Table 4.1 we notice that these two studies are actually the same study that use the two different types of catheter segment culture quantification. We also notice that these two entries have very large values for TN. The individual estimates for specificity for these two entries is 0.98, which are the two highest values observed in this meta-analysis. While this may have been an obvious finding even without the deletion diagnostics, the influence is quantified for the analyst. In cases where the data are not obvious outliers, the deletion diagnostics provide insight as to the influence of the cluster and/or observation.

With studies 18 and 20 identified as influential, we re-run the 3 models with these two studies removed. Table 4.3 displays the results, and we immediately notice the impact of studies 18 and 20. In general the removal of these two influential studies causes an expected increase in estimates of mean sensitivity, along with an associated decrease in mean specificity. Also of note is the decrease in the overdispersion estimates ϕ_0 for all 3 models, and a noticeable decrease in ρ_g. Given that we removed one study of each Type it is not surprising to observe that the 1 df Wald test still shows no significance for the separate effect of Type of sensitivity and specificity (p=0.928), and the reduced model for the common effect type is also not significant (p=0.763). For the model with no covariates we observe that comparing the mean estimates and their standard errors between Tables 4.2 and 4.3 that the impact of studies 18 and 20 is a 3.9% decrease in mean specificity and a 4.0% increase in mean sensitivity. This additional analysis serves to provide potentially important information about the impact of influential studies on the meta-analysis final recommendations, especially with regards to any clinical practice impact.

One other interesting observation pertains to the estimates of overdispersion in Tables 4.2 and 4.3. For correlated binomial data, such as these, there are natural boundary restrictions placed on the overdispersion parameter estimates: $\phi_0 \leq min(n_{i0+})$ and $\phi_1 \leq min(n_{i1+})$. In the current example $min(n_{i0+}) = 16$ (Study 28) and $min(n_{i1+}) = 4$ (Study 5). For the three models presented in Table 4.2 the estimates of ϕ_0 are all above 20 which violates the condition while for ϕ_1 all estimates are less than 4. When viewing the impact of studies 18 and 20 in Table 4.2, in addition to effects on the estimates of sensitivity and specificity we also notice an approximate halving of the estimates of ϕ_0 such that the boundary condition is satisfied. This finding seems to confirm that highly influential points such as Studies 18 and 20 contribute to substantial increase in variation, which the the overdispersion parameters are highlighting.

4.4.2 Data set 2: Lymph node metastases data

Data and Descriptive Study

The second example data set is a meta-analysis of 32 diagnostic accuracy studies previously analyzed in Klerkx et al. (2010). The diagnostic accuracy of gadolinium-enhanced MRI in detecting lymph node metastases using histopathologic test as the reference gold standard. The mean number (std. dev.) of diseased and non-diseased persons per study was 15(18.5) and 28 (30.4) respectively. Covariates for partial verification bias (PVB, 8 studies) and study design (case control, 6 studies or cohort, 26 studies) are available. The data are presented in Table 4.4.

Model and macro inputs

We defined three models of interest *a priori*: a full model, reduced model and no covariates model. The full model is as follows

$$g(\mu_{ij}) = \beta_{0j} + x_{ij}\beta_j,\ j = 0,1 \tag{4.11}$$

where $x_{i0} = I(j = 0)x_i$, $x_{i1} = I(j = 1)x_i$; and, $x_i = 1$ if PVB present and $x_i = 0$ if PVB absent. This example employs reference cell coding of the design matrix since the main interest is to estimate sensitivity and specificity when PVB is not present. The data structure is as follows:

```
StudyID  Y   N   SESP  PVB ONESP  SE  ONESPPVB  SEPVB
   1     3  12    1     1    1     0     1         0
   1     7  10    2     1    0     1     0         1
   2     1  41    1     0    1     0     0         0
   2     4   9    2     0    0     1     0         0
```

We interpret the full model parameters as follows: $\beta_{00} + \beta_0$ is the 1-specificity estimate for PVB=1 (partial verification bias present) on the link scale; and, β_{00} is the 1-specificity estimate for PVB=0 (partial verification bias not present) on the link scale. Similarly for β_{01} and β_1 we obtain estimates for sensitivity. For the reduced model we assume a common PVB effect for both 1-specificity and sensitivity where $\beta_0 = \beta_1 = \beta$. Finally, all estimates are then available on the data scale by applying the inverse link (inverse logit in the current example).

For the present data analysis we assume a logit link PA model and a covariate for whether verification bias was known to be present (PVB). For the analysis without the covariate we simply remove PVB. The following macro call was used to analyze the lymph node metastases data (Klerkx et al., 2010) using the logit link. We demonstrate the macro call that includes separate effects of PVB for each of sensitivity and specificity.

```
\%PAMETA(filenet = D:\Analysis\Paper3\KlerkxData\,
      sourcedat = klerkx2010,
yvar = ny,
nvar = nn,
covars = ONESP SE ONESPPVB SEPVB,
     studyid = ID,
obsid = SESP,
link = 3,
corr = 4,
outobsdiag = klerkxobsdiag,
outclusdiag = klerkxclusdiag,
outcsv=1);
```

Results

The results for the full dataset are presented in Table 4.5. The model with separate effects of PVB for each of sensitivity and specificity is displayed in the left column. Although we notice quite different estimates of mean specificity and sensitivity, this may be tempered by the fact there are only 8 studies with PVB=1. The single df Wald test is not significant for the separate effects of PVB (p=0.80). Upon reducing the model to PVB having a common effect we observe the adjusted means (for PVB present) for specificity equal to 0.882 (0.034) and mean sensitivity equal to 0.601 (0.079). For PVB not present specificity is 0.786 (0.022) and sensitivity equals 0.755 (0.025). The test for significance of the PVB effect based on the z-score (bias-corrected SE) is -2.37 with $p = 0.018$. This result is significant at the 5% level, and does caution us to interpret the parameter estimates of sensitivity and specificity separately for PVB present or absent. For completeness we also present the model without covariates, where the results for mean specificity and sensitivity, 0.802 (0.056) and 0.731 (0.033), respectively.

The deletion diagnostics are displayed in Figure 4.2. Study 28 stands out as a study that is influential. The 3 models were re-fit without study 28 and the results given in Table 4.6. The effects of removal of this study are noticeable immediately in estimates of ϕ_1 and ρ_g. When considering the source data for study we notice that the estimate for specificity of this study is 0.409 and sensitivity 0.850. We would expect an increase in mean specificity estimates, especially in the no covariate model. Comparing the results across Tables we do in fact notice a slight increase in specificity with removal of the influential study 28. Further we also notice a slight reduction in mean sensitivity for the models with removal of study 28's contribution (0.850).

Verification bias in meta-analysis is an important and potentially serious cause for concern (Ransohoff and Feinstein, 1978). Ma et al. (2010) investigate the impact of partial verification bias on the Klerkx et al. (2010) data using a hybrid Bayesian approach including a trivariate random effects model which includes prevalence estimates. In the current analysis we take a slightly different approach by accounting for presence of PVB as a covariate that potentially affects mean specificity and sensitivity. Concurrently, the analysis assumes correlation and overdispersion parameters are common across studies regardless of PVB status. Given that many meta-analyses will have small samples sizes (10-25), it may be argued that studies, which might otherwise be identified for removal for minor design flaws, still be included in the model in the way described in order to increase information for estimation. The concept of "serious" design flaw excluding the study from the meta-analysis versus "minor" would be an assumption that would be built into the literature search and filtering process.

4.5 Conclusions

The PA approach to diagnostic accuracy meta-analysis provides mean estimates of specificity and sensitivity, with or without adjustment for covariates. Further, influential observations may be evaluated using the available deletion diagnostics. This type of exploratory analysis may provide the analyst with key insights into the make-up of the meta-analysis study sample. We provide some simple analysis for two data examples along with implementation and interpretation guidelines.

The SAS software provided allows for easy implementation of model estimation as well as options for output datasets to create graphics. To our knowledge no other software allows for heterogeneous overdispersion parameter estimation for the analysis of correlated binomial data with GEE making this set of software a valuable resource for those wishing to undertake a PA analysis approach to their meta-analysis.

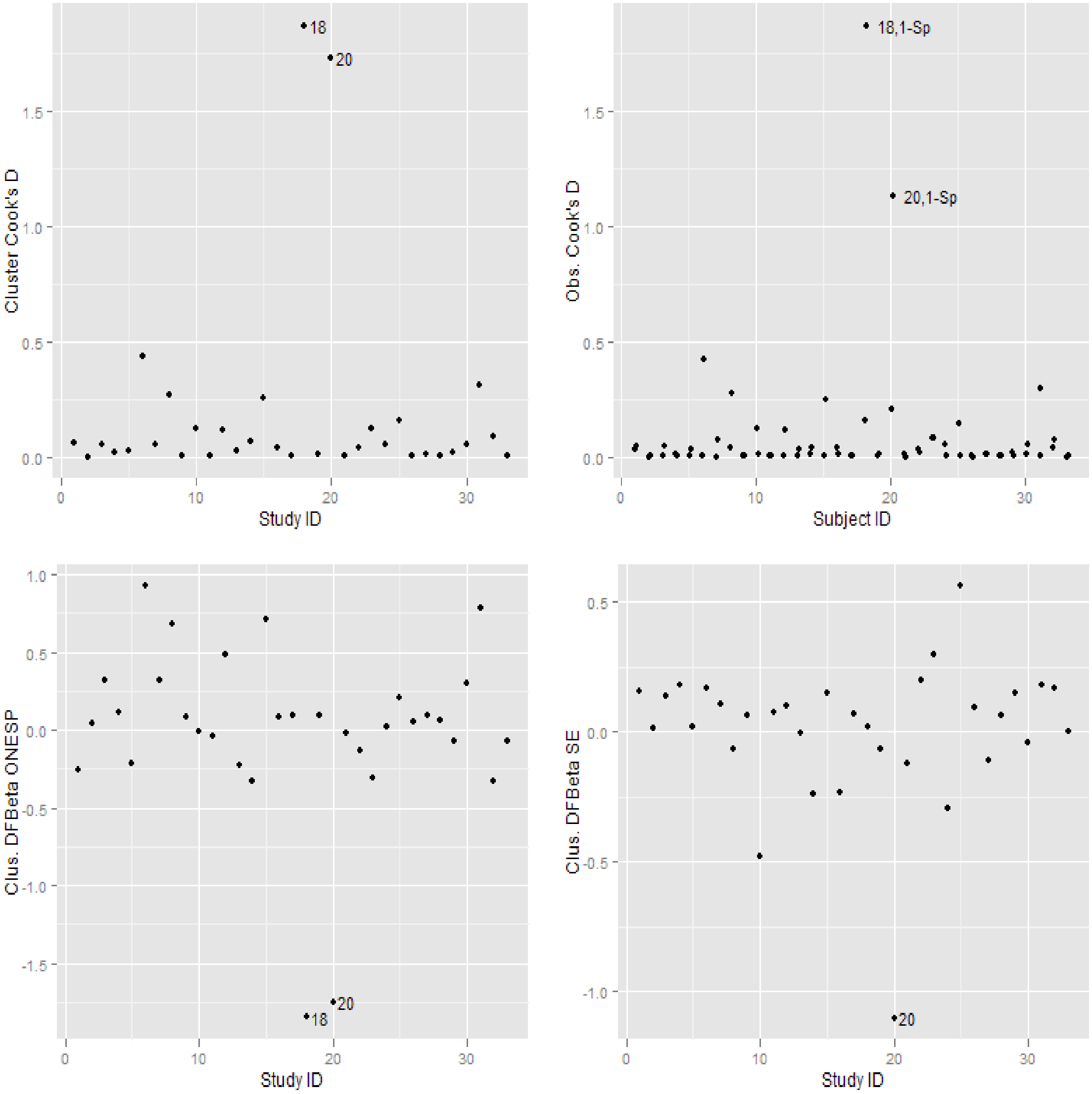

Figure 4.1: Deletion diagnostics for catheter segment culture data. Upper left panel: Cluster (Study) level Cook's D; upper right: Observation (sensitivity and 1- specificity within a cluster) Cook's D; lower left: DFBETAS for 1-specificity; and, lower right: DFBETAS for sensitivity.

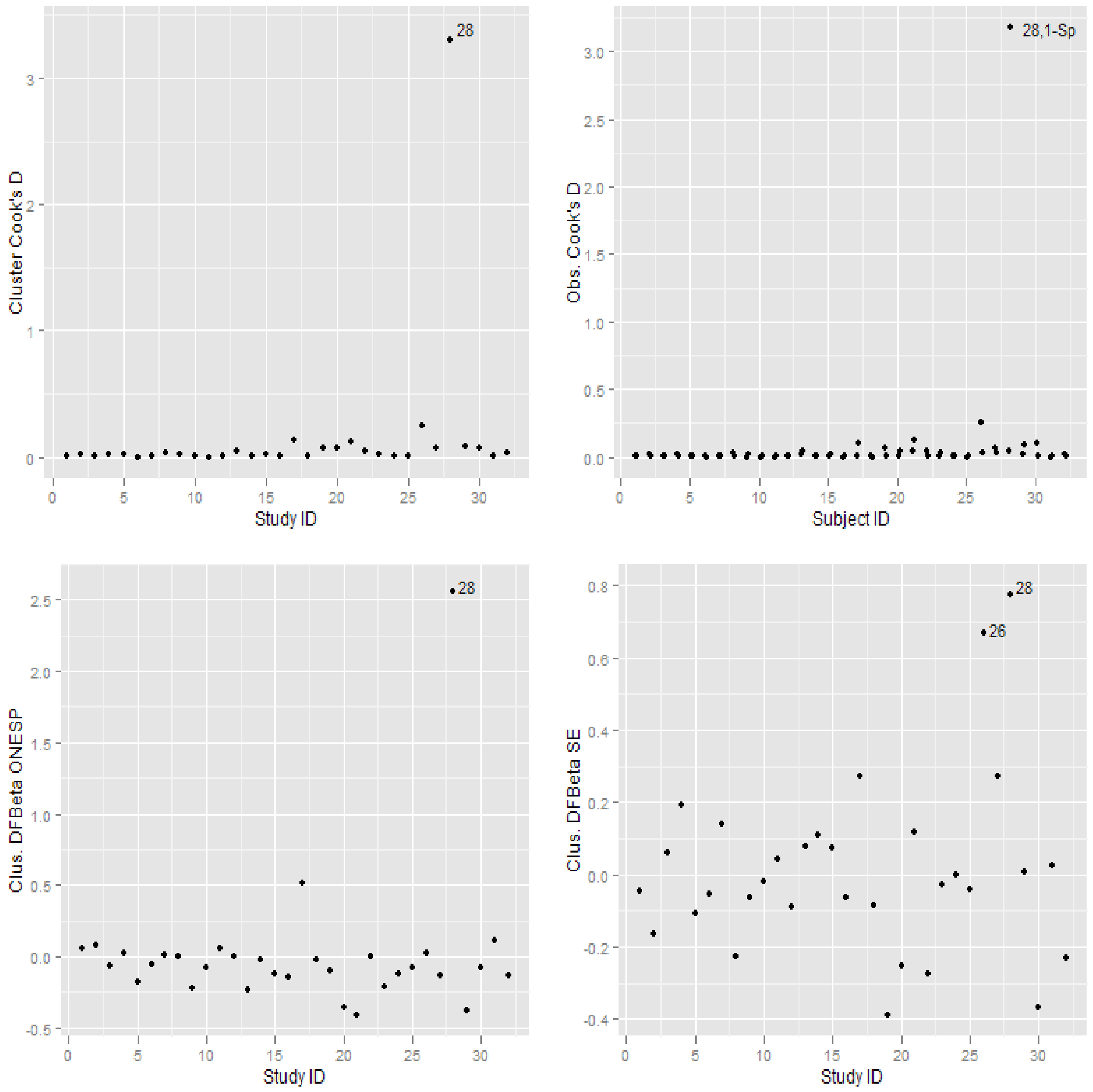

Figure 4.2: Deletion diagnostics for lymph node metastases data. Upper left panel: Cluster (Study) level Cook's D; upper right: Observation (sensitivity and 1- specificity within a cluster) Cook's D; lower left: DFBETAS for 1-specificity; and, lower right: DFBETAS for sensitivity.

Table 4.1: Data from a Meta-Analysis of Studies on Semi-Quantitative (Type=1) or Quantitative (Type=2) Catheter Segment Culture for Diagnosis of Intravascular Device-Related Bloodstream Infection. (Source: Chu et al. (2010)

Study	No. TP	No. FN	No. FP	No. TN	Type
1	12	0	29	289	1
2	10	2	14	72	1
3	17	1	36	85	1
4	13	0	18	67	1
5	4	0	21	225	1
6	15	2	122	403	1
7	45	5	28	34	1
8	18	4	69	133	1
9	5	0	11	34	1
10	8	9	15	96	1
11	5	0	7	63	1
12	11	2	122	610	1
13	5	1	6	145	1
14	7	5	25	342	1
15	10	1	93	296	1
16	5	5	41	271	1
17	5	0	15	53	1
18	55	13	19	913	1
19	6	2	12	30	1
20	42	26	19	913	2
21	5	3	5	37	2
22	13	0	11	135	2
23	20	0	24	287	2
24	7	6	13	72	2
25	48	2	15	47	2
26	11	1	14	72	2
27	15	5	32	170	2
28	68	13	5	11	2
29	13	1	5	72	2
30	8	3	66	323	2
31	13	1	98	293	2
32	14	1	0	155	2
33	8	2	4	60	2

Table 4.2: Comparison of Full and Reduced models for the catheter segment culture data (all data), for PA model with logit link (bias-corrected standard errors)

Parameter	Full Model	Reduced Model	No covar. Model
β_{00}(1-Sp)	-1.955 (0.271)	-1.948 (0.265)	-1.901 (0.223)
β_{01}(Se)	1.464 (0.209)	1.494 (0.198)	1.480 (0.197)
β_0(Type)	0.170 (0.271)	0.165 (0.245)	-
β_1(Type)	-0.003 (0.209)	-	-
ϕ_0	20.97	20.19	20.35
ϕ_1	2.965	2.950	2.794
ρ_g	0.324	0.310	0.330
Mean Spec. Type=1	0.856 (0.047)	0.856 (0.044)	-
Mean Sens. Type=1	0.812 (0.045)	0.840 (0.042)	-
Mean Spec. Type=2	0.893 (0.045)	0.892 (0.035)	-
Mean Sens. Type=2	0.813 (0.055)	0.791 (0.052)	-
Mean Spec. Overall	0.876 (0.029)	0.875 (0.029)	0.870 (0.025)
Mean Sens. Overall	0.812 (0.032)	0.817 (0.030)	0.815 (0.030)

Table 4.3: Studies 18 and 20 removed: Comparison of Full and Reduced models for the catheter segment culture data (all data),for PA model with logit link (bias-corrected standard errors)

Parameter	Full Model	Reduced Model	No covar. Model
β_{00}(1-Sp)	-1.670 (0.153)	-1.662 (0.143)	-1.629 (0.126)
β_{01}(Se)	1.696 (0.200)	1.724 (0.203)	1.450 (0.194)
β_0 Type	0.111 (0.153)	0.096 (0.134)	-
β_1 Type	-0.149 (0.200)	-	-
ϕ_0	11.06	10.65	10.59
ϕ_1	2.685	2.821	2.590
ρ_g	0.142	0.120	0.118
Mean Spec. Type=1	0.826 (0.031)	0.827 (0.028)	-
Mean Sens. Type=1	0.824 (0.041)	0.861 (0.029)	-
Mean Spec. Type=2	0.856 (0.033)	0.853 (0.025)	-
Mean Sens. Type=2	0.864 (0.041)	0.836 (0.033)	-
Mean Spec. Overall	0.842 (0.020)	0.841 (0.019)	0.836 (0.017)
Mean Sens. Overall	0.845 (0.026)	0.849 (0.026)	0.810 (0.030)

Table 4.4: Data from a Meta-Analysis of Studies on lymph node metastases. (Source: Klerkx et al. (2010)

Study	No. TP	No. FN	No. FP	No. TN	Ver. Bias	Design
1	7	3	3	6	0	Cohort
2	7	5	5	12	0	Cohort
3	18	6	3	19	0	Cohort
4	11	1	4	14	0	Cohort
5	9	5	3	29	1	Cohort
6	3	2	1	9	1	Cohort
7	6	0	2	7	0	Cohort
8	15	9	8	33	0	Cohort
9	15	7	1	25	0	Cohort
10	2	1	0	7	0	Cohort
11	5	1	2	3	1	Cohort
12	13	6	3	11	1	Cohort
13	7	1	1	23	1	Cohort
14	7	1	1	7	0	Cohort
15	9	2	4	26	1	Cohort
16	1	1	1	18	1	Cohort
17	10	0	17	20	0	Cohort
18	10	5	3	14	0	Cohort
19	5	8	2	17	0	Cohort
20	4	5	1	40	1	Cohort
21	12	2	2	41	0	Cohort
22	2	5	2	7	0	Cohort
23	3	1	1	22	0	Cohort
24	5	2	0	12	0	Cohort
25	6	3	1	11	0	Cohort
26	36	4	6	29	0	Cohort
27	16	2	2	22	0	Cohort
28	91	16	65	45	0	Case Control
29	4	0	1	31	0	Case Control
30	5	8	25	133	0	Case Control
31	18	6	8	22	0	Case Control
32	6	6	1	16	0	Case Control

Table 4.5: Comparison of Full and Reduced models for the lymph node metastases data (all data), for PA model with logit link (bias-corrected standard errors)

Parameter	Full Model	Reduced Model	No covar. Model
β_{00}(1-Sp)	-1.279 (0.134)	-1.300 (0.131)	-1.400 (0.352)
β_{01}(Se)	1.074 (0.135)	1.124 (0.133)	1.001 (0.167)
β_0(PVB)	-0.973 (0.449)	-0.714 (0.301)	-
β_1(PVB)	-0.455 (0.334)	-	-
ϕ_0	2.263	2.256	2.143
ϕ_1	5.735	5.557	5.925
ρ_g	0.387	0.372	0.408
Mean Spec. PVB=1	0.905 (0.056)	0.882 (0.034)	-
Mean Sens. PVB=1	0.650 (0.081)	0.601 (0.079)	-
Mean Spec. PVB=0	0.782 (0.066)	0.786 (0.022)	-
Mean Sens. PVB=0	0.745 (0.036)	0.755 (0.025)	-
Mean Spec. Overall	-	-	0.802 (0.056)
Mean Sens. Overall	-	-	0.731 (0.033)

Table 4.6: Study 28 removed: Comparison of Full and Reduced models for the lymph node metastases data (all data), for PA model with logit link (bias-corrected standard errors)

Parameter	Full Model	Reduced Model	No covar. Model
β_{00}(1-Sp)	-1.697 (0.188)	-1.713 (0.178)	-1.782 (0.176)
β_{01}(Se)	0.987 (0.204)	1.010 (0.188)	1.933 (0.172)
β_0 PVB	-0.550 (0.422)	-0.430 (0.322)	-
β_1 PVB	-0.315 (0.312)	-	-
ϕ_0	2.019	1.971	1.882
ϕ_1	2.507	2.412	2.450
ρ_g	0.181	0.179	0.186
Mean Spec. PVB=1	0.904 (0.040)	0.895 (0.035)	-
Mean Sens. PVB=1	0.662 (0.084)	0.641 (0.086)	-
Mean Spec. PVB=0	0.845 (0.025)	0.847 (0.023)	-
Mean Sens. PVB=0	0.729 (0.040)	0.733 (0.037)	-
Mean Spec. Overall	-	-	0.856 (0.022)
Mean Sens. Overall	-	-	0.718 (0.027)

Chapter 5

Conclusion

Methods for diagnostic accuracy analysis were presented for both the single study and meta-analysis frameworks. Specifically, population-averaged methods were employed in both cases using slightly different aspects of related methodology. In Chapter 2, the ROC-GLM was presented within the context of the deletion diagnostics by Preisser and Qaqish (1996). In this context we are able to identify potentially influential subjects as a sensitivity analysis within the ROC-GLM method. Chapters 3 presented a PA model for diagnostic accuracy meta-analysis which to our knowledge has not been done previously. Chapter 4 then presented the deletion diagnostics again, this time in the meta-analysis contexts. Overall, the common theme is population-averaged models along with the associated deletion diagnostics to enhance sensitivity analysis of diagnostic accuracy studies.

There are a number of future directions for extension of these methods in the diagnostic accuracy setting. In the single-study case, it is suggested in Chapter 2 that the deletion diagnostics may in fact be a useful tool in the identification of imperfect gold standard measurements. This finding was presented within the context of the neonatal audiology data where the original authors acknowledged that gold standard was difficult to measure accurately. Simulations and further study might investigate this further.

For PA methods related to diagnostic accuracy meta-analysis, certainly the production of an ROC curve via inversion of the equations presented in Chapter 3 would be important future work. An initial outline of a method is presented in Appendix II, and could be an interesting publication in and of itself. It was shown that the PA method performs well as measured with simulations against the GLMM of Chu and Guo (2010).

Finally, there is a growing body of literature in predictive values and incremental value in the context of biomarker development. The methods presented here may perhaps be extended to these situations.

Appendix I

Supplemental Tables to Chapter 3

Table S1. Data from a Meta-Analysis of Studies on Semi-Quantitative or Quantitative Catheter Segment Culture for Diagnosis of Intravascular Device-Related Bloodstream Infection. (Source: Chu et al. (2010).

Study	No. TP	No. FN	No. FP	No. TN
1	12	0	29	289
2	10	2	14	72
3	17	1	36	85
4	13	0	18	67
5	4	0	21	225
6	15	2	122	403
7	45	5	28	34
8	18	4	69	133
9	5	0	11	34
10	8	9	15	96
11	5	0	7	63
12	11	2	122	610
13	5	1	6	145
14	7	5	25	342
15	10	1	93	296
16	5	5	41	271
17	5	0	15	53
18	55	13	19	913
19	6	2	12	30
20	42	26	19	913
21	5	3	5	37
22	13	0	11	135
23	20	0	24	287
24	7	6	13	72
25	48	2	15	47
26	11	1	14	72
27	15	5	32	170
28	68	13	5	11
29	13	1	5	72
30	8	3	66	323
31	13	1	98	293
32	14	1	0	155
33	8	2	4	60

Table S2. Simulated data of 25-study meta-analysis. Study cluster sizes were fixed at (175,25) and generated via correlated binomial data methods with mean 1-specificity=0.7 and mean sensitivity=0.75.

Study	No. TP	No. FN	No. FP	No. TN
1	12	13	54	121
2	22	3	31	144
3	16	9	34	141
4	14	11	40	135
5	20	5	50	125
6	19	6	53	122
7	20	5	39	136
8	24	1	55	120
9	11	14	57	118
10	12	13	61	114
11	21	4	51	124
12	17	8	68	107
13	17	8	49	126
14	21	4	24	151
15	17	8	49	126
16	16	9	57	118
17	19	6	51	124
18	18	7	16	159
19	14	11	43	132
20	18	7	78	97
21	20	5	55	120
22	17	8	56	119
23	19	6	67	108
24	22	3	67	108
25	12	13	24	151

Table S3. Comparison of Logit and Probit link functions for the PA method with constant scale for clusters (with empirical standard errors

Parameter	Logit	Probit
β_0^*	-1.901 (0.205)	-1.127 (0.110)
β_1^*	1.500 (0.187)	0.906 (0.105)
ϕ	11.227	11.227
ρ_g	0.222	0.222
Mean Specificity	0.870 (0.023)	0.870 (0.023)
Mean Sensitivity	0.818 (0.028)	0.818 (0.028)

Table S4. Percent relative bias of simulated parameters based upon empirical standard errors after fitting PA model as well as SS-model marginal converted results

				Logit				Probit			
Clus.				β_0^*		β_1^*		β_0^*		β_1^*	
Size	Type	(n_0, n_1)	(μ_0, μ_1)	PA	SS	PA	SS	PA	SS	PA	SS
K=15	Fix.	(175,25)	(0.7,0.75)	0.0	0.3	-0.1	0.2	0.0	-0.1	-0.1	0.0
			(0.9,0.75)	0.0	-0.2	-0.2	0.2	0.0	0.0	-0.2	-0.2
			(0.7,0.95)	0.3	0.6	0.2	0.3	0.3	0.3	0.2	1.0
			(0.9,0.95)	0.1	-0.2	0.0	0.1	0.1	0.1	0.0	0.8
	Uneq.	MVN(1)	(0.7,0.75)	-1.6	-1.4	0.3	0.7	-1.6	-1.6	0.3	0.4
			(0.9,0.75)	-0.2	-0.3	-0.1	0.2	-0.2	-0.2	-0.1	-0.1
			(0.7,0.95)	-0.4	-0.6	0.3	-0.4	-0.8	-0.8	0.3	0.3
			(0.9,0.95)	0.0	-0.2	0.8	0.9	0.0	0.0	0.8	0.8
K=25	Fix.	(175,25)	(0.7,0.75)	0.0	0.3	0.0	0.3	0.0	-0.1	0.0	0.0
			(0.9,0.75)	0.1	0.3	-0.1	-0.2	0.1	-0.1	-0.1	0.1
			(0.7,0.95)	0.2	0.6	0.1	0.2	0.2	0.2	0.1	0.1
			(0.9,0.95)	0.1	-0.2	0.0	0.1	0.1	0.1	0.0	0.0
	Uneq.	MVN(1)	(0.7,0.75)	0.5	-1.1	0.0	-1.4	-1.4	-1.4	0.0	0.2
			(0.9,0.75)	-0.1	-0.3	-0.2	0.1	-0.1	-0.1	-0.2	-0.3
			(0.7,0.95)	0.4	-0.6	0.4	-0.9	-0.9	-0.9	0.4	0.4
			(0.9,0.95)	0.0	-0.2	0.8	0.9	0.0	0.0	0.8	0.8
K=50	Fix.	(175,25)	(0.7,0.75)	0.1	0.4	-0.1	0.2	0.1	0.1	-0.1	-0.1
			(0.9,0.75)	-0.1	-0.3	-0.1	0.2	-0.1	-0.1	-0.1	-0.1
			(0.7,0.95)	-0.2	0.2	-0.1	0.0	-0.2	-0.2	-0.1	-0.1
			(0.9,0.95)	0.0	-0.3	-0.1	0.0	0.0	-0.1	-0.1	-0.1
	Uneq.	MVN(1)	(0.7,0.75)	-1.1	-1.3	0.1	0.5	-1.6	-1.5	0.1	0.1
			(0.9,0.75)	-0.1	-0.2	-0.1	0.2	-0.1	-0.1	-0.1	-0.1
			(0.7,0.95)	-0.8	-0.6	0.3	0.4	-0.8	-0.8	0.3	0.3
			(0.9,0.95)	0.0	-0.1	0.8	0.9	0.0	0.0	0.8	0.8

Table S5. Convergence rates of simulated models based upon empirical standard errors after fitting PA model as well as SS-model marginal converted results

				Logit		Probit	
Size	Type	(n_0, n_1)	(μ_0, μ_1)	PA	SS	PA	SS
K=15	Fix.	(175,25)	(0.7,0.75)	100	96.8	100	98.8
			(0.7,0.95)	100	96.6	100	98.8
			(0.9,0.75)	100	97.6	100	99.2
			(0.9,0.95)	100	98.6	100	98.7
	Uneq.	MVN(1)	(0.7,0.75)	100	98.2	100	99.6
			(0.7,0.95)	100	97.2	100	98.1
			(0.9,0.75)	100	97.4.2	100	97.4
			(0.9,0.95)	100	98.5	100	98.5
K=25	Fix.	(175,25)	(0.7,0.75)	100	99.6	100	99.8
			(0.7,0.95)	100	99.2	100	99.6
			(0.9,0.75)	100	100	100	99.8
			(0.9,0.95)	100	97.6	100	98.2
	Uneq.	MVN(1)	(0.7,0.75)	100	98.1	100	100
			(0.7,0.95)	100	97.2	100	97.2
			(0.9,0.75)	100	100	100	100
			(0.9,0.95)	100	96.5	100	96.5
K=50	Fix.	(175,25)	(0.7,0.75)	100	100	100	100
			(0.7,0.95)	100	100	100	100
			(0.9,0.75)	100	99.8	100	100
			(0.9,0.95)	100	99.2	100	99.4
	Uneq.	MVN(1)	(0.7,0.75)	100	99.6	100	99.8
			(0.7,0.95)	99.8	98.1	100	99.6
			(0.9,0.75)	100	100	100	99.8
			(0.9,0.95)	100	97.6	100	98.2

Appendix II

Diag104.sas macro

```
\%GEE  (
  DATA = SAS dataset,                                   { syslast }
  YVAR = y-variable,                                    { required }
  XVAR = x-variables,                                   { required }
  ID   = id-variable,                                   { required }
  TIME = within cluster variable                        { }
  HET = indicator for heterogeneous phi                 { }
  LINK = link function,                                 { required }
  VARI = mean-variance relation,                        { required }
  CORR = correlation structure,                         { required }
  N    = binomial denominator variable,                 { }
  M    = dependence,                                    { 1 }
  R    = given correlation matrix,                      { }
  SCALE= scale parameter,                               { }
  BETA = initial estimate of beta,                      { }
  OFFS = offset variable,                               { }
  PROBITOVAR = variance estimation
                using observed information (probit only) { 0 }
  NCOVOUT = output dataset of beta, model-based se,
            and model-based covariance matrix,          { }
  RCOVOUT = output dataset of beta, empirical se,
            and empirical covariance matrix,            { }
```

```
   BCOVOUT = output dataset of beta, bias-corrected se,
             and bias-corrected covariance matrix,         { }
   OBSOUT = output dataset of observation diagnostics,     { }
   CLSOUT = output dataset of cluster diagnostics,         { }
   NORM = Cook's D and DFBETAs normed by model-based
          (1, default), Empirical (2), or bias-corrected (3)
          variance estimator                               { 1 }
   ITER     = maximum iterations,                          { 20 }
   MONITOR  = print out iterations (YES | NO)              { NO }
    CRITTYPE = type of convergence criterion (REL | ABS)    { REL }
   CRIT     = convergence criterion                        { 1E-5 }
    BINRANGE = binary range checks enforced (Y | N)         { Y }
 )
```

REQUIRED MACRO SPECIFICATIONS

To run the macro, the user is required to provide a SAS dataset (DATA) with response variable (YVAR), a list of independent variables (XVAR), cluster identifier variable (ID), link (LINK) and variance (VARI) function, and working correlation structure (CORR).

If no dataset name is given, the last working SAS dataset is used. Only one response variable may be given. If an intercept term is desired in the model, the intercept variable must be explicitly included with the covariate list. Options for LINK, VARI, and CORR are below.

```
     The following choices of link (LINK) function are available:
       1 - Identity
       2 - Logarithm
       3 - Logit
```

```
  4 - Reciprocal

  5 - Probit


The following choices of variance (VARI) function are available:

  1 - Gaussian

  2 - Poisson

  3 - Binomial

  4 - Gamma


The following choices of correlation (CORR) structures are available:

  1 - Independence

   2 - Stationary m-dependent

   3 - Non-stationary m-dependent

   4 - Exchangeable

   5 - Autoregressive(1)

   6 - Unstructured

   7 - User-defined, R must be given when macro is called
```

For m-dependent correlation structures, (M) should be specified. If it is not specified, the default is 1-dependence. For user-defined correlation structures, all elements of (R) must be given in one string without commas. The macro creates the square matrix.

OPTIONAL SPECIFICATIONS

A within-cluster ordering variable (TIME) can be specified if desired. The possible values of this variable must be positive consecutive integers starting with 1 or the macro will not work correctly. If a time variable is specified, the data will be pre-sorted by cluster ID and time. Otherwise, the data will be used as ordered in the dataset when the macro is called.

If a heterogeneous (e.g. time-varying) scale parameter is desired, then this may be done using the HET parameter. The default is 0 (common scale parameter). If HET is equal to 1, then a scale parameter for each value of TIME is estimated.

A denominator (N) is required for binomial data, but if not specified is assumed to equal 1. Scale is assumed to equal 1 for binary data, but is otherwise estimated. Optionally, it may be set to equal 1 (or some other value) with the (SCALE) option. [Note that the scale parameter is assumed to be constant across all observations.] The (BETA) option may be used to specify starting values for the regression coefficient estimates, otherwise GLiM estimates are used as starting values. The offset option (OFFS) specifies a SAS variable containing offsets (these are used for example in poisson regression with unequal exposure periods).

The user can control the maximum number of iterations allowed (ITER), the convergence criteria (CRIT), and can print out details of each iteration (MONITOR). Convergence is determined by the magnitude of the maximum absolute or relative (default) change in the betas between iterations. Checking for absolute or relative changes can be set by the user (CRITTYPE).

There are a number of output datasets that can be requested by the user. For a dataset that contains beta estimates, standard errors, and the covariance matrix, enter an output dataset name for any or all of the following options:

```
For the model-based (naive) SEs and covariance matrix         (NCOVOUT)
For the empirical sandwich (robust) SEs and covariance matrix (RCOVOUT)
For the bias-corrected SEs and covariance matrix              (BCOVOUT)
```

For datasets that contain regression diagnostics, including Cooks distance, DFBETA, DFBETAS (standardized), and leverage, enter an output dataset name for either or both of the following options:

```
    For cluster-level diagnostics     (CLSOUT)
    For observation-level diagnostics (OBSOUT)
```

Cooks distance and DFBETAS are standardized by either the model-based, empirical, or bias-corrected variance estimators. Use the (NORM) option to designate which to use: Model-based = 1 (default), Empirical = 2, Bias- corrected = 3.

As noted in Prentice (Biometrics 1988:44, 1033-1048), among other places, there are restrictions placed on the range of values that the correlation coefficients in R may take for binary response data. An option has been added to this macro (BINRANGE, when set to Y) to allow the user to enforce these ranges, which have the effect of ensuring non-negative joint probabilities for all within-cluster pairs of observations.

REGRESSION DIAGNOSTICS

This macro provides computational formulae for case-deletion regression diagnostics (Preisser and Qaqish, 1996). These diagnostics are generalizations of Cook's distance, DFBETA and leverage for linear regression, and their counterparts for generalized linear models. They are an approximation to the difference in the estimated regression coefficients that one would obtain upon deleting either one observation or one cluster. The diagnostics are sometimes called "one-step" diagnostics because they are equal to the procedure that upon convergence of the iteratively reweighted least squares algorithm to the GEE solution, applies one more iteration after deletion of an observation (or cluster). The difference in the regression coefficients one would obtain from such a procedure is equivalent to the value of the diagnostic. Because, however, of computational formula for the diagnostics, no additional iterations are required in the computing algorithm to obtain the full set of observation-deletion and cluster-deletion diagnostics.

Diagnostics are not provided automatically by the software, but may be requested with optional statements declared in the macro call. To request observation-deletion diagnostics, one set for each observation, use the statement OBSOUT. For example, OBSOUT=infobs will create a sas dataset “infobs” that will contain the influence diagnostics for the observations, including cook’s distance, DFBETA, DFBETAS (standardized), observation leverage, cluster size of the cluster to which the observation belongs, fitted value, raw residual, and standardized residual (raw residual divided by the variance function). The statement CLSOUT is used to obtain cluster-deletion diagnostics. For example, CLSOUT=infcls will create a sas dataset “infcls” that will contain the influence diagnostics for the clusters, including cook’s distance, DFBETA, DFBETAS (standardized), cluster leverage, cluster size, and a quadratic summary of the standardized residual vector of the cluster (the usefulness of this last statistic is yet to be investigated).

Unstandardized and standardized DFBETAs are produced by the macro. In the output dataset, the variables containing the unstandardized values take the name of the covariate with DFBETA as the prefix; the variables containing the standardized values take the name of the covariate with DFBETAS as the prefix. While these prefixes are long, they allow differentiation from the original covariates if the diagnostic datasets are merged back onto the raw data.

The user may compute standardized DFBETAs using a different norm–either model-based, empirical sandwich, or bias-corrected standard errors–by first requesting an output dataset using the NCOVOUT, RCOVOUT, of BCOVOUT options, respectively. These output data sets are also useful for constructing contrasts and hypothesis tests, for example, using SAS/IML.

OUTPUT DIAGNOSTICS DATASETS

```
The observation-level dataset specified with the OBSOUT option will

contain the following variables:

   I                    Sequential cluster number
```

IJ	Sequential observation number
NI	number of records in the cluster
FIT	Predicted value for obs
RES	Unstandardized residual
SRES	Standardized residual
QWOBS	Leverage: Diagonal element of H matrix for obs
COOKDOBS	Cooks D
DFBETA<xvars>	Unstandardized DFBETA values
DFBETAS<xvars>	Standardized DFBETA values

The cluster-level dataset specified with the CLSOUT options will contain the following variables:

I	Sequential cluster number
NI	number of records in the cluster
TRQWCLS	Leverage: Trace of H matrix for cluster
COOKDCLS	Cooks D
GCLS	Scalar summary of the residual vector Ei for cluster, tr(Ei)[var(Ei)]^{-1}Ei (similar to MCLS_i on p557 of Preisser and Qaqish, 1996, but without the Hi and p)
DFBETA<xvars>	Unstandardized DFBETA values
DFBETAS<xvars>	Standardized DFBETA values

Appendix III

Determination of ROC curve from PA model

To develop an ROC curve for the case of balanced cluster sizes using the PA model approach, the formula in equations (9), (12) and (14) are inverted, solving for the GLMM parameter $\theta_m^T = (\beta_0, \beta_1, \rho_m, \sigma_0, \sigma_1)$ from the PA model parameter $\theta_g^T = (\beta_0^*, \beta_1^*, \rho_g, \phi_0, \phi_1)$. In other words, defining the vector function $F(\cdot)$ such that $\theta_g^T = F(\theta_m^T)$, we solve $\hat{\theta}_m^T = F^{-1}(\hat{\theta}_g^T)$, where $\hat{\theta}_g$ is the GEE estimate of the PA model parameter. Then, leveraging the bivariate normal distributional structure of the GLMM model, $\hat{\theta}_m$ is plugged into formula (2) of Chu, Guo and Zhou (2009) obtaining the ROC curve

$$g(Se) = (\beta_1 - \rho_m \beta_0 \sigma_1/\sigma_0) + \rho_m \sigma_1/\sigma_0[g(1 - Sp)].$$

Fortunately, the required inversion can be broken down into three simple steps, as illustrated for the logit link:

1. Together, equation (9) expressed as $\beta_0^\star = f_1(\beta_0, \sigma_0^2)$ for $j = 0$, and equation (12) expressed as $\phi_0 = f_2(\beta_0, \sigma_0^2)$ give two equations in two unknowns. Specifically, equation (9) gives

   $$\beta_0 \approx [c^2\sigma_0^2 + 1]\beta_0^\star,$$

   which is then inserted into equation (12) giving an equation for σ_0^2 which is solved iteratively using Newton's method (details are provided below). The solution $\hat{\sigma}_0^2$ along with the GEE estimate $\hat{\beta}_0^\star$ are plugged into the equation above giving

   $$\hat{\beta}_0 \approx [c^2\hat{\sigma}_0^2 + 1]\hat{\beta}_0^\star.$$

2. Estimates $\hat{\beta}_1$ and $\hat{\sigma}_1^2$ are determined in steps analogous to (1).

3. Finally, it is easy to invert equation (14) giving

$$\rho_m = \frac{\rho_g \sqrt{[L_0^2\sigma_0^2 + \frac{\mu_0^*(1-\mu_0^*)}{n_{0+}}][L_1^2\sigma_1^2 + \frac{\mu_1^*(1-\mu_1^*)}{n_{1+}}]}}{L_0 L_1 \sigma_0 \sigma_1}, \tag{5.1}$$

   which provides an estimate $\hat{\rho}_m$ by plugging in $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_0$ and $\hat{\sigma}_1$ obtained in the first two steps, with $\hat{\sigma}_j = \sqrt{\hat{\sigma}_j^2}$ for $j = 0, 1$.

Details of application of Newton's method.

In step (1), let $x = \hat{\sigma}_0^2$, $k = c^2$ and $y = (kx+1)\beta_0^\star$. We wish to solve for the root x in the equation $f(x) = 0$ where (referencing equation (12)),

$$f(x) = \frac{n_{0+}[L_0(x)]^2 x}{\mu_0^*(1-\mu_0^*)} + 1 - \phi_0, \tag{5.2}$$

where $L_0(x) = \exp[y(x)]/\{1+\exp[y(x)]\}^2$ and $\mu_0^\star = \exp(\beta_0^\star)/[1+\exp(\beta_0^\star)]$. Given $x^{(t)}$, the estimate at the t-th iteration, the updated estimate at the $(t+1)$-th iterative step is given by Newton's method as

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

where, through application of the product, quotient and chain rules of differentiation

$$f'(x) = \partial f(x)/\partial x = \frac{k\beta_0^\star e^y(1-e^y)}{[1+e^y]^3}.$$

Convergence proceeds until $|x^{(t+1)} - x^{(t)}| < \epsilon$ for some small ϵ. A good starting value $x^{(0)}$ for σ_0^2 is necessary for convergence. The same basic procedure is used for step (2).

Bibliography

Alonzo, T. and Pepe, M. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3:421–432.

Arends, L., Hamza, T., van Houwelingen, H., Heijenbrok-Kal, M., Hunink, M., and Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*, 28:621–638.

Baker, R. and Jackson, D. (2008). A new approach to outliers in meta-analysis. *Health Care Management Science*, 11:121–31.

Brumback, L., Pepe, M., and Alonzo, T. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25:575–590.

Cai, T. and Moskowitz, C. (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics*, 5:573–86.

Cai, T. and Pepe, M. (2002). Semiparametric ROC analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97:1099–1107.

Cai, T. and Zheng, Y. (2007). Model checking for ROC regression analysis. *Biometrics*, 63(1):152–163.

Chu, H. and Cole, S. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59:1331–1332.

Chu, H. and Guo, H. (2009). Letter to the editor: a unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 10:201–203.

Chu, H. and Guo, Y. (2010). Diagnostic accuracy meta-analysis using generalized linear mixed models. *Medical Decision Making*, 8:385–403.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–45.

Devlin, S., Thomas, E., and Emerson, S. (2010). Robustness of approaches to ROC curve modeling under misspecification of the underlying probability model. *UW Bios Technical Paper Series*, 355:000–000.

Dorfman, D. and Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, 33:117–24.

Emrich, J. and Piedmonte, M. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45:302–304.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19:242–51.

Gatsonis, C. (1995). Random-effects models for diagnostic accuracy data. *Academic Radiology*, 2:S14–21.

Gonen, M. and Heller, G. (2010). Lehmann family of ROC curves. *Medical Decision Making*, 2010:509–17.

Gumedze, F. and Jackson, D. (2011). A random effects variance shift model for detecting and accomodating outliers in meta-analysis. *BMC Medical Research Methodology*, 11:19.

Hammill, B. and Preisser, J. (2006). A sas/iml software program for GEE and regression diagnostics. *Computational Statistics and Data Analysis*, 51:1197–1212.

Hanley, J. (1996). The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine*, 15:1575–85.

Hanley, J. and McNeil, B. (1982). The meaning and use of the area under the ROC curve. *Radiology*, 143:29–36.

Harbord, R., Deeks, J., Egger, M., Whiting, P., and Sterne, J. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8:239–251.

Heagerty, P. and Pepe, M. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Applied Statistics*, 48:533–51.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-analysis.* Academic Press, New York, first edition.

Hsieh, F. and Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24:25–40.

Ishwaran, H. and Gatsonis, C. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC curves. *The Canadian Journal of Statistics*, 28:731–50.

Janes, H., Longton, G., and Pepe, M. (2009). Accommodating covariates in receiver operating characteristic analysis. *The Stata Journal*, 9(1):17–39.

Klerkx, W., Bax, L., Veldhuis, W., Heintz, A., Mali, W., Peeters, P., and Moons, K. (2010). Detection of lymph node metastases by gadolinium-enhanced magnetic resonance imaging: systematic review and meta-anlaysis. *Journal of National Cancer Institute*, 102:244–253.

Krzanowski, W. and Hand, D. (2009). *ROC curves for continuous data*. CRC Press, United Kingdom, first edition.

Leeflang, M., Bossuyt, P., and Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology*, 62:5–12.

Leisenring, W., Pepe, M., and Longton, G. (1997). A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics in Medicine*, 16:1263–1281.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lu, B., Preisser, J., Qaqish, B., Suchindran, C., Bangdiwala, S., and Wolfson, M. (2007). Comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63:935–41.

Lusted, L. (1960). Logical analysis in roentgen diagnosis. *Radiology*, 74:178–93.

Ma, X., Chu, H., Chen, Y., and Cole, S. (2012). A hybrid bayesian hierarchical model combining cohort and case-control studies for meta-analysis of diagnostic tests: accounting for disease prevalence and partial verification bias. *Statistics in Medicine*, 999:Submitted.

Mancl, L. and DeRouen, T. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57:126–134.

Martus, P., Stroux, A., Junemann, A., Korth, M., Jonas, J., Horn, F., and Ziegler, A. (2004). GEE approaches to marginal regression models for medical diagnostic tests. *Statistics in Medicine*, 23:1377–1398.

Metz, C. (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21:720–33.

Metz, C., Herman, B., and Shen, J. (1998). Maximum likelihood estimation of reciever operating characteristic curves from continuously distributed data. *Statistics in Medicine*, 17:1033–53.

Moses, L., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12:1293–1316.

Norton, S., Gorga, M., Widen, J., Folsom, R., Sininger, Y., Cone-Wesson, B., Vohr, B., Mascher, K., and Fletcher, K. (2000). Identification of neonatal hearing impairment: evaluation of transient evoked otoacoustic emission, distortion product otoacoustic emission, and auditory brain stem response test performance. *Ear and Hearing*, 21(5):508–528.

Patton, M. (1978). *Utilization-focused Evaluation.* SAGE, Beverly Hills.

Pepe, M. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3):595–608.

Pepe, M. (1998). Three approaches to regression analysis of receiver operating characteristic curves in medical diagnostic testing. *Biometrics*, 54:124–35.

Pepe, M. (2000). An interpretation for the ROC curve and inference using glm pROCedures. *Biometrics*, 56:352–9.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, United Kingdom, first edition.

Pepe, M. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60:528–35.

Petitti, D. (2000). *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine.* Oxford University Press, United Kingdom, first edition.

Preisser, J., By, K., Perin, J., and Qaqish, B. (2012). Deletion diagnostics for alternating logistic regressions. *Biometrical Journal*, 54(5):701–715.

Preisser, J. and Qaqish, B. (1996). Deletion diagnostics for generalized estimating equations. *Biometrika*, 83:551–562.

Reitsma, J., Glas, A., Rutjes, A., Scholten, R., Bossuyt, P., and Zwinderman, A. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58:982–990.

Riley, R., Abrams, K., Sutton, A., Lambert, P., and Thompson, J. (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7:1–15.

Riley, R., Dodd, S., Craig, J., Thompson, J., and Williamson, P. (2008). Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine*, 27:6111–36.

Rutter, C. and Gatsonis, C. (2001). A hierarchical regression approach to meta-anlysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20:2865–2884.

SAS Institute (2008). *SAS/STAT 9.2 Users Guide: PROC NLMIXED*. United States, first edition.

Toledano, A. and Gatsonis, C. (1995). Regression analysis of correlated receiver operating characteristic data. *Academic Radiology*, 2:530–536.

Tosteson, A. and Begg, C. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making*, 3:204–215.

van Houwelingen, H., Arends, L., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21:589–624.

Vens, M. and Ziegler, A. (2012). Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: A case study. *Computational Statistics and Data Analysis*, 56:1232–1242.

Walter, S. (2002). Properties of the summary receiver operating characteristic (sROC) curve for diagnostic test data. *Statistics in Medicine*, 21:1237–56.

Wang, W., Davis, C., and Soong, S. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, 15:2215–2229.

Zeger, S., Liang, K., and Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–60.

Zhou, X., McClish, D., and Obuchowski, N. (2002). *Statistical methods in diagnostic medicine*. Wiley, New York, first edition.

Ziegler, A., Blettner, M., Kastner, C., and Chang-Claude, J. (1998). Identifying influential families using regression diagnostics for generalized estimating equations. *Genetic Epidemiology*, 15:341–353.

Zou, K., Hall, W., and Shapiro, D. (1997). Smooth non-parametric receiver operating characteristc (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–56.

Zweig, M. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–77.