STATISTICAL METHODS FOR EVALUATING THE DIAGNOSTIC ACCURACY OF INCOMPLETE MULTIPLE TESTS

Yi Zhang

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:

Donglin Zeng
Haitao Chu
Jianwen Cai
Michael Hudgens
Stephen Cole

Abstract

YI ZHANG: STATISTICAL METHODS FOR EVALUATING THE DIAGNOSTIC ACCURACY OF INCOMPLETE MULTIPLE TESTS
(Under the direction of Donglin Zeng and Haitao Chu)

The accurate diagnosis of a molecularly-defined subtype of cancer is often a very important step toward its effective prevention and treatment. For the diagnosis of some subtypes of certain cancers, a gold standard with perfect sensitivity and specificity may be unavailable. In those scenarios, the status of the tumor subtype commonly is measured by multiple imperfect diagnostic markers. In many such studies, some subjects are only measured by a subset of diagnostic tests and the missing probabilities may depend on the unknown disease status. In this research, we present novel statistical methods based on an EM algorithm to evaluate incomplete multiple imperfect diagnostic tests under conditional independence and conditional dependence assumptions. We applied the proposed methods to a set of real data from the NCI Colon Cancer Family Registry (C-CFR) on diagnosing microsatellite instability (MSI) for hereditary non-polyposis colorectal cancer (HNPCC) to estimate diagnostic accuracy (i.e., sensitivities and specificities) and prevalence for 11 biomarker tests. Simulations are conducted to evaluate the small-sample performance of our methods. The advantages and limitations of our methods are discussed. An R package was developed for easy implementation of our methods. Finally, a proposal for future research also was presented.

Acknowledgments

I especially thank Dr. Donglin Zeng and Dr. Haitao Chu for their mentorship, guidance, inspiration, support, and patience during the preparation of this dissertation and throughout my graduate study. Also, I convey my sincere thanks to my committee members, Dr. Jianwen Cai, Dr. Michael Hudgens, and Dr. Stephen Cole for their constant encouragement and helpful comments. Finally, this dissertation is a dedication to my wife Vicky Pan and my parents. It is impossible to have completed this journey without their unconditional love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Literature Review

1.1 Diagnostic Accuracy for Binary Tests

Accurate diagnosis of a disease or classification of a subtype of a disease is often the first step toward the treatment and prevention of the disease. A diagnostic test is expected to contribute to a reliable diagnosis of a patient's medical condition and aid in the health practitioner's development of an appropriate treatment plan. Misdiagnoses are likely to result in mislead health practitioners' initiating unnecessary or incorrect treatment plans. Therefore, evaluation the diagnostic accuracy of a test is pivotal to medical practices.

Conventional measures of diagnostic accuracy include sensitivity and specificity, predictive values, the area under the receiver operating characteristic (ROC) curve, and Youden's index. These measures address different aspects of a diagnostic test, such as its discriminative property or predictive ability. Our research considers only binary diagnostic tests with two possible outcomes, i.e., whether a subject has a certain disease or not. Sensitivity and specificity are basic measures of the performance of a binary test. Sensitivity is the probability of having a positive test result when the subject actually has the disease; specificity is the probability of having a negative test result when the subject does not have the disease. (1-specificity) and (1-sensitivity)

are often referred to as false positive and false negative error rates associated with a test. Neither sensitivity nor specificity is affected by the prevalence of the disease. This crucial property makes estimates of sensitivity and specificity from one study readily applicable to other studies in which the prevalence of the disease was different. Nevertheless, sensitivity and specificity may be affected by the stages of the disease or by patients' characteristics (e.g., different densities of fat tissue). Ledley and Lusted (1959) made some early contributions to the paradigm of diagnostic accuracy.

Ideally, the sensitivity and specificity of a new test are evaluated by comparing its results with the true status/condition of the disease (presence or absence) in a subject, which is the result of a test with a perfect ability to classify the disease; such a test is referred to as a 'gold standard'. In other words, a gold standard is an error-free reference standard with a sensitivity of 100% and a specificity of 100%. In practice however, a gold standard is often impossible to find, or it simply may not exist. Multiple imperfect diagnostic tests often are used in the absence of a gold standard. These tests are either applied simultaneously to one subject and interpreted altogether or applied sequentially in a prespecified order. The latter is usually more cost-effective but less efficient, since the decisions of whether to administer subsequent tests and when to administer them depend on the results of tests that already have been conducted. The next section introduces the common methods that are used to evaluate diagnostic accuracy in the absence of a gold standard.

1.2 Diagnostic Accuracy Evaluation without a Gold Standard

1.2.1 Discrepant Analysis

Discrepant analysis (also known as discordant analysis or discrepant resolution) applies a series of reference standards without statistical modeling to find the true

disease status. When two tests show discrepant results, a resolver test (often with better diagnostic accuracy but which is costly and/or invasive) is chosen to reconcile the discrepancy. This approach is subject to the error rate of the test used to reconcile the discrepancy as well as the error rates in the wrongfully concordant results of the initial tests. The resolver test is assumed to be independent of the preceding tests, which may not be the case. Hadgu (1996, 1997, 1999) posited that discrepant analysis cause serious bias in the estimates of diagnostic accuracy that are obtained and that such estimates are scientifically flawed.

1.2.2 Composite Reference Standard

Alonzo et al. (1999) discussed the composite reference standard (CRS) that combines the results of several imperfect reference tests to define a pseudo-gold standard based on some predefined rule. It is assumed that a composite reference standard works better than each single test by itself (Martin et al., 2004). The development of a CRS depends significantly on the target diseases. Investigators may adjust the threshold used to define a disease for the specific clinical problems encountered. A typical example is when any of the reference test results are positive; in that case, disease status will be labeled as present; conversely, if all test results are negative, the disease is considered as absent. Thus, any patient with positive results for the first reference test does not have to be retested by other reference tests. However, the simple decision rule is prone to misclassification bias. Ideally, different tests should be targeted on the same disease status with different error rates. It is problematic when these tests define the disease status differently, but the decision rule treats them equally.

1.2.3 Expert Review Panel

When no reference standard is generally accepted, an expert review panel may reach a consensus diagnosis concerning the status of a subject with respect to a specific disease using assorted information from different sources, including symptoms/signs, physical characteristics, medical history, clinical follow-ups, and imperfect reference tests. Test results that are undergoing evaluation usually are not presented to the panel for review in order to avoid the incorporation of bias that could lead to an overestimation of the accuracy of the diagnosis. Thornbury et al. (1993) demonstrated a gold standard panel of neurosurgeons, neurologists, and physicians experienced in technology assessment for diagnosis of patients receiving MRI and CT for acute low-back pain. They eliminated bias by using a diagnosis that was independent of the diagnostic test being evaluated. Another practice is to present the results of the diagnostic test that is being evaluated to the panel after they make a decision about the diagnosis, and then determine whether these results would change their opinion. The Delphi method is a formal procedure that collects and integrates the opinions of each individual panel member in an anonymous way to avoid influence among fellow members or from a dominating member (Jones and Hunter, 1995).

1.2.4 Latent Class Analysis (LCA)

Latent class analysis (LCA) is a group of methods that combine information from multivariate categorical data to investigate the existence of unobserved heterogeneity in groups or subtypes of cases. The latent variables can only be evaluated indirectly through information collected from observable measurements, which are referred to as manifest variables. LCA has application in many fields such as marketing/survey research, sociology, and psychology. It gained popularity in the evaluation of the diagnostic accuracy of multiple tests in the absence of a gold standard. The latent variable

in this case is the unobservable, true status of the disease, whereas the manifest variables are the test results.

Estimates of the parameters of diagnostic accuracy can be derived with maximum likelihood (ML) based methods involving iterative computation, such as expectation-maximization (EM) algorithms (Dempster et al., 1977), Fisher scoring (Espeland and Handelman, 1989), and the Newton-Raphson method (Qu and Hadgu, 1998). ML approaches provide a unified framework for various latent class models (LCMs), including the Hui-Walter reference-free method (Hui and Walter, 1980), a Gaussian random effects model (Qu et al., 1996), a marginal model (Yang and Becker, 1997), a joint model approach (Albert, 2009), a finite mixture model (Albert et al., 2004), a model that incorporates multiple latent variables and covariates (Huang and Bandeen-Roche, 2004), a probit latent class (PLC) model with a general correlation structure (Xu and Craig, 2009), and a Bayesian model using the Gibbs sampler algorithm to approximate marginal posterior densities of all parameters (Joseph et al., 1995). In the absence of a gold standard, LCMs have been well reviewed (Walter and Irwig, 1988; Goetghebeur et al., 2000).

Under the conditional independence assumption, multiple tests applied to the same subject are assumed to be independent conditional on his/her true disease status. In other words, if the true disease status is misclassified by one test, the probability that it will be misclassified by another test will not be affected. Conditional independence assumption is plausible when different tests are based on different scientific/technological grounds or when they measure different characteristics of the disease. In reality, though, this assumption is often impractical when results from multiple tests are similar due to some other latent effects other than disease status, e.g., similar severities/stages of the disease, similar biological basis of the tests, similar subject-specific characteristics (e.g.,

age and gender), and similar training/experience of those who rate the test. The earliest work on LCA relied on the conditional independence assumption (Hui and Walter, 1980; Walter and Irwig, 1988; Rindskopf and Rindskopf, 2006). Torrance-Rynard and Walter (1998) found that LCA under the conditional independence assumption often handled conditional correlated tests well and produced relatively unbiased estimates of diagnostic accuracy. However, for diseases with very low prevalence or for tests with very low specificity, the estimates could be seriously biased with slightly correlated tests. Many studies have shown that ignoring the correlation of misspecification errors between tests led to biased estimates of diagnostic accuracy when the conditional independence assumption does not hold (Thibodeau, 1981; Vacek, 1985; Hui and Zhou, 1998). Most of the recently developed LCA methods relax the conditional independence assumption (Joseph et al., 1995; Qu et al., 1996; Yang and Becker, 1997; Xu and Craig, 2009; Albert, 2009). Albert and Dodd (2004) showed that estimates of diagnostic accuracy and prevalence are sensitive to the choice of dependence structure. The correct dependence structure between tests often is hard to specify, and estimates of the parameters can be biased if the structure is misspecified. To better distinguish the dependence structures between different latent classes, they strongly recommended a large number of tests (ideally 10 or more).

One drawback of LCMs is model non-identifiability (Goodman, 1974), which occurs due to either poor specification of the model (intrinsic non-identifiability) or certain unexpected structures of the observed data (empirical non-identifiability). Empirical non-identifiability usually is concerned with small sample sizes and sparse data. Intrinsic non-identifiability occurs when the number of tests is small. Dendukuri and Joseph (2001) described the problem of intrinsic non-identifiability from "ill-defined" models with less than four tests, resulting in available degrees of freedom not being large enough to handle the number of parameters to be estimated. A simple way to

get around intrinsic non-identifiability is to add a plausible restriction to the model, e.g., setting equal sensitivities and specificities for the two tests. Bayesian methods are particularly useful for ill-defined models. The unknown parameters are treated as random variables with a prior distribution that incorporates pertinent information, such as estimates from similar studies, expert's opinions about the characteristics of the test, and demographic data about patients or the study population. The prior distribution is updated with new information from observed data to derive the posterior distribution for each parameter. Then, point estimates along with the highest posterior density (HPD) credible sets for diagnostic accuracy are obtained from the posterior distribution. The disadvantages of the ML approaches are that Bayesian approaches usually are intensive computationally, and they are sensitive to the prior distribution that is chosen.

Despite the popularity of LCA, it has been cautioned that estimates of diagnostic accuracy are subject to bias if the underlying assumptions of the model cannot be justified (Albert and Dodd, 2004; Pepe and Janes, 2007; Bertrand et al., 2005). Another concern with LCA is that the true disease status is defined mathematically rather than clinically, which results in scientific doubt among some clinicians about the meaning of the resulting estimates (Pepe, 2004).

1.3 Evaluation with a Partially-Missing Gold Standard

Even when a gold standard exists, it may be too invasive and/or costly to be applied to all study subjects. In practice, it is common that subjects who appear to have a high risk of disease are treated with the gold standard, whereas subjects who have lower risk of disease are not. One typical scenario that occurs with screening studies is that subjects whose results are negative in the screening test may forgo the more invasive/costly gold standard. Sometimes a less-desirable, imperfect reference test

(often less expensive/invasive) is feasible rather than the gold standard. This section covers the evaluation of diagnostic accuracy when only a subset of subjects receives the gold standard.

1.3.1 Case-Deletion Approach

The case-deletion approach excludes all subjects from analysis who were not tested with a gold standard. This method could drastically reduce the power and cause partial verification bias. Sensitivity tends to be overestimated, and specificity tends to be underestimated. The direction and magnitude of the biases are affected by multiple factors, such as the proportion of missing tests, the extent to which the true disease status is dependent on or independent of the missing tests, and the unobserved ratio of positive results to negative results for the missing tests. The case-deletion approach should be avoided unless the missing proportion is very small.

1.3.2 Correction Methods

Verification bias occurs when the decision concerning the use of the gold standard is influenced by test results or clinician/patient decisions. Correction methods apply a mathematical correction to the biased estimates of diagnostic accuracy using models based on certain assumptions related to missing data (Baker, 1995; Zhou, 1998; Schneeweiss, 2000; Hadgu et al., 2005; Alonzo, 2005). Conditional independence between tests and the true disease status are common assumptions for correction methods (Begg and Greenes, 1983). Brenner (1996) reinforced the correction method with consideration of correlated classification errors between tests conditional on the true disease status. The main limitation is overcorrection. Begg and Greenes (1983) found that their correction method underestimated sensitivity and overestimated specificity. Wacholder et al. (1993) examined the performance of correction methods when the correlation of

classification errors between tests was misspecified and found that the bias of adjusted estimates could be worse than the bias of unadjusted estimates.

1.3.3 Imputation Methods

Imputation methods replace missing values with substituted values, e.g., the arithmetic means of available cases, predicted values from regression equations, observations from subjects with similar response profiles, and observations that immediately precede dropout in a longitudinal design (a.k.a. last observation carried forward (LOCF)). More complex imputation models incorporate additional information, such as symptoms, morbidity, and the patient's characteristics. Single imputation generates a single replacement value for each missing observation, whereas multiple imputation (MI) replaces each missing value with a set of plausible values that represent the uncertainty about the true value that should be imputed. An inappropriate imputation model, which can result from an improper assumption about a missing pattern, will lead to biased estimates. Multiple imputation is generally preferred over single imputation because it adjusts the standard errors for missing data and is less likely to produce biased estimates of parameter. Harel and Zhou (2007) reviewed the theory and application of MI in a tutorial. The selection of an imputation model is subject to the missing data assumptions. When missing not at random (MNAR) is tenable, it is difficult to identify a suitable imputation model, and the models that are used often have questionable validity. Harrell et al. (1996) found that imputation methods as well as correction methods require large sample sizes to model the data. Harel and Zhou (2006) noted that multiple imputation methods are more robust than correction methods for small sample size.

1.3.4 Expectation-maximization Algorithm

The EM algorithm, formalized by Dempster et al. (1977), has been used in many fields, including diagnostic medicine. The DLR paper stimulated great interest in the use of finite mixture distributions for modeling heterogeneous data. Finite mixture or unobserved heterogeneity models are one type of LCM that assumes that the observations of one sample arise from a mixture of two or more unobserved classes of unknown proportions (Day, 1969). Determining ML estimates (MLEs) of mixture models with incomplete data is simplified substantially by the EM algorithm. McLachlan and Peel (2000) reviewed the application of finite mixture models. Dawid and Skene (1979) first used the EM algorithm to determine MLEs of observed error rates when a gold standard was not available, and they regarded the latent disease status as a missing value.

The EM algorithm has an inherent advantage over other methods in that it takes into consideration missing gold standards as well as missing imperfect diagnostic tests. To our knowledge, there is very little literature on evaluating multiple imperfect diagnostic tests with missing data and without a gold standard. For our case study, only a small proportion (6.3%) of the $3,487$ subjects has been tested with all 11 biomarker tests. The latent true disease status (due to the absence of a gold standard), along with very high proportions of missing tests, have posed statistical challenges related to the evaluation of their diagnostic accuracy and motivated our research utilizing EM algorithm-based approaches, which will be discussed fully in later chapters.

1.4 Outline for the Dissertation

The organization of the dissertation is as follows. In Chapter 2, we introduce a traditional latent class model based on the EM algorithm to evaluate incomplete multiple

imperfect diagnostic tests in the absence of a gold standard under the conditional independence assumption. We applied the proposed method to a real data set from the NCI Colon Cancer Family Registry (C-CFR) on diagnosing MSI for hereditary nonpolyposis colorectal cancer (HNPCC). Estimates of diagnostic accuracy, prevalence, and differential missing probabilities for the eleven biomarker tests were obtained. Simulations also were conducted to evaluate the small-sample performance of our methods and the advantages and limitations of our methods are discussed. In Chapter 3, we relaxed the conditional independence assumption and extended an improved probit latent class (PLC) model to evaluate incomplete multiple imperfect diagnostic tests under the conditional dependence assumption. We applied a parameter-expanded Monte Carlo EM (PX-MCEM) algorithm to the C-CFR data to derive point estimates of the model parameters, and we used the bootstrap method to obtain their standard errors. The validity of inference is demonstrated with extensive simulation studies. In Chapter 4, we present **DiagLCA** as the first R package for evaluation of multiple correlated diagnostic tests with abundant missing data and without a gold standard. Finally, Chapter 5 concludes with a brief summary, discussion, and recommendations for future research work.

Chapter 2

Conditional Independence Assumption

2.1 Introduction

Accurate diagnosis of a disease or classification of a sub-type of a disease is often the first step toward its treatment and prevention. Multiple imperfect biomarker tests may be used when a gold standard test does not exist. It can be considered a missing data issue where the gold standard (i.e. the true disease status) is always missing. A considerable methods are developed to assess the diagnostic accuracy (usually quantified by sensitivity and specificity) of "index tests" (the tests whose performance is under evaluation) in the absence of a gold standard (Alonzo et al., 1999; Hadgu et al., 2005; Enøe et al., 2000). These methods either resort to some imperfect (non-gold) reference standards, or utilize all index tests simultaneously in a unified manner if there is no accepted reference standard. Among them latent class analysis (LCA) methods are popular by treating unobservable true disease status as a latent variable. The parameter estimates of diagnostic accuracy can be derived through Bayesian approaches (Joseph et al., 1995; Dendukuri and Joseph, 2001), or through maximum likelihood (ML) approaches involving iterative computation such as the Expectation-maximization (EM)

algorithm (Dempster et al., 1977; Dawid and Skene, 1979), Fisher scoring (Espeland and Handelman, 1989), and Newton-Raphson method (Qu and Hadgu, 1998). The ML approaches provide a unified framework for various latent class models (LCMs) in a dispersed literature (Chu et al., 2009; Hui and Walter, 1980; Qu et al., 1996; Yang and Becker, 1997; Albert et al., 2004; Albert, 2009; Huang and Bandeen-Roche, 2004; Xu and Craig, 2009). Walter and Irwig (1988) and Goetghebeur et al. (2000) reviewed LCMs based on ML approaches.

Heretofore the LCMs merely handle latent disease status, whereas the imperfect tests typically have no missing value. However, missing data is ubiquitous in diagnostic medical settings where some subjects are only measured by a subset of tests. A commonly reported missing data issue deals with a partially missing gold standard (e.g. patients with negative results are more likely to skip the gold standard due to cost/invasiveness). General methods have been established to deal with missing data under different missing pattern assumptions (Little and Rubin, 1987). A missing data mechanism is called missing completely at random (MCAR) if the probability of missingness does not depend on any missing or observed observations. A mechanism is said to be missing at random (MAR) when the probability of missingness does not depend on any missing observation conditional on some observed observations. Both MCAR and MAR are considered "ignorable" missing data mechanism, as MCAR is a special case of MAR. When the probability of missing depends on some missing observations or latent disease status, it is known as a missing not at random (MNAR) mechanism or "non-ignorable" (NI) missing data mechanism. There are various references dealing with ignorable (Alonzo, 2005; Begg and Greenes, 1983; Harel and Zhou, 2007; He and McDermott, 2012; Lin et al., 2006; Yu et al., 2010; Zhou, 1998) and non-ignorable (Baker, 1995; Harel and Zhou, 2007; Kosinski and Barnhart, 2003b,a; Zhou, 1993) missing gold standards. When a gold standard is always present (i.e. true disease status

always known), Poleto et al. (2011) proposed a two-stage hybrid procedure (ML in the first stage; weighted least squares in the second stage) in estimating the diagnostic accuracy of three index tests with abundant missing data under the MCAR/MAR assumptions.

The aforementioned literature deals with two scenarios: (i) a gold standard does not exist whereas the index tests / imperfect reference standards have no missing value; (ii) a gold standard is available (partially or fully observed). To the best of our knowledge, when a gold standard does not exist, there are no studies on handling missing test results from multiple imperfect diagnostic tests to facilitate the estimation of their diagnostic accuracy. For the Colon Cancer Family Registry (C-CFR) study that we will consider (see Section 2), only a small proportion (6.3%) of the $3,487$ subjects has been tested for all 11 biomarkers. The latent true disease status along with very high proportions of missing test results have created statistical challenges for evaluating the population prevalence and the estimates of diagnostic accuracy for the 11 biomarkers. Motivated by the C-CFR case study, we develop a LCM to handle the general case where some subjects may only be tested by a subset of markers. It accounts for missing data under the MAR or MNAR assumptions and unobservable latent disease status simultaneously. Specifically, it allows for differential missing probability of each test to depend on latent disease status under one MNAR scenario. The C-CFR study is described in Section 2. Section 3 introduces the statistical methods. Full analysis of the case study (Section 4) and simulation studies (Section 5) are summarized next. Section 6 presents a brief discussion.

2.2 Colon Cancer Family Registry Study

Colorectal cancer is the fifth most common form of cancer in the United States and the third leading cause of cancer-related death in the Western world (based on

statistics from NCI and WHO websites). In the United States, about 15,000 new cases of colorectal cancer are diagnosed each year (Ford and Whittemore, 2006). About two to five percent of all colon cancer cases are attributed to hereditary nonpolyposis colorectal cancer (HNPCC), also called Lynch Syndrome after Dr. Henry Lynch. HNPCC is a hereditary syndrome that is caused by a mutation in genes involved in the DNA mismatch repair pathway. People with HNPCC have a much higher risk of developing colon cancer than the general population if they do not undergo early and regular screening. The average age of diagnosis of cancer in patients with HNPCC is 44 years, as compared to 64 years in people without the syndrome (Lynch and de la Chapelle, 1999; DeFrancisco and Grady, 2003).

Microsatellites are common and normal repeated sequences of DNA. Although the length of microsatellites is highly variable from person to person, each individual has microsatellites of set length. In cells with mutations in DNA repair genes, however, some of these sequences accumulate errors and become longer or shorter. The appearance of such long or short microsatellites in an individual's DNA is referred to as microsatellite instability (MSI). MSI is a key factor in several cancers including colorectal, endometrial, ovarian and gastric cancers. Cancers with MSI account for approximately 15% of all colorectal cancers and for HNPCC germline mutations (Boland et al., 1998; Umar et al., 2004; Lynch and de la Chapelle, 1999). The diagnosis of HNPCC may be determined if the cancer exhibits a high level of MSI. People with HNPCC have a much higher risk of developing colon cancer than the general population if they do not undergo early and regular screening.

Our methodology research is motivated by a real study from the NCI Colon Cancer Family Registry (C-CFR). It is an international consortium of six centers located in North America and Australia formed to support studies on the etiology, prevention and clinical management of colorectal cancer (Newcomb et al., 2007). The C-CFR

data includes diagnostic test results of eleven molecular biomarkers (*BAT25*, *BAT26*, *BAT40*, *BAT34C4*, *D10S197*, *D17S250*, *D18S55*, *D2S123*, *D5S346*, *ACTC* and *MYCL*) which are used to assess the level of MSI.

In this paper, we consider the $3,487$ subjects from families with a single subject per family. The observed missing proportions range from 7.26% for biomarker *BAT26* to 81.73% for biomarker *D2S123*. Table 2.2 summarizes the number of subjects by frequency of missing test results for the 11 biomarkers. Most of the subjects (93.72%) have at least one test result missing. Specifically, the "missing" category includes the following categories defined by the Colon CFR code book: a) quantity of DNA or tissue not sufficient (code 13); b) not tested, reason not specified (code 12); c) no amplification (code 11); d) equivocal (inconclusive, code 6); and e) normal DNA not used in test (code 9). The high proportions of missing and latent disease status have motivated our methodology research which we introduce in the next section.

2.3 Statistical Methods

The total number of subjects is $N = 3,487$ and the total number of biomarkers is $J = 11$. Let $D_i = d(d = 1, 0)$ denote latent disease status (whether the i^{th} subject has disease or not, 1=Yes, 0=No). Let $\pi_d = Pr(D_i = d)$ represent probability of disease/no disease ($\pi_1 = Pr(D_i = 1)$ is prevalence, $\pi_0 = 1 - \pi_1$); Se_j represent sensitivity of the j^{th} biomarker; Sp_j represent specificity of the j^{th} biomarker. Let $T_i = (t_{i1}, \cdots, t_{iJ})$ be the collection of all test results of the i^{th} subject ($t_{ij} = 1$ for positive and $t_{ij} = 0$ for negative). Under a conditional independence assumption, the multiple test results of T_i are independent given D_i. Let $\Delta_i = (\delta_{i1}, \cdots, \delta_{iJ})$ be the collection of all missing indicators of the i^{th} subject, where δ_{ij} is indicating whether the subject has been tested by the j^{th} biomarker ($\delta_{ij} = 1$ for tested and $\delta_{ij} = 0$ for not tested). Let T_i^{obs} denote the observed tests for subject i.

2.3.1 Diagnostic Performance under a MAR Assumption

We first assume MAR for the missing data mechanism. The probability of observing (T_i^{obs}, Δ_i) can be expressed as a finite mixture of two components

$$\begin{aligned} P(T_i^{obs}, \Delta_i) &= P(\Delta_i | T_i^{obs}) P(T_i^{obs}) \\ &\propto P(T_i^{obs}) \\ &= \sum_{d=0}^{1} \pi_d h_{id} \end{aligned}$$

where

$$\begin{aligned} h_{i1} &= \prod_{j=1}^{J} Se_j^{t_{ij}\delta_{ij}} (1 - Se_j)^{(1-t_{ij})\delta_{ij}} \\ h_{i0} &= \prod_{j=1}^{J} (1 - Sp_j)^{t_{ij}\delta_{ij}} Sp_j^{(1-t_{ij})\delta_{ij}}. \end{aligned}$$

The parameters $\theta = (\pi_1, Se_1, \cdots, Se_J, Sp_1, \cdots, Sp_J)$ can be estimated by the EM algorithm, where the complete data $Y_i = (T_i^{obs}, \Delta_i, D_i) \sim p_\theta(y_i) = (\pi_1 h_{i1})^{d_i} (\pi_0 h_{i0})^{1-d_i}$. The complete log-likelihood is $logL_c(\theta) = \sum_{i=1}^{N} (d_i log(\pi_1 h_{i1}) + (1 - d_i) log(\pi_0 h_{i0}))$. Thus the M-step is to solve the score equations

$$\begin{aligned} 0 &= \sum_{i=1}^{N} E[\{\frac{d_i}{\pi_1} - \frac{1 - d_i}{1 - \pi_1}\} | Y_i, \theta^{(n)}] \\ 0 &= \sum_{i=1}^{N} E[\{\frac{d_i t_{ij} \delta_{ij}}{Se_j} - \frac{d_i \delta_{ij} - d_i t_{ij} \delta_{ij}}{1 - Se_j}\} | Y_i, \theta^{(n)}] \\ 0 &= \sum_{i=1}^{N} E[\{\frac{(1 - d_i)(1 - t_{ij})\delta_{ij}}{Sp_j} - \frac{(1 - d_i) t_{ij} \delta_{ij}}{1 - Sp_j}\} | Y_i, \theta^{(n)}] \end{aligned}$$

where $\theta^{(n)} = (\pi_1^{(n)}, Se_1^{(n)}, \cdots, Se_J^{(n)}, Sp_1^{(n)}, \cdots, Sp_J^{(n)})$. The E-step computes $E[d_i | Y_i, \theta^{(n)}] = \frac{\pi_1^{(n)} h_{i1}^{(n)}}{\sum_{d=0}^{1} \pi_d^{(n)} h_{id}^{(n)}}$. Thus we can get closed-form solutions for $\pi_1^{(n+1)}$, $Se_j^{(n+1)}$, $Sp_j^{(n+1)}$

and iterate EM steps until convergence to get $\hat{\theta}$.

Because analytical evaluation of the second-order derivatives of the incomplete-data log-likelihood $logL(\theta)$ is difficult, Louis's formula (1982) can be used to obtain the observed information matrix of the MLE obtained via the EM algorithm. Once the information matrix $I(\hat{\theta}; T_i^{obs}, \Delta_i)$ is derived and inverted, the standard errors are the square root of the diagonal elements of the inverse matrix. Refer to Appendix 1 for more details on derivation of information matrices for Louis Formula.

2.3.2 An Extension to One MNAR Scenario

Now we consider the situation when missingness only depends on unobserved disease status. That is, we allow the missing pattern to differentiate between diseased and non-diseased populations. We note that under this situation, MAR no longer holds. However, our previous method can be easily generalized to estimate diagnostic measures under this MNAR scenario. Let r_{1j} be the missing probability of the j^{th} biomarker when the subject has disease; r_{0j} be the missing probability of the j^{th} biomarker when the subject is disease free. The probability of observing (T_i^{obs}, Δ_i) is

$$P(T_i^{obs}, \Delta_i) = \sum_{d=0}^{1} \pi_d h_{id} s_{id} \tag{2.1}$$

where π_d and h_{id} are same as before and s_{id} accounts for missing probabilities:

$$s_{id} \;=\; \prod_{j=1}^{J} r_{dj}^{\delta_{ij}} (1 - r_{dj})^{(1-\delta_{ij})}.$$

The complete data is $Y_i = (T_i^{obs}, \Delta_i, D_i) \sim p_\theta(y_i) = (\pi_1 h_{i1} s_{i1})^{d_i} (\pi_0 h_{i0} s_{i0})^{1-d_i}$. Its log-likelihood becomes $logL(y_i) = \sum_{i=1}^{N} (d_i log(\pi_1 h_{i1} s_{i1}) + (1 - d_i) log(\pi_0 h_{i0} s_{i0}))$. In the likelihood expression, s_{id} can be integrated with h_{id} so that $h_{id} s_{id}$ become our new h_{id}.

Hence estimates and 95% confidence intervals (CIs) of missing probabilities allowing for the MNAR assumption, r_{1j} and r_{0j}, can be derived as $(1 - Se_j)$ and Sp_j respectively.

We can further test the null hypothesis that the missingness does not depend on the disease status. Particularly, we apply the likelihood ratio test (LRT) to test the hypotheses $H_0 : r_{1j} = r_{0j}$ for $j = 1, \cdots, J$ vs. $H_1 : r_{1j} \neq r_{0j}$, for at least some of $j = 1, \cdots, J$. The log-likelihood under H_1 is $l(\hat{\theta}|T, \Delta) = \sum_{i=1}^{N} log(\pi_1 h_{i1} s_{i1} + (1-\pi_1) h_{i0} s_{i0})$, i.e. the log-likelihood of observed data (incomplete without knowing D) and $\hat{\theta}$ (the estimates for π_1, Se_j, Sp_j, r_{1j} and r_{0j} under H_1). Under H_0, the log-likelihood with observed data is $l(\tilde{\theta}|T, \Delta) = \sum_{i=1}^{N} log(\pi_1 h_{i1} s_i + (1-\pi_1) h_{i0} s_i)$, where $\tilde{\theta}$ are the estimates for π_1, Se_j, Sp_j, and r_j under H_0.

2.3.3 Model checking using Kappa statistics

We adopt the Kappa agreement plot (Chu et al., 2009) as a simple graphical method to quantitatively check the conditional dependence assumption based on the final model. The model based Kappa statistics for any two of the 11 tests are

$$\kappa_{ij} = \frac{P_{ij11} + P_{ij00} - (P_{ij11} + P_{ij10})(P_{ij11} + P_{ij01}) - (P_{ij00} + P_{ij10})(P_{ij00} + P_{ij01})}{1 - (P_{ij11} + P_{ij10})(P_{ij11} + P_{ij01}) - (P_{ij00} + P_{ij10})(P_{ij00} + P_{ij01})}$$

where for two tests i and j

$$\begin{aligned}
P_{ij11} &= \pi_1 Se_i Se_j + (1-\pi_1)(1-Sp_i)(1-Sp_j) \\
P_{ij10} &= \pi_1 Se_i (1-Se_j) + (1-\pi_1)(1-Sp_i) Sp_j \\
P_{ij01} &= \pi_1 (1-Se_i) Se_j + (1-\pi_1) Sp_i (1-Sp_j) \\
P_{ij00} &= \pi_1 (1-Se_i)(1-Se_j) + (1-\pi_1) Sp_i Sp_j.
\end{aligned}$$

Plug in $\hat{\pi}_1$, $\widehat{Se_j}$, $\widehat{Sp_j}$, we have the estimates $\hat{\kappa}_{ij}$ for model based Kappa statistics.

The Kappa agreement plot can be obtained by plotting $\hat{\kappa}_{ij}$ with 95% simultaneous confidence intervals (correcting possible agreement by chance) vs. the observed Kappa for each pair of tests. There is not enough evidence to reject the conditional independence assumption if the model based 95% simultaneous confidence intervals contain the observed Kappa statistics at close to the nominal rate.

2.4 Analysis of the Colon Cancer Family Registry

Applying the proposed method to the C-CFR data we obtained the point estimates for θ, which includes π_1, Se_j, Sp_j as well as r_{1j}, r_{0j} for all 11 biomarkers, and their standard errors (SE). The 95% CIs are then obtained from point estimates and SEs. The results for θ under different missing assumptions are presented in Table 2.3. The point estimates with 95% CIs under the MNAR assumption are almost the same with those under the MAR assumption. The slight difference is caused by introducing columns of 1 for missing indicator of δ_{ij} for our method.

The estimates for missing probabilities are presented in Table 2.6. Under the MNAR assumption, we can see that for some tests r_{1j} and r_{0j} are quite different while for others they are very close. Generally r_{1j} tends to be greater than r_{0j} (the only exception is for *MYCL*). The reason is probably that non-diseased patients are more likely to have negative test results, and patients with negative test results are more willing to take additional tests (hence smaller missing probabilities) to confirm they do not have the disease. As expected, r_j under the MAR assumption falls between r_{1j} and r_{0j}.

The log-likelihood ratio statistic (LRS) is found to be $-2(l(\tilde{\theta}|T,\Delta) - l(\hat{\theta}|T,\Delta))= -2 \times (-21407.85 - (-21166.83))= 482.04 > \chi^2_{0.95,11}$, where $\chi^2_{0.95,11} = 19.675$ is the 95^{th} percentile of the the χ^2 distribution with d.f.=11. Therefore we reject the null hypothesis and conclude that at least for some tests, the missing probabilities of those who have colon cancer are significantly different from those who don't have

colon cancer. To tell whether r_{1j} and r_{0j} are significantly different for each test j, we calculated p-values based on Wald statistic, which is calculated as $\frac{r_{1j}-r_{0j}}{\text{SE}_{r_{1j},r_{0j}}}$ where $\text{SE}_{r_{1j},r_{0j}} = \sqrt{\text{Var}_{r_{1j}} + \text{Var}_{r_{0j}} - 2\text{Cov}_{r_{1j},r_{0j}}}$. These nominal p-values can be adjusted for multiplicity using Bonferroni correction (multiply by $J = 11$). The results suggest that the MNAR assumption is more plausible than the MAR assumption for our case study.

Figure 2.1 plots the model based Kappa statistics versus observed Kappa statistics as a graphical check for conditional independence assumption. The model based Kappa statistics in Figure 2.2(a) are derived from diagnostic accuracy estimates under the MAR assumption, and the model based Kappa statistics in Figure 2.2(b) are from the MNAR assumption. Not surprisingly, these two figures look similar due to the resemblance of diagnostic accuracy estimates under the two assumptions. In both figures, all of the 95% simultaneous confidence intervals of model-based Kappa contain the observed Kappa statistics. Hence these figures fail to reject the null hypothesis of conditional independence assumption.

2.5 Simulation Studies

To further investigate the performance of the proposed methods, 10000 simulations were run. For each simulation, we generate $N = 3,500$ observations (chosen to be close to the sample size of the real data, $N = 3,487$) with $J = 5$ tests including missing indicators for each test under three different missing pattern assumptions. The true values of prevalence is set to be $\pi_1 = 0.2$, close to the estimates from the C-CFR study. The true values of sensitivity and specificity are set to be close to estimates for the five NCI-recommended microsatellite sequence panels (*BAT25*, *BAT26*, *D17S250*, *D2S123*, *D5S346*), i.e. we let $Se = (0.9, 0.9, 0.8, 0.8, 0.6)$ and $Sp = (0.9, 0.9, 0.9, 0.9, 0.9)$. For each subject, disease status D_i is randomly assigned $0, 1$ from a binomial distribution with π_1 as the binomial p. Given a subject's D_i, test results T_i are randomly assigned

$0, 1$ from a binomial distribution with Se and $(1 - Sp)$ as the binomial p under the assumption that T_i are independent given D_i. Finally, we randomly assign missing indicator Δ_i to each test results considering the three missing scenarios. Simulation results are listed in Table 2.1.

We first simulated the missingness of test results under the MCAR assumption. All five tests were randomly assigned a missing indicator ($1 = observed; 0 = missing$), through a binomial distribution, independent of any other missing or observed values. The missing probabilities of the five tests $(r_1, \cdots, , r_5)$ are 0.274, 0.090, 0.251, 0.023, 0.089 subsequently, which were randomly selected from a uniform distribution bounded by $(0, 0.35)$. Secondly, under the MAR assumption, we let the missing patterns of Test 4 and Test 5 depend on some fully-observed test results. Δ_{i1} and Δ_{i2} all equal to 1 (the first two tests have no missing value); Δ_{i3} are randomly generated from a binomial distribution with a missing probability of 0.2 (the third test is MCAR); Δ_{i4} is randomly generated from a binomial distribution with missing indicator Δ_{i4} depending on T_{i1} through equation $logit(P(\Delta_{i4} = 1|T_{i1})) = \alpha + \beta_1 T_{i1}$; and Δ_{i5} is randomly generated from a binomial distribution with missing indicator Δ_{i5} depending on T_{i1} and T_{i2} through equation $logit(P(\Delta_{i5} = 1|T_{i1}, T_{i2})) = \alpha + \beta_1 T_{i1} + \beta_2 T_{i2}$. The regression coefficients are set to $\alpha = -1.5$, $\beta_1 = -4$, $\beta_2 = 2$. Lastly, for the MNAR assumption, Test 4 and Test 5 depend on a subset of unobserved tests or disease status. Δ_{i1}, Δ_{i2} and Δ_{i3} are generated following the same settings as for the MAR assumption. Δ_{i4} is randomly generated from a binomial distribution with missing indicator Δ_{i4} depending on partially-observed T_{i3} through equation $logit(P(\Delta_{i4} = 1|T_{i1}, T_{i3})) = \alpha + \beta_3 T_{i3}$, and Δ_{i5} is randomly generated from a binomial distribution with missing indicator Δ_{i5} depending on both T_{i3} and unobservable disease status D through equation $logit(P(\Delta_{i5} = 1|T_{i3}, D)) = \alpha + \beta_3 T_{i3} + \beta_5 D$, where $\alpha = -1.5$, $\beta_3 = 0.75$, $\beta_5 = 0.7$.

Under the MCAR assumption, the EM algorithm is very robust with all coverage

Table 2.1: Summary of EM Estimates and Coverage of 95% CIs for Simulated Data under Different Missing Pattern Assumptions

			MCAR			MAR			MNAR		
Parameter	Marker	True	Mean (SD)	SE	CP	Mean (SD)	SE	CP	Mean (SD)	SE	CP
Prevalence		0.2	0.200 (0.0078)	0.0078	95.7%	0.200 (0.0077)	0.0076	95.4%	0.199 (0.0077)	0.0077	95.4%
Sensitivity	Test 1	0.9	0.899 (0.0167)	0.0168	95.5%	0.899 (0.0155)	0.0149	94.7%	0.902 (0.0150)	0.0148	95.6%
	Test 2	0.9	0.899 (0.0157)	0.0157	95.2%	0.900 (0.0143)	0.0140	95.0%	0.903 (0.0152)	0.0147	95.3%
	Test 3	0.8	0.800 (0.0201)	0.0203	95.6%	0.800 (0.0189)	0.0190	95.6%	0.802 (0.0189)	0.0192	96.3%
	Test 4	0.8	0.800 (0.0181)	0.0183	95.7%	0.800 (0.0174)	0.0173	95.2%	0.801 (0.0208)	0.0206	95.3%
	Test 5	0.6	0.601 (0.0216)	0.0212	95.2%	0.600 (0.0205)	0.0204	95.1%	0.593 (0.0271)	0.0270	94.4%
Specificity	Test 1	0.9	0.900 (0.0073)	0.0073	95.7%	0.900 (0.0061)	0.0061	95.9%	0.899 (0.0063)	0.0063	95.5%
	Test 2	0.9	0.900 (0.0067)	0.0066	95.8%	0.900 (0.0063)	0.0063	96.0%	0.899 (0.0063)	0.0063	95.7%
	Test 3	0.9	0.900 (0.0070)	0.0070	95.7%	0.900 (0.0067)	0.0067	96.0%	0.899 (0.0069)	0.0068	95.5%
	Test 4	0.9	0.900 (0.0063)	0.0062	96.0%	0.900 (0.0067)	0.0066	95.7%	0.900 (0.0067)	0.0067	95.7%
	Test 5	0.9	0.900 (0.0063)	0.0062	95.7%	0.900 (0.0066)	0.0066	95.8%	0.901 (0.0066)	0.0065	95.4%

True = True Parameter Values
SD = Standard Deviation
CP = Coverage Probability

Note: Each simulation study consists of $10,000$ simulations. Each simulation generated a simulated data set with $N = 3,500$ subjects and $J = 5$ tests.

probabilities above 95%. The SE estimates are also very close to the standard deviations from point estimates of π_1, Se and Sp. The average missing probability for Test 1 under MCAR assumption is the highest (almost 30% observations missing), yet the coverage probability is still very good ($> 95\%$) for its sensitivity and specificity. Under the MAR assumption, the coverage probabilities for π_1 and Se is a bit poorer than the MCAR assumption albeit the coverage probabilities for Sp is almost undifferentiable between the two tests. The missing probabilities under the MAR assumption vary between $[0.004, 0.182]$ for Test 4, and $[0.004, 0.622]$ for Test 5. The average missing probabilities under the MAR assumption are 0.136 for Test 4 and 0.183 for Test 5. Although the missing probabilities for Test 5 can be as high as 0.622, the coverage probability for Test 5 is very good ($> 95\%$) for both sensitivity and specificity. Our method is well suitable for both the MCAR and MAR assumptions. Under the MNAR assumption, the missing probabilities vary between $[0.182, 0.321]$ for Test 4, and $[0.182, 0.488]$ for Test 5. The average missing probabilities under the MNAR assumption are 0.217 for Test 4 and 0.250 for Test 5, both greater than those under the MAR assumption. Generally speaking the coverage probabilities under the MNAR assumption are not compromised much as all are above 94%. We do observe that the coverage probability for Sp is consistently smaller comparing to the MAR assumption. This may be explained by the missingness of Test 4 only depends on Test 3 (which is MCAR), while the missingness of Test 5 depends on latent disease status in addition to Test 3. As a sensitivity analysis, we did another simulation for the MNAR assumption with everything the same except that Δ_{i5} depends on T_{i3} and T_{i5} through equation $logit(P(\Delta_{i5} = 1|T_{i3}, T_{i5})) = \alpha + \beta_3 T_{i3} + \beta_5 T_{i5}$. Our method did not handle Test 5 estimation very well: although both Test 4 and Test 5 are MNAR, the coverage for Test 4 is barely affected while the coverage for Test 5 is seriously impacted (coverage probability 26.4% for sensitivity and 37.4% for specificity). These results imply that MNAR may not be a serious issue

when the cause of a missing test lies in the value of another test or the true disease status, whereas MNAR can cast doubt on our estimation if the cause of a missing is the value of the missing test itself.

We also conducted a simulation to assess estimation of missing probabilities under the new assumption for MNAR, where the missingness of each test depends on the unknown disease status. The missing probabilities for Test 1 through Test 5 were subsequently set to $(0.1, 0.1, 0.2, 0.2, 0.3)$ for r_1 and $(0.1, 0.2, 0.4, 0.6, 0.8)$ for r_0, while π_1, Se and Sp hold the same values as before. The estimates of r_1 and r_0 along with π_1, Se and Sp are obtained simultaneously with coverage probabilities varying from 94.8% to 95.8%. Our method is once more shown to handle different missing scenarios fairly well.

2.6 Discussion

In this paper we developed an EM algorithm based approach to evaluate the population prevalence and diagnostic accuracy of multiple imperfect tests in the absence of a gold standard, when the tests are assumed to be independent conditional on the true disease status. Under either a MAR assumption or one MNAR scenario, the proposed method can efficiently and precisely estimate population prevalence and diagnostic accuracy of each test with its associated missing probability. Simulations under different missing data mechanisms consistently result in fairly high coverage probabilities for the estimates. Although there is no established statistical test to assess the underlying missing data mechanism, our method has shown robustness to missing data assumptions as long as the missing percentages are not extremely high. Even though our estimates are biased under the MNAR scenario, the bias is confined to those specific tests affected and our estimates can still provide a reasonably good approximation to the "true" parameter values. On the contrary, many other methods tend to propagate

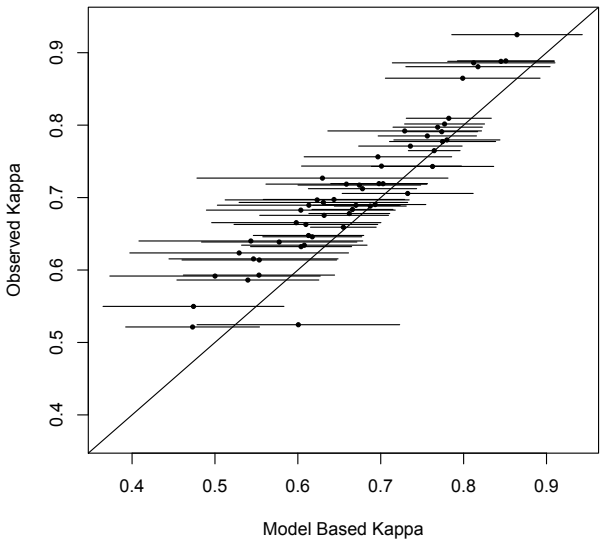
bias throughout all tests. In conclusion, our method is straightforward to comprehend and simple to implement for diagnostic studies involving multiple conditionally independent tests with moderate percentages of missing data and without a gold standard. It has the potential to improve public health by facilitating the diagnosis of cancer and other prevalent diseases.

Common methods such as case deletion, maximum likelihood estimation, multiple imputation, etc. are valid for MCAR and MAR but cannot handle MNAR without explicitly modeling the missing pattern. Two possible models to account for MNAR data are selection models (Heckman, 1979) and pattern mixture models (Little, 1993). These models are complicated and require substantial statistical knowledge and software experience, yet their validity is not easily justifiable and sometimes questionable. Methodological development to cope with missing data under the MNAR assumption is beyond our current scope but would be considered for future research.

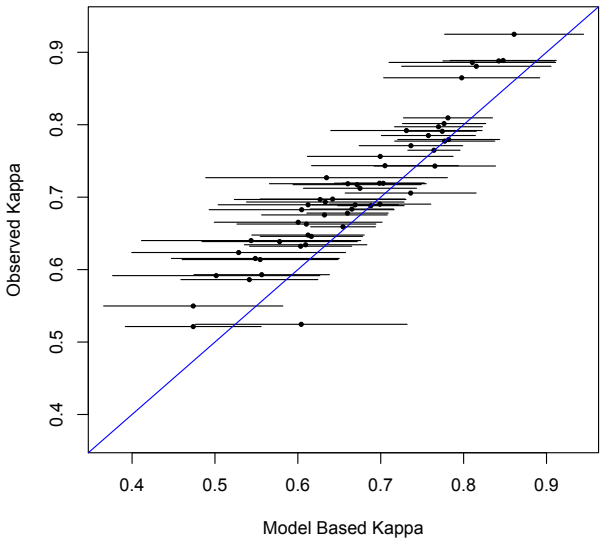
One limitation of our method is that we assumed conditional independence between tests, which is difficult to verify in practice. We cannot be absolutely certain of the conditional independence assumption, although our Kappa agreement plots fail to reject it. Some degrees of dependence may exist due to the similarity of biological basis. The effects of conditional dependence on the estimation of diagnostic accuracy and prevalence have been reviewed (Vacek, 1985; Dendukuri and Joseph, 2001). It has been shown that the latent class models under the independence assumption can produce relatively unbiased estimates when the degree of dependence is not too strong (Torrance-Rynard and Walter, 1998; Black and Craig, 2002; Georgiadis et al., 2003; Monti et al., 2005). To assess the robustness of our method, the conditional independence model is applied to data simulated under the conditional dependence assumption. It is able to estimate the parameters fairly well albeit the estimates are biased and the confidence intervals have worse coverage probability. Several methods have been developed for estimation

of diagnostic accuracy under the assumption of conditional dependence (Qu et al., 1996; Black and Craig, 2002; Xu and Craig, 2009; Shih and Albert, 2004; Dendukuri and Joseph, 2001; Yang and Becker, 1997). These methods produce superior point estimates for diagnostic accuracy than methods based on the conditional independence assumption, especially when the tests are indeed highly correlated. For future investigations, we would relax the conditional independence assumption by extending the application of such methods to data abundant with missing values.

Figure 2.1: Plot of Observed vs. Model Based Kappa



(a) Under the MAR Assumption



(b) Under the MNAR Assumption

Note: Dots and lines are model based Kappa statistics with their corresponding 95% simultaneous confidence intervals. There are a total of 55 dots and lines for 11 tests.

Table 2.2: Number of Subjects by Frequency of Missing Test Results

Number of Missing Tests	0	1	2	3	4	5	≥ 6
Number of Subjects	219	871	1202	584	216	203	192

Table 2.3: Estimates and 95% CIs of Sensitivity, Specificity, and Prevalence from Different Models

	MAR[a]		**MNAR**[b]	
Marker	Sensitivity (95%CI)	Specificity (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
ACTC	0.7111 (0.665, 0.753)	0.9700 (0.963, 0.976)	0.7098 (0.663, 0.752)	0.9703 (0.963, 0.976)
BAT25	0.9405 (0.915, 0.959)	0.9962 (0.993, 0.998)	0.9365 (0.909, 0.956)	0.9961 (0.993, 0.998)
BAT26	0.9308 (0.903, 0.951)	0.9985 (0.996, 0.999)	0.9280 (0.900, 0.949)	0.9985 (0.996, 1.000)
BAT40	0.9301 (0.902, 0.951)	0.9867 (0.981, 0.991)	0.9251 (0.896, 0.947)	0.9868 (0.981, 0.991)
BAT34C4	0.8499 (0.812, 0.882)	0.9980 (0.995, 0.999)	0.8490 (0.811, 0.881)	0.9980 (0.995, 0.999)
D10S197	0.8353 (0.796, 0.868)	0.9816 (0.975, 0.986)	0.8354 (0.796, 0.868)	0.9816 (0.975, 0.986)
D17S250	0.8121 (0.762, 0.854)	0.9564 (0.946, 0.965)	0.8110 (0.761, 0.853)	0.9569 (0.946, 0.966)
D18S55	0.8098 (0.770, 0.844)	0.9842 (0.978, 0.988)	0.8104 (0.771, 0.844)	0.9846 (0.979, 0.989)
D2S123	0.8493 (0.747, 0.915)	0.9769 (0.960, 0.987)	0.8495 (0.746, 0.916)	0.9773 (0.961, 0.987)
D5S346	0.6455 (0.600, 0.688)	0.9923 (0.988, 0.995)	0.6435 (0.598, 0.686)	0.9923 (0.988, 0.995)
MYCL	0.7587 (0.716, 0.797)	0.9360 (0.926, 0.945)	0.7547 (0.712, 0.793)	0.9366 (0.926, 0.946)
Prevalence		0.1482 (0.137, 0.160)		0.1506 (0.139, 0.163)

[a] Missingness depends on observed test results without modeling missing probabilities.
[b] Missingness depends on latent disease status when modeling missing probabilities.

Table 2.4: Estimates and 95% CIs of r_{1j}, r_{0j}, and r_j under Different Missing Data Assumptions

	MAR[a]	MNAR[b]			
Marker	r_j	r_{1j}	r_{0j}	$r_{1j} - r_{0j}$ (SE)	p-value
ACTC	0.1394	0.2248	0.1242	0.1006 (0.0194)	< 0.0001
BAT25	0.0771	0.0908	0.0747	0.0161 (0.0138)	0.2450
BAT26	0.0726	0.1077	0.0663	0.0414 (0.0147)	0.0048
BAT40	0.1408	0.1535	0.1386	0.0149 (0.0173)	0.3884
BAT34C4	0.1397	0.2200	0.1254	0.0946 (0.0194)	< 0.0001
D10S197	0.1792	0.2231	0.1715	0.0516 (0.0198)	0.0091
D17S250	0.4193	0.4554	0.4129	0.0425 (0.0237)	0.0724
D18S55	0.1342	0.1508	0.1313	0.0195 (0.0172)	0.2573
D2S123	0.8173	0.8606	0.8097	0.0509 (0.0168)	0.0025
D5S346	0.0849	0.1301	0.0769	0.0532 (0.0158)	0.0008
MYCL	0.1663	0.1636	0.1668	-0.0032 (0.0177)	0.8566

[a] Missingness does not depend on latent disease status.

[b] Missingness depends on latent disease status.

Note: r_j denotes missing probability for each test; r_{1j} and r_{0j} denote missing probabilities for diseased and non-diseased subjects respectively.

Chapter 3

Conditional Dependence Assumption

3.1 Introduction

Diagnostic accuracy, commonly quantified by sensitivity and specificity, plays a key role in the development of new diagnostic binary tests. A gold standard with perfect sensitivity and specificity may not always be administered due to the invasiveness, cost, or other limitations. For example, patients with negative test results are more likely to forgo the gold standard for a definitive diagnosis. This is considered to be a missing data problem involving partially missing gold standards. Following the language of Little and Rubin (1987), the mechanism that led to the missing gold standard is said to be missing completely at random (MCAR) if the probability of missingness is independent of any missing or observed data; the mechanism is called missing at random (MAR) when the probability of missingness depends only on observed data; the mechanism is called missing not at random (MNAR) when the probability of missing depends on some missing data. Both MCAR and MAR are "ignorable" missing data mechanisms whereas MNAR is a "non-ignorable" (NI) missing data mechanism. A considerable literature exists dealing with partially missing gold standards under the ignorable (Alonzo, 2005;

Begg and Greenes, 1983; Harel and Zhou, 2007; He and McDermott, 2012; Lin et al., 2006; Yu et al., 2010; Zhou, 1998) and the non-ignorable (Baker, 1995; Geloven et al., 2012; Harel and Zhou, 2007; Kosinski and Barnhart, 2003b,a; Zhou, 1993) missing data assumptions.

In many instances a gold standard does not exist, which causes another type of missing data problem in which the gold standard (i.e. the true disease status) is always missing. Various methods have been developed to assess the diagnostic accuracy of new tests in the absence of a gold standard (Alonzo et al., 1999; Hadgu et al., 2005; Enøe et al., 2000; Reitsma et al., 2009; Goetghebeur et al., 2000). One practice is to compare the new tests to some imperfect reference standard and attempt to correct imperfect reference bias. When there is no acceptable reference standard, the latent class analysis (LCA) treats the unobservable true disease status as a latent variable and utilizes all tests simultaneously in a unified manner. Early latent class models were based on the conditional independence assumption, which states that multiple tests are independent conditional on the true disease status (Hui and Walter, 1980; Walter and Irwig, 1988; Rindskopf and Rindskopf, 2006). These models often are referred to as traditional latent class (TLC) models. However, the conditional independence assumption usually does not hold when multiple tests have a similar basis, e.g. measuring a similar biological attribute, or when they are influenced by some subject-specific characteristics other than the disease status. Several maximum likelihood (ML) approaches that allow for conditional dependence between tests have been developed to estimate diagnostic accuracy, including a finite mixture model (Albert et al., 2004) that uses a quasi-Newton method, a latent class joint cell probability log-linear model that uses a Fisher scoring algorithm (Espeland and Handelman, 1989), a marginal latent class model that uses an accelerated EM gradient algorithm (Yang and Becker, 1997), and probit latent class (PLC) models (Qu et al., 1996; Qu and Hadgu, 1998; Uebersax, 1999; Chib and

Greenberg, 1998; Xu and Craig, 2009).

A PLC model is a version of LCA in which specified threshold locations discretize a latent continuous variable into different regions that correspond to observed response levels. Qu et al. (1996) developed a special PLC model, called the Gaussian random effects model, because conditional dependence between tests is addressed by subject-specific random effects in a standard Gaussian distribution that links the observed test results to the latent disease status through a probit model. Qu and Hadgu (1998) extended the Gaussian random effects model to a generalized linear mixed model with a hybrid algorithm that combined the EM algorithm and the Newton-Raphson method for its ML estimates. Dendukuri and Joseph (2001) presented a Bayesian approach similar to the random effects model of Qu et al. by imposing a prior distribution to summarize the uncertainty about each parameter. However, these models simply assume that the dependence between tests is based on their having the same distribution, which is hard to justify. To relax this assumption, Uebersax (1999) proposed a PLC model that assumes a multivariate-normal distribution within each latent class so that the correlation structure can be modeled flexibly. Employing the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990), Chib and Greenberg (1998) obtained ML estimates for multivariate probit models with a general covariance structure. Xu and Craig (2009) further developed a PLC model to estimate diagnostic accuracy while accommodating a general correlation structure between tests using a parameter-expanded Monte Carlo EM (PX-MCEM) algorithm, which was motivated by the MCEM algorithm (Chib and Greenberg, 1998) and the parameter-expanded EM (PX-EM) algorithm (Liu et al., 1998). A TLC model is a special version of the PLC model when the two covariance matrices are restricted to be diagonal.

Heretofore we have discussed two types of missing data problems regarding gold standard: (i) a gold standard is available but not applied on all subjects (partially

missing); (ii) a gold standard is never applied or does not exist (totally missing). Poleto et al. (2011) presented a scenario in which all subjects are evaluated with a gold standard and one of the three imperfect tests under evaluation, whereas the other two imperfect tests are not always performed. They used a two-stage hybrid approach (i.e., ML in stage one and weighted least squares in stage two) to estimate the diagnostic accuracy of the three imperfect tests. To our knowledge, when a gold standard is totally missing, there is no study that has evaluated the diagnostic accuracy of multiple correlated diagnostic tests with excessive missing data. Motivated by the Colon Cancer Family Registry (C-CFR) study (Section 2), we extended the PLC models under the conditional dependence assumption to evaluate the prevalence and diagnostic accuracy of multiple imperfect diagnostic tests with high proportions of missing data and without a gold standard. In addition, we also evaluated the correlation between these tests through a general correlation structure of the PLC model.

The remainder of this paper is organized as follows. Section 3.2 introduces the C-CFR study that motivated our research. Section 3.3 describes the PLC-based methodology of estimating diagnostic accuracy and correlation matrices, as well as the bootstrap approach for estimating their standard errors. Section 3.4 summarizes the results of simulation studies with different missing data assumptions and examines the finite sample properties of the proposed model. Section 3.5 presents the preliminary analyses of the C-CFR data. Finally, section 3.6 concludes with an extensive discussion.

3.2 Motivating Example

Hereditary nonpolyposis colorectal cancer (HNPCC), also known as Lynch syndrome (1999), is the most common familial colorectal cancer syndrome accounting for two to five percent of all colorectal cancer. HNPCC is a genetic disease caused by a deleterious germline mutation in genes involved in the DNA mismatch repair (MMR)

pathway that repairs the mismatches in the genome that occur during cell duplication. It is estimated that 600,000 individuals in the United States have HNPCC. These individuals have a substantially increased (up to 80%) lifetime risk of developing cancer in the colorectum and other sites when compared to the general public. Mutation analysis of the MMR genes may be considered a gold standard for HNPCC diagnosis. However its high cost ($2,000 - $3,000 per individual) precludes its broad use in HNPCC screening. A relatively inexpensive alternative ($200 - $300 per individual) (Thibodeau, 1981) seeks to identify a high level of microsatellite instability (MSI), the amplification or deletion within microsatellites (common and normal repeated sequences of DNA). Since the establishment of a consensus definition of MSI and unifying criteria for its measurement in 1998 (Boland et al., 1998), MSI biomarker tests have been used regularly as part of the international guidelines for HNPCC diagnosis (Umar et al., 2004). Therefore it is of great interest from a public health perspective to evaluate the diagnostic accuracy of MSI biomarkers for early detection and prevention of HNPCC.

The NCI C-CFR study is an international consortium of six centers in North America and Australia that was formed to support studies on the etiology, prevention, and clinical management of colorectal cancer (Newcomb et al., 2007). The C-CFR data include test results of 11 MSI biomarkers (namely *BAT25*, *BAT26*, *BAT40*, *BAT34C4*, *D10S197*, *D17S250*, *D18S55*, *D2S123*, *D5S346*, *ACTC*, and *MYCL*). A total of 3,487 subjects from families with a single subject are included in this research. Only a small proportion (6.3%) of the 3,487 subjects has been tested for all 11 MSI biomarkers. The observed missing proportions range from 7.3% for biomarker *BAT26* to 81.7% for biomarker *D2S123*. No gold standard was used for the C-CFR study. Furthermore, the 11 MSI biomarkers share similar biological bases so it is reasonable to assume that they are conditionally dependent. The high percentages of missing tests and the latent disease status, together with the conditional dependence between tests have motivated

our methodology research, which is to be introduced in the next section.

3.3 Statistical Methods

3.3.1 PLC Model Parameters Specification and Expansion

Suppose that we have a total of N subjects and J binary tests. Let $D_i = d(d = 1, 0)$ represent the latent variable for disease status (1=Yes, 0=No) of the i^{th} subject ($i = 1, \cdots, N,$). Let $\pi_d = Pr(D_i = d)$ denote the probability of disease/no disease of the the i^{th} subject (π_1 is the prevalence). Let t_{ij} be the result of the j^{th} test of the i^{th} subject. Notice that due to missing values, the test results are not in the typical "binary" fashion with three possible values, i.e., 1=positive, 0=negative, 99=missing. Let δ_{ij} be the indicator of whether the i^{th} subject has been tested by the j^{th} test (1=tested, 0=not tested). $T_i = (t_{i1}, \cdots, t_{iJ})$ and $\Delta_i = (\delta_{i1}, \cdots, \delta_{iJ})$ represent all the test results and missing indicators of the i^{th} subject. Under the conditional dependence assumption, the similarity between tests of the i^{th} subject is explained by some Gaussian latent variable $Z_i = (z_{i1}, \cdots, z_{iJ})'$, which has a multivariate normal distribution conditional on D_i, i.e. $Z_i|D_i = d \sim N_J(\mu_d, \Sigma_d)$ with mean vector μ_d and variance-covariance matrix $\Sigma_d = \{\sigma_{ij}^{(d)}\}$. We assume $z_{ij} > 0$ when $t_{ij} = 1$; $z_{ij} <= 0$ when $t_{ij} = 0$; z_{ij} could take any value within $(-\infty, \infty)$ when t_{ij} is missing. The probability of observing (T_i, Δ_i) is obtained by integrating over Z_i:

$$P(T_i, \Delta_i|D_i = d, \mu_d, \Sigma_d) = \int_{B_{i1}} \cdots \int_{B_{iJ}} \phi_J(Z_i; \mu_d, \Sigma_d) dz_{i1} \cdots dz_{iJ}$$

where the integration interval of each test is

$$B_{ij} = \begin{cases} (-\infty, 0] & \text{if } t_{ij}=0 \\ (0, \infty) & \text{if } t_{ij}=1 \\ (-\infty, \infty) & \text{if } t_{ij}=99, \text{ i.e. } \delta_{ij} = 0 \end{cases} \tag{3.1}$$

However we are unable to find a fixed solution for the model parameters (μ_d, Σ_d) since they are not identifiable. One way to overcome this challenge is to restrict the variance-covariance matrix Σ_d to correlation matrix R_d with all diagonal elements equal to 1 and all off-diagonal elements between $[-1, 1]$, and then re-parameterize μ_d to a_d (Chib and Greenberg, 1998). It can be shown that the sensitivity and specificity of the j^{th} test are $Se_j = \Phi(a_{1j})$ and $Sp_j = \Phi(-a_{0j})$, respectively. Let $\theta = (\pi_1, a_1, R_1, a_0, R_0)$ denote the vector of all $1+2J+2\times\frac{J(J-1)}{2} = J^2+J+1$ unique parameters. Assume that we know Z_i and D_i in addition to the observed data (T_i, D_i), and assume that, for any pair of two tests, there are at least some subjects with both test results present. (For the C-CFR data set, this is of no concern because 219 subjects have all 11 test results present.) By Bayes' theorem, the log-likelihood of complete data $Y_i = (T_i, \Delta_i, Z_i, D_i)$ is

$$\begin{aligned} \log L_c(\theta) &= \log L(\theta | T_i, \Delta_i, Z_i, D_i) \\ &= \log \prod_{i=1}^{N} \{P(T_i, \Delta_i | Z_i, D_i, \theta)\} \\ &= \log \prod_{i=1}^{N} \{P(D_i|\pi_1) P(Z_i|D_i, a_1, R_1, a_0, R_0) P(T_i, \Delta_i | Z_i, a_1, R_1, a_0, R_0)\} \end{aligned}$$

Since $P(Z_i|D_i, a_1, R_1, a_0, R_0) = \phi_J(Z_i; a_1, R_1)^{d_i} \phi_J(Z_i; a_0, R_0)^{(1-d_i)}$, $P(T_i, \Delta_i | Z_i, a_d,$

$R_d) = \prod_{i=1}^{J} I(z_{ij} \in B_{ij})$, $P(D_i|\pi_1) = \pi_1^{d_i}(1-\pi_1)^{(1-d_i)}$, the log-likelihood becomes

$$\begin{aligned}
\log L_c(\theta) &= \log \prod_{i=1}^{N} \{\pi_1^{d_i}(1-\pi_1)^{(1-d_i)} \phi_J(Z_i; a_1, R_1)^{d_i} \phi_J(Z_i; a_0, R_0)^{(1-d_i)} \\
&\quad \prod_{i=1}^{J} I(z_{ij} \in B_{ij})\} \\
&= \sum_{i=1}^{N} \{d_i \log(\pi_1) + (1-d_i)\log(1-\pi_1) + d_i \log(\phi_J(Z_i; a_1, R_1)) \\
&\quad + (1-d_i)\log(\phi_J(Z_i; a_0, R_0)) + \sum_{i=1}^{J} \log(I(z_{ij} \in B_{ij}))\}
\end{aligned}$$

3.3.2 ML Estimation Using the Monte Carlo EM Algorithm

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is an iterative computation approach that has been used extensively to find maximum likelihood estimates (MLEs) of the parameters of an underlying distribution in two general incomplete data problems: (i) the data contains missing values; (ii) the model can be simplified by assuming the existence of additional unobserved (latent) variables. The EM algorithm involves both integration methods (for the E-step) and optimization methods (for the M-step). When one or both steps are analytically intractable, as frequently is encountered in the high-dimensional integration for the E-step, approximation methods (e.g., Laplace approximation and Taylor series expansions), numerical methods (e.g., Gauss-Hermite and Newton-Cotes), and Monte Carlo methods have been used. With increasingly powerful computing resources, Monte Carlo methods have gained in popularity for high-dimensional integrations without closed-form solutions. The Monte Carlo EM (MCEM) algorithm was introduced by Wei and Tanner (1990) to compute expectation in the E-step using Monte Carlo simulations. The Monte Carlo sample size does not depend as much on dimensionality as it does on numerical methods, such as Gaussian quadrature (Evans and Swartz, 1995). It allows for an easy assessment of

the approximation error since one can increase the Monte Carlo sample size until the desired accuracy is obtained.

The MCEM algorithm is well suited for our C-CFR case study: first, we are dealing with missing test results as well as the latent disease status; secondly, our PLC model involves an intractable E-step due to the high-dimensional integration incurred by the 11 tests. However, the M-step does not have a closed-form solution when the variance-covariance matrices are restricted to be correlation matrices R_d. Thus we need to expand the parameters as $\mu_d = V_d^{\frac{1}{2}} a_d$ and $\Sigma_d = V_d^{\frac{1}{2}} R_d V_d^{\frac{1}{2}}$ following the PX-EM algorithm by Liu et al. (1998). V_d is a $J \times J$ diagonal matrix with all diagonal elements positive. The parameter vector becomes $\beta = (\pi_1, \mu_1, \Sigma_1, \mu_0, \Sigma_0)$ and the log-likelihood becomes

$$\begin{aligned}
\log L_c(\beta) &= \log \prod_{i=1}^{N} \{\pi_1^{d_i}(1-\pi_1)^{(1-d_i)} \phi_J(Z_i;\mu_1,\Sigma_1)^{d_i} \phi_J(Z_i;\mu_0,\Sigma_0)^{(1-d_i)} \\
&\quad \prod_{i=1}^{J} I(z_{ij} \in B_{ij})\} \\
&= \sum_{i=1}^{N} \{d_i \log(\pi_1) + (1-d_i)\log(1-\pi_1) + d_i \log(\phi_J(Z_i;\mu_1,\Sigma_1)) \\
&\quad + (1-d_i)\log(\phi_J(Z_i;\mu_0,\Sigma_0)) + \sum_{i=1}^{J} \log(I(z_{ij} \in B_{ij}))\}
\end{aligned}$$

Substituting the joint density functions of the multivariate normal distribution $\phi_J(Z_i;\mu_d,\Sigma_d) = \frac{1}{(2\pi)^{J/2}|\Sigma_d|^{1/2}} \exp\{-\frac{1}{2}(Z_i-\mu_d)'\Sigma_d^{-1}(Z_i-\mu_d)\} (d = 1, 0)$ into $\log L_c(\beta)$ we have the final log-likelihood function of the complete data

$$\begin{aligned}
\log L_c(\beta) &= \log \prod_{i=1}^{N} \{\pi_1^{d_i}(1-\pi_1)^{(1-d_i)} \phi_J(Z_i;\mu_1,\Sigma_1)^{d_i} \phi_J(Z_i;\mu_0,\Sigma_0)^{(1-d_i)} \\
&\quad \prod_{i=1}^{J} I(z_{ij} \in B_{ij})\}
\end{aligned}$$

$$
\begin{aligned}
= & \sum_{i=1}^{N}\{d_i\log(\pi_1)+(1-d_i)\log(1-\pi_1) \\
& -\frac{1}{2}\log|\Sigma_1|d_i-\frac{1}{2}d_i(Z_i-\mu_1)'\Sigma_1^{-1}(Z_i-\mu_1)-\frac{J}{2}\log(2\pi)d_i \\
& -\frac{1}{2}\log|\Sigma_0|d_i-\frac{1}{2}d_i(Z_i-\mu_0)'\Sigma_0^{-1}(Z_i-\mu_0)-\frac{J}{2}\log(2\pi)d_i \\
& +\sum_{i=1}^{J}\log(I(z_{ij}\in B_{ij}))\} \qquad (3.2)
\end{aligned}
$$

The M-step is to solve the score equations after taking expectations conditional on the complete data:

$$
\begin{aligned}
0 &= \sum_{i=1}^{N} E[\{\frac{d_i}{\pi_1}-\frac{1-d_i}{1-\pi_1}\}|T_i,\Delta_i,\beta^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{d_iZ_i-d_i\mu_1\}|T_i,\Delta_i,\beta^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{(1-d_i)Z_i-(1-d_i)\mu_0\}|T_i,\Delta_i,\beta^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{-\frac{1}{2}\Sigma_1^{-1}d_i+\frac{1}{2}\Sigma_1^{-1}(Z_i-\mu_1)(Z_i-\mu_1)'\Sigma_1^{-1}d_i\}|T_i,\Delta_i,\beta^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{-\frac{1}{2}\Sigma_0^{-1}(1-d_i)+\frac{1}{2}\Sigma_0^{-1}(Z_i-\mu_0)(Z_i-\mu_0)'\Sigma_0^{-1}(1-d_i)\}|T_i,\Delta_i,\beta^{(n)}]
\end{aligned}
$$

The solutions are:

$$\begin{aligned}
\pi_1^{(n+1)} &= \frac{1}{N}\sum_{i=1}^{N} E[d_i|T_i,\Delta_i,\beta^{(n)}] \\
\mu_1^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i Z_i|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i|T_i,\Delta_i,\beta^{(n)}]} \\
\mu_0^{(n+1)} &= \frac{\sum_{i=1}^{N} E[(1-d_i) Z_i|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[(1-d_i)|T_i,\Delta_i,\beta^{(n)}]} \\
\Sigma_1^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i (Z_i-\mu_1)(Z_i-\mu_1)'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i|T_i,\Delta_i,\beta^{(n)}]} \\
\Sigma_0^{(n+1)} &= \frac{\sum_{i=1}^{N} E[(1-d_i) (Z_i-\mu_0)(Z_i-\mu_0)'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[(1-d_i)|T_i,\Delta_i,\beta^{(n)}]}
\end{aligned}$$

They are simplified for $d = 1, 0$ as

$$\begin{aligned}
\pi_1^{(n+1)} &= \frac{1}{N}\sum_{i=1}^{N} E[d_i|T_i,\Delta_i,\beta^{(n)}] \\
\mu_d^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d} Z_i|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} \\
\Sigma_d^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d} (Z_i-\mu_d)(Z_i-\mu_d)'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} \\
&= \frac{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d} Z_i Z_i'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d (1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} - \mu_d^{(n+1)}(\mu_d^{(n+1)})' \qquad (3.3)
\end{aligned}$$

The estimates for a_d and R_d can be derived by reducing the expanded parameter through U_d (a diagonal matrix with diagonal elements equal to those of $\Sigma_d^{(n+1)}$):

$$\begin{aligned}
a_d^{(n+1)} &= U_d^{-\frac{1}{2}} \mu_d^{(n+1)} \\
R_d^{(n+1)} &= U_d^{-\frac{1}{2}} \Sigma_d^{(n+1)} U_d^{-\frac{1}{2}} \qquad (3.4)
\end{aligned}$$

.

For the E-step we compute the conditional expectations of the expanded complete data sufficient statistics, i.e. $\sum_{i=1}^{N} E[d_i|T_i, \Delta_i, \beta^{(n)}]$, $\sum_{i=1}^{N} E[d_i Z_i|T_i, \Delta_i, \beta^{(n)}]$, $\sum_{i=1}^{N} E[d_i Z_i Z_i'|T_i, \Delta_i, \beta^{(n)}]$, $\sum_{i=1}^{N} E[(1-d_i)Z_i|T_i, \Delta_i, \beta^{(n)}]$, and $\sum_{i=1}^{N} E[(1-d_i)Z_i Z_i'|T_i, \Delta_i, \beta^{(n)}]$ via Markov chain Monte Carlo (MCMC) routines such as the Gibbs and Metropolis-Hastings samplers. For computation efficiency, we grouped subjects by K distinctive response profiles with n_k subjects for each response profile. (Subjects with the same results for all J tests are said to have the same response profile.) With three possible test results (1, 0, missing), the total number of possible profiles for J tests is $3^J - 1$. It is unlikely for the observed data to contain all possible profiles since $3^J - 1$ increases exponentially with J. For the C-CFR data, the actual number of profiles observed is $K = 887$ but the total number of possible profiles is $3^{11} - 1 = 177,146$. It is reasonable to speculate that the $177,146 - 887 = 176,259$ missing profiles are due to ignorable missing data mechanism, i.e. there is no clinical or other logical consideration that precludes a response profile from occurring.

We adopt Xu and Craig's (2009) sampling algorithm, which reduces a truncated multivariate normal distribution to a computationally much easier problem involving a series of univariate truncations. We proceed as follows:

- Begin with a set of arbitrary starting values for the parameter $\theta^{(0)} = \beta^{(0)} = (\pi_1^{(0)}, a_1^{(0)}, a_0^{(0)}, R_d^{(0)}, R_0^{(0)})$ and the latent variable $Z_k^{(0)} = (z_{k1}, \cdots, z_{kJ})'(k = 1, \cdots, K)$. For the first MC sample $m = 1$, generate $d_k^{(0)}$ from Bernoulli$(p_k^{(0)})$, where $p_k^{(0)} = \frac{\pi_1^{(0)}}{\pi_1^{(0)} + (1-\pi_1^{(0)})r}$ and $r = \frac{\phi_J(z_k^{(0)}; a_0^{(0)}, R_0^{(0)})}{\phi_J(z_k^{(0)}; a_1^{(0)}, R_1^{(0)})}$.
- Generate $Z_k^{(1)}$ given $d_k^{(0)} = d$ from a truncated normal distribution $TN(\mu^*, \sigma^{*2})$ where the integration interval is B_{kj}, $\sigma^{*2} = \frac{1}{(R_d^{-1})_{j,j}}$, $\mu^* = a_{dj} - \sigma^{*2}(R_d^{-1})_{j,-j}(Z_{k,-j} - a_{d,-j})$. Draw each $z_{kj}(j = 1, \cdots, J)$ from the distribution of z_{kj} conditioned on all other variables, making use of the most recent values and updating z_{kj} with its new value as soon as it has been drawn, i.e. draw $z_{k1}^{(1)}$ from $[z_{k1}|z_{k2}^{(0)}, \cdots, z_{kJ}^{(0)}]$,

draw $z_{k2}^{(1)}$ from $[z_{k2}|z_{k1}^{(1)}, z_{k3}^{(0)}, \cdots, z_{kJ}^{(0)}]$, $\cdots$, draw $z_{kJ}^{(1)}$ from $[z_{kJ}|z_{k1}^{(1)}, \cdots, z_{k,(J-1)}^{(1)}]$.

- Repeat the above simulation steps for $m = \{2, \cdots, M\}$ to generate M samples of d_k and Z_k. The conditional expectations of the expanded complete data sufficient statistical are estimated by averaging over the M Monte Carlo samples, e.g., $\sum_{i=1}^{N} E[d_i Z_i|T_i, \Delta_i, \beta^{(n)}] = \sum_{i=1}^{K} E[n_k d_k Z_k|T_i, \Delta_i, \beta^{(n)}] = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} n_k d_k^{(m)} Z_k^{(m)}$. Substituting the estimated conditional expectations into 3.3, we derive parameter estimates $\beta^{(1)} = (\pi_1^{(1)}, \mu_1^{(1)}, \mu_0^{(1)}, \Sigma_1^{(1)}, \Sigma_0^{(1)})$, and subsequently $\theta^{(1)} = (\pi_1^{(1)}, a_1^{(1)}, a_0^{(1)}, R_1^{(1)}, R_0^{(1)})$. This completes the first iteration $l = 1$. $\theta^{(1)}$ and $Z_k^{(M)}$ (the last MC sample) will be starting values for the next iteration.

- Repeat the above PX-MCEM algorithm steps for $l = 2, 3, \cdots$ until the convergence at $l = L$, i.e. the difference between $\theta^{(L)}$ and $\theta^{(L-1)}$ is consecutively less than a preset tolerance for several iterations. Then, we have the final estimates $\theta^{(L)}$, which converge to the true parameter values θ by the law of large numbers.

To control the between-simulation variability known as Monte Carlo error (MCE), a large MC sample size M (typically at least $10,000$) is preferred but it may quickly become computationally burdensome. It is advisable to implement a small M for the first few iterations when $\theta^{(l)}$ is far from the true parameter values θ, and increase M for later iterations when $\theta^{(l)}$ moves closer to θ (Wei and Tanner, 1990). For example, McCulloch developed MCEM algorithms that increase M linearly (1994) and nonlinearly (1997) with the number of iterations. We used a more efficient cumulative MCEM algorithm (Kou et al., 1998) by fixing M as a relatively small number (e.g. $M = 1500$) for all iterations. It utilizes MC samples from the current iteration as well as an adaptively increasing number of previous iterations, so that simulations from previous iterations are not wasted.

As with any MCMC method, each MC sample is correlated with nearby samples.

We thinned the chain by saving every o^{th} simulated sample from each sequence. Another issue arises when the MC samples at the beginning of the chain do not represent the desired distribution accurately. We discarded the initial MC samples of each iteration for early EM iterations during the burn-in period.

The adequacy of the model can be checked by plotting all $\frac{J(J-1)}{2}$ pairwise correlation residuals (Qu et al., 1996). The correlation coefficient between any two tests t_{ij} and $t_{ij'}$ is $\frac{P(t_{ij}=1,t_{ij'}=1)-P(t_{ij}=1)P(t_{ij'}=1)}{\sqrt{P(t_{ij}=1)(1-P(t_{ij}=1))P(t_{ij'}=1)(1-P(t_{ij}=0))}}$. For observed pairwise correlation coefficients, $P(t_{ij}=1)=\frac{\sum_{i=1}^{N} t_{ij}}{N}$ and $P(t_{ij}=1,t_{ij'}=1)=\frac{\sum_{i=1}^{N} t_{ij}t_{ij'}}{N}$. For model-based pairwise correlation coefficients, $P(t_{ij}=1)=\sum_0^1 \pi_d \int_0^\infty \phi_j(Z_i;\mu_d,\Sigma_d)dz_{ij}$ and $P(t_{ij}=1,t_{ij'}=1)=\sum_0^1 \pi_d \int_0^\infty \int_0^\infty \phi_{j,j'}(Z_i;\mu_d,\Sigma_d)dz_{ij}dz_{ij'}$. The residuals are the differences between the observed and model-based pairwise correlation coefficients.

3.3.3 Starting Values for the PX-MCEM Algorithm

When the loglikelihood function is concave and unimodal over the entire parameter space, the PX-MCEM algorithm converges to the unique MLE $\theta^{(L)}$ from any set of starting values. In that sense $\theta^{(0)}=(\pi_1^{(0)},a_1^{(0)},a_0^{(0)},R_1^{(0)},R_0^{(0)})$ can be selected arbitrarily as long as $R_d^{(0)}$ is positive definite. In reality it always helps to choose starting values that are likely to be close to the true values of θ instead of making a wild guess. For example, Walter and Irwig (1988) used starting values based on the majority opinion among three radiologists. Staquet et al. (1981) used "the most probable value" based on medical and biological knowledge about the PCR and ME tests. The choice of starting values is more crucial when missing data occur, especially if the proportion of missing data is high for certain tests. Little and Rubin (1987) outlined a few choices for the starting values of parameters assuming that the missing-data mechanism was ignorable, such as the complete-case solution and the available-case solution.

However, multiple maxima often exist, and the PX-MCEM algorithm is not guaranteed to converge to a unique global maximum. One suggestion is to try a variety of starting values to examine whether a global maxima is reached rather than a local maximum. This becomes impractical due to the computational intensity of the PX-MCEM algorithm, as well as the large number of tests and subjects for the C-CFR data. Conversely, the TLC model is much more efficient computationally. It has been shown that the TLC model is adequate even with conditionally dependent tests when the accuracies of the tests are high or when the tests are weakly dependent (Hui and Zhou, 1998; Georgiadis et al., 2003). In this paper we extended the TLC model under the conditional independence assumption to allow for missing data, and we used the model's estimates of the parameter as our starting values. Assuming that the ignorable missing data mechanism is tenable, the probability of observing (T_i, Δ_i) for subject i is

$$p_\theta(T_i, \Delta_i) = \sum_{d=0}^{1} \pi_d h_{id} \tag{3.5}$$

where

$$\begin{aligned} h_{i1} &= \prod_{j=1}^{J} Se_j^{t_{ij}\delta_{ij}} (1 - Se_j)^{(1-t_{ij})\delta_{ij}} \\ h_{i0} &= \prod_{j=1}^{J} (1 - Sp_j)^{t_{ij}\delta_{ij}} Sp_j^{(1-t_{ij})\delta_{ij}}. \end{aligned}$$

Let $\gamma = (\pi_1, Se_1, \cdots, Se_J, Sp_1, \cdots, Sp_J)$. Start with some arbitrary starting values $\gamma^{(0)}$ for the PX-MCEM algorithm. Given $\gamma^{(n)}$ for the $(n)^{th}$ iteration, the M-step solves

the score equations as:

$$
\begin{aligned}
0 &= \sum_{i=1}^{N} E[\{\frac{d_i}{\pi_1} - \frac{1-d_i}{1-\pi_1}\}|Y_i, \gamma^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{\frac{d_i t_{ij}\delta_{ij}}{Se_j} - \frac{d_i\delta_{ij} - d_i t_{ij}\delta_{ij}}{1-Se_j}\}|Y_i, \gamma^{(n)}] \\
0 &= \sum_{i=1}^{N} E[\{\frac{(1-d_i)(1-t_{ij})\delta_{ij}}{Sp_j} - \frac{(1-d_i)t_{ij}\delta_{ij}}{1-Sp_j}\}|Y_i, \gamma^{(n)}]
\end{aligned}
$$

where $Y_i = (T_i, \Delta_i, D_i)$ are the complete data. We compute $E[d_i|Y_i, \gamma^{(n)}]$ in the E-step and derive $\gamma^{(n+1)}$. Iterating EM steps until convergence, we get the final estimates $\hat{\gamma}$. After transformation $\hat{a}_{1j} = \Phi^{-1}(\hat{Se}_j)$ and $\hat{a}_{0j} = -\Phi^{-1}(\hat{Sp}_j)$ are our starting values for a_{1j} and a_{0j} to initiate the PX-MCEM algorithm aforementioned. Simulation studies indicate great coverage probability for $\hat{Se}_j$ and $\hat{Sp}_j$ when the missing rate of test j is not too high and the conditional independence assumption holds. When the conditional independence assumption is relaxed to the conditional dependence assumption, $\hat{Se}_j$ and $\hat{Sp}_j$ are still reasonably close to the true values. The TLC model converts otherwise arbitrary starting values to the best available starting values. Nevertheless, the TLC model does not estimate R_1 and R_0, so their starting values have to be selected arbitrarily.

3.3.4 Bootstrap Method for Standard Errors

Many methods have been developed to estimate standard error (SE) in the context of the EM algorithm with missing data (Tanner, 1991; Little and Rubin, 1987). In modern high-performance computing environment, resampling methods such as the bootstrap method (Efron, 1979) and the jackknife method (Miller, 1974) are used broadly to derive asymptotic SE estimates using just the data at hand. The bootstrap method is shown to be robust in many situations, i.e., it provides large-sample SE estimates of

MLEs with good coverage, even if the model is misspecified or if the model assumptions, such as the ignorable missing-data assumption, are invalid (Efron, 1994).

The bootstrap method is employed to estimate the SE of the ML estimate of θ. We randomly drew B bootstrap samples of size N from observed data $T_1, \cdots, T_N$ with replacement. Then, the PLC model was applied to each bootstrap sample to get the ML estimates $(\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)})$. The bootstrap estimate of θ is $\hat{\theta}_{boot} = \frac{\sum_{b=1}^{B} \hat{\theta}^{(b)}}{B}$ and the SE estimate is $\hat{SE}_{boot} = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2}{B-1}}$. It has been demonstrated that $B = 200$ is required if the bootstrap distribution is approximately normal (Efron, 1994).

3.4 Simulation Studies

We assessed the performance of the PLC model when fitting simulated data sets from different missing data mechanisms. Each simulation study consists of $S = 500$ simulations. Each simulation generated a simulated data set with $N = 3,500$ subjects and $J = 5$ tests. The true parameter values were set to $\pi_1 = 0.2$, $Se = (0.7, 0.9, 0.8, 0.6, 0.75)$, and $Sp = (0.9, 0.85, 0.9, 0.9, 0.8)$. All test results are conditional dependent, i.e., the correlation coefficients between any two tests are 0.6 for diseased subjects and 0.45 for non-diseased subjects. For simulated data sets under the MCAR mechanism, all tests were assigned randomly $\delta_{ij} = 1$ or 0 with $P(\delta_{ij} = 1) = (0.95, 0.9, 0.8, 0.5, 0.1)$. For simulated data sets under the MAR and MNAR mechanisms, t_{i1} and t_{i2} are always observed, whereas t_{i3} has $P(\delta_{i3} = 1) = 0.8$. Under the MAR mechanism, t_{i4}'s missing probability depends on t_{i1} through $logit(P(\delta_{i4} = 1|t_{i1})) = \alpha + \beta_1 t_{i1}$; t_{i5}'s missing probability depends on t_{i1} and t_{i2} through $logit(P(\delta_{i5} = 1|t_{i1}, t_{i2})) = \alpha' + \beta_1' t_{i1} + \beta_2 t_{i2}$. Under the MNAR mechanism, t_{i4}'s missing probability depends on t_{i1} and t_{i3} through $logit(P(\delta_{i4} = 1|t_{i1}, t_{i3})) = \alpha + \beta_1 t_{i1} + \beta_3 t_{i3}$; t_{i5}'s missing probability depends on t_{i1}, t_{i2}, t_{i3} and unobservable disease status D_i through $logit(P(\delta_{i5} = 1|t_{i1}, t_{i2}, t_{i3}, D_i)) = \alpha' + \beta_1' t_{i1} + \beta_2 t_{i2} + \beta_3' t_{i3} + \beta_4 D_i$. The coefficients of logit

models are set to $\alpha = 1.5$, $\beta_1 = -2.5$, $\beta_2 = -2$, $\beta_3 = -1.5$, $\beta_4 = -0.7$, $\alpha' = 2$, $\beta'_1 = -2$, $\beta'_3 = -3$. Under the MAR mechanism, the missing probabilities vary from 0.18 to 0.73 for t_{i4} and from 0.12 to 1.00 for t_{i5}. The average missing probabilities are about 30% for T_4 and 40% for T_5. Under the MNAR mechanism, the missing probabilities vary from 0.18 to 0.92 for t_{i4} and from 0.12 to 1.00 for t_{i5}. The average missing probabilities for T_4 and T_5 are 40% and 50% respectively.

The PLC model's estimates were obtained from each simulated data set, and the mean and standard deviation (SD) of the estimates were calculated. For each EM iteration, 3000 Gibbs sampler iterations were simulated. A burn-in of 1000 MC samples was implemented for the initial 10 EM iterations. Thinning of the chains was performed by saving every 2^{nd} MC sample, which resulted in 1500 MC samples for each EM iteration. Thinning decreases the correlation of the chain at the cost of increasing the number of samples required to obtain the same MC sample size. Hence, we refrained from over-thinning in the interest of computation time. We assessed convergence by visual examination. The value chosen for burn-in appears to be reasonable as it cut off all the early fluctuations. Trace plots of MC samples (e.g. $\sum_{k=1}^{K} n_k d_k^{(m)} Z_k^{(m)}$) versus m do not exhibit any pattern or poor mixing of MCMC. We also plotted cumulative parameter estimates $\theta^{(l)}$ against the PX-MCEM algorithm iteration l, which stabilized (leveled off to a flat line) within 100 EM iterations. According to the convergence checks, our simulation settings will likely suffice.

In addition, each simulated data set has $B = 200$ bootstrap samples drawn for SE estimation. Then, the mean of these SE estimates were calculated. Both the MCMC simulation and bootstrap resampling are computationally intensive. Parallel computing was employed to subdivide a simulation task into sub-tasks that can undergo analysis simultaneously using multi-processors in a high-performance computing environment.

Histograms of the bootstrap samples demonstrated a bell-shaped curve that was approximately Gaussian. We also ran the Shapiro-Wilk test for each bootstrap sample and failed to reject the null assumption of normality, with almost all p-values being less than 0.05. The SE estimates closely match the SD estimates. All these results indicate good behavior of the bootstrap method. To explore the effect of increasing simulation size S and/or bootstrap sample size B, we did sensitivity studies with $S = 400$ and $B = 400$. The results of these studies do not show much improvement. To save computational time, we stick with $S = 200$ and $B = 200$ for all simulation studies.

Table 3.1 summarizes the simulation results when the true parameter values as used as the starting values. Under both the MCAR and MAR mechanisms, the PLC model gave unbiased estimates with good coverage probabilities for all parameters (all above 90% and most around 95%). Under the MNAR mechanism, the coverage probabilities are slightly worse for Se_5 (88.2%). The PLC model is robust to data sets with abundant missing values (e.g., the missing probabilities of t_{i5} are 90% for MCAR, 40% for MAR, and 50% for MNAR) when the starting values are very close to the true values.

For our proposed method, we fit the TLC model first and used the parameter estimates as starting values for prevalence and diagnostic accuracy. For R_1 and R_0, the starting values were set to 0.5 for all off-diagonal elements. The estimates of the PLC model are considerably more accurate than the estimates of the TLC model as they move closer towards the true values. Under the MCAR mechanism, coverage probabilities for prevalence, sensitivities, and specificities are around 95% for all tests except for t_{i5}, due to its high percentage of missing values. Under the MAR and MNAR mechanism, coverage probabilities for sensitivities and specificities are a little worse but the majority are still above 90%. Not surprisingly, the coverage probabilities are much worse for R_1 and R_0 due to their arbitrary starting values. It is noteworthy that the coverage probabilities for R_0 usually are better than for R_1, possibly due to the fact

that more data are available from non-diseased subjects for estimation of R_0.

Additional simulation studies were conducted with different starting values for the PX-MCEM algorithm to illustrate their effects on parameter estimates. For example, we obtained starting values following the method of Walter and Irwig (1988), which was based on the majority opinion among multiple radiologists. The starting value of prevalence was set to the proportion of subjects with at least three positive tests among all subjects with at least three non-missing tests; the starting value of sensitivity for each test was set to the proportion of subjects with a positive result for this test among all subjects with at least three positive tests and a non-missing result for this test; the starting value of specificity for each test was set to the proportion of subjects with a negative result for this test among all subjects with no more than two positive tests and a non-missing result for this test. These starting values deviate further away from the true values than do the TLC estimates, and hence result in much poorer coverage probabilities. Essentially, the closer the starting values are to the true values, the better the PLC model performs in terms of coverage probabilities.

Figure 3.2 presents correlation residual plots under different missing data mechanisms. The true parameter values for π_d, μ_d, and Σ_d $(d = 1, 0)$ were used for the model-based correlation coefficients. Under the MCAR mechanism, all pairwise correlation residuals are close to zero, and there is no noticeable pattern, which means a good fit. Under the MAR mechanism, the correlation residuals between the two MAR tests (t_{i4} and t_{i5}) and the three non-MAR tests tend to be negative, suggesting an overestimation of such correlations. Under the MNAR mechanism, the correlation residuals between the two MNAR tests (t_{i4} and t_{i5}) and the three non-MNAR tests indicate more serious overestimation of their correlations. All three plots are supported by the simulation results.

Table 3.1: Summary of Simulation Results

		MCAR			**MAR**			**MNAR**		
	True	Mean (SD)	SE	CP	Mean (SD)	SE	CP	Mean (SD)	SE	CP
Prevalence	0.25	0.2510 (0.0083)	0.0082	95.8	0.2513 (0.0083)	0.0084	96.2	0.2504 (0.0087)	0.0085	95.6
Se 1	0.70	0.7012 (0.0186)	0.0187	95.8	0.7003 (0.0196)	0.0188	94.2	0.7017 (0.0203)	0.0191	94.0
Se 2	0.90	0.9008 (0.0111)	0.0106	94.8	0.9001 (0.0126)	0.0116	92.4	0.9005 (0.0125)	0.0119	95.0
Se 3	0.80	0.8000 (0.0160)	0.0156	95.2	0.7993 (0.0171)	0.0163	94.2	0.7989 (0.0166)	0.0167	95.0
Se 4	0.60	0.6016 (0.0249)	0.0241	94.6	0.5997 (0.0276)	0.0268	95.0	0.5909 (0.0297)	0.0295	96.2
Se 5	0.75	0.7514 (0.0235)	0.0227	95.0	0.7477 (0.0246)	0.0243	94.0	0.7355 (0.0221)	0.0215	88.2
Sp 1	0.90	0.9008 (0.0069)	0.0068	95.4	0.9006 (0.0067)	0.0068	95.4	0.9004 (0.0069)	0.0068	95.4
Sp 2	0.85	0.8505 (0.0077)	0.0079	97.0	0.8506 (0.0076)	0.0077	96.2	0.8498 (0.0073)	0.0079	98.0
Sp 3	0.90	0.9009 (0.0070)	0.0072	95.2	0.9008 (0.0070)	0.0073	96.8	0.9005 (0.0068)	0.0074	96.6
Sp 4	0.90	0.9001 (0.0090)	0.0088	95.2	0.9007 (0.0079)	0.0077	95.0	0.9017 (0.0081)	0.0078	94.6
Sp 5	0.80	0.8009 (0.0152)	0.0154	96.0	0.7998 (0.0103)	0.0098	95.6	0.8041 (0.0102)	0.0100	92.6
R1(2,1)	0.60	0.5970 (0.0328)	0.0337	95.8	0.5983 (0.0357)	0.0372	96.2	0.5998 (0.0399)	0.0391	96.0
R1(3,1)	0.60	0.5978 (0.0409)	0.0406	93.6	0.5981 (0.0433)	0.0427	94.2	0.5979 (0.0446)	0.0439	95.6
R1(3,2)	0.60	0.5977 (0.0291)	0.0293	96.8	0.5981 (0.0319)	0.0329	95.8	0.5970 (0.0360)	0.0360	95.0
R1(4,1)	0.60	0.5950 (0.0545)	0.0525	94.6	0.5999 (0.0551)	0.0539	94.4	0.5942 (0.0550)	0.0535	94.0
R1(4,2)	0.60	0.5957 (0.0330)	0.0344	96.4	0.5984 (0.0360)	0.0367	95.8	0.5949 (0.0382)	0.0376	94.0
R1(4,3)	0.60	0.5963 (0.0444)	0.0448	95.6	0.5967 (0.0462)	0.0447	95.6	0.5935 (0.0481)	0.0463	95.2
R1(5,1)	0.60	0.5996 (0.0393)	0.0393	95.2	0.5983 (0.0484)	0.0495	95.6	0.6039 (0.0410)	0.0403	95.0
R1(5,2)	0.60	0.5985 (0.0289)	0.0286	94.8	0.5979 (0.0360)	0.0359	95.2	0.5977 (0.0367)	0.0351	95.4
R1(5,3)	0.60	0.6001 (0.0349)	0.0337	94.8	0.5982 (0.0392)	0.0399	96.4	0.5979 (0.0399)	0.0388	94.2
R1(5,4)	0.60	0.5978 (0.0382)	0.0377	96.0	0.5985 (0.0505)	0.0472	94.6	0.5984 (0.0410)	0.0392	94.0
R0(2,1)	0.45	0.4468 (0.0313)	0.0312	94.8	0.4466 (0.0312)	0.0322	95.2	0.4464 (0.0316)	0.0328	95.8
R0(3,1)	0.45	0.4469 (0.0372)	0.0370	95.6	0.4457 (0.0396)	0.0387	95.2	0.4450 (0.0398)	0.0395	96.2
R0(3,2)	0.45	0.4458 (0.0270)	0.0263	94.0	0.4464 (0.0284)	0.0293	96.6	0.4476 (0.0275)	0.0290	96.2
R0(4,1)	0.45	0.4484 (0.0400)	0.0389	95.0	0.4480 (0.0378)	0.0382	95.2	0.4435 (0.0377)	0.0382	95.4
R0(4,2)	0.45	0.4480 (0.0356)	0.0349	94.6	0.4469 (0.0378)	0.0372	93.8	0.4464 (0.0373)	0.0377	95.4
R0(4,3)	0.45	0.4468 (0.0375)	0.0364	95.2	0.4473 (0.0379)	0.0401	96.6	0.4463 (0.0367)	0.0389	96.2
R0(5,1)	0.45	0.4485 (0.0338)	0.0322	94.0	0.4453 (0.0428)	0.0422	94.6	0.4373 (0.0438)	0.0431	93.8
R0(5,2)	0.45	0.4477 (0.0310)	0.0308	95.8	0.4458 (0.0369)	0.0379	96.2	0.4342 (0.0374)	0.0397	95.6
R0(5,3)	0.45	0.4463 (0.0319)	0.0304	94.6	0.4442 (0.0404)	0.0410	95.2	0.4358 (0.0366)	0.0383	96.0
R0(5,4)	0.45	0.4478 (0.0310)	0.0306	94.4	0.4482 (0.0440)	0.0429	95.6	0.4413 (0.0445)	0.0444	95.4

3.5 Results

We applied the proposed method to the C-CFR study to obtain estimates of the diagnostic accuracy under the conditional dependence assumption. For the C-CFR data, we have $N = 3,487$ subjects and $J = 11$ tests. First, the TLC model under the conditional independence assumption was fit, and the estimates from the TLC model were used as the starting values for the PLC model under the conditional dependence assumption. The starting values for R_1 and R_0 are all set to 0.5. An MC sample of size $M = 1,500$ was simulated for each EM iteration with a burn-in of 1000 for the initial 10 EM iterations. We thinned a chain by keeping every other simulated draw. Trace plots of MCMC iterates $\sum_{k=1}^{K} n_k d_k^{(m)}$ and $\sum_{k=1}^{K} n_k d_k^{(m)} Z_k^{(m)}$ show that the chains have reached good mixing. To monitor the convergence of the PX-MCEM algorithm, we plotted the estimates of the parameters of prevalence, sensitivity, and specificity versus the number of iterations of the PX-MCEM algorithm. The convergence plots also indicate good convergence since all parameter estimates fluctuate randomly around the $\theta = \hat{\theta}$ line and stabilize (converge) after 100 iterations.

The SE estimates used for 95% CIs are based on $B = 1,000$ bootstrap samples randomly sampled with replacement from the original data. Figure 3.3 shows histograms of bootstrap samples for prevalence and diagnostic accuracy. Most of the histograms are well approximated by a Gaussian bell curve, which indicates that their sampling distributions are close to normal. Tests *BAT26* and *BAT34C4* have histograms that were cut off to the right boundary 1 due to their very high specificities. The QQ plots in Figure 3.4 also support the approximate normality of the bootstrap samples.

Figure 3.1 plots all $\frac{11\times(11-1)}{2} = 55$ pairwise correlation residuals for both the TLC and PLC models. In general, the deviation of the correlation residuals from the zero reference line is smaller for the PLC model, indicating that the model provided a better fit than the TLC model. The largest deviation involves tests *D17S250* and *D2S123*

because they have the highest missing probabilities (41.9% and 81.7%).

Table 2.3 presents the estimates of prevalence and diagnostic accuracy with 95% CIs from the TLC and PLC models. As expected, the estimates provided by the TLC and PLC models are very similar. Test *BAT25* has the highest sensitivity of 0.9394 while *D5S346* has the lowest sensitivity of 0.6294. All of the tests have high specifities (> 0.93) with the highest sensitivity of 0.9986 for test *BAT26*. *BAT25* and *BAT26* are shown to be the two best biomarkers (*BAT25* has the highest sensitivity and the third highest specificity; *BAT26* has the highest specificity and the second highest sensitivity.) They also happen to be two of the five biomarkers for the NCI-recommended microsatellite sequence panel, with the other three being *D2S123*, *D5S346*, and *D17S250*. The pairwise correlation coefficients R_1 and R_0 are listed in Table 3.3. Estimates of R_1 vary from 0.0953 to 0.6166, and estimates of R_0 vary from 0.2730 to 0.4863. The estimates of R_d must be viewed with caution since simulation studies have suggested that they are less accurate due to their much poorer starting values. We test H_0: all correlation coefficients of R_1 are equal to zero vs. H_1: at least some correlation coefficients of R_1 are greater than zero. Let $\hat{\theta}_{R_1}$ denote estimates of R_1 and $SE_{\hat{\theta}_{R_1}}$ denote standard errors of $\hat{\theta}_{R_1}$). The nominal p-values $\Phi(-|\frac{\hat{\theta}_{R_1}}{SE_{\hat{\theta}_{R_1}}}|)$ are adjusted by Bonferroni correction, i.e. multiplied by the total number of tests $\frac{J(J-1)}{2} = 55$. 40 out of the 55 p-values are less than 0.05. Therefore we reject H_0 and conclude that at least some test results are conditionally dependent for diseased subjects. Similarly, testing for R_0 supports conditional dependence assumption for non-diseased subjects.

3.6 Discussion

In this article we extended the use of the PLC model to the analysis of diagnostic tests for which many results are missing and for which there is no gold standard. The application we was concerned with the estimation of the prevalence and diagnostic

accuracy of 11 biomarker tests from the C-CFR study. A parameter expanded cumulative MCEM algorithm was implemented to facilitate the computation of the orthant probabilities of multivariate normal distributions. The PX-MCEM algorithm provides an analytically tractable M-step and also eases the complexity of evaluating conditional expectations in the E-step. We applied the TLC model under the conditional independence assumption to derive the starting values for the PLC model. Then, we assumed conditional dependence among the tests for the PLC model, which is more plausible due to the biologically similar basis of these tests. The estimates provided by the PLC model were fairly close to the estimates provided by the TLC model. This is consistent with our simulation studies that also showed that the two models produced similar results. We are confident that the estimates from the TLC model are useful as first approximations for subsequent iterations of the PX-MCEM algorithm. The PLC model generated estimates that are even closer to the true values than those of the TLC model. Although there is no established way to assess the underlying missing data mechanism, when the starting values used in the PLC model are close to the true values and the missing percentages are moderate, the model is robust irrespective of the underlying assumptions concerning missing data. In cases in which the percentages of missing data are high, multiple imputation can be used, but this is beyond the scope of this paper.

Another advantage of the PLC model is that it readily can be extended to explicitly model the effects of the characteristics of an individual (including, but not limited to, gender, race, age at diagnosis, and stage of colon cancer) on prevalence and diagnostic accuracy. Let X_i denote the covariate vector of the characteristic of subject i. Then, we assume that $Z_i|D_i = d \sim N_J(X_i\mathbf{B}_d, \Sigma_d)$, where $d = 1$ or 0. This will be fully investigated in our future research.

It is important to note that the final estimates are highly sensitive to the starting

values. We also found that starting values affect the speed of convergence, i.e., it took much longer to reach convergence with poor starting values, which would be computationally prohibitive if we repeat the method on a large amount of bootstrap samples to get SE estimates. Therefore, it is critical to find even better starting values. For prevalence and diagnostic accuracy, we can resort to other models that allow for conditional dependence, such as the Gaussian random effects model (Qu et al., 1996). Unfortunately, no method is available for achieving viable starting values for correlation coefficients. We cannot use the observed correlations between T_{ij} and $T_{ij'}$ because they are very different from the correlations between the latent variables Z_{ij} and $Z_{ij'}$, due to loss of information from the dichotomization of continuous Z_i to binary T_i. For this article we used an arbitrary starting value of 0.5, which results in much less accurate estimates of R_1 and R_0. It will be beneficial to identify starting values for R_d that are sufficiently reliable. One potential strategy would be to assume certain simplified structures for R_1 and R_0 that could be justified scientifically and/or on the basis of expert opinion, e.g., it could be assumed that the tests are dependent for the diseased subjects but independent for the non-diseased subjects, i.e., all of the off-diagonal entries of R_0 are zero.

One limit of the PLC model is that the number of tests must be $J \geq 5$. The reason for this is that the total number of model parameters $(J^2 + J + 1)$ cannot exceed the degrees of freedom of the observed data $(2^J - 1)$ for a model to be identifiable. For studies that involve less than 5 tests, other methods must be considered, e.g., Bayesian methods that incorporate non-identifiability in the likelihood (Dendukuri and Joseph, 2001), fixing the values of certain parameters (Hui and Walter, 1980), and methods that address partial identifiability (Jones et al., 2010).

Simulation studies demonstrate compelling closeness of bootstrap SE estimates to SD of the PLC model parameter estimates, even under the nonignorable missing data

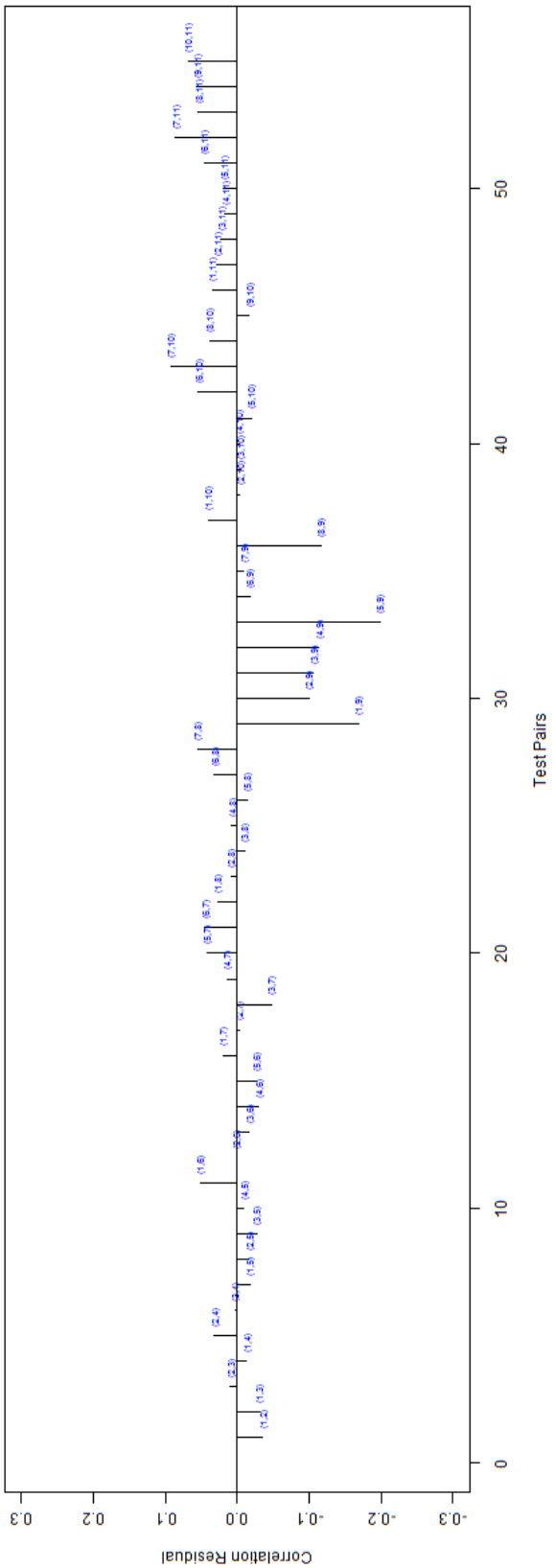
assumption. Nevertheless, Efron (1994) considered the bootstrap estimates to be invalid under the nonignorable missing data assumption. Another drawback of the bootstrap method is that it can be extremely time-consuming for studies that have a large number of subjects N and/or a large number of tests J. Then, it may be essential to investigate alternative methods for SE estimation, e.g., the Louis formula (1982) and supplemented EM (SEM) algorithm (Meng and Rubin, 1991). For the C-CFR data, the Louis formula did not work out probably due to the high sensitivities and specificities. We expect the Louis formula will perform better with moderate sensitivities and specificities in further research.

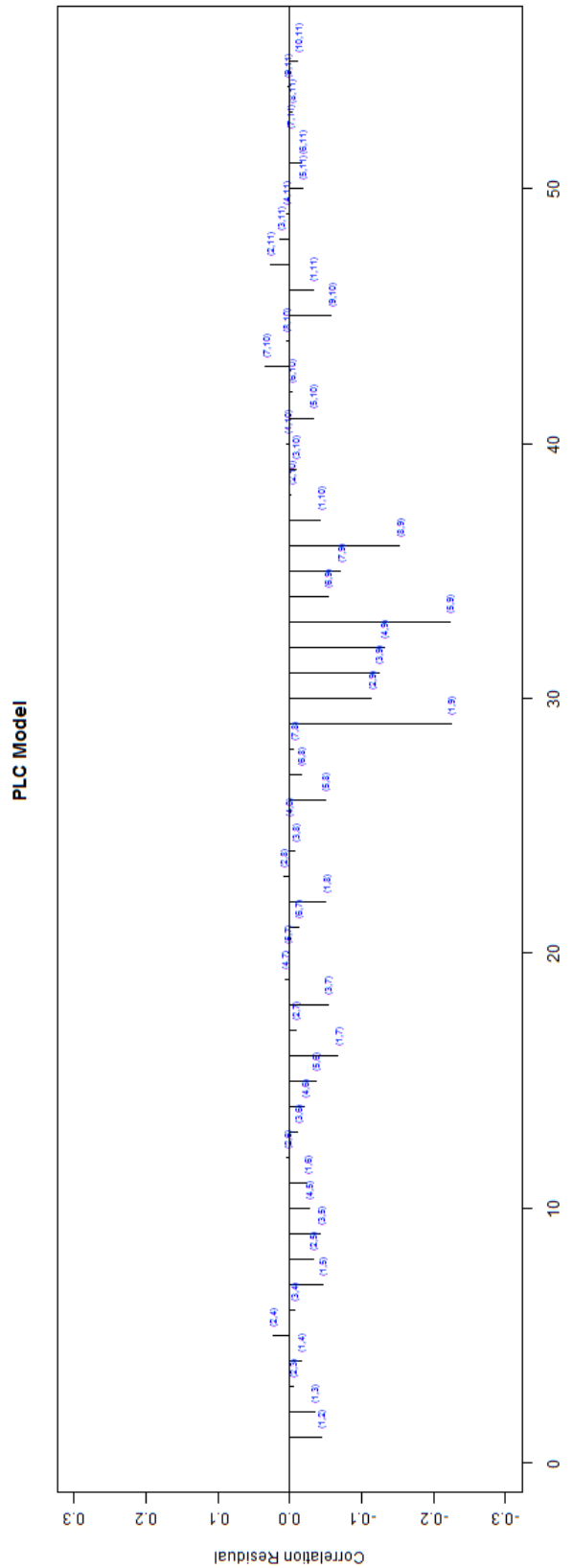
Computation challenges are posed by both bootstrap resampling and MCMC simulations during the E-steps. Not to mention that the number of parameters increases quadratically for high-dimensional correlation matrices. For a modest number of tests, Gaussian-Hermite quadrature can be a more efficient alternative for approximating integrals over a finite number of quadrature points. But the C-CFR data require much more computing resources with the large number of quadrature points required for high dimensions. Embarrassingly parallel computing was employed for data simulation and bootstrap. By contrast, MCMC simulations are based on correlated sampling, i.e., the input of the next iteration is dependent on the output from the previous iteration. Such iterations are not naturally convertible into parallel code because they cannot be executed concurrently. To render the PX-MCEM algorithm more computationally feasible, we will explore methods for parallelization of MCMC, for instance, by distributing entire chains or parts of chains to different processors (Feng et al., 2003).

One concern over latent class models is the lack of clinical definition of disease, prevalence, and diagnostic accuracy, because disease is an implicitly defined random variable (Pepe and Janes, 2007). There are certain cases in which the nature of disease has a spectrum of severity rather than being binary. For our case study, while it

was expected that the PLC model would provide a reasonable latent structure for the biology in HNPCC, additional caution must be exercised in the interpretation of the study results. It also is worthwhile to note that there is no substitute for a gold standard test when one is available, even if the gold standard is applied to only a small fraction of the subjects. Another area for potential future research is to determine whether the previously mentioned mutational analysis (a costly gold standard) can be applied to at least some patients to facilitate the evaluation of the 11 MSI biomarker tests. Interdisciplinary collaboration between statisticians, clinicians, and laboratory scientists is vital for the achievement of these research goals.

Figure 3.1: Correlation Residual Plots for C-CFR Data





Note: 1=ACTC, 2=BAT25, 3=BAT26, 4=BAT40, 5=BAT34C4, 6=D10S197, 7=D17S250, 8=D18S55, 9=D2S123, 10=D5S346, 11=MYCL

Figure 3.2: Correlation Residual Plots for Simulated Data Sets

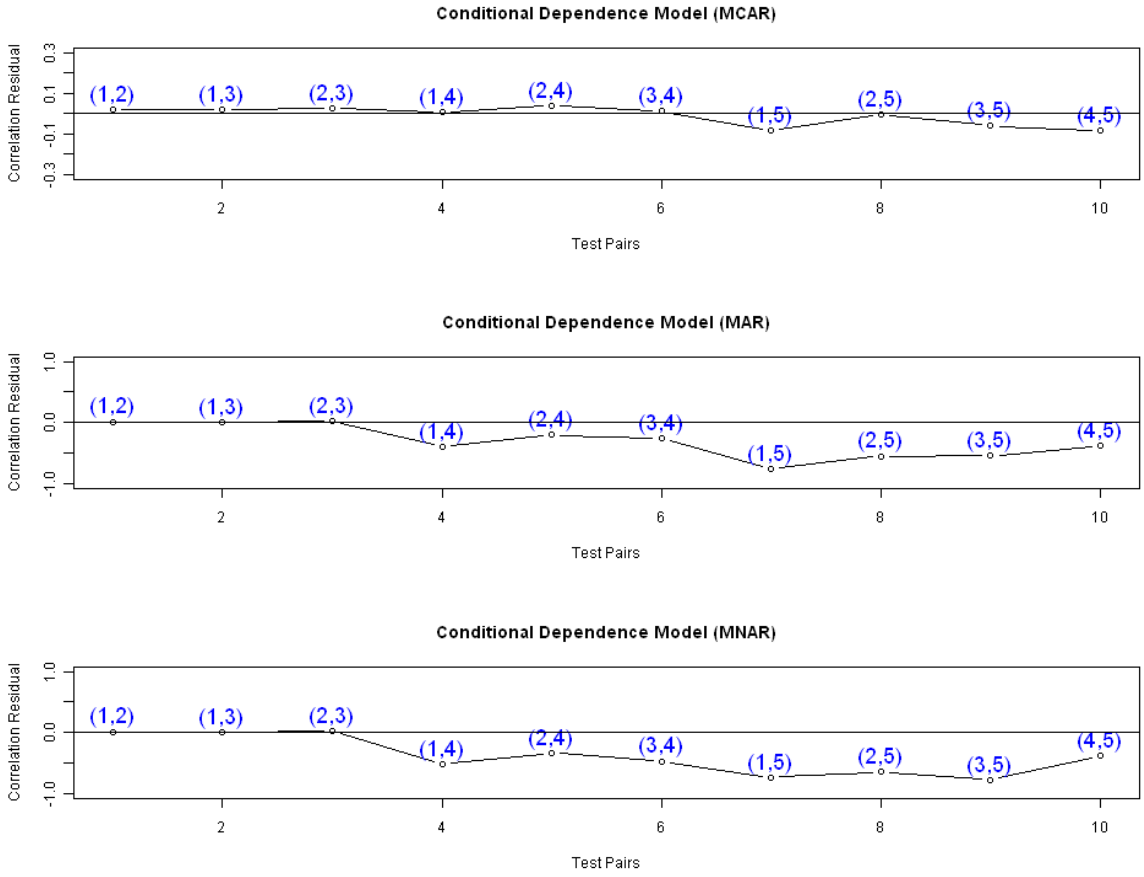
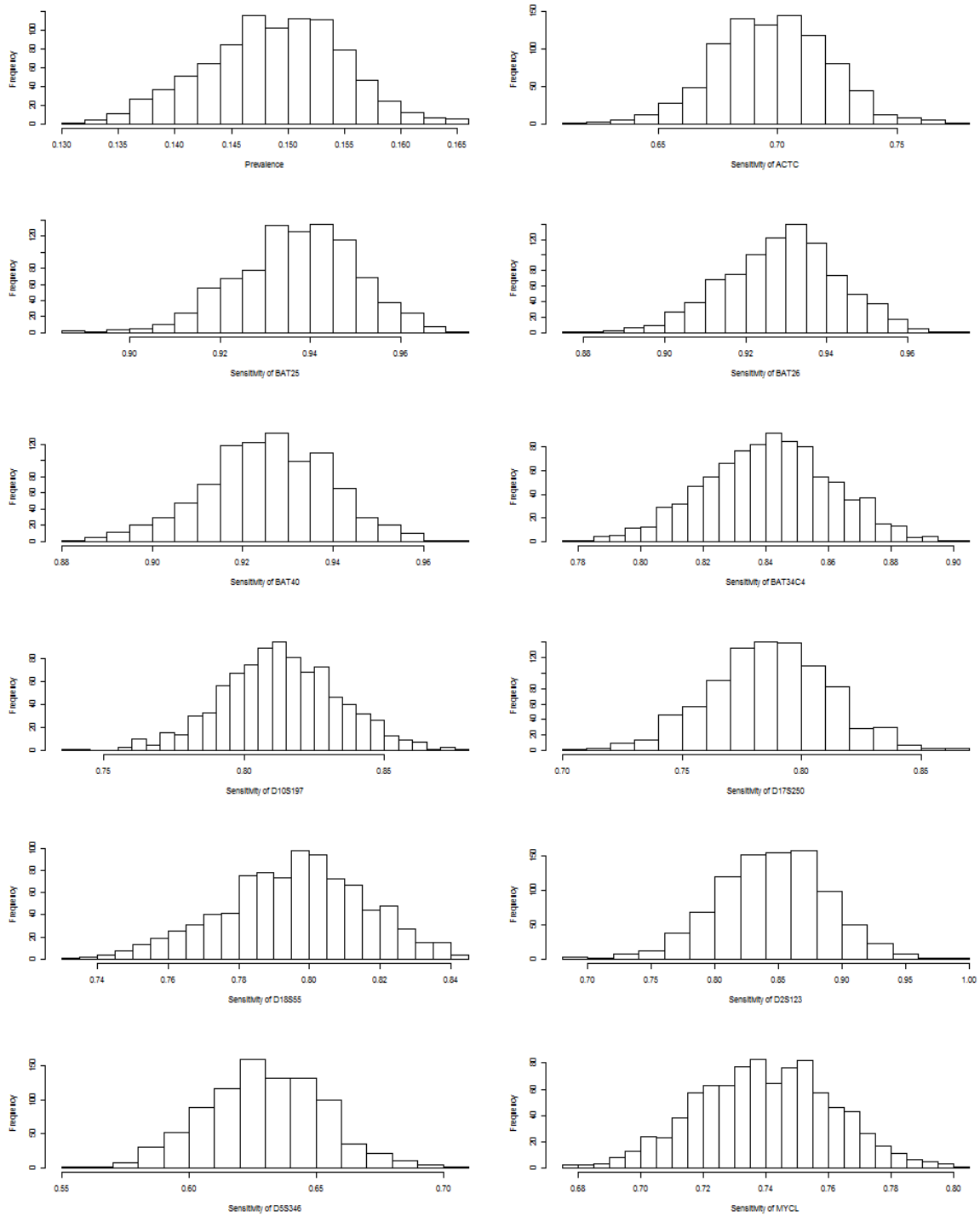
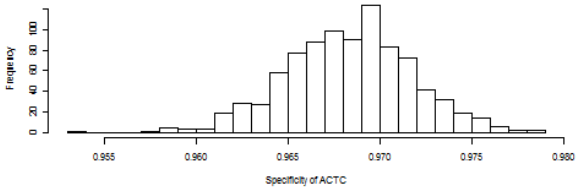
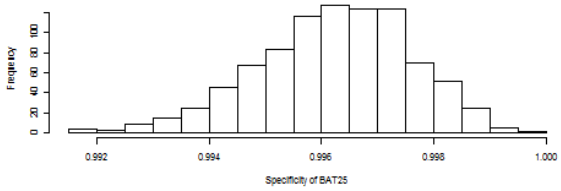
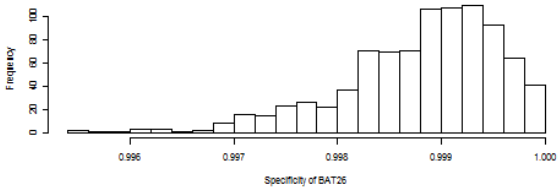


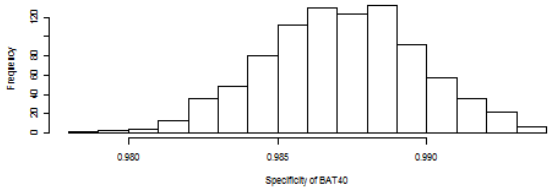
Figure 3.3: Histograms of Bootstrap Samples (B=1000) for Prevalence and Diagnostic Accuracy

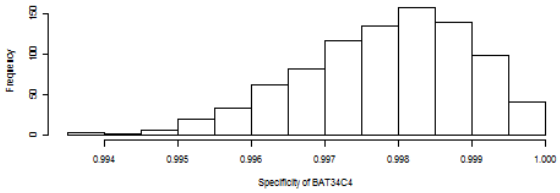


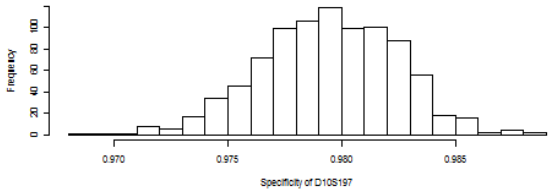

Frequency
0
20
40
60
80
100
0.955
0.960
0.965
0.970
0.975
0.980
Specificity of ACTC

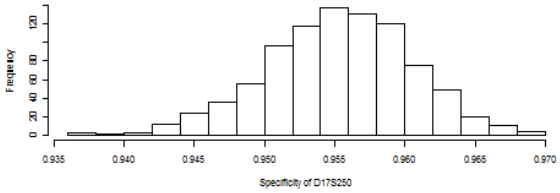

Frequency
0
20
40
60
80
100
0.992
0.994
0.996
0.998
1.000
Specificity of BAT25

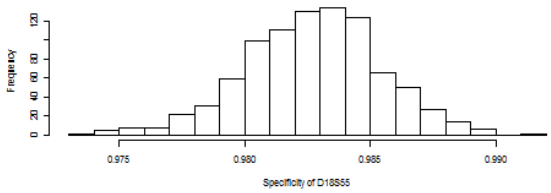

Frequency
0
20
40
60
80
100
0.996
0.997
0.998
0.999
1.000
Specificity of BAT26

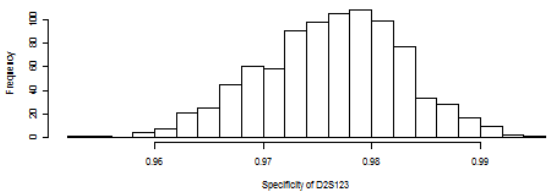

Frequency
0
20
40
60
80
100
120
0.980
0.985
0.990
Specificity of BAT40

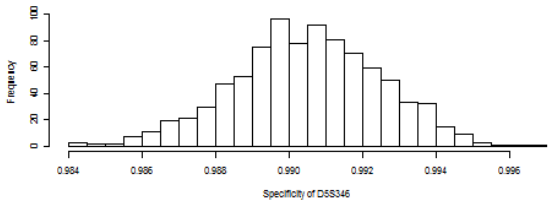

Frequency
0
50
100
150
0.994
0.995
0.996
0.997
0.998
0.999
1.000
Specificity of BAT34C4


Frequency
0
20
40
60
80
100
0.970
0.975
0.980
0.985
Specificity of D10S197


Frequency
0
20
40
60
80
100
120
0.935
0.940
0.945
0.950
0.955
0.960
0.965
0.970
Specificity of D17S250


Frequency
0
20
40
60
80
100
120
0.975
0.980
0.985
0.990
Specificity of D18S55


Frequency
0
20
40
60
80
100
0.96
0.97
0.98
0.99
Specificity of D2S123


Frequency
0
20
40
60
80
100
0.984
0.986
0.988
0.990
0.992
0.994
0.996
Specificity of D5S346

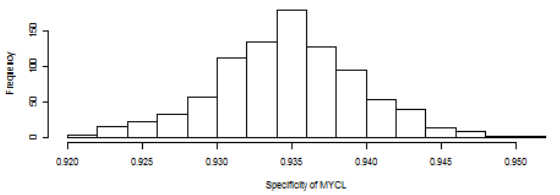
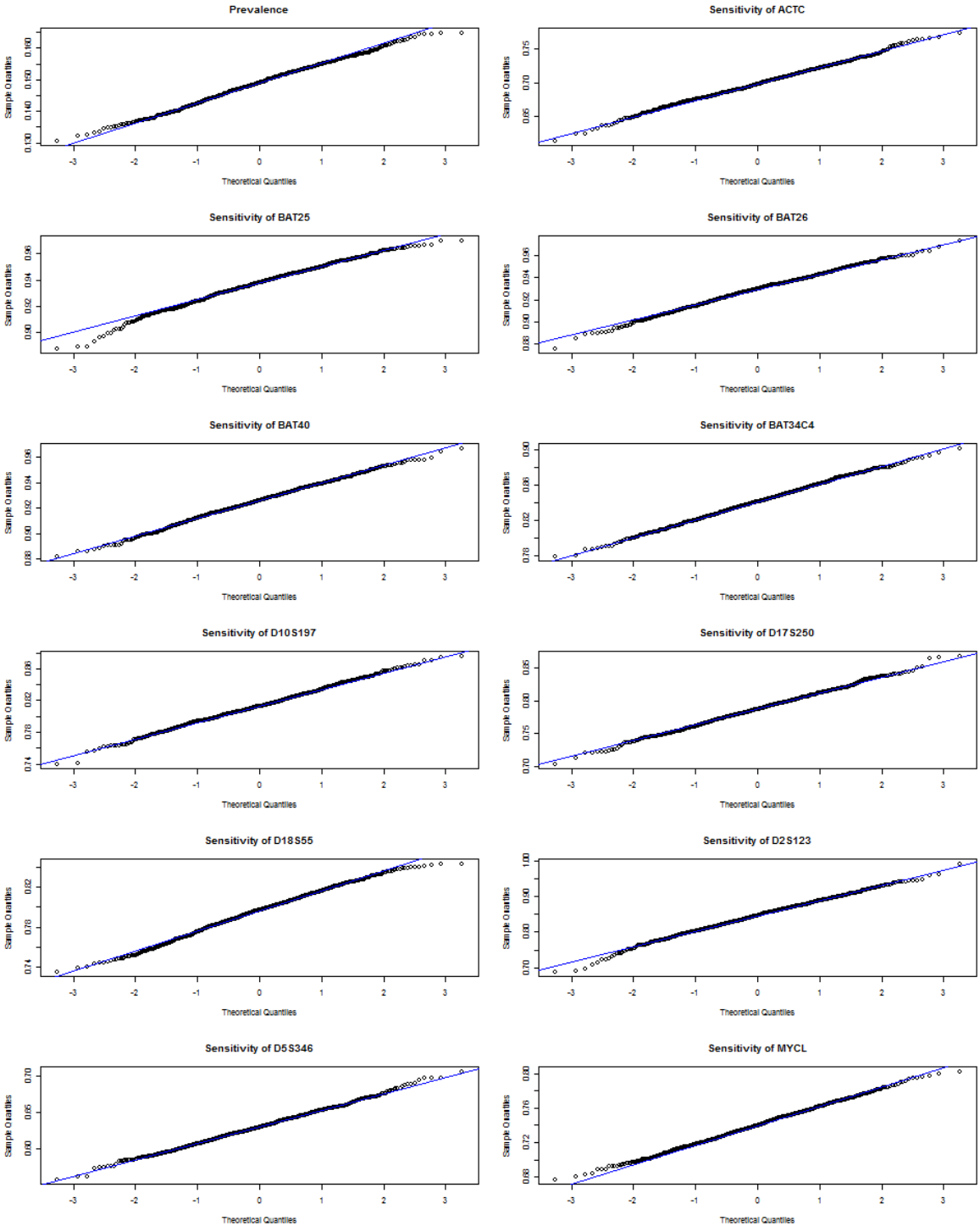
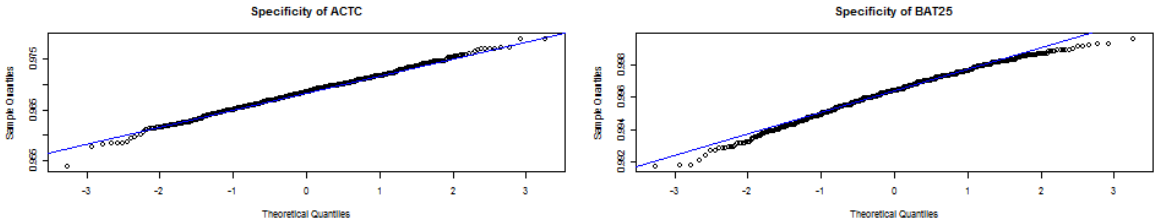
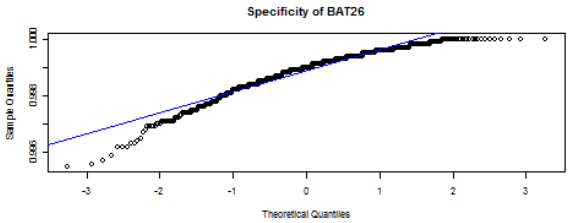

Frequency
0
50
100
150
0.920
0.925
0.930
0.935
0.940
0.945
0.950
Specificity of MYCL

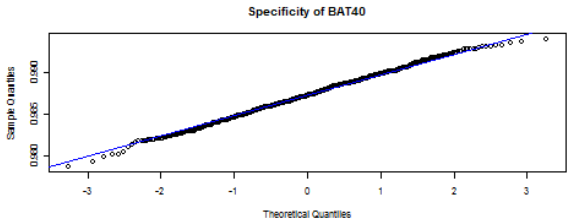
Figure 3.4: QQ Plots of Bootstrap Samples (B=1000) for Prevalence and Diagnostic Accuracy

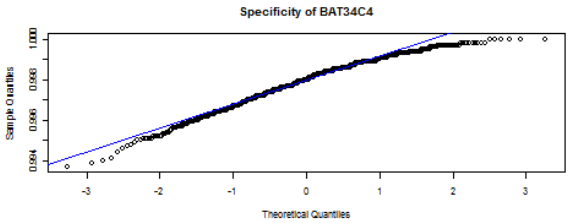


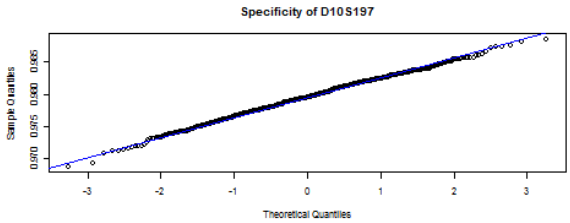

Specificity of ACTC
Sample Quantiles
Theoretical Quantiles

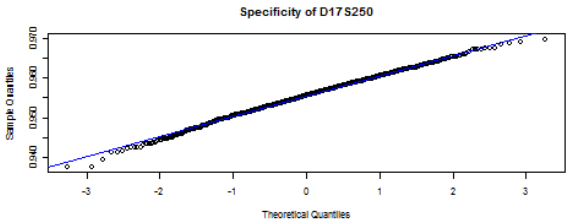
Specificity of BAT25
Sample Quantiles
Theoretical Quantiles

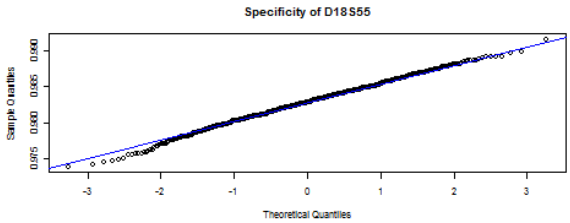

Specificity of BAT26
Sample Quantiles
Theoretical Quantiles

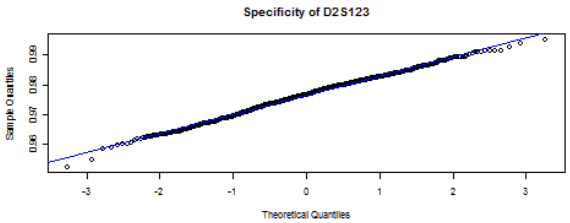

Specificity of BAT40
Sample Quantiles
Theoretical Quantiles

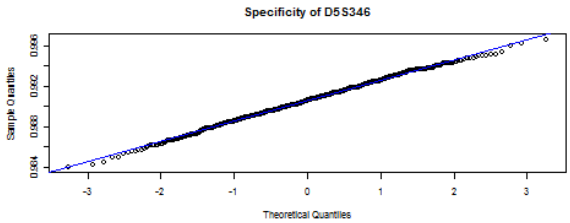

Specificity of BAT34C4
Sample Quantiles
Theoretical Quantiles


Specificity of D10S197
Sample Quantiles
Theoretical Quantiles


Specificity of D17S250
Sample Quantiles
Theoretical Quantiles


Specificity of D18S55
Sample Quantiles
Theoretical Quantiles


Specificity of D2S123
Sample Quantiles
Theoretical Quantiles


Specificity of D5S346
Sample Quantiles
Theoretical Quantiles

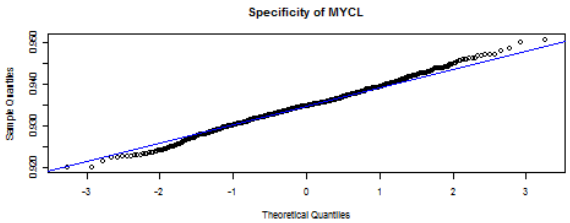

Specificity of MYCL
Sample Quantiles
Theoretical Quantiles

Table 3.2: Estimates and 95% CIs of Se, Sp, and Prevalence from Different Models

	TLC Model		**PLC Model**	
	Se (95%CI)	Sp (95%CI)	Se (95%CI)	Sp (95%CI)
ACTC	0.7111 (0.665, 0.753)	0.9700 (0.963, 0.976)	0.6982 (0.650, 0.743)	0.9680 (0.961, 0.974)
BAT25	0.9405 (0.915, 0.959)	0.9962 (0.993, 0.998)	0.9394 (0.907, 0.961)	0.9958 (0.992, 0.998)
BAT26	0.9308 (0.903, 0.951)	0.9985 (0.996, 0.999)	0.9314 (0.897, 0.955)	0.9986 (0.996, 1.000)
BAT40	0.9301 (0.902, 0.951)	0.9867 (0.981, 0.991)	0.9268 (0.895, 0.950)	0.9869 (0.981, 0.991)
BAT34C4	0.8499 (0.812, 0.882)	0.9980 (0.995, 0.999)	0.8434 (0.799, 0.879)	0.9977 (0.994, 0.999)
D10S197	0.8353 (0.796, 0.868)	0.9816 (0.975, 0.986)	0.8153 (0.771, 0.853)	0.9787 (0.972, 0.984)
D17S250	0.8120 (0.762, 0.854)	0.9564 (0.946, 0.965)	0.7881 (0.735, 0.833)	0.9543 (0.943, 0.963)
D18S55	0.8098 (0.770, 0.844)	0.9842 (0.978, 0.988)	0.7967 (0.754, 0.834)	0.9822 (0.976, 0.987)
D2S123	0.8493 (0.747, 0.915)	0.9769 (0.960, 0.987)	0.8444 (0.740, 0.912)	0.9766 (0.960, 0.987)
D5S346	0.6455 (0.600, 0.688)	0.9923 (0.988, 0.995)	0.6294 (0.583, 0.673)	0.9900 (0.985, 0.993)
MYCL	0.7587 (0.716, 0.797)	0.9360 (0.926, 0.945)	0.7419 (0.697, 0.782)	0.9348 (0.925, 0.944)
Prevalence		0.1482 (0.137, 0.160)		0.1485 (0.137, 0.161)

Table 3.3: Estimates and 95% CIs of R_1 and R_0 from the PLC Model

	R_1 (95% CI)	R_0 (95% CI)		R_1 (95% CI)	R_0 (95% CI)
(1,2)	0.3210 (0.154, 0.550)	0.3885 (0.333, 0.448)	(1,9)	0.4213 (0.260, 0.602)	0.4275 (0.367, 0.490)
(1,3)	0.2516 (0.091, 0.531)	0.3885 (0.337, 0.442)	(2,9)	0.3470 (0.209, 0.517)	0.4537 (0.404, 0.504)
(2,3)	0.4446 (0.266, 0.639)	0.4548 (0.411, 0.499)	(3,9)	0.4111 (0.259, 0.582)	0.4513 (0.407, 0.496)
(1,4)	0.1147 (0.017, 0.488)	0.3881 (0.320, 0.461)	(4,9)	0.2856 (0.158, 0.460)	0.4322 (0.387, 0.479)
(2,4)	0.3370 (0.185, 0.533)	0.4196 (0.376, 0.465)	(5,9)	0.3759 (0.230, 0.548)	0.4670 (0.423, 0.511)
(3,4)	0.3346 (0.176, 0.543)	0.4500 (0.405, 0.495)	(6,9)	0.4498 (0.300, 0.609)	0.4257 (0.371, 0.483)
(1,5)	0.3643 (0.213, 0.549)	0.3923 (0.343, 0.444)	(7,9)	0.5443 (0.402, 0.679)	0.4786 (0.417, 0.540)
(2,5)	0.4630 (0.287, 0.649)	0.4625 (0.420, 0.506)	(8,9)	0.3715 (0.217, 0.557)	0.4561 (0.404, 0.509)
(3,5)	0.4141 (0.233, 0.622)	0.4604 (0.418, 0.503)	(1,10)	0.4657 (0.330, 0.607)	0.4236 (0.360, 0.490)
(4,5)	0.4651 (0.284, 0.656)	0.4325 (0.389, 0.477)	(2,10)	0.2448 (0.088, 0.521)	0.4559 (0.407, 0.505)
(1,6)	0.6008 (0.462, 0.725)	0.4180 (0.344, 0.496)	(3,10)	0.2904 (0.116, 0.560)	0.4804 (0.432, 0.529)
(2,6)	0.3202 (0.153, 0.552)	0.4472 (0.391, 0.505)	(4,10)	0.1011 (0.011, 0.529)	0.4432 (0.393, 0.495)
(3,6)	0.2847 (0.115, 0.551)	0.4298 (0.382, 0.479)	(5,10)	0.2575 (0.122, 0.464)	0.4421 (0.398, 0.487)
(4,6)	0.0953 (0.011, 0.492)	0.4053 (0.354, 0.458)	(6,10)	0.5068 (0.369, 0.644)	0.4819 (0.411, 0.554)
(5,6)	0.3567 (0.202, 0.549)	0.4446 (0.396, 0.495)	(7,10)	0.4654 (0.322, 0.614)	0.4863 (0.410, 0.563)
(1,7)	0.6048 (0.471, 0.725)	0.3868 (0.313, 0.466)	(8,10)	0.3232 (0.186, 0.499)	0.4546 (0.395, 0.516)
(2,7)	0.4777 (0.302, 0.659)	0.4148 (0.362, 0.470)	(9,10)	0.3758 (0.220, 0.562)	0.4643 (0.414, 0.515)
(3,7)	0.4692 (0.282, 0.666)	0.4440 (0.394, 0.496)	(1,11)	0.4589 (0.321, 0.604)	0.2730 (0.199, 0.362)
(4,7)	0.3659 (0.203, 0.566)	0.3794 (0.322, 0.440)	(2,11)	0.2615 (0.114, 0.494)	0.3593 (0.299, 0.424)
(5,7)	0.5990 (0.457, 0.726)	0.4191 (0.366, 0.474)	(3,11)	0.4034 (0.220, 0.619)	0.3832 (0.333, 0.436)
(6,7)	0.5878 (0.451, 0.712)	0.4629 (0.389, 0.539)	(4,11)	0.2955 (0.141, 0.518)	0.3430 (0.273, 0.420)
(1,8)	0.5552 (0.422, 0.681)	0.4138 (0.347, 0.484)	(5,11)	0.5000 (0.344, 0.656)	0.3665 (0.315, 0.421)
(2,8)	0.2486 (0.100, 0.497)	0.4159 (0.369, 0.464)	(6,11)	0.6089 (0.472, 0.731)	0.3122 (0.244, 0.390)
(3,8)	0.1848 (0.051, 0.491)	0.4604 (0.414, 0.507)	(7,11)	0.6166 (0.483, 0.735)	0.3583 (0.282, 0.443)
(4,8)	0.2261 (0.078, 0.502)	0.4488 (0.385, 0.514)	(8,11)	0.4739 (0.330, 0.622)	0.3854 (0.311, 0.466)
(5,8)	0.4853 (0.328, 0.645)	0.4377 (0.392, 0.485)	(9,11)	0.4171 (0.275, 0.575)	0.4076 (0.348, 0.470)
(6,8)	0.5287 (0.385, 0.668)	0.4689 (0.400, 0.539)	(10,11)	0.5228 (0.395, 0.648)	0.3567 (0.293, 0.426)
(7,8)	0.5735 (0.428, 0.707)	0.4441 (0.373, 0.518)			

Note: 1=ACTC, 2=BAT25, 3=BAT26, 4=BAT40, 5=BAT34C4, 6=D10S197, 7=D17S250, 8=D18S55, 9=D2S123, 10=D5S346, 11=MYCL

Chapter 4

DiagLCA - An R Package for the Evaluation of Binary Tests

4.1 Introduction

In diagnostic medicine, observed signs, symptoms, or test results are commonly dichotomized into two possible outcomes, i.e., "positive" or "negative." Evaluation of the diagnostic accuracy (sensitivity and specificity) of such binary tests is of great importance because reliable diagnoses of patients' medical conditions are critical in health practitioners' treatment plans. Ideally, diagnostic accuracy of a new test could be evaluated by comparing its results with the test results of a gold standard, which would definitively separate those subjects with disease from those without disease. In other words, a gold standard is, by definition, error-free with both sensitivity and specificity equal to 1. In practice, a gold standard may not exist or is too costly/invasive to apply. Latent class analysis (LCA) is a group of popular methods that assess the diagnostic accuracy of multiple imperfect tests in the absence of a gold standard. It treats the unobserved true disease status as a latent variable with binary classification (present or absent). The latent variables can only be evaluated indirectly through observable measurements called manifest variables, e.g. observed test results and/or

patient characteristics (e.g., gender and age).

LCA has a sound theoretical basis in maximum likelihood (ML) or Bayesian methodologies. For the ML approach, the estimation of parameters by latent class models usually involve iterative methods, such as the Fisher scoring algorithm (Espeland and Handelman, 1989), the Newton-Raphson method (Qu and Hadgu, 1998), and the expectation-maximization (EM) algorithm (Dempster et al., 1977; Dawid and Skene, 1979). For the Bayesian approach, the parameters often are estimated by Markov Chain Monte Carlo (MCMC) methods via Gibbs sampling (Joseph et al., 1995; Dendukuri and Joseph, 2001). The Bayesian approach is particularly useful by incorporating prior information to address non-identifiability situations when the number of parameters to be estimated exceeds the available degrees of freedom. However, it is sensitive to the prior distribution that is chosen. Great caution must be taken to avoid any bias when collecting prior information. For this article, we focused on ML-based LCA approaches. The traditional latent class (TLC) model assumes the tests are conditionally independent given the true disease status, known as the conditional independence assumption (Hui and Zhou, 1998; Walter and Irwig, 1988; Rindskopf and Rindskopf, 2006). When the tests share the same attribute, the conditional independence assumption no longer holds. Latent class models relaxed for the conditional dependence assumption have been developed in a dispersed literature (Albert et al., 2004; Espeland and Handelman, 1989; Yang and Becker, 1997; Qu et al., 1996; Qu and Hadgu, 1998; Uebersax, 1999; Chib and Greenberg, 1998; Xu and Craig, 2009), including the probit latent class (PLC) model proposed by Uebersax (1999) and introduced to diagnostic accuracy estimation by Xu and Craig (2009). The PLC model accommodates conditional dependence among tests with a general correlation structure assuming a multivariate-normal distribution within each latent class. A parameter expanded cumulative MCEM (PX-MCEM) algorithm is implemented to facilitate an analytically tractable M-step and an computationally

manageable E-step (Liu et al., 1998).

Several R packages have been developed for evaluation of the diagnostic accuracy of binary tests. **DiagnosisMed** is applied in evaluation of an index test with comparison to a gold standard (Brasil, 2010). **DTComPair** computes the accuracy of two binary diagnostic tests in a paired study design that requires a gold standard (Stock et al., 2013). R packages utilizing LCA in the absence of a gold standard are also available. **lcmr** estimates latent class models with random effects with the Bayesian approach (Wang and Dendukuri, 2012). **TAGS** generates ML estimates using the Newton-Raphson method and EM algorithms based on Hui and Walter's model (1980) under the conditional independence assumption (Pouillot et al., 2002). **randomLCA** utilizes Qu's random effects model (1996) to allow for condition dependence between tests (Beath, 2011). However, none of the aforementioned packages can handle missing tests, which are very common in diagnostic medicine, i.e., a test kit may be out of stock during a patient's visit; the doctor may decide to withhold the test; and the patient may decline the test. Essentially, there are three types of missing data mechanisms (Little and Rubin, 1987): missing completely at random (MCAR) when the probability of missing does not depend on any missing or observed observations; missing at random (MAR) when the probability of missing does not depend on any other missing observations, but can depend on some observed observations; missing not at random (MNAR) when the probability of missing depends on some missing observations or latent disease status. MCAR and MAR are both "ignorable" missing data mechanisms whereas MNAR is a "non-ignorable" (NI) missing data mechanism. Research that deal with missing tests are available when a gold standard is present (Kosinski and Barnhart, 2003b; Zhou, 1993; Alonzo, 2005; Lin et al., 2006; He and McDermott, 2012; Yu et al., 2010; Harel and Zhou, 2006, 2007; Harrell et al., 1996). But to our knowledge, there is no literature on missing tests evaluation in the absence of a gold standard.

In this paper, we introduce the **DiagLCA** package to address the missing data problem in the evaluation of multiple correlated diagnostic tests without a gold standard. Section 4.2 provides a methodological background of the TLC and PLC models. Section 4.3 gives an overview of the **DiagLCA** package. Section 4.4 presents a step-by-step demonstration of the usage of the package through a real-world example. Section 4.5 concludes with a brief summary.

4.2 Methodology

The **DiagLCA** package extends Xu and Craig's (2009) PLC model to estimate the diagnostic accuracy of conditionally dependent tests when some results are missing. It also has the capability of fitting a TLC model under the conditional independence assumption. We give a brief overview of the TLC and PLC models and introduce all parameters needed as follows:

- N is the total number of subjects.
- J is the total number of tests.
- t_{ij} is the test result of the j^{th} test for the i^{th} subject.
- $T_i = (t_{i1}, \cdots, t_{iJ})$ is for all J test results of the i^{th} subject.
- $Z_i = (z_{i1}, \cdots, z_{iJ})$ is a vector of Gaussian latent variables with a multivariate normal distribution $N_J(\mu_d, \Sigma_d)$. $z_{ij} > 0$ when $t_{ij} = 1$; $z_{ij} <= 0$ when $t_{ij} = 0$; z_{ij} could take any value within $(-\infty, \infty)$ when t_{ij} is missing.
- δ_{ij} indicates whether the i^{th} subject was tested by the j^{th} test (1=Yes).
- $\Delta_i = (\delta_{i1}, \cdots, \delta_{iJ})$ is all non-missing indicators of the i^{th} subject.
- $D_i = d(d = 1, 0)$ is latent disease status (1=Yes).

- $\pi_d = Pr(D_i = d)$ is probability of disease/no disease (π_1 is prevalence).
- Se_j and Sp_j are sensitivity and specificity of the j^{th} test.
- a_d $(d = 1, 0)$ are mean vectors of $N_J(\mu_d, \Sigma_d)$. Notice the relationship $Se_j = \Phi(a_{1j})$ and $Sp_j = \Phi(-a_{0j})$.
- R_d $(d = 1, 0)$ are correlation matrices of $N_J(\mu_d, \Sigma_d)$.
- $\theta = (\pi_1, Se_1, \cdots, Se_J, Sp_1, \cdots, Sp_J)$ is the vector of all parameters for the TLC model.
- $\eta = (\pi_1, a_{11}, \cdots, a_{1J}, a_{01}, \cdots, a_{0J}, R_{1,12}, \cdots, R_{1,(J-1)J}, R_{0,12}, \cdots, R_{0,(J-1)J})$ is the vector of all parameters for the PLC model.

Assuming the unobserved disease status D_i is known for each subject i, the log-likelihood of complete data $Y_i = (T_i, \Delta_i, D_i)$ under the conditional independence assumption is $logL_c(\theta) = \sum_{i=1}^{N}(d_i log(\pi_1 h_{i1}) + (1 - d_i) log(\pi_0 h_{i0}))$, where

$$\begin{aligned} h_{i1} &= \prod_{j=1}^{J} Se_j^{t_{ij}\delta_{ij}}(1 - Se_j)^{(1-t_{ij})\delta_{ij}} \\ h_{i0} &= \prod_{j=1}^{J}(1 - Sp_j)^{t_{ij}\delta_{ij}} Sp_j^{(1-t_{ij})\delta_{ij}}. \end{aligned}$$

θ is then estimated using the EM algorithm. Assume $\theta^{(n)}$ is known at iteration n for $n = 0, 1, \cdots$, ($\theta^{(0)}$ is starting value). Solving the score equations during the M-step and computing the conditional expectations of the complete-data sufficient statistics during the E-step, we get closed-form solutions for $\theta^{(n+1)}$:

$$
\begin{aligned}
\pi_1^{(n+1)} &= \frac{\sum_{i=1}^{N} \frac{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{N} \\
Se_j^{(n+1)} &= \frac{\sum_{i=1}^{N} \frac{t_{ij}\delta_{ij}\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{\sum_{i=1}^{N} \frac{\delta_{ij}\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}} \\
Sp_j^{(n+1)} &= \frac{\sum_{i=1}^{N} \frac{(1-t_{ij})\delta_{ij}(1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{\sum_{i=1}^{N} \frac{\delta_{ij}(1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}
\end{aligned}
$$

We iterate EM steps until convergence to get the ML estimate $\hat{\theta}$ for the TLC model. Now we move on to describe the PLC model. The PX-MCEM algorithm is used to estimate η. To ensure a closed-form solution for the M-step, the parameter vector η is expanded to $\beta = (\pi_1, \mu_1, \Sigma_1, \mu_0, \Sigma_0)$ where $\mu_d = V_d^{\frac{1}{2}} a_d$ and $\Sigma_d = V_d^{\frac{1}{2}} R_d V_d^{\frac{1}{2}}$ (Liu et al., 1998). Z_i is introduced to account for the similarity between tests of subject i under the conditional dependence assumption. The log-likelihood of complete data $Y_i = (T_i, \Delta_i, Z_i, D_i)$ is

$$
\begin{aligned}
\log L_c(\beta) &= \log \prod_{i=1}^{N} \{\pi_1^{d_i}(1-\pi_1)^{(1-d_i)} \phi_J(Z_i; \mu_1, \Sigma_1)^{d_i} \phi_J(Z_i; \mu_0, \Sigma_0)^{(1-d_i)} \prod_{i=1}^{J} \\
&\quad I(z_{ij} \in B_{ij})\} \\
&= \sum_{i=1}^{N} \{d_i \log(\pi_1) + (1-d_i)\log(1-\pi_1) + d_i \log(\phi_J(Z_i; \mu_1, \Sigma_1)) \\
&\quad + (1-d_i)\log(\phi_J(Z_i; \mu_0, \Sigma_0)) + \sum_{i=1}^{J} \log(I(z_{ij} \in B_{ij}))\}
\end{aligned}
$$

The closed-form solutions for $\beta^{(n+1)}$ is:

$$
\begin{aligned}
\pi_1^{(n+1)} &= \frac{1}{N}\sum_{i=1}^{N} E[d_i|T_i,\Delta_i,\beta^{(n)}] \\
\mu_d^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}Z_i|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} \\
\Sigma_d^{(n+1)} &= \frac{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}(Z_i-\mu_d)(Z_i-\mu_d)'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} \\
&= \frac{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}Z_iZ_i'|T_i,\Delta_i,\beta^{(n)}]}{\sum_{i=1}^{N} E[d_i^d(1-d_i)^{1-d}|T_i,\Delta_i,\beta^{(n)}]} - \mu_d^{(n+1)}(\mu_d^{(n+1)})'
\end{aligned}
$$

Reduce the expanded parameter through U_d (a diagonal matrix with diagonal elements equal to those of $\Sigma_d^{(n+1)}$), we get the estimates for a_d and R_d:

$$
\begin{aligned}
a_d^{(n+1)} &= U_d^{-\frac{1}{2}}\mu_d^{(n+1)} \\
R_d^{(n+1)} &= U_d^{-\frac{1}{2}}\Sigma_d^{(n+1)}U_d^{-\frac{1}{2}}
\end{aligned}
$$

.

Now, we can compute the above conditional expectations of the expanded complete data sufficient statistics via a Markov chain Monte Carlo (MCMC) method. As for standard errors of point estimates, **DiagLCA** employs the Louis formula (1982) for the TLC model and the bootstrap method for the PLC model (Efron, 1979). **DiagLCA** provides a graphic check for the conditional independence assumption (Qu et al., 1996; Chu et al., 2009). If the graphic check or other resources (clinician opinions, biology knowledge, etc.) dictate the conditional independence assumption is appropriate in practice, one may proceed with the TLC model to get diagnostic accuracy. Otherwise if the conditional independence assumption does not hold, users can go with the PLC option.

Starting values for the parameters of the PLC model are needed in order to initiate the iterations of the PX-MCEM algorithm. The final estimates of the PLC model are highly sensitive to the starting values. One common practice for the EM algorithm is to try different sets of starting values to ensure that a global maximum is reached. We do not recommend this approach due to the computational intensity of the PX-MCEM algorithm and the bootstrap method. Users should enter the best set of starting values available from clinician opinions, other studies for the same tests, or preliminary model estimates from other statistical methods. When users do not have any reliable information on starting values, we propose fitting the computationally efficient TLC model first and using the TLC model estimates as starting values for the PLC model. Our simulation results have proven that the TLC model's estimates are good starting values for the PLC model to generate estimates closer to the true values.

The distinctive advantage of **DiagLCA** over other available R packages is its ability to handle missing data. Although the assumption of the missing data mechanism cannot be checked/tested without additional information on the missing data, sensitivity analyses have shown that the PLC model is robust to the underlying missing data assumptions as long as the starting values are close to the true values and the missing percentages are not too high.

4.3 The R Package DiagLCA

4.3.1 Function `indTLC`

`indTLC`, one of the two main functions, applies the TLC model under the conditional independence assumption. The synopsis for `indTLC` is:

```
indTLC(data, nTest=11, iniSize=10, prev=prev0, sens=sens0,
spec=spec0, thresh=0.001, stable=5, print=FALSE)
```

The arguments of `indTLC` are described as follows:

- `data`: The original data set name. The data is expected to be in a specific format. Each record corresponds to one subject. The first column is for subject ID. Columns 2 to $J+1$ correspond to T; columns $J+2$ to $2J+1$ correspond to Δ.
- `nTest`: Number of tests, i.e. J.
- `iniSize`: The number of different sets of initial values. In order to ensure a global maxima is reached, we recommend no less than 5.
- `prev`: A vector of length `iniSize` for initial values of prevalence.
- `sens`: A matrix with `iniSize` rows and J columns for initial values of sensitivity.
- `spec`: A matrix with `iniSize` rows and J columns for initial values of specificity.
- `thresh`: The threshold to decide when EM iterations converge. The sum of absolute changes for all parameter estimates should be less than this threshold for a prespecified number of consecutive iterations (`stable`). Default value is 0.001.
- `stable`: The number of consecutive iterations that a threshold (`thresh`) has been reached. Default value is 5.
- `print`: If TRUE, `depPLC` will print in the output window the outputs resulted. Default value is TRUE.

4.3.2 Function `depPLC`

The other main function is `depPLC`, which fits the PLC model under the conditional dependence assumption. It is used as:

```
depPLC(data, count=TRUE, nTest=11, iniSize=1, prev=prev0, a1=a1.0,
a0=a0.0, R1=R1.0, R0=R0.0, ndraw=1500, burniter=10, burndraw=1000,
thin=2, nboot=200, bootseed=1:200, thresh=0.001, stable=5, print=FALSE)
```

Many arguments for `depPLC` are the same as `indTLC`. The diffrent/new ones are:

- `count`: Default value is TRUE. It converts the original data input into a data frame with each record corresponding to one data pattern. Columns 1 to J correspond to T, column $J+1$ represents number of subjects sharing the same pattern. If the input data is already in this format, the value should be set to FALSE.

- `iniSize`: Here we recommend using the TLC model estimates as starting values, i.e. `iniSize` is 1.

- `a1`: A matrix with `iniSize` rows and J columns for initial values of a_1.

- `a0`: A matrix with `iniSize` rows and J columns for initial values of a_0.

- `R1`: An array of `iniSize` matrices (each with $J \times J$ dimension) for initial values of R_1. Defaul value is a diagonal matrix with 1 for all diagonal elements and 0.5 for all off-diagonal elements.

- `R0`: An array of `iniSize` matrices (each with $J \times J$ dimension) for initial values of R_0. Defaul value is a diagonal matrix with 1 for all diagonal elements and 0.5 for all off-diagonal elements.

- `ndraw`: Number of draws for each Monte Carlo sample. Default value is 1500.

- `burniter`: Number of early MC samples with burn-in draws. Default value is 10.

- `burndraw`: Number of burn-in draws to be discarded during initial portion of a MC sample. Default value is 1000.

- `thin`: Number of thinning. Default vaule is 2, i.e. keep every 2^{nd} simulated draw from each sequence.

- `nboot`: The number of bootstrap samples for standard error calculation. Default value is 200.

- `bootseed`: A vector of length `nboot` for random seeds of bootstrap samples. Defalut values are the sequence number of each bootstrap sample $(1, 2, \cdots, 200)$.

The rest of the functions provide diagnostic plots for the PX-MCEM algorithm and bootstrap method, produce visual checks for model assumptions, and create simulated data sets for simulation studies.

4.3.3 Function `modelcheck`

`modelcheck` generates correlation residual plots proposed by Qu et al. (1996), or plots observed vs. model based Kappa statistics proposed by Chu et al. (2009).

```
modelcheck(x, type = "l", xlab="Model Based Kappa",
ylab="Observed Kappa", method="kappa", ...)
```

Most of the arguments here are the same to graphic function `plot`, e.g., `type` denotes whether to plot symbols, lines, or both; `xlab` and `ylab` define labels of axises. (The x-axis label is typically blank.) The special arguments are:

- `x`: A depPLC object containing the input values needed: all pairwise differences between observed and model based correlation coefficients for method "kappa"; a vector of all pairwise differences between observed and model based Kappa statistics for method "kappa".

- `method`: Model checking method. Possible values are "kappa" and "corr".

4.3.4 Function `traceplot`

`traceplot` plots MC iterations vs. sampled values for specified PX-MCEM iteration(s), with a separate plot for each parameter.

```
traceplot(x, iter=1:10, type = "l", xlab = "Iterations",
ylab = "", ...)
```

Refer to `plot` for other arguments except for `x` and `iter`:

- `x`: A depPLC object containing the input values needed: cumulative MC samples during each PX-MCEM iteration.
- `iter`: Specify the PX-MCEM iteration(s) where MC samples are plotted. Possible values are 1, $1:10$, etc.

4.3.5 Function `convergplot`

`convergplot` plots PX-MCEM algorithm iterations vs. the PLC model parameter estimates (prevalence, diagnostic accuracy, and correlation coefficients).

```
convergplot(x, type = "l", xlab = "Iterations", ylab = "", ...)
```

Refer to `plot` for other arguments except for `x`:

- `x`: A depPLC object containing the input values needed: cumulative parameter estimates over all PX-MCEM iterations until convergence is reached.

4.3.6 Function `histgram`

`histgram` generates histograms of all bootstrap samples (used for standard error estimation) for each parameter.

```
histgram(x, xlab = colnames(x), breaks=20, ...)
```

The arguments are the same as the standard R function `hist` except for `x`:

- `x`: A depPLC object containing the input values needed: parameter estimates for all bootstrap samples.

4.3.7 Function `qqplot`

`qqplot` generates QQ plots of all bootstrap samples for each parameter.

```
qqplot(x, main =colnames(x), ...)
```

The arguments are the same as the standard R function `qqnorm` except for `x`:

- `x`: A depPLC object containing the input values needed: parameter estimates for all bootstrap samples.

4.3.8 Function `simudata`

`simudata` produces simulated data sets under different missing data mechanisms. Each data record corresponds to one subject. The first column is for subject ID; columns 2 to $J+1$ correspond to T; columns $J+2$ to $2J+1$ correspond to Δ, the last column is for true disease status. `simudata` can be integrated with `indTLC` and `depPLC` to carry out simulation studies.

```
simudata(seed=1:200, nTest=5, nSubj=3500, nData=200,
prev=prev.0, a1=a1.0, a0=a0.0, R1=R1.0, R0=R0.0,
CDA=TRUE, miss=rep(1,5), missprob=missprob)
```

The arguments are defined as:

- `seed`: Random seed.
- `nTest`: Number of tests, i.e. J.
- `nSubj`: Number of subjects, i.e. N.
- `nData`: Number of data sets simulated for each simulation study.
- `prev`: A vector of length `nStudy` for initial values of prevalence.
- `a1`: A matrix with `nStudy` rows and J columns for initial values of a_1.
- `a0`: A matrix with `nStudy` rows and J columns for initial values of a_0.
- `R1`: An array of `nStudy` matrices (each with $J \times J$ dimension) for initial values of R_1.
- `R0`: An array of `nStudy` matrices (each with $J \times J$ dimension) for initial values of R_0.
- `CDA`: Specify the conditional dependence assumption among tests (value TRUE) or conditional independence assumption (value FALSE). Default value is TRUE.
- `miss`: Specify missing data algorithms. Possible values are: 0 for no missing values, 1 for MCAR; 2 for MAR; 3 for MNAR.
- `missprob`: A vector for missing probabilities. Only to be used for MCAR tests.
- `missmodel`: A vector for regression coefficients of a logistic regression model for missing data probability. Only to be used for MAR or MNAR tests.

4.4 Implementation

4.4.1 Example Data

Colorectal cancer is the third most common cancer in the world. It is more common in developed countries, where around 60% of cases are diagnosed. Hereditary nonpolyposis colorectal cancer (HNPCC), also known as Lynch Syndrome, is an autosomal dominant genetic condition caused by mutations that impair DNA mismatch repair. HNPCC patients have an 80% lifetime risk of developing colon cancer compared to 5% for the general population. Thus, early detection of the disease is highly important for timely treatment. Microsatellite instability (MSI) biomarker tests have been used to diagnose HNPCC (Boland et al., 1998; Umar et al., 2004; Lynch and de la Chapelle, 1999). Our data set comes from NCI Colon Cancer Family Registry (C-CFR). To this end, we have $J = 11$ MSI tests measured on $N = 3,487$ individuals from single-subject families. The data set is included in the **DiagLCA** package named `msi`. After installation of **DiagLCA**, load the data set:

```
R> library(DiagLCA)
R> data(msi)
```

4.4.2 Initial Exploration

We first examine the data structure:

```
R> head(msi)
           SUBJID T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 M1 M2 M3 M4 M5 M6
[1,] 110030000020  0  0  0 99  0  0 99  0  0   0   0  1  1  1  0  1  1
[2,] 110030000244  0  1  1 99 99  0 99  0 99   1   0  1  1  1  0  0  1
[3,] 110030000392 99  0  0 99 99 99  0 99  0   0  99  0  1  1  0  0  0
```

```
[4,] 110030000491 99  0  0  0  0  0  0  0  0  99   0  0  1  1  1  1  1
[5,] 110030000624  0 99  0  0  0  0  0  0  0   0   0  1  0  1  1  1  1
[6,] 110030000665  0  0 99 99 99 99  1 99 99   0  99  1  1  0  0  0  0
     M7 M8 M9 M10 M11
[1,]  0  1  1   1   1
[2,]  0  1  0   1   1
[3,]  1  0  1   1   0
[4,]  1  1  1   0   1
[5,]  1  1  1   1   1
[6,]  1  0  0   1   0
```

The first column is for subject ID, followed by 11 columns of test results (T) and 11 columns of non-missing indicators (Δ). Tests T_1 to T_{11} correspond to *BAT25*, *BAT26*, *BAT40*, *BAT34C4*, *D10S197*, *D17S250*, *D18S55*, *D2S123*, *D5S346*, *ACTC* and *MYCL*, respectively. Notice that the column names are only used in the default labels on figures. Users may freely choose column names as long as the columns follow this standard structure. All missing values are coded to 99, corresponding to 0 for non-missing indicators.

```
R> J <- 11
R> miss <- round((1 - colMeans(msi[, 1+J+1:J])), 4)
R> miss
    M1     M2     M3     M4     M5     M6     M7     M8     M9
0.1394 0.0771 0.0726 0.1408 0.1397 0.1792 0.4193 0.1342 0.8173
   M10    M11
0.0849 0.1663

R> A=msi[,1+J+1:J]
```

```
R> a=rowSums(A)
R> table(a)
a
   1    2    3    4    5    6    7    8    9   10   11
  17   19   16   52   88  203  216  584 1202  871  219
```

The missing probabilities range from 7.26% for T_3 to 81.73% for T_9. Only 219 subjects have all 11 test results. And only 192 subjects have 6 or more tests missing.

4.4.3 Fitting a TLC Model

First, a TLC model is fit using `indTLC`. We tried different sets of starting values and ended up with very similar results. For showcasing purposes, we use one set of arbitrary starting values by specifying `iniSize=1`.

```
R> prev0 <- 0.1
R> sens0 <- rep(0.8, J)
R> spec0 <- rep(0.9, J)
R> results1 <- indTLC(data=msi, nTest=11, iniSize=1, prev=prev0,
sens=sens0, spec=spec0, thresh=0.001, stable=5, print=FALSE)

R> results1$Estimates
 [1] 0.1482364 0.7110802 0.9405087 0.9308248 0.9301345 0.8499049
 [7] 0.8352653 0.8120392 0.8097522 0.8492710 0.6454502 0.7586832
[13] 0.9699805 0.9961616 0.9985180 0.9866882 0.9979755 0.9815781
[19] 0.9564028 0.9841752 0.9769292 0.9922828 0.9360068

R> results1$StdErrs
```

```
 [1] 0.0060 0.0226 0.0111 0.0121 0.0123 0.0178 0.0184 0.0234 0.0187
[10] 0.0423 0.0225 0.0207 0.0034 0.0012 0.0008 0.0023 0.0009 0.0027
[19] 0.0049 0.0025 0.0064 0.0017 0.0049

R> results1$UpperLimits
 [1] 0.160 0.753 0.959 0.951 0.951 0.882 0.868 0.854 0.844 0.915
[11] 0.688 0.797 0.976 0.998 0.999 0.991 0.999 0.986 0.965 0.988
[21] 0.987 0.995 0.945

R> results1$LowerLimits
 [1] 0.137 0.665 0.915 0.903 0.902 0.812 0.796 0.762 0.770 0.747
[11] 0.600 0.716 0.963 0.993 0.996 0.981 0.995 0.975 0.946 0.978
[21] 0.960 0.988 0.926
```

`indTLC` is highly efficient. For the C-CFR data, the EM algorithm converges in nine iterations (from `results1$iter`) taking only a few seconds. All results are contained in a list object named `results1`. We specify `print=FALSE` in the interest of conserving space. Only a selected set of results is presented: point estimates, standard errors, and upper/lower limits of 95% confidence intervals. (The order of paramaters is prevalence, sensitivities for T_1 to T_{11}, and specificities for T_1 to T_{11}.) Users may also call `results1$Prevalence`, `results1$Sensitivity`, `results1$Specificity` to get the exact estimates. Other results, such as complete data information matrix (`results1$ComMatrix`), observed data information matrix (`results1$ObsMatrix`), model based Kappa statistics (`results1$ModKappa`), and observed Kappa statistics (`results1$ObsKappa`) also are accessible. According to the TLC model's estimates, the two best tests are T_2 with the highest sensitivity (0.9405) and third highest specificity, and T_3 with the highest specificity (0.9985) and second highest sensitivity.

The TLC model is based on the conditional independence assumption. We can run the following code to check the validity of this assumption:

```
R> modelcheck(results1$Kappa, method="kappa", xlab="Model Based Kappa",
ylab="Observed Kappa", xlim=c(0.37, 0.94), ylim=c(0.37, 0.94), err="x",
slty=1, cex=0.5, pch=19, sfrac=0, abline=c(0, 1), ablcol="black")
```

`results1$ModKappa` is a matrix object including observed Kappa statistics and model based Kappa statistics with upper and lower limits. All of the 95% simultaneous confidence intervals of model-based Kappa contain the observed Kappa statistics. We fail to reject the null hypothesis of conditional independence assumption. However, there is no conclusive test for examination of the conditional independence assumption. Users are always encouraged to consult healthcare practitioners for clinical perspective on the assumption. For the C-CFR data, it is reasonable to suspect the tests are correlated due to their similar biological basis. Thus we proceed with the PLC model fitting next.

4.4.4 Fitting a PLC Model

The diagnostic accuracy estimates from `indTLC` provide good starting values for `a1` and `a0`. As for `R1` and `R0`, there is no feasible way of deriving good starting values so a diagonal matrix with 1 for all diagonal elements and 0.5 for all off-diagonal elements is used.

```
R> a1.0 <- as.matrix(round(qnorm(results1$Sensitivity),4))
R> a0.0 <- as.matrix(round(-qnorm(results1$Specificity),4))
R> R1.0 = matrix(0.5, J, J, byrow=TRUE) + diag(0.5, J)
R> R0.0 = matrix(0.5, J, J, byrow=TRUE) + diag(0.5, J)
```

```
R> results2 <- depPLC(data=msi, count=FALSE, nTest=11, iniSize=1,
prev=results1$Prevalence, a1=a1.0, a0=a0.0, R1=R1.0, R0=R0.0,
ndraw=1500, burniter=10, burndraw=1000, thin=2, nboot=1000,
bootseed=1:200, thresh=0.001, stable=5, print=FALSE)
```

By specifying `count=FALSE`, the data input is reconstructed to group subjects by response profiles, so that subjects with all the same test results are processed together to ease the computation burden of the PX-MCEM algorithm. The reconstructed data set is stored in the list object as `results2$countdata` along with other results.

```
R> msi2 <- results2$countdata
R> head(msi2)
  T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11   n
1  0  0  0  0  0  0  0  0  0   0   0 166
2  0  0  0  0  0  0  0  0  0   0   1  11
3  0  0  0  0  0  0  0  0  0   0  99  28
4  0  0  0  0  0  0  0  0  0   1   0   1
5  0  0  0  0  0  0  0  0  0   1   1   1
6  0  0  0  0  0  0  0  0  0  99   0   6


R> tail(msi2)
    T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 n
882 99 99 99 99 99 99  0  0 99  99  99 1
883 99 99 99 99 99 99  0 99 99  99  99 1
884 99 99 99 99 99 99  1  1 99   1   0 1
885 99 99 99 99 99 99  1 99 99  99  99 1
886 99 99 99 99 99 99 99 99 99   0  99 1
887 99 99 99 99 99 99 99 99 99  99   0 3
```

There is a total of 887 response profiles for the C-CFR data. The total number of possible profiles is $3^{11} - 1 = 177146$ with three possible test results (1, 0, 99). Thus, many profiles have very few subjects, and many more profiles are not observed. We assume ignorable missing data assumptions for the missing profiles. The final results are:

```
R> results2$iter
[1] 100

R> results2$Prevalence
[1] 0.1485005

R> results2$Sensitivity
           [,1]
 [1,] 0.6982015
 [2,] 0.9393666
 [3,] 0.9314466
 [4,] 0.9267739
 [5,] 0.8433803
 [6,] 0.8153116
 [7,] 0.7881319
 [8,] 0.7967001
 [9,] 0.8444460
[10,] 0.6294338
[11,] 0.7419136

R> results2$Specificity
```

```
           [,1]
 [1,] 0.9679827
 [2,] 0.9958260
 [3,] 0.9986460
 [4,] 0.9868597
 [5,] 0.9977364
 [6,] 0.9787428
 [7,] 0.9543197
 [8,] 0.9822086
 [9,] 0.9766155
[10,] 0.9899994
[11,] 0.9348244
```

```
R> results2$R1
           [,1]      [,2]      [,3]       [,4]      [,5]       [,6]
 [1,] 1.0000000 0.3210300 0.2516105 0.11474140 0.3642791 0.60076510
 [2,] 0.3210300 1.0000000 0.4445768 0.33700505 0.4629772 0.32020180
 [3,] 0.2516105 0.4445768 1.0000000 0.33464877 0.4140617 0.28471884
 [4,] 0.1147414 0.3370051 0.3346488 1.00000000 0.4651138 0.09528937
 [5,] 0.3642791 0.4629772 0.4140617 0.46511382 1.0000000 0.35668883
 [6,] 0.6007651 0.3202018 0.2847188 0.09528937 0.3566888 1.00000000
 [7,] 0.6048265 0.4777211 0.4692013 0.36592877 0.5990494 0.58777314
 [8,] 0.5551875 0.2486382 0.1848360 0.22614256 0.4852895 0.52871200
 [9,] 0.4213350 0.3469771 0.4110545 0.28561179 0.3758858 0.44980459
[10,] 0.4657090 0.2448462 0.2903905 0.10111881 0.2574552 0.50680537
[11,] 0.4588924 0.2614890 0.4034129 0.29552066 0.4999776 0.60891431
```

```
           [,7]      [,8]      [,9]     [,10]     [,11]
 [1,] 0.6048265 0.5551875 0.4213350 0.4657090 0.4588924
 [2,] 0.4777211 0.2486382 0.3469771 0.2448462 0.2614890
 [3,] 0.4692013 0.1848360 0.4110545 0.2903905 0.4034129
 [4,] 0.3659288 0.2261426 0.2856118 0.1011188 0.2955207
 [5,] 0.5990494 0.4852895 0.3758858 0.2574552 0.4999776
 [6,] 0.5877731 0.5287120 0.4498046 0.5068054 0.6089143
 [7,] 1.0000000 0.5734948 0.5443290 0.4654053 0.6166252
 [8,] 0.5734948 1.0000000 0.3714795 0.3231993 0.4738745
 [9,] 0.5443290 0.3714795 1.0000000 0.3757601 0.4171032
[10,] 0.4654053 0.3231993 0.3757601 1.0000000 0.5228118
[11,] 0.6166252 0.4738745 0.4171032 0.5228118 1.0000000

R> results2$R0
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
 [1,] 1.0000000 0.3885253 0.3885319 0.3880731 0.3922833 0.4179660
 [2,] 0.3885253 1.0000000 0.4548361 0.4195803 0.4625262 0.4471870
 [3,] 0.3885319 0.4548361 1.0000000 0.4500090 0.4604344 0.4297889
 [4,] 0.3880731 0.4195803 0.4500090 1.0000000 0.4325155 0.4052659
 [5,] 0.3922833 0.4625262 0.4604344 0.4325155 1.0000000 0.4446123
 [6,] 0.4179660 0.4471870 0.4297889 0.4052659 0.4446123 1.0000000
 [7,] 0.3868106 0.4147975 0.4439540 0.3794102 0.4190582 0.4628957
 [8,] 0.4138077 0.4158910 0.4603657 0.4488452 0.4376806 0.4689404
 [9,] 0.4274973 0.4536899 0.4513213 0.4322316 0.4669834 0.4257487
[10,] 0.4235997 0.4559045 0.4804309 0.4432340 0.4420838 0.4818607
[11,] 0.2730062 0.3593206 0.3832185 0.3429687 0.3665439 0.3122027
```

```
           [,7]      [,8]      [,9]     [,10]     [,11]
 [1,] 0.3868106 0.4138077 0.4274973 0.4235997 0.2730062
 [2,] 0.4147975 0.4158910 0.4536899 0.4559045 0.3593206
 [3,] 0.4439540 0.4603657 0.4513213 0.4804309 0.3832185
 [4,] 0.3794102 0.4488452 0.4322316 0.4432340 0.3429687
 [5,] 0.4190582 0.4376806 0.4669834 0.4420838 0.3665439
 [6,] 0.4628957 0.4689404 0.4257487 0.4818607 0.3122027
 [7,] 1.0000000 0.4440690 0.4786316 0.4862578 0.3583231
 [8,] 0.4440690 1.0000000 0.4561418 0.4546393 0.3854386
 [9,] 0.4786316 0.4561418 1.0000000 0.4643155 0.4075752
[10,] 0.4862578 0.4546393 0.4643155 1.0000000 0.3566871
[11,] 0.3583231 0.3854386 0.4075752 0.3566871 1.0000000
```

The PX-MCEM algorithm converges after 100 iterations. The PLC model estimates are similar to the TLC model estimates but closer to the true parameter values as simulation studies have shown. The results confirm T_2 and T_3 as the two best tests with superior diagnostic accuracy. The estimates for R_1 and R_0 are less accurate due to the lack of good starting values. Hence they must be viewed with caution.

We assess the convergence of Markov chains with the following example for MC samples of T_1, which is denoted by $\sum_{k=1}^{K} n_k d_k^{(m)} z_{k1}^{(m)}$. `results2$cumZ1` is an array containing all MC samples. The 1^{st} dimension of the array is set to 1 for T_1, the 2^{nd} dimension is left blank as all MC samples during the PX-MCEM iteration are utilized, and the 3^{rd} dimension is set to 100, meaning we only include MC samples during the final PX-MCEM iteration.

```
R> traceplot(results2$cumZ1[1,,100], iter=100, type = "l",
xlab = "Iteration", ylab = "")
```

Figure 4.2 shows the chain is mixing very well. We also plot PX-MCEM algorithm iterations vs. all sensitivity estimates with the statement below. `results2$cumSens` is a matrix with rows for PX-MCEM algorithm iterations and columns for sensitivities at each PX-MCEM algorithm iteration.

```
R> convergplot(results2$cumSens, type = "l", xlab = "Iterations",
ylab = "", ylim=c(0.62, 0.94), main="Sensitivity", xlab="Iteration",
col=1:11, abline=results2$Sensitivity, ablcol=1:11)
```

In Figure 4.3, all sensitivity estimates stabilize after about 50 PX-MCEM iterations with tiny fluctuation around final estimates before convergence. Finally, we assess standard error estimation by checking the QQ plot of all bootstrap estimates for the prevalence. `results2$bootMLE` is a matrix with columns corresponding to model parameters (first column is for prevalence) and rows corresponding to bootstrap samples.

```
R> qqplot(results2$bootMLE[,1], main ="Prevalence")
```

Figure 4.4 indicates approximate normality of bootstrap samples for prevalence.

4.4.5 Simulate Data Sets for Simulation Studies

We demonstrate how to simulate data sets for three simulation studies under the conditional dependence assumption. First, we simulate 200 data sets under the MCAR mechanism. Each data set contains $J = 5$ tests and $N = 3,500$ subjects. `miss=rep(1,5)` defines the missing data mechanism for all 5 tests as MCAR.

```
R> prev.0 <- 0.25
R> se.0 <- c(0.7, 0.9, 0.8, 0.6, 0.75)
R> sp.0 <- c(0.9, 0.85, 0.9, 0.9, 0.8)
R> R1.0 <- matrix(0.6, J, J, byrow=TRUE) + diag(0.4, J)
```

```
R> R0.0 <- matrix(0.45, J, J, byrow=TRUE) + diag(0.55, J)
R> missprob <- c(0.05, 0.1, 0.2, 0.5, 0.9)

R> simMCAR <- simudata(seed=1:200, nTest=5, nSubj=3500, nData=200,
prev=prev.0, se=se.0, sp=sp.0, R1=R1.0, R0=R0.0, CDA=TRUE,
miss=rep(1,5), missprob=missprob)
```

`simMCAR` is an array object. The first 6 records of the first data set (`simMCAR[,,1]`) is showcased below. Notice that the simulated data sets contain one extra column D for true disease status.

```
R > head(simMCAR[,,1])
        SUBJID T1 T2 T3 T4 T5 M1 M2 M3 M4 M5 D
[1,]  001-0001  1  1  1  1 99  1  1  1  1  0 1
[2,]  001-0002  0 99 99 99 99  1  0  0  0  0 0
[3,]  001-0003  1  1  1 99 99  1  1  1  0  0 1
[4,]  001-0004  0  0  0 99  0  1  1  1  0  1 0
[5,]  001-0005  0  0  1 99 99  1  1  1  0  0 0
[6,]  001-0006  0  0  0  0 99  1  1  1  1  0 0
```

For the second simulation study, we simulate 200 data sets under the MAR mechanism. The true parameter values are the same as above. `miss=c(0,0,1,2,2)` denotes that T_1 and T_2 have no missing values; T_3 is MCAR; T_4 and T_5 are MAR. Missing probabilities of T_4 depend on T_1 whereas missing probabilities of T_5 depend on T_1 and T_2. `missprob=2` defines the missing probability for T_3. `parameter4` and `parameter4` define the logistic regression model parameters for missing probabilities of T_4 and T_5.

```
R> parameter4 <- c(1.5, -2.5, -2, -1.5, 0, -0.7)
R> parameter5 <- c(2, -2, 0, -3, 0, 0)
```

```
R> simMAR <- simudata(seed=1:200, nTest=5, nSubj=3500, nData=200,
prev=prev.0, a1=a1.0, a0=a0.0, R1=R1.0, R0=R0.0, CDA=TRUE,
miss=c(0,0,1,2,2), missprob=0.2, missmodel=c(parameter4,parameter5))
```

Here we list the last 6 records of the 200^{th} data set (`simMAR[,,200]`).

```
R> tail(simMAR[,,200])
          SUBJID T1 T2 T3 T4 T5 M1 M2 M3 M4 M5 D
[3495,] 200-3495  1  1  1 99 99  1  1  1  0  0 1
[3496,] 200-3496  0  0  0  0  0  1  1  1  1  1 0
[3497,] 200-3497  0  0  0  0  0  1  1  1  1  1 0
[3498,] 200-3498  1  1 99 99 99  1  1  0  0  0 1
[3499,] 200-3499  0  0  0  0  0  1  1  1  1  1 0
[3500,] 200-3500  1  1  1 99 99  1  1  1  0  0 1
```

Finally we simulate 200 data sets under the MNAR mechanism. Same as the 2^{nd} simulation study, T_1 and T_2 have no missing values and T_3 is MCAR. The logistic regression model parameters `parameter4` and `parameter4` are also unchanged. The only difference is that `miss` becomes `c(0,0,1,3,3)`, which means T_4 and T_5 are MNAR: missing probabilities of T_4 depend on T_1 and T_3; missing probabilities of T_5 depend on T_1, T_2, T_3, as well as the latent disease status D. Again we showcase the last six records of the 200^{th} data set.

```
R> simMNAR <- simudata(seed=1:200, nTest=5, nSubj=3500, nData=200,
prev=prev.0, a1=a1.0, a0=a0.0, R1=R1.0, R0=R0.0, CDA=TRUE,
miss=c(0,0,1,3,3), missprob=0.2, missmodel=c(parameter4,parameter5))


R> tail(simMNAR[,,200])
```

```
          SUBJID T1 T2 T3 T4 T5 M1 M2 M3 M4 M5 D
[3495,] 200-3495  0  0  0  0  0  1  1  1  1  1 0
[3496,] 200-3496  0  0  0  0  0  1  1  1  1  1 0
[3497,] 200-3497  0  0  0  0 99  1  1  1  1  0 1
[3498,] 200-3498  1  1  1 99 99  1  1  1  0  0 1
[3499,] 200-3499  0  1  1  0 99  1  1  1  1  0 0
[3500,] 200-3500  0  1  1  1 99  1  1  1  1  0 1
```

We can also simulate data sets under the conditional independence assumption by specifying `CDA=FALSE` and dropping `R1` and `R0`. For example, the statement below simulates data sets under the MCAR mechanism and the conditional independence assumption.

```
R> simMCAR <- simudata(seed=1:200, nTest=5, nSubj=3500, nData=200,
prev=prev.0, se=se.0, sp=sp.0, miss=rep(1,5), missprob=missprob)
```

4.5 Summary

Evaluation of multiple diagnostic tests without a gold standard yet with abundant missing data is important in diagnostic medicine but no software handling missing tests has been known to authors. Our **DiagLCA** is the first R package to this end. We first introduced the two useful latent class models, the traditional latent class (TLC) model for conditionally independent tests and the probit latent class (PLC) model for conditionally dependent tests. The utility of **DiagLCA** package is demonstrated in detail using a real-world example data set in the diagnosis of HNPCC. We fit a TLC model under the conditional independence assumption, and subsequently fit a PLC model under the conditional dependence assumption using the TLC model estimates as starting values for the PX-MCEM algorithm. We further examined how to simulate

data sets under different missing data mechanisms.

The current version has its limitations. We are committed to continuous improvement of the package to make it more accessible to medical researchers and applied statisticians. For example, the runtime for `depPLC` is quite extensive with a large number of tests due to the time-consuming nature of the PX-MCEM algorithm. For future work, we aim to reduce runtime by improving computing efficiency.

Figure 4.1: Observed vs. Model Based Kappa for Model Checking

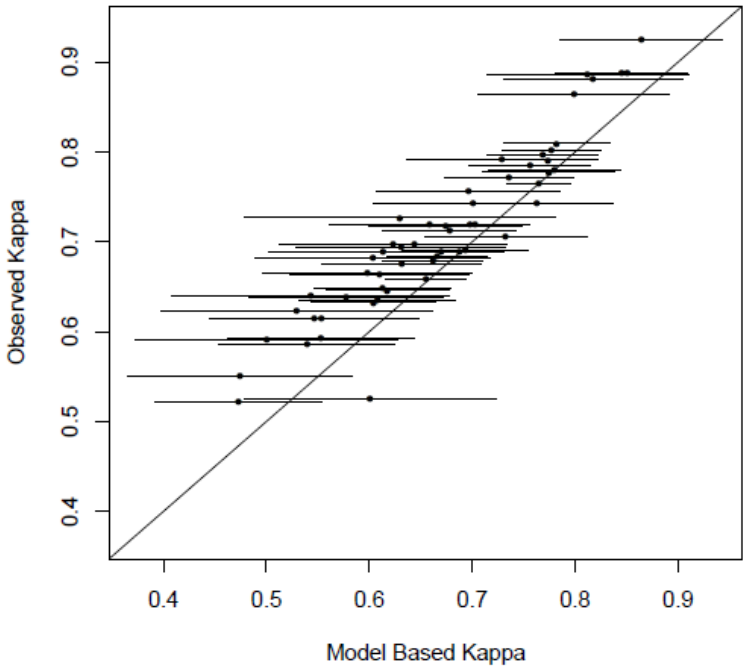


Figure 4.2: Trace Plot of $\sum_{k=1}^{K} n_k d_k^{(m)} z_{k1}^{(m)}$ over the Last PX-MCEM Algorithm Iteration

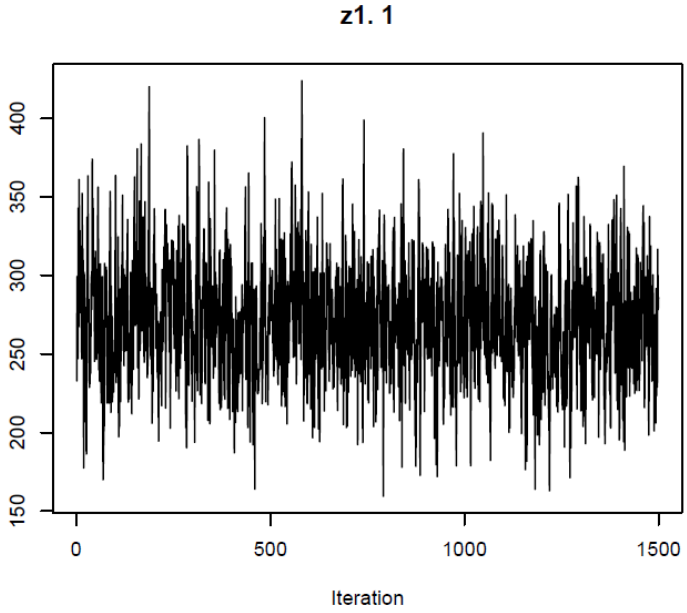


Figure 4.3: Convergence Plot of Sensitivity Estimates

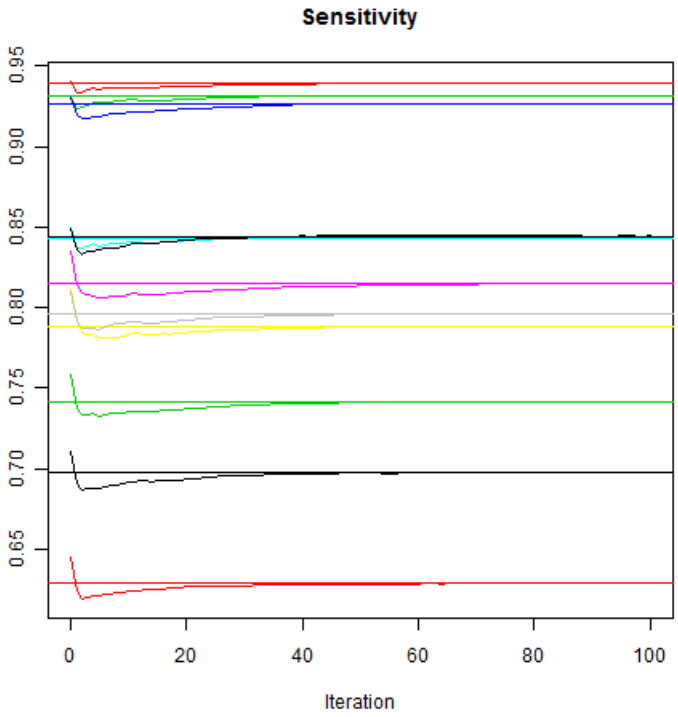
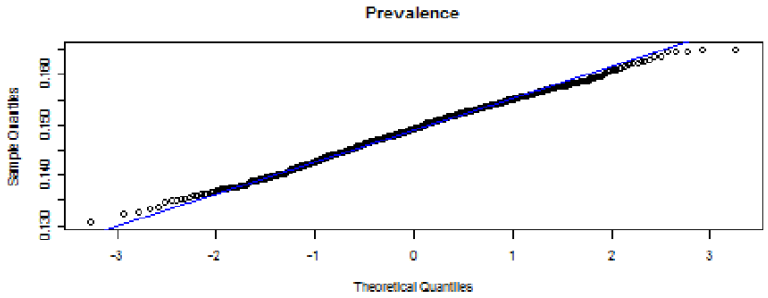


Figure 4.4: QQ Plot of Prevalence Estimates from 1000 Bootstrap Samples



Chapter 5

Future Research

Clustered data are common in diagnostic medicine when the clusters that are tested for the targeted condition are positively correlated, such as the data from multiple lesions of the same patient, from multiple teeth of the same mouth, and from multiple subjects in the same family. For the C-CFR data, there are a total of $6,131$ subjects from $4,494$ families. Table 5.1 summarizes the distribution of the number of subjects per family. For our research, we considered only the $3,487$ subjects from families with a single subject per family. However, 22.4% families have two or more subjects, and 1.3% families have five or more subjects. Here, the families constitute the cluster, and the subjects constitute the diagnostic unit of study within a cluster. In another word, we have $N = 6,131$ subjects distributed in $C = 4,494$ clusters. The clusted data created a unique statistical challenge for future research.

In the analysis of such clustered data, if observations from the same cluster are assumed independent, point estimates can still be derived without adjusting for correlation within clusters. However, the standard errors likely are biased and lead to incorrect test statistics and confidence intervals. For our study, if correlation within clusters is ignored, the standard error for the diagnostic accuracy is likely to be overestimated when subjects from the same family are in different disease status, while it

is likely to be underestimated when subjects have the same disease status. Thus to avoid biased statistical inference, we must take the correlation within clusters into consideration. To the best of our knowledge, there is no particular method that addresses clustering data in estimating the diagnostic accuracy of medical tests. We propose to extend the methods in Chapter 3 by adding another level of random effect for each cluster in our diagnostic models. In other words, there are two levels of random effects, i.e., one is an individual-level random effect accounting for within-subject correlation among multiple tests, which is denoted by $Z_i = (Z_1, \cdots, Z_J)$ for the i^{th} subject as in Chapter 3, and the other is a cluster-level random effect accounting for within-cluster correlation among multiple subjects, which can be denoted as W_i for the i^{th} subject and $W_i|D_i = d \sim N(\mu_d, \Sigma_d)$.

Other methods, such as the Bayesian approach, may also be explored. We can specify the joint prior distribution of Θ as a multivariate normal distribution with a certain mean vector and variance-covariance matrix based on expert knowledge or previous study results. The joint posterior distribution can be introduced as well as latent variables, and then we can use Gibbs sampling to draw samples from the posterior distribution. Based on Bayes' rule, we may further derive the predictive probability of colon cancer for an individual subject. The predictive 95% credible intervals for the predictive probability of having cancer given the test results can be computed via MCMC methods.

Table 5.1: The Distribution of the Number of Subjects per Family

Number of Subjects	Frequency	Percentage (%)
1	3487	77.59
2	668	14.86
3	208	4.63
4	73	1.62
≥ 5	58	1.29

Appendix

Derivation of Information Matrices for Louis Formula

For simplicity, we will first derive the components of information matrices for each individual subject, and then sum them up. For subject i, the complete-data log-likelihood is $logL_c(\theta) = logL(\theta|Y_i) = logL(\theta|T, D_i) = d_i log(\pi_1 h_{i1}) + (1 - d_i) log(\pi_0 h_{i0})$, where $h_{i1} = \prod_{j=1}^{J} Se_j^{t_{ij}\delta_{ij}}(1 - Se_j)^{(1-t_{ij})\delta_{ij}}$; $h_{i0} = \prod_{j=1}^{J}(1 - Sp_j)^{t_{ij}\delta_{ij}} Sp_j^{(1-t_{ij})\delta_{ij}}$.

Thus $S_c(Y_i; \theta)$ is

$$\frac{\partial logL_c(\theta)}{\partial \pi_1} = \frac{d_i}{\pi_1} - \frac{1 - d_i}{1 - \pi_1}$$

$$\frac{\partial logL_c(\theta)}{\partial Se_j} = \frac{d_i t_{ij}\delta_{ij}}{Se_j} - \frac{d_i\delta_{ij} - d_i t_{ij}\delta_{ij}}{1 - Se_j}$$

$$\frac{\partial logL_c(\theta)}{\partial Sp_j} = \frac{(1 - d_i)(1 - t_{ij})\delta_{ij}}{Sp_j} - \frac{(1 - d_i)t_{ij}\delta_{ij}}{1 - Sp_j}$$

and $I_c(\theta; Y_i)$ is

$$-\frac{\partial S_c(Y_i; \theta)}{\partial \pi_1} = \frac{d_i}{\pi_1^2} + \frac{1 - d_i}{(1 - \pi_1)^2}$$

$$-\frac{\partial S_c(Y_i; \theta)}{\partial Se_j} = \frac{d_i t_{ij}\delta_{ij}}{Se_j^2} + \frac{d_i\delta_{ij} - d_i t_{ij}\delta_{ij}}{(1 - Se_j)^2}$$

$$-\frac{\partial S_c(Y_i; \theta)}{\partial Sp_j} = \frac{(1 - d_i)(1 - t_{ij})\delta_{ij}}{Sp_j^2} + \frac{(1 - d_i)t_{ij}\delta_{ij}}{(1 - Sp_j)^2}$$

Notice that we took the conditional expectation of $S_c(Y_i; \theta)$ given T to derive the

point estimates (1) – (3):

$$\pi_1^{(n+1)} = \frac{\sum_{i=1}^{N} \frac{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{N} \tag{5.1}$$

$$Se_j^{(n+1)} = \frac{\sum_{i=1}^{N} \frac{t_{ij}\delta_{ij}\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{\sum_{i=1}^{N} \frac{\delta_{ij}\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}} \tag{5.2}$$

$$Sp_j^{(n+1)} = \frac{\sum_{i=1}^{N} \frac{(1-t_{ij})\delta_{ij}(1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}}{\sum_{i=1}^{N} \frac{\delta_{ij}(1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}{\pi_1^{(n)} \prod_{j=1}^{J} (Se_j^{(n)})^{t_{ij}\delta_{ij}} (1-Se_j^{(n)})^{(1-t_{ij})\delta_{ij}} + (1-\pi_1^{(n)}) \prod_{j=1}^{J} (1-Sp_j^{(n)})^{t_{ij}\delta_{ij}} (Sp_j^{(n)})^{(1-t_{ij})\delta_{ij}}}} \tag{5.3}$$

Now take the conditional expectation of $I_c(\theta; Y_i)$ given T_i and Δ_i, we have

$$I_c(\pi_1; T_i, \Delta_i) = \frac{E[d_i|T_i, \Delta_i]}{\pi_1^2} + \frac{N - E[d_i|T_i, \Delta_i]}{(1-\pi_1)^2} \tag{5.4}$$

$$I_c(Se_j; T_i, \Delta_i) = \frac{E[d_i|T_i, \Delta_i] t_{ij}\delta_{ij}}{Se_j^2} + \frac{E[d_i|T_i, \Delta_i]\delta_{ij} - E[d_i|T_i, \Delta_i] t_{ij}\delta_{ij}}{(1-Se_j)^2} \tag{5.5}$$

$$I_c(Sp_j; T_i, \Delta_i) = \frac{(1 - E[d_i|T_i, \Delta_i])(1-t_{ij})\delta_{ij}}{Sp_j^2} + \frac{(1 - E[d_i|T_i, \Delta_i]) t_{ij}\delta_{ij}}{(1-Sp_j)^2} \tag{5.6}$$

where

$$E[d_i|T_i,\Delta_i] = \frac{\pi_1 h_{i1}}{\sum_{d=0}^{1}\pi_d h_{id}}$$

$$1 - E[d_i|T_i,\Delta_i] = \frac{\pi_0 h_{i0}}{\sum_{d=0}^{1}\pi_d h_{id}}$$

Using EM algorithm we already obtained the point estimates of θ. Plug these estimates and T, Δ_i into $I_c(\theta;Y_i)$, we can get diagonal elements of $I_c(\hat{\theta};T_i,\Delta_i)$ derived above. It is obvious that all off-diagonal elements are zero since $-\frac{\partial logL_c(\theta)}{\partial\pi_1\partial Se_j} = 0$, $-\frac{\partial logL_c(\theta)}{\partial\pi_1\partial Sp_j} = 0$, and $-\frac{\partial logL_c(\theta)}{\partial Se_j\partial Sp_j} = 0$. Therefore $I_c(\hat{\theta};T_i,\Delta_i)$ is a $(2J+1)\times(2J+1)$ diagonal matrix in which the entries outside the main diagonal are all zero.

We have $I_m(\theta;T_i,\Delta_i) = \mathrm{cov}_\theta\{S_c(Y_i;\theta)|T_i,\Delta_i\}$. Thus

$$\begin{aligned} I_m(\pi_1;T_i,\Delta_i) &= \mathrm{var}_{\pi_1}\{S_c(Y_i;\pi)|T_i,\Delta_i\} \\ &= \frac{1}{\pi_1^2(1-\pi_1)^2}\mathrm{var}\{d_i|T_i,\Delta_i\} \end{aligned}$$

Since we are assuming indepent observations, $\mathrm{cov}\{d_i,d_j|T_i,\Delta_i\} = 0$ for $i \neq j$. Thus above becomes

$$I_m(\pi_1;T_i,\Delta_i) = \frac{1}{\pi_1^2(1-\pi_1)^2}\mathrm{var}\{d_i|T_i,\Delta_i\} \tag{5.7}$$

where

$$\begin{aligned} \mathrm{var}\{d_i|T_i,\Delta_i\} &= E[d_i^2|T_i,\Delta_i] - (E[d_i|T_i,\Delta_i])^2 \\ &= \frac{\prod_{d=0}^{1}\pi_d h_{id}}{(\sum_{d=0}^{1}\pi_d h_{id})^2} \end{aligned}$$

For diagonol elements, we have

$$\begin{aligned} I_m(Se_j; T_i, \Delta_i) &= \text{var}_{Se_j}\{S_c(Y_i; Se_j)|T_i, \Delta_i\} \\ &= \frac{1}{Se_j^2(1-Se_j)^2}[t_{ij}\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} \\ &\quad +Se_j^2\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} - 2Se_j t_{ij}\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\}] \end{aligned} \tag{5.8}$$

$$\begin{aligned} I_m(Sp_j; T_i, \Delta_i) &= \text{var}_{Sp_j}\{S_c(Y_i; Sp_j)|T_i, \Delta_i\} \\ &= \frac{1}{Sp_j^2(1-Sp_j)^2}[(1-t_{ij})\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} \\ &\quad +Sp_j^2\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} - 2Sp_j(1-t_{ij})\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\}] \end{aligned} \tag{5.9}$$

$$\begin{aligned} I_m(\pi_1, Se_j; T_i, \Delta_i) &= \text{cov}_{\pi_1, Se_j}\{S_c(Y_i; \pi_1, Se_j)|T_i, \Delta_i\} \\ &= \frac{1}{\pi_1(1-\pi_1)Se_j(1-Se_j)}t_{ij}\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} \\ &\quad -\frac{1}{\pi_1(1-\pi_1)(1-Se_j)}\delta_{ij}\text{var}\{d_i|T_i, \Delta_i\} \end{aligned} \tag{5.10}$$

For off-diagonal elements, we have

$$
\begin{aligned}
I_m(\pi_1, Sp_j; T_i, \Delta_i) &= \mathrm{cov}_{\pi_1, Sp_j}\{S_c(Y_i; \pi_1, Sp_j)|T_i, \Delta_i\} \\
&= -\frac{1}{\pi_1(1-\pi_1)Sp_j}(1-t_{ij})\delta_{ij}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad +\frac{1}{\pi_1(1-\pi_1)(1-Sp_j)}t_{ij}\delta_{ij}\mathrm{var}\{d_i|T_i, \Delta_i\}
\end{aligned} \tag{5.11}
$$

$$
\begin{aligned}
I_m(Se_j, Se_{j'}; T_i, \Delta_i) &= \mathrm{cov}_{Se_j, Se_{j'}}\{Sc(Y_i; Se_j, Se'_j)\} \\
&= \frac{1}{Se_j(1-Se_j)Se_{j'}(1-Se_{j'})}t_{ij}\delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad -\frac{1}{Se_j(1-Se_j)(1-Se_{j'})}t_{ij}\delta_{ij}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad -\frac{1}{(1-Se_j)Se_{j'}(1-Se_{j'})}\delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad +\frac{1}{(1-Se_j)(1-Se_{j'})}\delta_{ij}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\}
\end{aligned} \tag{5.12}
$$

$$
\begin{aligned}
I_m(Sp_j, Sp_{j'}; T_i, \Delta_i) &= \mathrm{cov}_{Sp_j, Sp_{j'}}\{Sc(Y_i; Sp_j, Sp'_j)\} \\
&= \frac{1}{Sp_jSp_{j'}}(1-t_{ij})\delta_{ij}(1-t_{ij'})\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad -\frac{1}{Sp_j(1-Sp_{j'})}(1-t_{ij})\delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad -\frac{1}{(1-Sp_j)Sp_{j'}}t_{ij}\delta_{ij}(1-t_{ij'})\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad +\frac{1}{(1-Sp_j)(1-Sp_{j'})}t_{ij}\delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\}
\end{aligned} \tag{5.13}
$$

$$
\begin{aligned}
I_m(Se_j, Sp_{j'}; T_i, \Delta_i) &= \mathrm{cov}_{Se_j, Sp_{j'}}\{Sc(Y_i; Se_j, Sp'_j)\} \\
&= -\frac{1}{Se_j(1-Se_j)Sp_{j'}} t_{ij}\delta_{ij}(1-t_{ij'})\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad + \frac{1}{Se_j(1-Se_j)(1-Sp_{j'})} t_{ij}\delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad + \frac{1}{(1-Se_j)Sp_{j'}} \delta_{ij}(1-t_{ij'})\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \\
&\quad - \frac{1}{(1-Se_j)(1-Sp_{j'})} \delta_{ij}t_{ij'}\delta_{ij'}\mathrm{var}\{d_i|T_i, \Delta_i\} \qquad (5.14)
\end{aligned}
$$

Bibliography

Albert, P. S. (2009). Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine 28*(5), 780–97.

Albert, P. S. and L. E. Dodd (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics 60*(2), 427–435.

Albert, P. S., L. M. McShane, and J. H. Shih (2004). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics 57*(2), 610–619.

Alonzo, T. A. (2005). Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. *Statistics in Medicine 24*(3), 403–417.

Alonzo, T. A., M. S. Pepe, et al. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine 18*(22), 2987–3003.

Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics 51*(1), 330–337.

Beath, K. (2011). randomlca: Random effects latent class analysis. *R package version 0.7-4, URL http://CRAN. R-project. org/package= randomLCA*.

Begg, C. B. and R. A. Greenes (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics 39*(1), 207–215.

Bertrand, P., J. Benichou, P. Grenier, and C. Chastang (2005). Hui and walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. *J Clin Epidemiol 58*(7), 688–700.

Black, M. A. and B. A. Craig (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine 21*(18), 2653–69.

Boland, C. R., S. N. Thibodeau, S. R. Hamilton, D. Sidransky, J. R. Eshleman, R. W. Burt, S. J. Meltzer, M. A. Rodriguez-Bigas, R. Fodde, G. N. Ranzani, et al. (1998). A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer research 58*(22), 5248–5257.

Brasil, P. (2010). Diagnosismed: diagnostic test accuracy evaluation for medical professionals. *R package version 0.2-3, URL http://CRAN.R-project.org/src/contrib/Archive/DiagnosisMed*.

Brenner, H. (1996). Correcting for exposure misclassification using an alloyed gold standard. *Epidemiology 7*(4), 406–410.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*(2), 347–361.

Chu, H., S. Chen, and T. A. Louis (2009). Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *J Am Stat Assoc 104*(486), 512–523.

Chu, H., S. R. Cole, Y. Wei, and J. G. Ibrahim (2009). Estimation and inference for case-control studies with multiple non-gold standard exposure assessments: with an occupational health application. *Biostatistics 10*(4), 591–602.

Dawid, A. P. and A. M. Skene (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics 28*(1), 20–28.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474.

DeFrancisco, J. and W. Grady (2003). Diagnosis and management of hereditary non-polyposis colon cancer. *Gastrointestinal Endoscopy 58*(3), 390–408.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Dendukuri, N. and L. Joseph (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics 57*(1), 158–167.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics 7*(1), 1–26.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association 89*(426), 463–475.

Enøe, C., M. P. Georgiadis, W. O. Johnson, et al. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive veterinary medicine 45*(1-2), 61–81.

Espeland, M. A. and S. L. Handelman (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics 45*(2), 587–599.

Evans, M. and T. Swartz (1995). Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical Science 10*(3), 254–272.

Feng, X., D. A. Buell, J. R. Rose, and P. J. Waddell (2003). Parallel algorithms for bayesian phylogenetic inference. *Journal of Parallel and Distributed Computing 63*(7), 707–718.

Ford, J. M. and A. S. Whittemore (2006). Predicting and preventing hereditary colorectal cancer. *JAMA 296*(12), 1521–3.

Geloven, N., K. A. Broeze, B. C. Opmeer, B. W. Mol, and A. H. Zwinderman (2012). How to deal with double partial verification when evaluating two index tests in relation to a reference test? *Statistics in Medicine 31*(28), 3787–3788.

Georgiadis, M. P., W. O. Johnson, I. A. Gardner, and R. Singh (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 52*(1), 63–76.

Goetghebeur, E., J. Liinev, M. Boelaert, and P. Van der Stuyft (2000). Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Statistical methods in medical research 9*(3), 231–248.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Hadgu, A. (1996). The discrepancy in discrepant analysis. *The Lancet 348*(9027), 592–593.

Hadgu, A. (1997). Bias in the evaluation of dna-amplification tests for detecting chlamydia trachomatis. *Statistics in Medicine 16*(12), 1391–1399.

Hadgu, A. (1999). Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *Journal of clinical epidemiology 52*(12), 1231–1237.

Hadgu, A., N. Dendukuri, and J. Hilden (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test. *Epidemiology 16*(5), 604–612.

Harel, O. and X. H. Zhou (2006). Multiple imputation for correcting verification bias. *Statistics in Medicine 25*(22), 3769–86.

Harel, O. and X. H. Zhou (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine 26*(16), 3057–77.

Harrell, F., K. L. Lee, and D. B. Mark (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*, 361–387.

He, H. and M. P. McDermott (2012). A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics 13*(1), 32–47.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society 47*(1), 153–161.

Huang, G.-H. and K. Bandeen-Roche (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika 69*(1), 5–32.

Hui, S. L. and S. D. Walter (1980). Estimating the error rates of diagnostic tests. *Biometrics 36*(1), 167–171.

Hui, S. L. and X. H. Zhou (1998). Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research 7*(4), 354–370.

Jones, G., W. O. Johnson, T. E. Hanson, and R. Christensen (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics 66*(3), 855–863.

Jones, J. and D. Hunter (1995). Consensus methods for medical and health services research. *BMJ: British Medical Journal 311*(7001), 376.

Joseph, L., T. W. Gyorkos, and L. Coupal (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology 141*(3), 263–272.

Kosinski, A. S. and H. X. Barnhart (2003a). Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics 59*(1), 163–171.

Kosinski, A. S. and H. X. Barnhart (2003b). A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Statistics in Medicine 22*(17), 2711–21.

Kou, S., C. Liu, and Y. Wu (1998). Cumulative implementation of monte carlo em. Technical report, Technical report, Department of Statistics, University of Michigan.

Ledley, R. S. and L. B. Lusted (1959). Reasoning foundations of medical diagnosis. *Science 130*(3366), 9–21.

Lin, C. Y., H. X. Barnhart, and A. S. Kosinski (2006). The weighted generalized estimating equations approach for the evaluation of medical diagnostic test at subunit level. *Biometrical journal 48*(5), 758–771.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*(421), 125–134.

Little, R. J. and D. B. Rubin (1987). *Statistical analysis with missing data*, Volume 4. Wiley New York.

Liu, C., D. B. Rubin, and Y. N. Wu (1998). Parameter expansion to accelerate em: the px-em algorithm. *Biometrika 85*(4), 755–770.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 226–233.

Lynch, H. T. and A. de la Chapelle (1999). Genetic susceptibility to non-polyposis colorectal cancer. *Journal of medical genetics 36*(11), 801–818.

Martin, D. H., M. Nsuami, J. Schachter, r. Hook, E. W., D. Ferrero, T. C. Quinn, and C. Gaydos (2004). Use of multiple nucleic acid amplification tests to define the infected-patient "gold standard" in clinical trials of new diagnostic tests for chlamydia trachomatis infections. *J Clin Microbiol 42*(10), 4749–58.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*(425), 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association 92*(437), 162–170.

McLachlan, G. and D. Peel (2000). *Finite mixture models*, Volume 299. Wiley-Interscience.

Meng, X.-L. and D. B. Rubin (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Miller, R. G. (1974). The jackknife-a review. *Biometrika 61*(1), 1–15.

Monti, G. E., K. Frankena, B. Engel, W. Buist, H. D. Tarabla, and M. C. M. de Jong (2005). Evaluation of a new antibody-based enzyme-linked immunosorbent assay for the detection of bovine leukemia virus infection in dairy cattle. *Journal of Veterinary Diagnostic Investigation 17*(5), 451–457.

Newcomb, P. A., J. Baron, M. Cotterchio, S. Gallinger, J. Grove, R. Haile, D. Hall, J. L. Hopper, J. Jass, L. Le Marchand, P. Limburg, N. Lindor, J. D. Potter, A. S. Templeton, S. Thibodeau, and D. Seminara (2007). Colon cancer family registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev 16*(11), 2331–43.

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Pepe, M. S. and H. Janes (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics 8*(2), 474–484.

Poleto, F. Z., J. M. Singer, and C. D. Paulino (2011). Comparing diagnostic tests with missing data. *Journal of Applied Statistics 38*(6), 1207–1222.

Pouillot, R., G. Gerbier, and I. A. Gardner (2002). tags, a program for the evaluation of test accuracy in the absence of a gold standard. *Preventive veterinary medicine 53*(1), 67–81.

Qu, Y. and A. Hadgu (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association 93*(443), 920–928.

Qu, Y., M. Tan, and M. H. Kutner (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics 52*(3), 797–810.

Reitsma, J. B., A. W. Rutjes, K. S. Khan, A. Coomarasamy, and P. M. Bossuyt (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of clinical epidemiology 62*(8), 797–806.

Rindskopf, D. and W. Rindskopf (2006). The value of latent class analysis in medical diagnosis. *Statistics in Medicine 5*(1), 21–27.

Schneeweiss, S. (2000). Sensitivity analysis of the diagnostic value of endoscopies in cross-sectional studies in the absence of a gold standard. *International journal of technology assessment in health care 16*(03), 834–841.

Shih, J. H. and P. S. Albert (2004). Latent model for correlated binary data with diagnostic error. *Biometrics 55*(4), 1232–1235.

Staquet, M., M. Rozencweig, Y. J. Lee, and F. M. Muggia (1981). Methodology for the assessment of new dichotomous diagnostic tests. *Journal of chronic diseases 34*(12), 599–610.

Stock, C., T. Hielscher, and M. C. Stock (2013). Package dtcompair: comparison of binary diagnostic tests in a paired study design. *R package version 0.9-3, URL http://CRAN.R-project.org/package=DTComPair*.

Tanner, M. A. (1991). *Tools for statistical inference: observed data and data augmentation methods*. Springer-Verlag New York.

Thibodeau, L. (1981). Evaluating diagnostic tests. *Biometrics 37*, 801–804.

Thornbury, J., D. Fryback, P. Turski, M. Javid, J. McDonald, B. Beinlich, L. Gentry, J. Sackett, E. Dasbach, and P. Martin (1993). Disk-caused nerve compression in patients with acute low-back pain: diagnosis with mr, ct myelography, and plain ct. *Radiology 186*(3), 731–738.

Torrance-Rynard, V. L. and S. D. Walter (1998). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine 16*(19), 2157–2175.

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models. *Applied Psychological Measurement 23*(4), 283–297.

Umar, A., C. R. Boland, J. P. Terdiman, S. Syngal, A. d. l. Chapelle, J. Ruschoff, R. Fishel, N. M. Lindor, L. J. Burgart, R. Hamelin, S. R. Hamilton, R. A. Hiatt, J. Jass, A. Lindblom, H. T. Lynch, P. Peltomaki, S. D. Ramsey, M. A. Rodriguez-Bigas, H. F. A. Vasen, E. T. Hawk, J. C. Barrett, A. N. Freedman, and S. Srivastava (2004). Revised bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *JNCI Journal of the National Cancer Institute 96*(4), 261–268.

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics 41*(4), 959–968.

Wacholder, S., B. Armstrong, and P. Hartge (1993). Validation studies using an alloyed gold standard. *American Journal of Epidemiology 137*(11), 1251–1258.

Walter, S. D. and L. M. Irwig (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of clinical epidemiology 41*(9), 923–937.

Wang, L. and N. Dendukuri (2012). lcmr: An r package for bayesian estimation of latent class models with random effects.

Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Xu, H. and B. A. Craig (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics 65*(4), 1145–55.

Yang, I. and M. P. Becker (1997). Latent variable modeling of diagnostic accuracy. *Biometrics 53*(3), 948–958.

Yu, Q., W. Tang, S. Marcus, Y. Ma, H. Zhang, and X. Tu (2010). Modeling sensitivity and specificity with a time-varying reference standard within a longitudinal setting. *Journal of Applied Statistics 37*(7), 1213–1230.

Zhou, X.-H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics-Theory and Methods, 22*(11), 3177–3198.

Zhou, X.-H. (1998). Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical methods in medical research 7*(4), 337–353.