Diagnostic research in the absence of a gold standard

Maarten van Smeden

Diagnostic research in the absence of a gold standard

PhD thesis. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands.

ISBN	978-94-6233-226-3
Author	Maarten van Smeden
Printed by	Drukkerij Gildeprint, Enschede

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (project 918.10.615). Financial support by the Julius Center for Health Sciences and Primary Care for the publication of this thesis is gratefully acknowledged.

Diagnostic research in the absence of a gold standard

Diagnostisch onderzoek in afwezigheid van een gouden standaard

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 10 maart 2016
des middags te 4.15 uur

door

Maarten van Smeden

geboren op 16 mei 1986
te Delft

Promotor:	Prof. dr. K.G.M. Moons
Copromotoren:	Dr. J.B. Reitsma
	Dr. J.A.H. de Groot

Manuscripts based on the studies presented in this thesis

Chapter 2	van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent class models in diagnostic studies when there is no reference standard-a systematic review. *American Journal of Epidemiology 2014;179(4):423-31.*
Chapter 3	van Smeden M, Oberski DL, Reitsma JB, Vermunt JK, Moons KGM, de Groot JAH. Detecting misfit of latent class models in diagnostic research without a gold standard. *Journal of Clinical Epidemiology 2015* [Epub ahead of print].
Chapter 4	van Smeden M[†], Schumacher SG[†], Dendukuri N, Joseph L, Nicol MP, Pai M, Zar HJ. Diagnostic test accuracy in childhood pulmonary tuberculosis: a Bayesian latent class analysis.
Chapter 5	van Smeden M, Dendukuri N, Joseph L, Moons KGM, Reitsma JB, de Groot JAH. Revisiting the caution on conditional dependence latent class models for estimating diagnostic error without a gold standard.
Chapter 6	Naaktgeboren CA, Bertens LCM, van Smeden M, de Groot JAH, Moons KGM, Reitsma JB. Value of composite reference standards in diagnostic research. *British Medical Journal (BMJ) 2013;347:f5605.*
Chapter 7	Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. *Statistics in Medicine 2015* [Epub ahead of print].
Chapter 8	van Smeden M, de Groot JAH, Moons KGM, Collins, GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion when considering sample size for binary logistic regression analysis.
Chapter 9	van Smeden M, de Groot JAH, Nikolakopoulos S, Bertens LCM, Moons KGM, Reitsma JB. Guidance to construct and use a nomogram for multinomial logistic regression models.

[†]authors have contributed equally

Contents

Chapter 1

General introduction

In the diagnostic work-up of a patient, physicians rely on multiple sources of diagnostic information. Such information may come from patient history, or signs and symptoms presented by the patient or from more elaborate and invasive diagnostic testing such as imaging tests, laboratory tests or biopsies. A single piece of diagnostic information is referred to as a (diagnostic) test result. For physicians, knowledge about the diagnostic accuracy of such tests, in isolation or preferably in combination with other tests [119, 120], is of vital importance to value test results when diagnosing a patient.

Ideally, the accuracy of a test or combination of tests is evaluated cross-sectionally by subjecting all study patients suspected of the condition of interest (i.e. the target condition) to the test(s) under study (i.e. the index test(s)) and a reference standard [27,104,123]. The term 'reference standard' describes the best available single method or combination of methods for verifying a patient's target condition status. In literature, one often encounters the use of the term 'gold standard' instead of 'reference standard'. Although both terms are used to describe the prevailing best available method for target condition verification, the use of the term reference standard has been advocated [28, 104, 143]. This term better reflects that the majority of existing best methods for target condition verification are prone to at least some level of target condition misclassification or ambiguous results (i.e. the best available reference standard is almost never truly 'gold'). Following this principle, in this thesis we only use the term 'gold standard' to describe methods for target condition verification that have perfect accuracy for diagnosing the target condition.

For most diagnostic studies, obtaining a classification that reliably distinguishes between study patients with and without the target condition is challenging [135, 143, 145, 166]. For instance, the best available reference standard may be too invasive or too costly to apply in all study patients, or some patients may even refuse to undergo the reference standard. In these cases, the disease status is not verified for all study subjects who underwent the index test(s). This problem is known as partial verification [16, 48, 49, 145]. In other situations, some of the studied patients may only be verified by a reference standard that is inferior to the prevailing best available reference standard [125, 146]. This is known as differential verification. Finally, there are situations where the prevailing best available reference standard is subject to classification error [143, 166].

Errors in the classification of the target condition can lead to a biased evaluation of test accuracy [29, 112, 176]. Commonly used measures of test accuracy, such as the index test's sensitivity, specificity and area under the receiver operating characteristic curve that are derived

directly from the cross-classification of index test and an imperfect reference standard are likely to be biased [32, 112, 159, 173]. Estimates of target condition prevalence and index test's predictive values are similarly affected [145, 173]. These biases can cause a series of problems for clinical practice, including: misguided treatment decisions, unrecognized under- and overtreatment and delayed acceptance of new and better tests [78].

A variety of alternative disease verification strategies have been suggested to overcome the problem of absence of a gold standard in diagnostic research. These methods have in common that they combine the outcomes of multiple tests to improve disease verification. One may consider the use of an expert (or consensus) panel diagnosis [19, 65, 179], in which a group of clinicians reach consensus based on the available test results for individual patients. Results from multiple tests can also be combined through a fixed rule as in a composite reference standard (i.e., diagnostic decision rule [9, 79]). Finally, statistical models that account for uncertainty about the true disease status have been suggested. These probabilistic models originated in the social sciences in the work of Lazarsfeld [107] in the 1950s, and Lazarsfeld and Henry [108] in the 1960s. In more recent years these models have found application in modeling of multiple tests results to estimate the target condition prevalence and index test accuracy in the absence of a gold standard [64, 94, 95, 144]. These models are commonly referred to as latent class models.

Objectives

The primary aim of the work in this thesis is to explore and improve the methodology for the evaluation of diagnostic test accuracy in the absence of a gold standard. In particular, we focus on two commonly used methods found in diagnostic test evaluation: latent class modeling and composite reference standards.

Outline of this thesis

Chapters 2 through 5 focus on the use of latent class models in diagnostic studies without a gold standard. In **Chapter 2** we present a systematic review of studies that used latent class models to estimate diagnostic accuracy. We focus on the statistical methods used and their reporting. In **Chapter 3** we evaluate the performance of goodness-of-fit testing to detect violations of the criticized conditional independence assumption underlying the standard 2-class latent class model. Power and Type-I error rates are evaluated based on three empirical examples. In **Chapter 4** we present a Bayesian latent class analysis evaluating the accuracy of five diagnostic

tests for childhood pulmonary tuberculosis. Data from a study of hospitalized children in South Africa are used. In **Chapter 5** we revisit the evidence presented in an influential paper by Albert and Dodd [6] that cautioned against the use of latent class analysis in the absence of a gold standard.

The use of composite reference standards in diagnostic studies without a gold standard is discussed in Chapters 6 and 7. In **Chapter 6** we explore the rationale for using composite reference standards and make recommendations for reporting their results. In **Chapter 7** we provide insight into the workings of the composite reference standard. We evaluate the driving factors of bias in estimators of index test accuracy and disease prevalence that is due to target condition misclassification which remains present when using composite reference standards.

Chapter 8 and 9 concern logistic regression models, which are often used in the context of (multivariable) diagnostic research. In **Chapter 8** we investigate the role of the number of events per variable (i.e. the well-known 1 in 10 rule) as a factor in the performance of binary logistic regression analysis. We explore the potential reasons for the varying minimal events per variable recommendations from earlier simulation studies. In **Chapter 9** we present a nomogram that can be used to improve the reporting of multinomial logistic regression models.

Chapter 10 provides a general discussion of methods to improve diagnostic research in the absence of a gold standard. Future challenges and directions for new research are discussed.

Chapter 2

Latent class models in the absence of a gold standard: a systematic review

Abstract

Latent class models (LCMs) combine the results of multiple diagnostic tests through a statistical model to obtain estimates of disease prevalence and diagnostic test accuracy in situations where there is no single, accurate reference standard. We performed a systematic review on the methodology and reporting of LCMs in diagnostic accuracy studies. This review shows that the use of LCMs in such studies increased sharply in the past decade, notably in the domain of infectious diseases (overall contribution: 59%). The reviewed studies (n = 64) used a range of differently specified parametric latent variable models, applying Bayesian and frequentist methods. The critical assumption underlying the majority of LCM applications (61%) is that the test observations must be independent within two classes. As violations of this assumption can lead to biased estimates of accuracy and prevalence, performing and reporting checks whether assumptions are met is essential. Unfortunately, our review shows that 28% of the included studies failed to report any information that enables verification of model assumptions or performance. Due to the lack of information on model fit and adequate evidence external to the LCMs, it is often difficult for readers to judge the validity of LCM based inferences and conclusions reached.

An essential step in the evaluation of a diagnostic test or biomarker is to obtain valid estimates of its accuracy, that is, its ability to discriminate between patients who have the disease of interest and those who do not [104, 145]. Typically, the accuracy of a single or a set of diagnostic index tests is analyzed by examining the results of index tests in relation to the outcome of the reference standard in patients suspected of a disease of interest. A single and error-free reference standard is preferred, but for many diseases such a reference standard, also known as a "gold standard," does not exist [16, 123, 143]. The use of an imperfect reference standard will often lead to misclassification of the disease status in a substantial portion of subjects, which can lead to biased estimates of index test performance and disease prevalence.

One approach to reducing these misclassifications is to combine multiple pieces of diagnostic information to determine the disease status among study patients. Multiple tests may, for example, be used in expert panels in which a group of clinicians reach consensus based on the available test results of patients [65, 104, 123, 179]. Results from multiple tests can also be combined through a fixed rule as in a composite reference standard (i.e., diagnostic decision rule [9]). Finally, as a probabilistic alternative, a latent variable approach may be adopted by combining multiple diagnostic tests using a latent class model (LCM).

In the past decades latent class modeling (i.e., latent class analysis) has been applied in medical and veterinary sciences, particularly in test accuracy research [64, 67, 95, 144, 182]. The use of LCMs appears attractive as it avoids the time-consuming process of reaching consensus diagnoses or the inherent difficulty in defining a diagnostic decision rule a-priori in case a single reference standard for the target disease is lacking. LCMs can produce valid estimates of accuracy even in the absence of a perfectly accurate disease status classification (an accurate reference standard) and can be estimated in popular statistical software packages such as SAS software (SAS Institute Inc., Cary, NC) and R (R Foundation, Vienna, Austria) as well as specialized software such as Latent Gold [170].

Latent class modeling refers to a heterogeneous group of statistical models. Differently specified LCMs can be fitted to the same set of test results, which in turn can lead to relevant differences in disease prevalence and test accuracy estimates [6, 23, 39, 52, 180]. Researchers, therefore, need to inform the reader how their LCM(s) were specified. Additionally, as with any statistical technique or model, the validity of its results are jeopardized when assumptions are not met. Hence, performing and reporting checks whether assumptions are met is essential to readers to appraise the validity of the reported results when LCMs are used.

To explore the methodology and reporting of LCM applications in diagnostic research, we performed a systematic review of test accuracy studies that applied such a model. This review will provide an overview of the history of LCM applications in test accuracy research, reveal variations in the models used across studies and provide clues on how to improve the reporting and methodology of these studies. Before presenting the results of the review, we will first describe the key characteristics of LCMs.

Introduction to Latent Class Models

Diagnostic studies that apply LCMs treat the target disease status as an unmeasured (”latent”) categorical variable with K classes, reflecting the levels of the underlying disease. The manifest variables, the outcomes of R (binary) diagnostic tests, are considered to be imperfect classifiers of the disease status. The LCM describes a statistical model relating the manifest variables to the latent disease status. For the mathematical underpinning we refer the reader to Appendix 1.

When $K = 2$, it is assumed that the two latent classes correspond to a class of subjects in which the target disease is present and a class of subjects in which the target disease is absent. Parameter estimates obtained from the 2-class LCM are interpreted as estimates of the sensitivity of each test (i.e., the probability of a positive test result when the target disease is present) and specificity of each test (i.e., the probability of a negative test result when the target disease is absent) and the prevalence of the target disease (i.e., the (prior) probability that the target disease is present). Two important issues that are encountered when applying LCMs are identification of the LCM and the assumption of conditional independence.

Estimation and identifiability of Latent Class Models

Maximum likelihood estimates of the LCM parameters can be obtained using a variety of estimation methods, including EM or Newton-Raphson algorithms [170]. However, LCMs may not always be identified which implies that the maximum likelihood estimates are not unique; a different set of parameter estimates exists for which the likelihood (value) is the same.

A necessary condition for the LCM to be identified is that the number of freely estimated parameters does not exceed the number of unique diagnostic test patterns. For example, an unconstrained ”basic” 2-class LCM is not identified with $R \leq 2$ diagnostic tests and is ”just identified” with $R = 3$, resulting in an LCM with zero degrees of freedom. Non-negative degrees

of freedom, though, is not a sufficient condition for identification as is evidenced by an LCM with $R = 4$ and $K = 3$ which has 1 degree of freedom but is not locally identified [72]. In practice, local identifiability of LCMs can be explored by examining the rank of the Jacobian matrix [72, 118].

One solution to non-identifiability is imposing constraints to the parameters [72]. For example, if the conditional test outcome probabilities of a particular diagnostic test can be assumed to be known a-priori, e.g., based on theory [172], degrees of freedom can be gained by fixing the parameters to their true value, allowing the remaining LCM parameters to be estimated freely.

Another strategy is adopting a Bayesian approach. Since the true values of conditional diagnostic test outcome probabilities are often not exactly known in advance, the use of fixed parameters may be invalid. Instead of constraining the parameters to a fixed value, "informative" prior distribution can be defined for those parameters for which prior knowledge is available. With substantive prior information, estimates from the posterior densities of unidentified LCMs can be obtained by a Gibbs sampler [101]. Detailed discussions on estimation and identification are found elsewhere [72, 152].

Conditional independence

The important assumption that underlies LCM estimation is that of conditional independence (i.e., local independence). In its most basic form, this assumption reduces to independence of observations conditional on the presence or absence of the disease of interest. This assumption is central to the 2-class independence LCM (defined in equation 3 in the Appendix 1). It results in the model being identified with only 3 binary diagnostic tests. However, violations of local independence assumptions are known to lead to bias in estimates of accuracy and prevalence, while the assumptions may not be warranted in many practical situations [153, 160, 164, 174].

One way to relax the conditional independence assumption is by increasing the number of latent classes to be estimated. We will refer to these LCMs with more than two classes as multi-class independence LCMs. Alternatively, other LCMs that do not require the independence assumption of observations conditional on (or "within") the classes have been suggested. For example, when independence among observations in a K-class LCM is not met due to bivariate dependence among a pair or a subset of pairs of tests, these bivariate associations can be modeled directly by defining an additional parameter for each bivariate relation [59, 81, 83]. Other strategies to account for dependence within classes include defining marginal models

[181], estimating a multiple latent variable model [52], and adding random effects [53, 140]. We will refer to these models as dependence LCMs. For a detailed discussion on common LCM specifications in diagnostic research, we refer to literature [95].

Evidently, an increased number of parameters are estimated when the independence assumption is relaxed, at the expense of degrees of freedom. Hence, a higher number of diagnostic tests are needed for extensions to the independence 2-class LCMs. For example, at least 5 binary diagnostic tests are needed for an unconstrained 3-class independence LCM to be identified. Estimating multi-class independence LCMs or dependence LCMs is therefore not always feasible when the number of available diagnostic tests (R) is limited.

Latent class model verifications

To verify LCM assumptions, measures of model fit can provide important information. Preferred is an LCM that provides a superior, or at least equivalent, fit to the data compared to alternative LCMs (e.g., with more classes) and that has an adequate global fit (e.g., differences in observed vs. expected number of patients with specific patterns of test results should be small). Particularly useful are also residual dependence diagnostics that can pinpoint sources of misfit due to bivariate dependence between diagnostic tests [140, 148, 170].

Clearly, evaluating model fit can provide important information regarding potential misspecification of the LCMs used. The reporting of the model evaluation steps taken and results obtained from alternative models provides therefore valuable information for readers to judge the credibility of the results presented. Nonetheless, since the latent variable is unobserved, LCM assumptions cannot be tested directly. Even with a large sample size it may be hard to distinguish between LCMs [6].

The credibility of LCM based inferences may therefore also rely on the use of external data. For example, it is sometimes possible to compare the latent class outcomes to outcomes derived using a proxy measure for disease status, e.g., disease status measured using an adequate reference standard in a subset of patients [7].

Systematic Review of diagnostic studies applying Latent Class Models

Our aim was to identify diagnostic studies that reported diagnostic test accuracy or disease prevalence estimates derived directly or as a function of LCM parameter estimates. Hence, techniques that use assigned clusters (e.g., cluster analysis) to derive accuracy or prevalence

estimates fall outside of the scope of this review. EMBASE and PubMed databases were searched for the following free text search terms: latent class OR latent classes OR finite mixture OR finite mixtures on November 8th, 2011. Papers published in English, Dutch or Spanish were considered for inclusion.

Information was extracted on clinical context, study characteristics, model specifics and diagnostics, model comparisons and software. The extraction form was pilot tested by four researchers (MvS, CN, JdG and JR). One researcher (MvS) screened and evaluated all studies. Parallel screening and data-extraction were performed independently by a second researcher (CN, JdG or JR). Disagreements were resolved by discussion by the first and second researcher, and in case of remaining doubt, by a research-group discussion. To evaluate reported evidence on LCM performance in diagnostic studies, we differentiated between four categories of model fit criteria: (a.) Assessment of relative fit, which ideally results in the selection of one LCM among a set of theoretically plausible LCMs, e.g. by significance testing or information criteria (e.g., Akaike information criterion, AIC); (b.) Goodness-of-fit testing, evaluating the fit of an LCM to the observed data using a significance test (e.g., Pearson χ^2 test [63]); (c.) Dependence diagnostics to pinpoint residual dependence under an LCM [148], e.g., correlation residual plot [72] or BiVariate Residual Statistics [170]; (d.) A table of expected versus observed frequencies.

Model performance can also be evaluated by a leaving-one-out comparison, in which a manifest variable is excluded and its results are compared to the results obtained from another LCM in which the manifest variable is included. Finally, model performance can be evaluated using external evidence, e.g., by comparing LCM estimates with similar estimates derived using an imperfect reference standard.

Results

Figure 2.1 depicts the results of the search and inclusion of papers. Of the 1704 papers screened for eligibility on title and abstract, 242 met the eligibility criteria. One publication could not be retrieved in full-text format. After full-text reading of 241 publications, another 91 were excluded. Reasons for exclusion were falling outside the scope of this paper; publications that reported LCMs which only estimated (rater) agreement [1] and models that included continuous test results [105] or covariates [110]. Screening of reference list of the included papers yielded an additional 30 publications, resulting in a total of 180 publications, of which 179 were published in English and 1 in Spanish.

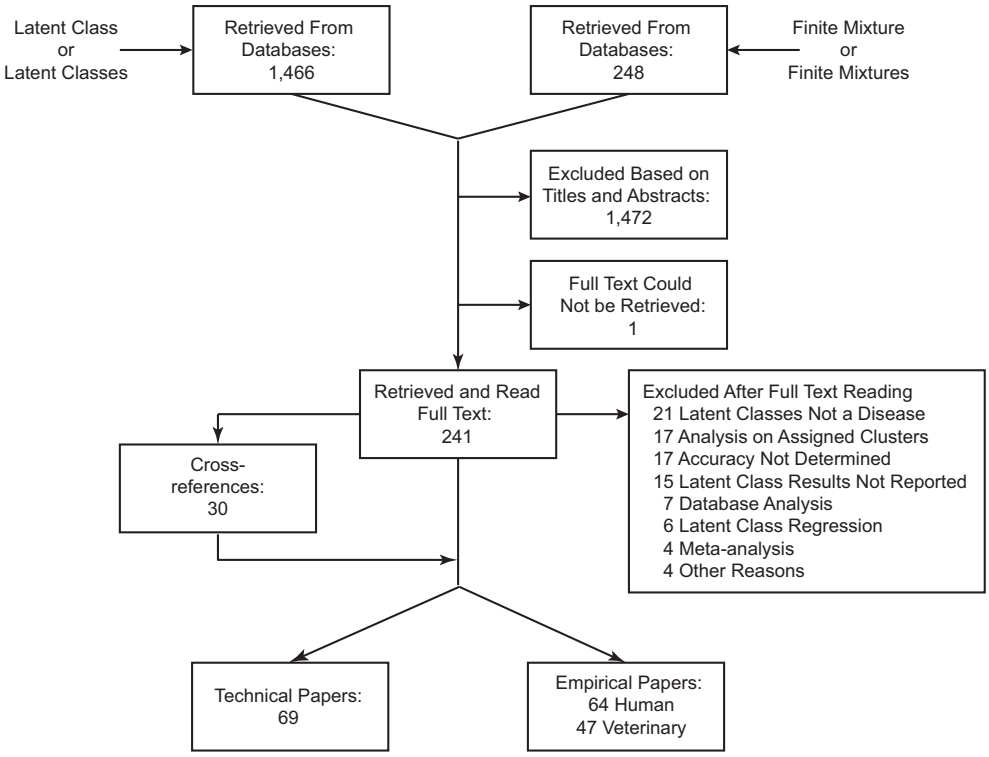

Figure 2.1: Inclusion of studies for systematic review on latent class methodology in diagnostic studies. Other reasons for exclusion were the use of longitudinal analysis, being identical to an included paper, or the use of nonparametric models.

The included papers were classified as theoretical (n=69), when focused on latent class model methodology, or empirical (n=111), when the focused on analyzing original data ("original papers"). The empirical studies were further divided into animal studies (n=47) and human studies (n=64). In the remainder of this paper we will only focus on the empirical studies involving human subjects. A short description of the veterinary studies is found in Appendix 2.

General characteristics

The first applications of LCMs in diagnostic studies involving human subjects originate around 1990. A steep increase in statistical and methodological publications has occurred since 2000, followed by an increase in empirical studies starting approximately 5 years later (Figure 2.2). In total, 64 studies were identified of which approximately half (n=34) were published between 2007 and 2011.

Table 2.1: Characteristics of 64 studies that used latent class models to estimate accuracy of diagnostic tests or prevalence of a disease.

	Frequency
Latent condition of interest[a]	
Infectious and parasitic diseases	38
Mental and behavioural disorders	6
Diseases of the musculoskeletal system	4
Diseases of the digestive system	3
Neoplasms	2
Diseases of the respiratory system	2
Other	9
Year of publication	
2007 - 2011	34
2002 - 2006	19
< 2002	11
Main goal of publication	
Test accuracy	51
Expert accuracy	8
Disease prevalence	4
Unknown	1

[a] Based on *International Classification of Diseases, Tenth Edition*, codes.

The primary goal of the vast majority of publications (n=59; 92%) was the evaluation of accuracy of diagnostic tests (n=51), or accuracy of a diagnosis made by clinicians (n=8) (Table 2.1). All of these studies reported test sensitivity and specificity estimates; predictive values of the tests were additionally reported in 11 of these studies (17%). The predictive values of each particular test pattern combination were reported in 7 studies. In 4 studies the primary goal was to estimate disease prevalence. In one study the primary goal could not be determined.

LCM applications were primarily found in the field of infectious diseases. In 38 publications (59%) the disease of interest was an infectious disease. Other diseases were mental or behavioral problems (n=6), diseases of the musculoskeletal system (n=4), diseases of the digestive system (n=3) and neoplasms (n=2).

Thirty-three papers (52%) had either 3 (n=19) or 4 (n=14) diagnostic tests (i.e., manifest variables) included in their latent class model(s). The median value of included diagnostic tests was 4 (inter quartile interval: [3 to 6]). Reported sample sizes ranged from 34 to 4708 with a median value of 315 (inter quartile interval: [171 to 737]).

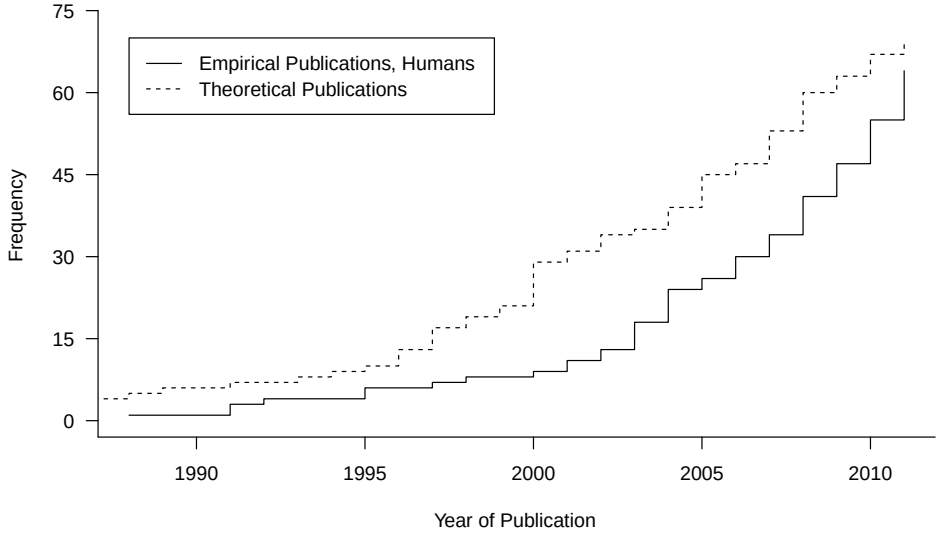

Figure 2.2: Cumulative number of empirical diagnostic studies in humans using a latent class model and theoretical studies published.

Variety of LCMs

The majority of studies (n=39; 61%) reported analyses based solely on 2-class independence LCMs (Table 2.2). These studies did not report LCM estimates or model fit statistics from multi-class LCMs or dependence LCMs. These studies also did not report comparisons between LCMs, of which at least one was not a 2-class independence LCM. Analyses originating from 3- and 4-class LCMs were reported in 12 papers (19%). LCMs with more than 4 classes were not found in the surveyed studies. Fifteen papers (23%) reported results based on dependence LCMs, e.g., by including random effects. Two papers could not be classified in Table 2.2 due to insufficient information on the number of classes estimated.

Assessment of relative fit

In 24 studies (38%), all results originated from a single LCM. The other 40 studies reported parameter estimates or model diagnostics of two or more differently specified LCMs. In 15 studies of the 40 studies presenting more than 1 LCM, information on relative fit was reported, e.g., by significance testing or information criteria.

Table 2.2: Properties of latent class models in absolute numbers as applied in 64 empirical diagnostic accuracy studies.

		Single model (n=23)			> 1 model (n=39)			Not classified[a]
	All	$K = 2$[b]	$K > 2$[c]	Dep[d]	$K = 2$[b]	$K > 2$[c]	Dep[d]	
Frequency	64	18	1	4	21	7	11	2
Model fit measures								
Relative model fit	15	NA	NA	NA	3	3	9	0
Goodness-of-fit testing	26	3	0	1	11	4	7	0
Dependence diagnostics	4	1	0	0	1	0	2	0
Exp/obs frequencies[e]	8	1	0	0	1	2	4	0
None of the above	35	14	1	3	10	3	2	2
Leave-one-out comparison[f]	5	NA	NA	NA	4	0	1	0
No model fit nor leave-one-out	33	14	1	3	8	3	2	2
External evidence	33	12	0	2	9	2	7	1
No model fit nor leave-one-out nor external evidence	18	6	1	1	4	3	2	1
Bayesian estimation	15	3	0	4	4	1	3	0

Abbreviations: LCM(s), latent class model(s); NA, not applicable.

[a] Number of latent classes modelled could not be extracted from publications. One of these papers reported analyses based on a single latent class model while the other reported multiple models of which the number of latent classes could not be determined.

[b] Studies using the 2-class independence LCM.

[c] Studies using the multiclass independence LCM.

[d] Four publications presenting results of dependence LCMs also presented results from multi-class independence LCMs or multi-class dependence LCMs.

[e] Comparison of expected versus observed frequencies.

[f] Leave-one-out comparison refers to the situation in which a manifest variable is excluded and its results are compared to the results obtained from another LCM in which the manifest variable is included.

Model fit and leave-one-out comparison

Goodness-of-fit evaluation of the model(s) based on significance testing was reported in 26 studies (41%), a table of observed against expected number of test patterns was present in 8 (13%) and residual dependence diagnostics were reported in 4 studies. Thirty-five (55%) studies did not report any of the model fit criteria on relative fit, goodness-of-fit, dependence diagnostics or expected versus observed test pattern frequencies.

Five studies reported a leave-one-out comparison. These studies evaluated the stability of obtained LCM estimates, using varying subsets of the available diagnostic tests in the LCMs (i.e., reported analyses originating from multiple LCMs differing in the diagnostic tests that were included in the model). In total, 33 studies (52%) did not report any information model fit criteria or an evaluation of estimates stability.

External evidence to support LCM findings

Thirty-three studies (52%) reported additional results that were not directly derived from an LCM, which enables a comparison with the estimates obtained from the LCM(s). For example, by using one of the diagnostic tests as a reference standard, apparent accuracy and prevalence estimates were determined for the other tests in 17 studies (27%). Eight studies (13%) applied a fixed rule that combined the results of diagnostic tests to determine the target disease status of patients (i.e., a composite reference standard). In 5 studies (8%) the disease status was determined by a single or a panel of experts. 3 studies used a combination of the above. Eighteen studies (28%) failed to report any information that could be used to verify the validity of latent class model estimates.

Estimation

Optimization was obtained using maximum likelihood estimation in most studies (n=49; 77%); 14 studies (22%) used a Bayesian optimization approach. One study reported estimates obtained from frequentist and Bayesian models. A total of 14 different software programs were reported in 52 studies; LatentGold (n=10), WinBUGS (n=9), LEM (n=8), Latent1 (n=7) and SAS (n=4) were most frequently reported.

Discussion

Our systematic review shows that the use of latent class models (LCMs) in diagnostic studies has increased considerably in the past decade. This is probably a reflection of increased awareness that a gold standard does not exist for many conditions [16, 123, 143, 145]. LCMs may seem appealing because they combine the results of multiple tests to improve the classification in an objective way. Our review, however, revealed several problematic issues related to the methodology and reporting of studies using LCMs that deserve further attention. The majority of studies used a 2-class independence LCM to estimate test sensitivity, specificity and target disease prevalence. The strong assumption made in these studies is that conditional on the binary target disease status, test results are independent (uncorrelated). This conditional independence assumption can easily be violated, for instance when some individuals without the disease of interest have another condition in common that increases the likelihood of two (or more) tests to become false positive as they are based on a comparable biological principle [52]. Another cause for conditional dependence among test results could arise if there is a subgroup of individuals with an early or less severe stage of the disease of interest, and if these individuals are more likely to be missed by different tests [31].

Examining whether the LCM used for inferences is appropriate for the data at hand is critical, as violating the independence assumptions can lead to biased estimates of accuracy [6,153,160,164]. Several approaches exist, such as examining residual correlations, comparing with alternative LCMs that assume different dependence structures (dependence LCMs) or evaluating the goodness-of-fit. Unfortunately, more than half of studies (52%) fail to present information that is related to the fit of the model or stability of the estimates. Due to this absence of model performance information, readers are often left in the dark about the appropriateness of the models and the validity of the results.

One contributing factor for the limited number of studies comparing the results from the 2-class independence model with more complex models is the limited number of diagnostic tests (i.e., manifest variables) available in the surveyed studies. This is especially true for studies that have data available on only 3 diagnostic tests. The possibilities for evaluating model fit and model comparisons are then limited, unless parameter constraints can be imposed.

The parameter constraints that could be imposed can take the form of fixing parameters to a known value or equality constraints. A Bayesian alternative is defining informative prior distributions on the parameters for which prior information is available, which creates an opportunity of estimating and comparing LCMs that, from a frequentists perspective, are not identified [101]. Of course, the validity of a Bayesian approach relies on proper values (use) of prior information. When prior information is lacking, researchers might want to collect data on additional diagnostic tests in order to be able to verify LCM assumptions.

Performing and reporting on the checks of assumptions is a major step forward in the critical appraisal of the results of LCMs. However, we do acknowledge the limed power of performance criteria to detect misfit. In the absence of explicit criteria, researchers can incorporate partial information on the disease status into the analysis, e.g. an adequate reference standard measured for a subset of patients [7], or use model averaging techniques [23]. Applications of model averaging or incorporation of partial information on the disease status were, however, not found in our systematic search.

To substantiate the face validity of inferences, some of the studies alternatively reported external evidence to enable comparison of estimates derived using the LCM. For example, some studies compared LCM estimates with estimates derived using a composite reference standard or a panel diagnosis.

More often, though, external evidence was based on apparent disease prevalence and test sensitivity and specificity estimates obtained by using one of the available tests as the (imperfect) reference standard. A comparison of these estimates with LCM parameter estimates rarely contributes to a satisfying conclusion regarding the credibility of LCM based accuracy and prevalence estimates since the reason for applying LCMs is the absence of an accurate reference standard. The comparison of the naïve estimates obtained from the imperfect reference standard with the LCM estimates might not be used for evaluation of the LCM, but rather as a sensitivity analysis for the obtained naïve estimates, taking into account the imperfect nature of the standard. Surely, the validity of this comparison also relies on the assumptions of the LCM.

We recognize that some publications of diagnostic studies that use latent class models may not have been identified in our review. However, as this is a review of reported methods, complete coverage is not critical as, for example, in an intervention review. Our broad search strategy captured a large representative sample and the risk of missing relevant publications was reduced by checking the reference lists of included publications. As our goal was to evaluate the LCM reporting and methodology in peer reviewed journals, research that has not been published falls outside the scope of our review. Therefore, the potential for bias due to unpublished work (i.e., publication bias) is not relevant.

The recent increase in use of LCMs in test accuracy research has coincided with an increase in attention for problems that can be encountered when facing a reference standard which is imperfect. It should be recognized that despite all the attention given to the potential problems, valid LCMs can reduce the risk of reference standard bias. However, the reporting of latent class analyses is often insufficient for readers to be able to critically appraise the obtained results. To improve reporting, we suggest that future studies provide detailed information about the exact specification of their LCM(s). Additionally, we suggest that all studies describe in detail how the assumption of conditional independence was verified, as well as report information that can help the reader appraise the validity of the obtained results.

Appendix

Mathematical Underpinning of latent class models (Appendix 1)

Suppose that data are collected from a single diagnostic test X in a sample of patients suspected of a specific target disease. Assuming that patients are either diseased ($D = 1$) or not diseased ($D = 0$) and that test results can be classified as either positive ($X = 1$) or negative ($X = 0$), it follows that the probability of observing a positive test result is the sum of probabilities of a true-positive and a false-positive test result,

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 1 \cap D = 1) + \mathbb{P}(X = 1 \cap D = 0) = \theta\alpha + (1 - \theta)(1 - \beta), \tag{1a}$$

where θ is the prevalence of the target disease, and α and β are the sensitivity and specificity of the diagnostic test respectively. Similarly, for the probability of a negative test result,

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 0 \cap D = 0) + \mathbb{P}(X = 0 \cap D = 0) = (1 - \theta)\beta + \theta(1 - \alpha). \tag{1b}$$

A generalization of 1a and 1b can be written as,

$$\mathbb{P}(X = x) = \theta\alpha^{x}(1 - \alpha)^{1-x}(1 - \theta)\beta^{1-x}(1 - \beta)^{x}. \tag{2}$$

Assuming that a sample of subjects suspected of a target disease of size N is drawn and every sampled subject receives R diagnostic tests, equation2 can be generalized by assuming independence of observations conditional on the target disease status as follows:

$$\mathbb{P}(X = [x_1, ..., x_R]) = \theta \prod_{r=1}^{R} \alpha_r^{x_r}(1 - \alpha_r)^{1-x_r}(1 - \theta) \prod_{r=1}^{R} \beta_r^{1-x_r}(1 - \beta_r)^{x_r}. \tag{3}$$

Positive and negative predictive values (probability of disease presence (or absence) when test is positive (or negative)) can be obtained using Bayes theorem.

By letting $\tau_{r|k}$ denote the probability of a positive test r in class k and θ_k denote the probability of a random individual belonging to class k, a generalization of equation 3 to a latent class model with K classes is obtained by,

$$\mathbb{P}(X = [x_1, ..., x_R]) = \sum_{k=1}^{K} \theta_k \prod_{r=1}^{R} \tau_{r|k}^{x_r}(1 - \tau_{r|k})^{1-x_r}, \quad \sum_{k=1}^{K} \theta_k = 1. \tag{4}$$

This model has degrees of freedom, $df = 2^R - P - 1$, where P is the number of freely estimated parameters. Note that when $K = 2$, the model in equation 4 is equivalent to model in equation 3.

Let $n = [n_1, ..., n_{2^R}]$ denote the frequencies of 2^R patterns of test results, where $\sum_{s=1}^{2^R} n_s = N$. The likelihood of the data may be written as,

$$\mathcal{L}(n|\theta_1, ..., \theta_K, \tau_{1|1}, ..., \tau_{R|1}, ..., \tau_{R|K}) \propto \prod_{s=1}^{S} \left[\sum_{k=1}^{K} \theta_k \prod_{r=1}^{R} \tau_{r|k}^{x_r} (1 - \tau_{r|k})^{1-x_r} \right]^{n_s}. \quad (4)$$

Optimization of the likelihood function, e.g. by an EM-algorithm, yields maximum likelihood estimates of the parameters and (asymptotic) variances and covariance can be obtained by the inverse of the expected Fisher information matrix. Alternatively, Bayesian estimation techniques (e.g., Gibbs sampling, [101]) and analytic solutions have been proposed to obtain the parameter and variance estimates.

Characteristics of Veterinary Publications (Appendix 2)

We identified 47 applications of latent class models (LCMs) in veterinary publications that evaluated the accuracy of diagnostic tests or disease prevalence. All of these studies targeted infectious diseases and 34 were published between 2007 and 2011. Interestingly, more than 3 out of 4 (n= 37) publications apply Bayesian models. Often, veterinary publications have test observations that are nested in herds varying in infection prevalence. Using the method introduced by Hui and Walter [30, 94], herd-level prevalences are estimated while assuming that the test accuracies are constant across herds. This can be viewed as a special case of a multi-group LCM and was used in 34 of the veterinary studies.

Chapter 3

Detecting misfit of latent class models in the absence of a gold standard

Abstract

Objectives

The objective of this study was to evaluate the performance of goodness-of-fit testing to detect relevant violations of the assumptions underlying the criticized standard 2-class latent class model. Often used to obtain sensitivity and specificity estimates for diagnostic tests in the absence of a gold reference standard, this model relies on assuming that diagnostic test errors are independent. When this assumption is violated, accuracy estimates may be biased: goodness-of-fit testing is often used to evaluate the assumption and prevent bias.

Study Design and Setting

We investigated the performance of goodness-of-fit testing by Monte Carlo simulation. The simulation scenarios were based on three empirical examples.

Results

Goodness-of-fit tests lack power to detect relevant misfit of the standard 2-class latent class model at sample sizes that are typically found in empirical diagnostic studies. The goodness-of-fit tests that are based on asymptotic theory are not robust to the sparseness of data. A parametric bootstrap procedure improves the evaluation of goodness-of-fit in the case of sparse data.

Conclusion

Our simulation study suggests that relevant violation of the local independence assumption underlying the standard 2-class latent class model may remain undetected in empirical diagnostic studies, potentially leading to biased estimates of sensitivity and specificity.

A key step in the evaluation of a diagnostic test (e.g., imaging test, electrophysiology or biomarker test) is the assessment of its accuracy, commonly measured in terms of sensitivity and specificity. To assess the accuracy of the diagnostic test under study it is necessary to obtain information on the true target disease status of study subjects that is preferably obtained from a reliable source with perfect accuracy: a gold reference standard. Often, however, the best available reference standard is not completely free of error [143, 145]. Using such a reference standard while disregarding these problems leads to a biased assessment of accuracy of the diagnostic (index) test [29, 112, 176].

Latent class analysis has been proposed to circumvent this bias [94, 101, 144, 173]. The latent class model combines the information from multiple, generally three or more, imperfect diagnostic tests to uncover the unobserved disease structure. This approach has, for example, been used to study the diagnostic value of immunohistochemical assays of bladder tumors [8], to evaluate diagnostic tests to detect *Visceral Leishmaniasis* [24, 25], to estimate diagnostic accuracy of test for acute maxillary sinusitis [47], and the accuracy of surgeons classifications of bone fracture types [12].

The standard 2-class latent class model that accounts for the majority of applications in diagnostic accuracy and disease prevalence studies [168], relies on making two interrelated assumptions: i) existence of two classes representing groups of true target disease positive subjects and true target disease negative subjects, and ii) local independence with respect to the imperfect diagnostic test used in the latent class analysis [118]: the outcomes of the diagnostic test are stochastically independent conditional on class membership. Together, these assumptions have been criticized for being unrealistic for the majority of diagnostic studies (e.g., see [9, 31, 136, 174]), potentially leading to severely biased assessments of sensitivity, specificity and disease prevalence [71, 79, 160, 164]. Suggested alternative latent class models that prevent this bias by accounting for dependence in diagnostic test errors have been developed and have found application in more recent literature (for reviews and mathematical underpinnings, see [41, 64, 95, 118]).

In practice, a justification for the local independence latent class model is often sought in testing its goodness-of-fit. However, no studies to date have examined whether this approach yields sufficient power to detect local dependence and prevent biases at sample sizes typical of diagnostic studies. A recent systematic review [168] of latent class applications in diagnostic accuracy and prevalence studies estimated a median sample size of approximately 350 subjects in such studies. One may therefore question whether the sample sizes of these studies are

indeed sufficient to detect relevant deviations from assumptions. Second, while commonly used measures of latent class model fit approach a chi-square distribution under the null hypothesis as the sample size increases, in finite samples this distribution may not be chi-square [142]. Especially when there is large agreement between diagnostic tests, leading to some combinations of diagnostic test outcomes to be observed only rarely, these test statistics may not approach their theoretical distribution. It is therefore paramount to study the behavior of the model fit test under realistic sample size conditions.

In this paper we study the performance of testing the goodness-of-fit of latent class models based on asymptotic theory and parametric bootstrap procedures [170]. We study power to detect misfit of the standard 2-class latent model in scenarios where there is a relevant violation of the local independence assumption. We will also study the false rejection rates (Type-I error) for scenarios where diagnostic test outcomes are locally dependent. First, we consider the basic theory and assumptions of latent class analysis. We subsequently describe 3 (large sample) case studies obtained from literature that have presented latent class models for dependent diagnostic test outcomes. The reported results from these publications will be used as the data generating mechanisms in a Monte Carlo simulation study to evaluate the performance of the goodness-of-fit tests in realistic settings.

Latent class model

The latent class model for the joint density of diagnostic test outcomes $f(x)$ can be written as,

$$f(x) = \sum_d \pi_d g(x|d),$$

where $\pi_d = \mathbb{P}(D = d)$ is an estimator of the prevalence of disease stratum d, and $g(x|d)$ is a model for the joint density of diagnostic test outcomes within stratum d. In the following, we shall limit our discussion to the common case in which diagnostic binary test data are available on N subjects, taking on the values $x_j = 1$ for a positive test result on test j, and $x_j = 0$ when negative.

The 2-class latent class model that has become the standard in applications in the field of diagnostic research, is based on the assumption that the outcomes of the diagnostic test, $j = 1, ..., J$, are mutually independent given the latent variable. This latent variable is assumed to have two classes, here denoted by $d = 0, 1$. Hereafter, we refer to this model by 2-class local

independence model (in short: 2-class LI model) that can be written as,

$$f(x) = \sum_d \pi_d g(x|d),$$
$$g(x|d) = \prod_{j=1}^{J} \pi_{x_j|d}^{x_j} (1 - \pi_{x_j|d})^{1-x_j}.$$

These parameters are estimators of the sensitivities of J diagnostic tests $\pi_{x_j|d=1} = \mathbb{P}(X_j = 1|D = 1)$, the specificities of J diagnostic tests $1 - \pi_{x_j|d=0} = \mathbb{P}(X_j = 0|D = 0)$ and the prevalence of the target disease $\pi_{d=1} = 1 - \pi_{d=0} = \mathbb{P}(D = 1)$.

Crucially, the parameters of the latent class model must be identifiable to obtain meaningful estimates [72, 100]. For estimating the 2-class LI model data must be available on at least 3 binary diagnostic tests. The other latent class models we consider require data on at least 4 (models described in case study II and III) or 5 diagnostic tests (model described in case study I).

Goodness-of-fit

The overall goodness-of-fit of the latent class model is evaluated by comparing the frequencies of the observed diagnostic test outcome patterns to expected pattern frequencies under the estimated latent class model. The prevailing test statistics that formalize this comparison are the Pearson Chi-square statistic, denoted by X^2, and likelihood ratio statistic, denoted by G^2. Both statistics are asymptotically chi-square distributed under the null hypothesis of perfect model fit, with degrees of freedom (*df*) equal to the residual degrees of freedom, that is, the number of independent diagnostic test outcome patterns (with binary tests given by $2^J - 1$) minus the number of freely estimated parameters. For details, see [2, 63].

In practice, behavior of these statistics under the null hypothesis can be poor when combinations of test outcomes are only rarely observed, i.e., when expected frequencies are low for particular diagnostic test outcome combinations [117, 142]. This sparseness of data is influenced by several factors, including: the sample size, the number of diagnostic tests (sparseness exponentially increasing with number of tests), the target disease prevalence and operating characteristics of the diagnostic tests under study. In this study, sparseness is expected to occur most pronounced in simulation scenario I, due to the combination of a high number of diagnostic tests ($J = 10$), low target disease prevalence of the target disease (also in simulation scenario II and III) and high specificities of tests (also in simulation scenario III).

Table 3.1: Case study I: Depression example Garrett et al. (2002), a 3-class LI model, with classes labeled as none, mild and severe depression. Estimated values are the within class probabilities of a positive outcome on the 10 symptoms.

	Class 1	Class 2	Class 3
Prevalence	.88	.09	.03
Movement	.01	.17	.73
Appetite/weight	.04	.31	.78
Sleep	.03	.49	.77
Morbid thoughts	.02	.24	.54
Guilt/sin	.01	.08	.37
Self-esteem	.01	.10	.56
Concentration	.01	.15	.78
Fatigue	.01	.25	.64
Loss of interest	.02	.32	.91
Depressed mood	.02	.30	.78

In our simulation study we evaluate power and Type-I error rates of these asymptotic X^2 and G^2 goodness-of-fit tests to detect relevant deviations from the local independence assumption. We additionally study the performance of bootstrapped X^2 and G^2 goodness-of-fit testing which has been suggested to be especially useful in situations where data are sparse [42, 106, 170].

Case studies

We review 3 previously published studies that developed latent class models for locally dependent diagnostic test outcomes. The reported models and parameter point-estimates are used as the data generating mechanisms in the simulation study that follows.

Case study I: Depression, Garrett et al. [67]

Garrett and colleagues have evaluated the diagnostic value of ten symptoms relating to depression ($N = 1322$). The authors concluded that the 2-class LI model showed poor fit to the data as compared to the 3-class local independence latent class model. Hence, Garret and colleagues hypothesized that the local independence assumption of the 2-class model was violated due to the existence of a third depression stratum.

The reported parameter estimates obtained from the 3-class LI model are found in Table 3.1. Class 1 is labeled as the no depression class, has the highest estimated prevalence (88%), and

Table 3.2: Case study II: Sensitivity and specificity estimates for 5 dentists from Finite Mixture model as published in Albert and Dodd (2004). Prevalence of caries was estimated to be .17.

	Sensitivity	Specificity
Dentist 1	.45	.97
Dentist 2	.74	.88
Dentist 3	.66	.98
Dentist 4	.51	.96
Dentist 5	.92	.67

positive outcome probabilities are $< .05$ for each of the ten symptom groups. Class 2 is labeled as mild depression with a prevalence of 9%, and the probabilities of a positive test outcome that are between .08 and .49. Class 3 corresponds to severe depression (3%), with highest positive outcome probabilities, ranging from .37 to .91.

Case study II: Dentistry data, Albert and Dodd [6]

Albert and Dodd re-analyzed Handelmans dentistry data [59, 82]. Data are available from 5 dentists classifying X-ray images on $N = 3869$ teeth as sound or carious. Fitting the 2-class LI model showed poor goodness-of-fit, $G^2 = 129.85, df = 20, p < .001$.

To relax the local independence assumption, Albert and Dodd [6] proposed a Finite mixture (FM) model that can be conceptualized as a 4-class LI latent class model with fixed value constraints on the $\pi_{x_j|d}$ parameters in two of the classes. We can write, $d = 0, ..., 3, \pi_{x_j|d=2} = 1 - \pi_{x_j|d=3} = 1, j = 1, ..., J$. Classes $d = 2, 3$ represent cases that are clinically obvious: teeth that are unambiguously sound ($d = 2$) or unambiguously carious ($d = 3$) are assumed to have zero probability for being mistaken for sound or carious respectively. Assume that $d = 0$ correspond to cases that are truly sound and $d = 1$ denotes teeth that are carious, however, are not cases that are clinically obvious and are thus subject to error.

Table 3.2 lists the estimated sensitivity and specificity for each dentist. The estimated prevalences for the four classes are: $\pi_{d=0} = .56, \pi_{d=1} = .16, \pi_{d=2} = .27$ and $\pi_{d=3} = .01$. Hence, the estimated caries prevalence is 17%.

Case study III: Chlamydia trachomatis, Dendukuri et al. [52]

Dendukuri et al. [52] developed a Bayesian multiple latent variable model based on the notion that diagnostic tests developed to detect similar biological mechanisms measure a

Table 3.3: Case study III: Sensitivity and specificity estimates with respect to *Chlamydia Trachomatis* DNA (DNA) and CT as published in Dendukuri et al. 2009. Prevalence of *Chlamydia Trachomatis* and DNA positive .099, *Chlamydia Trachomatis* negative and DNA positive .023, *Chlamydia Trachomatis* and DNA negative .878.

	Sensitivity	Specificity	DNA = 1	DNA = 0
LCR	.88	.97	.88	.99
PCR	.83	.97	.83	.99
DNAP	.81	.99		
CULT	.96	.99		

latent variable that is not the target disease status. In an evaluation of the accuracy of 4 diagnostic tests for *Chlamydia Trachomatis*, 2 of these tests: a ligase chain reaction (LCR) and a polymerase chain reaction (PCR) test are designed to measure the presences of *Chlamydia Trachomatis* DNA. For these data ($N = 3551$), the authors developed a latent class model in which the LCR and PCR tests measure the latent variable DNA which is related to the target disease. The other tests, a DNA probe test (DNAP) and culture (CULT) are direct and imperfect measures of the target disease status, which are less sensitive to non-viable *Chlamydia Trachomatis* DNA. Further, it is assumed that in the absence of *Chlamydia Trachomatis* DNA there is a zero probability that the target disease, *Chlamydia Trachomatis*, is present. As only 4 diagnostic tests are available, these parameter constraints are necessary for the remaining parameters to be identifiable. When such strict constraints are not plausible, the authors suggested the use of a Bayesian latent class analysis using informative prior distributions on some of the parameters on which reliable prior information is available.

Table 3.3 presents the results as published by Dendukuri and colleagues. The LCR and PCR tests, both the estimated sensitivity and specificity with respect to the disease and to *Chlamydia Trachomatis* DNA are given. The estimated specificities with respect to the target disease of the tests are high $\geq .97$. The sensitivities range from .81 for DNAP to .96 for culture.

Design simulation study

Data are generated under the models that were discussed in the 3 case studies using the parameter values reported in the original publications, resulting in 3 distinct simulation scenarios (I-III). To study the role of sample size, we evaluate a sequence of equally spaced sample size conditions ranging from $N = 200$ to $N = 2000$, with incremental steps of 200. For every step, 2000 simulation samples of size N are drawn. On each sample, the data-generating model is estimated to evaluate Type-I error rates of the G^2 and X^2 tests and the 2-class LI

model is estimated to approximate the power of these tests. We also perform a parametric bootstrap with 1000 bootstrap samples on each simulation sample to obtain a bootstrapped p-value for the G^2 and X^2 statistics.

To evaluate the relevance of detecting misfit of the 2-class LI model in our scenarios, we first study the bias and root mean square error in the estimators of sensitivity and specificity of the tests under evaluation, and the target disease prevalence. Bias is defined as the difference between the mean of the estimates under the 2-class LI model over simulation samples and its true value (Tables 3.1 to 3.3). To evaluate bias in scenario I, we let class $d = 2$ (mild depression) and class $d = 3$ (severe depression) merge into one category consisting of true depression positives, such that the sensitivity of test can be computed by $(\pi_{x_j|d=2}\pi_{d=2} + \pi_{x_j|d=3}\pi_{d=3})/(\pi_{d=2} + \pi_{d=3})$. Root mean square error, which reflects both bias and variability, is defined by the square root of the sum of the bias squared and the empirical standard deviation of the estimates over simulation samples squared [36].

Simulations are performed in the software Latent GOLD 5.0 [170]. The MLVM (scenario III) that was originally proposed in a Bayesian framework is estimated by maximum likelihood using the standard EM routine in Latent GOLD.

Results simulation study

Bias and root mean square error

Figure 3.1 depicts the bias in target disease prevalences, sensitivities and specificities due to violation of the assumptions of the 2-class LI model. In scenario I (upper panel Figure 3.1) where the data generating model is the 3-class LI depression model (Table 3.1), the 2-class LI model overestimates the sensitivity of each of these symptoms by a substantial amount. Bias in the estimators of sensitivity ranges from $+.06$ for symptom Guilt/sin (s5) to $+.16$ for symptom Loss of interest (s9). Bias in specificity (range between symptoms: $[-.01, -.002]$) and depression prevalence $(-.04)$ is of a smaller magnitude and in the opposite direction. Root mean square error (Table 3.4) in the prevalence and sensitivity parameters is substantially increased under the 2-class LI model as compared to the 3-class LI depression model.

The middle panel of Figure 3.1 shows the estimated bias in scenario II. Sensitivities of the dentists were underestimated, with bias ranging from close to zero $(-.001)$ to $-.06$ for dentist 3. Bias in prevalence $(+.02)$ and specificity (range: $[+.005, +.02]$) are in opposite direction. Increased root mean squared error under the 2-class LI model as compared to the true FM

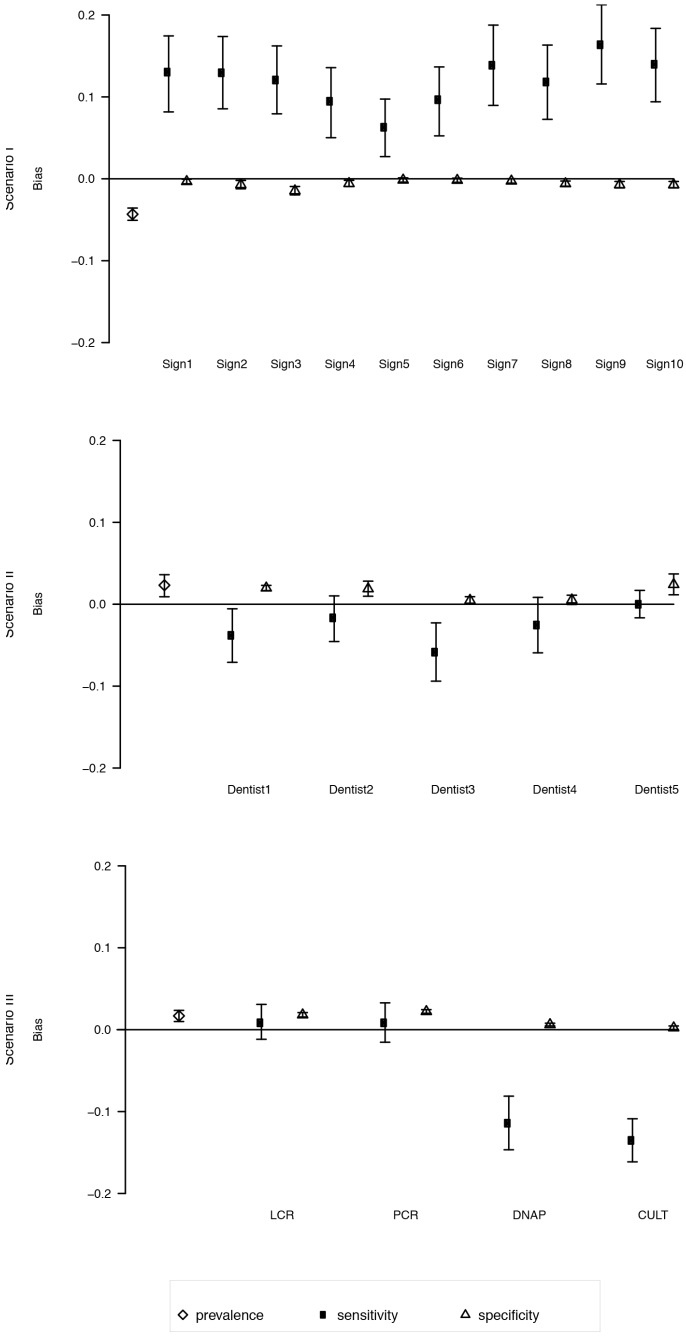

Figure 3.1: Absolute bias and interquartile range of sensitivity, specificity and prevalence in scenarios I-III. Based on 2000 simulation realizations, evaluated at a sample size of 1000.

Table 3.4: Simulation results based on 2000 simulation realizations, evaluated at a sample size of 1000. Root mean square error for Scenarios I, II and III under the data generating model (True model) and 2-class local independence model (2-class LI model).

Parameter	Model	Scenario I	Scenario II	Scenario III
Prevalence	True model	0.0226	0.0241	0.0094
	2-class LI model	0.0446	0.0298	0.0196
Sensitivity	True model	0.0594	0.0513	0.0342
	2-class LI model	0.1366	0.0518	0.0833
Specificity	True model	0.0053	0.0119	0.0058
	2-class LI model	0.0079	0.0142	0.0128

model is most pronounced in the prevalence and specificity parameters (Table 3.4).

In scenario III the data are generated under the multiple latent variable model for *Chlamydia Trachomatis*. For the indirect measures of *Chlamydia Trachomatis* (lower panel of Figure 3.1), the LCR and PCR tests have bias that is close to zero (bias: $< +.01$); the sensitivity of the direct measures of CT, DNAP test $(-.11)$ and culture $(-.14)$, are substantially underestimated under the 2-class LI model. Prevalence of *Chlamydia Trachomatis* is estimated with bias close to zero. Bias in specificity is close to zero $(< +.01)$. Under the 2-class LI model, root mean squared error is substantially increased in prevalence, sensitivity and specificity parameters as compared to the true multiple latent variable model.

Type-I error rates

The left panel of Figure 3.2 illustrates the finite sample behavior of the considered test statistics under the data-generating model. Type-I error rates are evaluated at typical nominal alpha = 5%. Thus, the number of rejections of the test statistic under the data-generating model should approach 5%.

In scenario I, the (asymptotic) G^2 statistic never exceeds the critical value to reach significance (i.e., in none of the simulation samples the p–value was lower than $.05$). The X^2 test has Type-I error rates close to the nominal alpha level only for the smallest sample size $(N = 200)$, while reaching significance in approximately $20 - 25\%$ of the simulation samples for the larger sample sizes considered $(800 \leq N \leq 2000)$. The X^2 test based on the bootstrap reaches close-to nominal level for all sample sizes considered, while the G^2 test based on the bootstrap performs similar to the asymptotic G^2 test, substantially below nominal alpha.

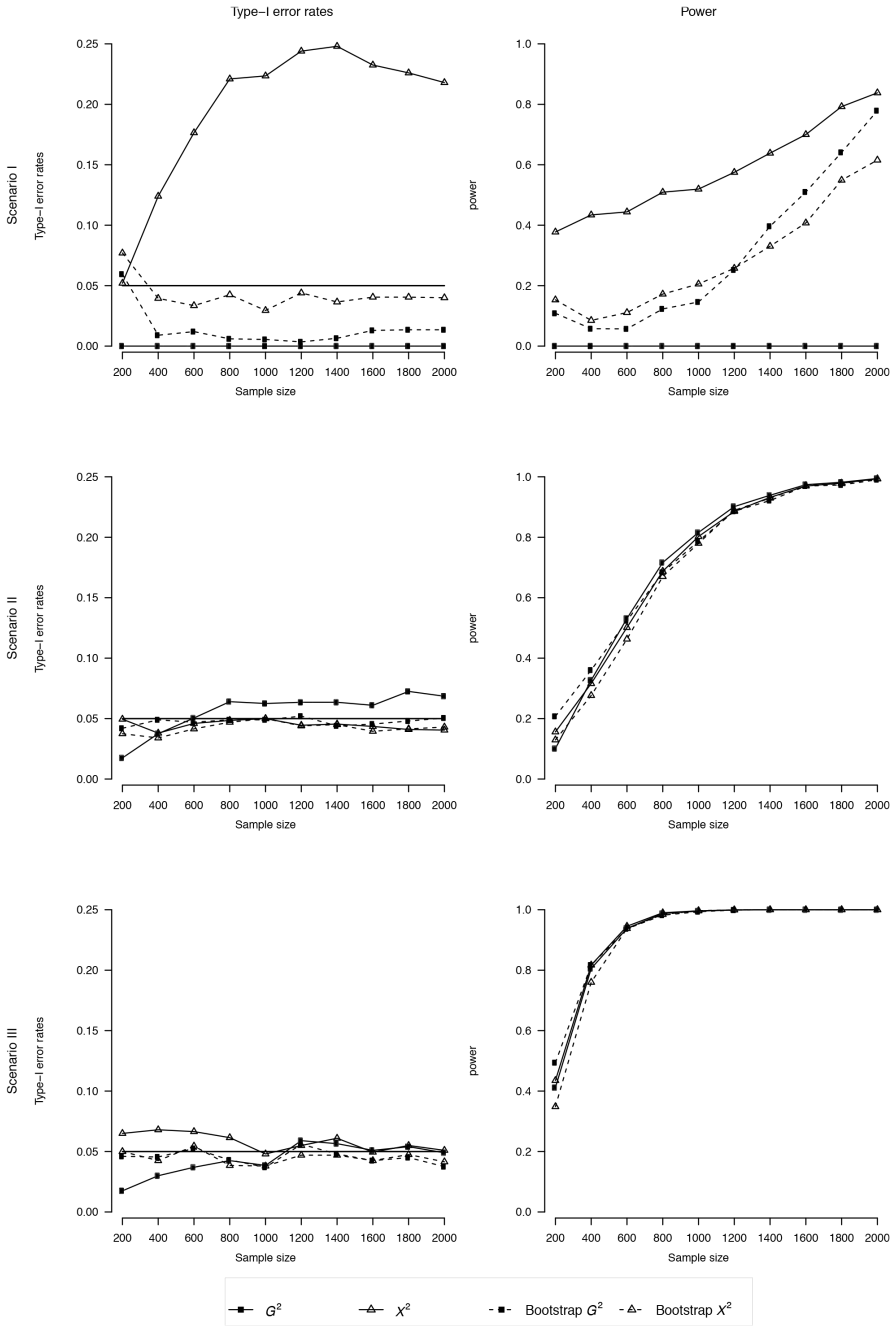

Figure 3.2: Type-I error rates (left panels) and Power (right panels) at $\alpha = .05$ in scenarios I (upper panel) to III (lower panel) for asymptotic and bootstrapped G^2 and X^2 tests.

Type-I error rates in scenario II are satisfactory for the asymptotic X^2 and bootstrapped X^2 and G^2 tests for the whole range of sample sizes considered. At $N = 200$, the asymptotic G^2 reaches lower than nominal alpha (1.8%).

In scenario III, the bootstrapped G^2 and X^2 tests show satisfying Type-I error rates at all sample sizes considered. Figure 3.2 suggests, however, that the asymptotic tests show some deviations at the smaller sample size, similar to scenario II, at $N = 200$ a Type-I error rate for the asymptotic G2 test is 1.7%. The Type-I error rates for the asymptotic test for sample sizes where $N > 200$ are satisfactory.

Power

Next, we evaluate the proportion of samples in which the misspecified model was rejected. Some caution should be exercised with interpreting the simulation results with respect to statistical power, which are summarized in the right panel of Figure 3.2. Clearly, not all tests have satisfactory Type-I error rates. We will focus on those results where tests have shown to have Type-I error rates close to nominal alpha (5%).

In scenario I, Type-I error rates are only satisfactory over the whole range of sample sizes considered ($200 \leq N \leq 2000$) for the bootstrapped X^2 test. The power of this test never exceeds 62%. Incidentally, at $N = 200$ both the asymptotic X^2 and bootstrapped G^2 have close to nominal alpha, reaching 38% and 11% power respectively.

To achieve a power of at least 80% in scenario II a sample size of approximately 1000 is necessary. $> 90\%$ power is achieved for $N > 1200$ for all asymptotic and bootstrapped tests. For sample sizes $N < 1000$ we observe unsatisfactory power. At the smallest sample size $N = 200$, the estimated power ranges from 10% for the asymptotic G^2 test (also reaches lower than nominal alpha) to 21% for the bootstrapped G^2 test.

In scenario III, the four significance tests perform equally well with respect to their achieved power at sample sizes of 600 and larger, reaching $> 90\%$ significance. At the smallest sample size considered, $N = 200$ power ranges from 35% (Bootstrap X^2 test) to 49% (Bootstrap G^2 test).

Discussion

We have illustrated the shortcomings of goodness-of-fit testing to evaluating the fit of the standard 2-class local independence (LI) latent class model. This approach is commonly found in the diagnostic studies that lack an accurate reference standard to justify the criticized local independence assumption [168]. We have evaluated three distinct simulation scenarios based on real-data examples. When local independence is falsely assumed in these scenarios, estimators of test sensitivity and specificity and disease prevalence are biased and mean square error is increased. Despite what could be considered important violations of the local independence assumption, power of the goodness-of-fit tests was unsatisfactory in at least two (scenarios I and II) out of three scenarios at sample sizes typically found in practice. This deficiency in power to detect conditional dependence when present is in accordance with earlier research [6, 157]. Our results also suggest that a substantial difference between the X^2 and G^2 tests can be found in practice.

The Type-I error rates were unsatisfactory in one scenario (scenario I) with data on 10 (dichotomous) diagnostic tests. The high number of possible diagnostic test outcome patterns ($2^{10} = 1024$) in conjunction with test outcome pattern probabilities that are near zero contributed to sparsely filled cells and low corresponding expected test outcome pattern frequencies. Under these circumstances the G^2 and X^2 tests do not approximate their limiting distribution, which explains the poor Type-I error performance of the asymptotic G^2 and X^2 tests [2].

Goodness-of-fit testing based on parametric bootstrap yielded better finite sample behavior in sparse data, with the bootstrap X^2 test yielding on average better results than the bootstrap G^2 test. In non-sparse data the asymptotic tests and testing based on the parametric bootstrap procedure performed equally well. Alternative measures to distinguish between latent class models have not been investigated. Instead of using global fit tests, an analysis of residuals may be more helpful to pinpoint causes of local dependence in practice. Graphical [140] and numerical [130, 167] techniques have been described. Future research may investigate the utility of these measures.

For each of the simulated scenarios, a higher number of subjects was needed to reach acceptable power levels than are often found in practice. In contrast to the three case studies we have presented, with sample sizes ranging from $N = 1355$ to $N = 3869$, the majority of diagnostic studies that use goodness-of-fit testing to find support for the local independence assumption may have lacked power to detect misfit under the standard 2-class latent class

model due to insufficient sample sizes. Repeating the simulations in the three scenarios at the sample sizes of original publication, we find that the power of the preferred bootstrap X^2 test is: .3325 (scenario I at N = 1322), 1.000 (scenario II at N = 3869) and 1.000 (scenario III at N = 3551).

We conclude that goodness-of-fit tests lack power to detect relevant misfit of the standard 2-class latent class model at sample sizes that are typically found in empirical diagnostic studies. Furthermore, the goodness-of-fit tests that are based on asymptotic theory are not robust to the sparseness of data. While a parametric bootstrap procedure improves the evaluation of the fit of the latent class model in the case of sparse data, violations of the local independence assumption are likely to be missed in studies as power is low at typical sample sizes. Our study re-emphasizes the relevance of obtaining an adequate sample size in diagnostic studies of diseases that lack a gold reference standard when using latent class analysis.

Chapter 4

Diagnostic test accuracy in childhood pulmonary tuberculosis: a Bayesian latent class analysis

Abstract

The evaluation of tests for the diagnosis of childhood pulmonary tuberculosis (CPTB) is complicated by the absence of an accurate reference test. We present a Bayesian latent class analysis evaluating the accuracy of five diagnostic tests for CPTB in the absence of a gold standard. We used data from a study of 749 hospitalized South African children suspected of CPTB, where test results were available from culture, smear microscopy, Xpert MTB/RIF (Xpert), tuberculin skin test (TST) and chest radiography. Using the latent class model, we estimated the prevalence of CPTB (with 95% Credible intervals) to be 27% (21%, 35%). Sensitivity of culture, Xpert and smear microscopy were estimated respectively at 60% (46%, 76%), 49% (38%, 62%) and 22% (16%, 30%); specificities of these tests were estimated in accordance with prior information, close to 100%. Chest radiography was estimated to have sensitivity of 64% (55%, 73%) and specificity of 78% (73%, 83%). Sensitivity of TST was estimated at 75% (61%, 84%), decreasing substantially among children who were malnourished and HIV-infected (56%); specificity of TST was 69% (63%, 76%). Using the latent class model it was estimated that 46% (42%; 49%) of true CPTB negative cases received anti-TB treatment, indicating substantial overtreatment.

Tuberculosis in children is an important global health problem with an estimated 0.5 to 1 million new cases each year [97, 178] amongst which childhood pulmonary tuberculosis (CPTB) is the most common form of pediatric TB. One of the major challenges is the lack of sensitive tests for diagnosing CPTB [46,73,158,183]. In clinical practice, the diagnosis of CPTB therefore relies on a combination of imperfect tests, which gives rise to unknown degrees of under- or overtreatment [126, 137].

In recent years, new tests for CPTB have been developed whose accuracy have been evaluated using mycobacterial culture as a reference standard [128, 183]. While culture is currently considered the best available reference standard, its sensitivity for detecting CPTB is acknowledged to be imperfect [56, 158, 183]. The culture reference standard thus inevitably leads to true CPTB cases to be misclassified as negative for CPTB. If these misclassifications by the reference standard are ignored, then the assessment of the test accuracy can be biased [112, 143, 145, 176].

To address the problem of lack of an accurate reference standard, multivariable diagnostic algorithms for CPTB have been proposed to combine information from multiple imperfect diagnostic tests (including tests for TB infection and clinical data) in a systematic manner. While more than a dozen of these algorithms have been described to date, estimates of CPTB prevalence derived from them vary widely [86, 90]. None of these algorithms have relied on statistical modeling approaches that take into account the imperfect nature and relative weight of each of the diagnostic tests.

In this study, we re-analyzed the results of a study on hospitalized children suspected of CPTB for which data had been prospectively collected on commonly used tests for CPTB [128]. The tests include three microbiological tests, the tuberculosis skin test (TST), and chest radiography. We use Bayesian latent class analysis to simultaneously estimate the accuracy of the five tests to detect CPTB, prevalence of CPTB and the degree of under- and overtreatment in the cohort. Latent class analysis has successfully been used in other diagnostic test accuracy studies in the absence of a gold standard [41, 145, 161, 168]. However, we present here one of the first applications of latent class analysis on prospectively collected data on CPTB.

Methods

Data were obtained from a study of hospitalized South African children suspected of CPTB [128]. Details on the design of the study are available from the original publications [128,185].

Table 4.1: Characteristics of 749 children suspected of pulmonary Tuberculosis (CPTB), South Africa (2009 to 2014).

No. (%)	749 (100)
Female sex - no. (%)	347 (46)
Age - months	
Median (IQR)	22 (12, 50)
Range	1, 120
HIV infection - no. (%)	154 (21)
Weight	
KG median (IQR)	10 (8, 14)
Z-score[a] median (IQR)	-1.1 (-2.2, 0.2)
Malnutrition[b]- no. (%)	211 (28)
Diagnostic test positive	
Culture - no. (%)	122 (16)
Xpert - no. (%)	106 (14)
Microscopy - no. (%)	42 (6)
Radiography - no. (%)	249 (33)
TST - no. (%)	321 (43)
Household TB contact - no. (%)	409 (55)
Treated for PTB - no. (%)	436 (58)

[a] Weight for age Z-score, calculated according to World Health Organization Child Growth Standards

[b] Malnutrition was defined by weight for age Z-score lower than 2

Briefly, children were consecutively enrolled when presenting to a hospital in Cape Town, South Africa, between February 2009 and June 2014 with signs or symptoms suggestive of PTB. Inclusion criteria were: (i) cough and at least one additional factor suggestive of CPTB [128, 185]; (ii) age under 15 years (iii) able to obtain informed consent from a parent or legal guardian. Children were excluded if: (i) they had received TB treatment or prophylaxis for more than 72 hours; (ii) their place of residence precluded follow up. Patient characteristics are shown in Table 4.1. In total, 749 children were included in our analysis.

Written informed consent for enrollment in the study was obtained from a parent or legal guardian. The Research Ethics Committee of the Faculty of Health Sciences, University of Cape Town approved the study. Renewed approval for the current analysis was not required, anonymized data were used.

Study procedures

Up to three induced sputum samples per child were each tested with three different microbiological tests: liquid culture (mycobacterial growth indicator, BACTEC MGIT, Becton

Table 4.2: Diagnostic tests for Childhood Pulmonary Tuberculosis (CPTB) and expected accuracy of these tests for diagnosis of CPTB.

Test		Expected accuracy
Microbiological[a]	Sensitivity	Not perfect, concentration of *Mycobacterium tuberculosis* is usually very low in sputum of children: true CPTB cases can be missed
	Specificity	Near perfect for Culture and Xpert, microscopy may give false positive results with a recent history of prior TB treatment or with BCG disease
Radiography	Sensitivity	Not perfect, limited pathology may not be visible; subject to inter- and intra-observer variability
	Specificity	Not perfect, positive reading may be caused by other respiratory diseases and past PTB infection, subject to inter- and intra-observer variability
TST	Sensitivity	Not perfect, due to anergy (HIV or malnutrition) or other reasons related to limited immune response; in adults decreased in cases with severe disease
	Specificity	Not perfect, false positives expected due to latent TB infection, Bacillus Calmette-Gurin immunization and nontuberculous Mycobacterial infections

[a] Culture, Xpert and microscopy

Dickinson Microbiology Systems, Cockeysville, MD, USA), hereafter referred to as culture, a molecular nucleic acid amplification test (Xpert MTB/RIF, Cepheid Inc, USA), hereafter referred to as Xpert, and sputum smear microscopy. A tuberculin skin test (TST) was administered and read according to standard procedures by measuring transverse induration in response to purified protein derivative (PPD; 2TU, PPD RT23, Staten Serum Institute, Denmark, Copenhagen). Radiographs of the chest were judged as consistent with CPTB or not consistent with CPTB by two independent readers, blinded to all other investigations and based on a standardized reporting format. Disagreement between readers was resolved by a third reader. The studied tests are complimentary; none of these tests in isolation is expected to yield perfect diagnostic accuracy (Table 4.2).

HIV testing (HIV rapid test in all children, followed by a confirmatory PCR for children younger than 18 months or HIV ELISA for children aged 18 months or older) was done in all children with unknown HIV infection status. Weight of the child was transformed to a standardized score (Z-score) as a measure of malnutrition according to World Health Organization Child Growth Standards [50]. Parents provided information about the child's date of birth and any household contact treated for TB in the last three months. Anti-TB treatment decisions were at the discretion of the treating doctor based on all available routinely collected information.

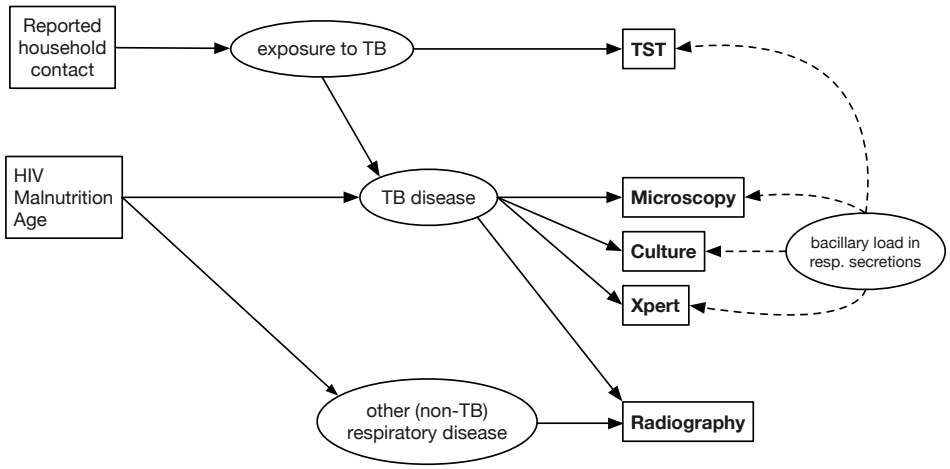

Figure 4.1: Heuristic model specifying prior beliefs of relations between latent and manifest variables.

CPTB model

Prior to undertaking our statistical analyses, we defined a heuristic CPTB model representing our assumptions about the pathophysiology of CPTB and the biological mechanisms that are believed to have given rise to the test results. This model is graphically depicted by Figure 4.1.

The mechanisms of the three microbiological tests under study (culture, Xpert and microscopy) are based on highly similar biological principles, via directly visualizing TB bacilli (microscopy), detecting their growth (culture) or amplifying and detecting bacterial DNA (Xpert). Among children with CPTB, we anticipate a positive relation between bacillary burden and the probability of a positive test outcome. This is because children with a higher bacillary burden are more easily detected by all three microbiological tests, while those with a very low bacillary burden are more likely to be missed by all tests.

We also anticipated conditional dependence between the TST results and results from the microbiological tests. For the case of adult TB it has been reported that TST may be less sensitive in severe disease, which can in turn be associated with high bacillary burden [14, 92, 93]. Since little is known about the exact functional form of this relationship in children, we allowed for the possibility of this association to be non-linear.

Further, based on the literature and clinical expertise of our team members, we expect that certain covariates influence sensitivity and specificity of the different tests and CPTB

Table 4.3: Covariates potentially associated with test accuracy and disease prevalence parameters.

Test	Parameter	HIV	Age	Malnu-trition	Household TB contact	Bacillary load[b]
	PTB prev[a]	•	•	•	•	
Culture	Sensitivity	○	○			•
	Specificity					
Xpert	Sensitivity	○	○			•
	Specificity					
Microscopy	Sensitivity	○	○			•
	Specificity					
Radiography	Sensitivity	○	○	○		
	Specificity	○				
TST	Sensitivity	•		•		○
	Specificity					

[a] PTB prevalence
[b] Bacillary burden is represented by a random effect
• clinical experts indicated strong belief that the covariate should be included in the model
○ clinical experts indicated there are reason to believe that covariate can exhibit an effect on the parameter
empty cells: clinical experts indicated no reason to believe that covariate should be included

prevalence. For example, we expect that the sensitivity of TST is systematically lower for HIV-infected children as compared to HIV-uninfected children [137, 158]. We also expect that some covariates influence CPTB prevalence. Table 4.3 lists the covariates and associations of interest. A distinction is made between those associations that are well established and those that remain to be studied.

Latent class models

Based on the heuristic CPTB model, we determined that the available tests would allow us to classify subjects into one of two latent classes representing CPTB true positive and CPTB true negative subjects. It should be noted that each class is in fact a mixture of subjects, e.g. the CPTB negative subjects will include those who have different respiratory diseases, and the CPTB positive subjects may also include children with co-infections. We assume conditional independence of test outcomes for CPTB negative subjects.

Our analyses proceeded stepwise through four different latent class models of increasing complexity. Our purpose was to study improvement in model fit and changes in parameter estimates as we proceeded to the model that most closely represented the model illustrated in Figure 4.1. Latent class model 1 (M1) is a 2-class latent class model based on the assumption of conditional independence of test outcomes within both classes. Latent class model 2 (M2)

adds to M1 a continuous random effect representing the unobserved true bacillary burden and its association with the microbiological tests. Latent class model 3 (M3) adds 9 established covariate effects on the sensitivity, specificity and/or class prevalence (i.e. CPTB prevalence) to M2. It also includes the association between the random effect and the sensitivity of TST. We consider M3 to be our main model.

We considered one more elaborate latent class model (M4) which includes the 10 additional covariate effects whose associations with test sensitivity, specificity and prevalence are not well established but are of potential interest (Table 4.3). As explained further in the Results section we considered M4 to be an exploratory analysis, given the large number of covariate effect parameters considered (19 in total) relative to the available data. Details about the specification of these four latent class models are found in Appendix A.

Similar to earlier studies [80, 140], we assume that the random effect is a Gaussian random variable whose effect on the sensitivity of each of the three microbiological tests is equal. The covariates and the random effect influence the sensitivities and specificities of individual tests through a probit model (see Appendix A for details). In order to model the possibly non-linear relation between TST sensitivity and bacillary burden, we used a quadratic function; all other associations with covariates and random effects are assumed to be linear and additive. The latent class prevalence parameter is allowed to vary with covariates (household contact, age, HIV and malnutrition) through a linear binary logistic function. To simplify modeling, age of the child was dichotomized at 24 months of age separating very young and young children.

The probability of CPTB for each combination of tests was estimated. The model was also used to estimate the proportion of overtreatment (proportion of those receiving anti-TB treatment in the latent class representing truly CPTB disease negative subject) and undertreatment (proportion not receiving anti-TB treatment in the latent class CPTB disease positive) in the cohort.

Estimation and prior distributions

We fitted the latent class models to the data using a Bayesian approach. Appendix A provides details of the form of the likelihood and prior distributions. We used informative prior distributions only on the specificity parameters of the culture test and Xpert. The specificities of these tests are widely acknowledged to be near perfect [158, 183]. We selected hyperparameters that let the 95% prior credible interval of the specificity parameters for

the culture test and Xpert to range from 99-100% and 98-100%, respectively. For all other parameters we used non-informative prior distributions (Appendix A).

Using the statistical software package JAGS [139] called from R [141] for all models we ran three parallel Markov Chains Monte Carlo chains each with 50,000 iterations. The first 10,000 iterations of each chain were discarded. Convergence was assessed by visual inspection and by checks of the Brooks-Gelman-Rubin statistic [70]. No convergence problems were identified. To avoid label switching problems between MCMC chains [96], the parameters associated with random effect were constrained to positive values among the microbiological tests.

Estimation of exploratory model M4 with informative priors only on the sensitivity of Xpert and culture test (as in models M1-M3) yielded at least one covariate parameter estimated with extremely wide credible intervals. This suggests that with the current priors some parameters of model M4 are not identifiable or only weakly identifiable with these data. In a Bayesian context, defining additional informative priors may help overcoming this problem [75,76,101]. We therefore adopted a Bayesian LASSO approach [132] to estimate the 10 additional covariate associations modeled in M4 (Table 4.3). This is implemented by placing zero-centered Laplace prior distributions with a diffuse prior on the scale parameter (for details see Appendix A). The shrinkage is adaptive in the sense that it is proportional to the variance of the parameter estimate such that parameters that are estimated with poor precision are more likely to be shrunk to the null effect.

Sensitivity analyses

We conducted a series of sensitivity analyses to explore alternative modeling choices. Firstly, to consider the impact of an alternative conditional dependence structure, we fit a 3-class latent class model resulting from treating the random effect as a dichotomous rather than as a Gaussian variable. In this model, children in the CPTB class belong to one of two groups: CPTB disease with TB detectable in respiratory secretions or CPTB disease with TB not detectable in respiratory secretions. In the latter class, we assume that sensitivity of each of the three microbiological tests to be 0% and sensitivity of TST is assumed equal for the two true CPTB class. This model (M2b) is compared for differences in sensitivity, specificity and prevalence estimates with model M2.

Further, the informative prior distributions for the specificity parameters of the culture and Xpert tests were replaced by non-informative priors. To test the effect of relaxing

the assumption that the random effects equally affect the sensitivity of each of the three microbiological tests, we conducted an analysis where this constraint was removed.

Results

Model fit

Figure 4.2 shows the pairwise residual correlations [140] between the test outcomes for the four latent class models considered. For models M1 (the conditional independence model) and M2 (conditional dependence assumed only between microbiological tests), substantial residual correlation was found. In comparison, models M3 and M4 have low residual correlation. From Table 4.4 we can see that for M1 and M2, the expected frequencies of test outcome patterns substantially deviated from the observed frequencies. For model M3, expected frequencies of test outcome patterns are close to the observed frequencies, together with low residual correlation this suggests satisfactory fit of M3 to the data. Expected frequencies of test outcome patterns for exploratory model M4 were similar to those for M3 (not shown) despite the addition of 10 covariate effect parameters to the latent class model.

Estimates of disease prevalence and diagnostic test accuracy

Table 4.5 summarizes the estimates of test accuracy and CPTB prevalence based on Models M1 to M3. Estimates from model M3 are marginalized over the covariates and random effect. The corresponding estimates for model M4 were very similar to those from M3 and are not shown.

When ignoring the conditional dependence (i.e. model M1), sensitivity and specificity of the culture test were estimated close to 100 per cent; the estimates for the other tests and CPTB prevalence are therefore close to those obtained from a nave analysis that assumes culture is a perfectly accurate reference standard. Accounting for conditional dependence between the microbiological tests (M2) provided lower estimates of culture sensitivity. Adjusting for the conditional dependence between TST and the microbiological tests caused the estimate of TST sensitivity to increase.

From Table 4.5 we see that by model M3, prevalence of CPTB was estimated (posterior median) at 26.7% (95% Credible interval (CrI): 20.8%, 35.2%). The average sensitivities of the microbiological tests were: 60.0% (95% CrI: 45.8%, 75.5%), 49.4% (95% CrI: 37.7%,

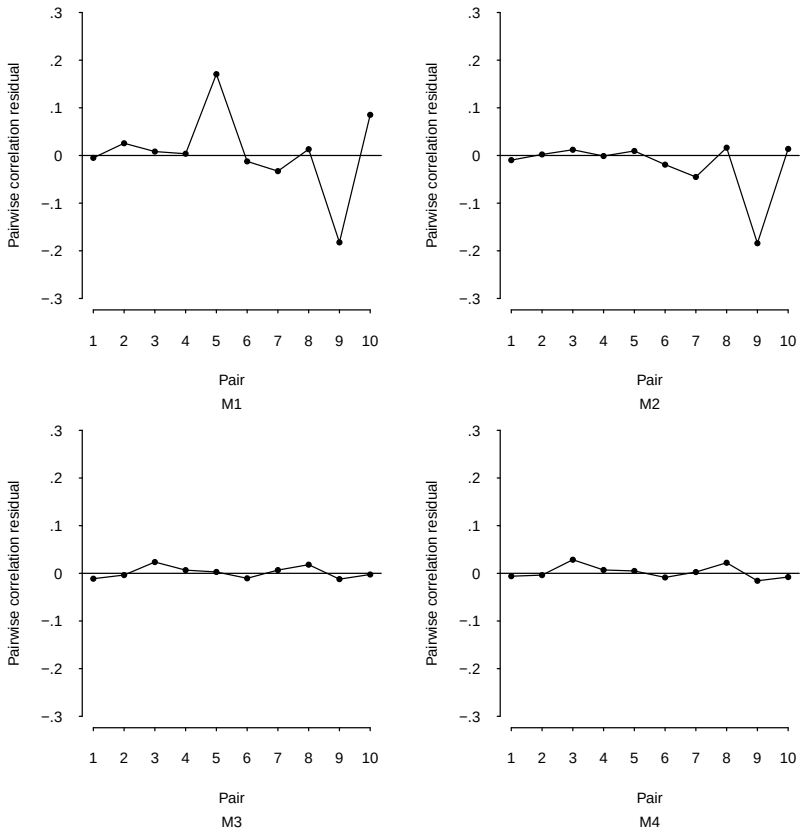

Figure 4.2: Residual correlation plots Model 1 to Model 4 (M1 to M4). Residual correlations are computed as the difference between observed and model-predicted correlation between each pair of tests (1: Culture-Xpert, 2: Culture-microscopy, 3: Culture-radiography, 4: Culture-TST, 5: Xpert-microscopy, 6: Xpert-radiography, 7: Xpert-TST, 8: Microscopy-radiography, 9: Microscopy-TST, 10: Radiography-TST).

62.2%) and 22.3% (95% CrI: 15.6%, 30.3%) for culture, Xpert and microscopy, respectively. These sensitivities strongly depend on the random effect which represents the unobserved bacillary burden in the sputum of the child (Figure 4.3, left panel). In accordance with our prior beliefs, the specificities of culture and Xpert were estimated near 100 per cent; the specificity of microscopy was estimated at 99.7% (95% CrI: 99.0, 100). The sensitivity of diagnosis by radiography was estimated at 64.2% (95% CrI: 54.9, 72.8) and specificity: 78.0% (95% CrI: 73.4%, 83.4%). For TST, overall sensitivity was estimated at 75.2% (95% CrI: 61.2%, 83.8%) and specificity 69.3% (95% CrI: 63.2%, 75.9%).

Table 4.4: Posterior median expected frequency of each combination of test results for models M1 to M3 and predicted probability of CPTB based on M3.

Test outcome pattern						Posterior median expected frequency			
Cu	Xp	Mi	Ra	TS	Frequency	M1	M2	M3	Probability of CPTB based on M3
0	0	0	0	0	296	278	292	294	2 (0; 7)
0	0	0	0	1	149	168	155	151	16 (5; 33)
0	0	0	1	0	87	102	90	89	9 (0; 34)
0	0	0	1	1	78	62	73	77	52 (26; 74)
0	0	1	0	1	1	0	0	0	11 (0;100)
0	1	0	0	0	5	5	4	4	4 (0; 40)
0	1	0	0	1	7	3	4	4	56 (0;100)
0	1	0	1	0	2	2	2	1	12 (0;100)
0	1	0	1	1	2	2	4	4	88 (50;100)
1	0	0	0	0	3	3	4	2	23 (0;100)
1	0	0	0	1	8	5	7	10	93 (62;100)
1	0	0	1	0	1	4	6	1	54 (0;100)
1	0	0	1	1	20	9	13	17	99 (90;100)
1	1	0	0	0	1	6	5	3	100 (100;100)
1	1	0	0	1	17	14	12	15	100 (100;100)
1	1	0	1	0	4	12	10	5	100 (100;100)
1	1	0	1	1	27	26	22	26	100 (100;100)
1	1	1	0	0	8	3	4	10	100 (100;100)
1	1	1	0	1	5	7	10	5	100 (100;100)
1	1	1	1	0	21	6	8	18	100 (100;100)
1	1	1	1	1	7	13	18	9	100 (100;100)

Cu = Culture, NA = Xpert, Mi = Microscopy, Ra = Radiography, TS = TST

Sensitivity analyses

Results of our sensitivity analyses are shown in Appendix B. Replacing the Gaussian random effects model (M2) between the microbiological tests by a 3-class latent class model (Model M2b, defined in Appendix A) did not affect the results substantially. Therefore, we concluded that our results seem robust to the choice of the conditional dependence structure and retained the Gaussian random effect in more complex models M3 and M4. Also, relaxing the equal random effects assumption (Model M3a) and replacing the informative priors on culture and Xpert specificity parameters by non-informative priors (Models M3b and M3c) has little effect on the model parameters.

Table 4.5: Posterior median estimates (95% Credible Interval) of marginalized sensitivity, specificity and CPTB prevalence (given in %) for models M1-M3.

Test	Parameters	M1	M2	M3
	Prevalence	16.6	28.7	26.7
		(15.6; 18.0)	(22.2; 36.3)	(20.8; 35.2)
Culture	Sensitivity	96.7	57.2	60.0
		(87.8; 99.8)	(44.8; 73.5)	(45.7; 75.5)
	Specificity	99.8	99.9	99.6
		(98.9;100.0)	(99.3;100.0)	(98.7;100.0)
Xpert	Sensitivity	74.4	46.7	49.4
		(66.0; 82.2)	(37.1; 59.1)	(37.7; 62.2)
	Specificity	98.3	98.9	98.6
		(97.0; 99.4)	(97.3; 99.9)	(97.3; 99.5)
Microscopy	Sensitivity	33.3	20.4	22.3
		(25.3; 42.1)	(14.6; 27.9)	(15.6; 30.3)
	Specificity	99.8	99.7	99.7
		(99.2;100.0)	(99.0;100.0)	(99.0;100.0)
Radiography	Sensitivity	65.4	64.7	64.2
		(56.5; 73.8)	(56.0; 73.0)	(54.9; 72.8)
	Specificity	73.1	79.4	78.0
		(69.6; 76.6)	(74.2; 84.9)	(73.4; 83.4)
TST	Sensitivity	69.0	69.3	75.2
		(60.5; 76.7)	(61.1; 76.8)	(61.2; 83.8)
	Specificity	62.4	67.8	69.3
		(58.5; 66.1)	(62.6; 73.4)	(63.2; 75.9)

Covariate effects

The estimated coefficients under model M3 show reduced TST sensitivity for HIV-infected children. A graphical presentation of the estimated covariate effect and random effect on TST sensitivity is found in the right panel of Figure 4.3. Sensitivity of TST dropped significantly at higher and low levels of the random effect, and for malnourished and HIV-infected children. Average sensitivity of TST under model M3 for children that are both malnourished and HIV-infected was estimated at 55.8% (95% CrI: 30.8%, 79.2%); for only HIV positive at 61.7% (95% CrI: 41.5%, 84.9%); only malnourished: 74.3% (95% CrI: 58.6%, 86.1%); for not malnourished and HIV-uninfected children: 80.9% (95% CrI: 69.0%, 89.5%).

The (marginal) covariate effects estimated by M4 are tabulated in Appendix C. Based on this exploratory analysis, only the effect age on sensitivity of radiography stands out. For children > 24 months, sensitivity of radiography was estimated at 52.5% (95% CrI: 39.4%, 66.0%), while

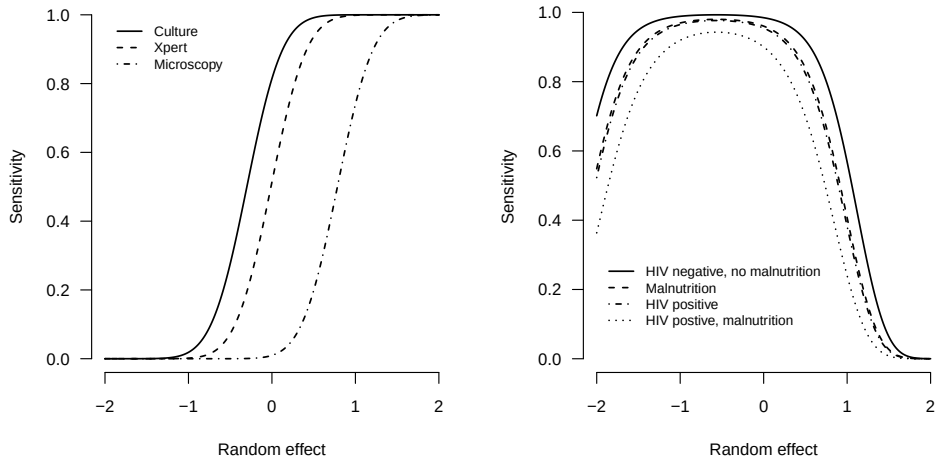

Figure 4.3: Estimated sensitivity as a function of the random effect (M3). Left panel: microbiological tests; right panel: TST.

for children < 24 months: 75.0% (95% CrI: 62.6%, 85.8%).

Posterior probability of CPTB

In addition to the estimates of prevalence and test accuracy, the Bayesian latent class model was used to estimate the posterior probability of CPTB for a given set of test outcomes. The estimated posterior probability of CPTB per test pattern under model M3 is given in Table 4.4. Not surprisingly, test patterns that included a positive culture generally had a predicted probability of CPTB of 1, with very high precision. When culture was negative, the highest posterior probability was obtained when Xpert, radiology and TST were positive. Two other patterns associated with a greater than 50% predicted probability were those where Xpert and radiology alone are positive, and where radiology and TST are positive. However, these estimates were accompanied by wide credible intervals, illustrating the difficulty in diagnosing individual culture negative children based on the 4 other tests we have considered.

Under- and overtreatment

Based on model M3, we evaluated potential overtreatment and undertreatment in the cohort. Details about these calculations are found in Appendix A. Subjects were classified into quintiles based on their posterior probability of CPTB. Within each quintile we estimated the mean probability of CPTB and the proportion receiving anti-TB treatment. The relation between

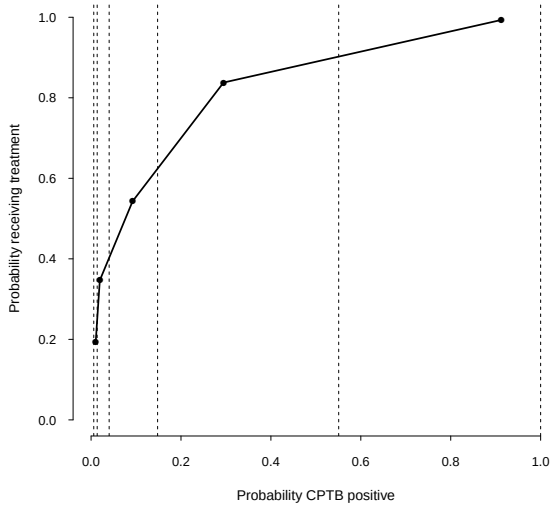

Figure 4.4: The probability of CPTB and the probability of treatment were estimated within quantiles of the posterior mean probabilities of CPTB for each child based on M3 (Dashed lines are boundaries of quantiles).

these two variables is depicted by Figure 4.4. The steep initial rise of the curve reflects the low treatment threshold applied by clinicians suggesting that the probability of receiving anti-TB treatment exceeds 80% even among subjects with probability of CPTB as low as 30%.

The proportion of CPTB positive children within the group of children that were receiving anti-TB treatment was estimated at 45.8% (95% CrI: 42.4%, 48.7%), reflecting the level of overtreatment; the proportion of children with CPTB within the group not receiving anti-TB treatment was estimated at 7.0% (95% CrI: 1.8%, 17.8%), reflecting the level of undertreatment. Additionally, the probability of not receiving treatment for CPTB negative children was estimated at 42.4% (95% CrI: 34.4%, 51.9%); conversely, the probability of receiving treatment for CPTB positive children was 95.5% (85.6%,99.0%), suggesting nearly all CPTB positive children did receive treatment.

Discussion

We presented a Bayesian latent class analysis in the context of CPTB. Using prospectively collected data from hospitalized children in South Africa suspected of CPTB, we estimated the accuracy of five commonly used diagnostic tests and provided estimates of under- and

overtreatment in this cohort. The predefined latent class models that incorporated conditional dependence between the three microbiological tests and TST showed good fit to the data. Through sensitivity analyses we showed our estimates of accuracy and CPTB prevalence are robust to changes to the prior distributions and the assumed dependence structure.

Our results are in agreement with pre-existing reports that the sensitivity of confirmatory tests for CPTB are low [158]. We estimated that a single mycobacterial culture test - generally regarded as the most sensitive confirmatory test for CPTB currently available and often the preferred reference standard - fails to detect almost 40 per cent of all true CPTB positive cases. The number of true CPTB positive cases missed by Xpert (missing about 50 per cent) and microscopy (missing about 77 per cent) are larger. Sensitivity of the TST and chest radiography based diagnosis is somewhat higher than the culture test, though the specificities of these tests are much lower.

We also found evidence that the sensitivity of the microbiological tests depends on the bacterial load in respiratory secretions. Our estimates of sensitivity may therefore not be generalizable to ambulatory settings as children with true CPTB presenting in outpatient clinics may be expected to be on average less severely diseased [184], hence, (on average) having lower bacillary burden in their respiratory secretions. Sensitivity of the microbiological tests in outpatient settings may thus be lower [184]. TST sensitivity is strongly dependent on the immune status and thus decreased in HIV-infected and malnourished children. Our results additionally indicated that TST sensitivity varied with the random effect, providing some evidence that - as in adult TB - TST sensitivity is reduced in more severe CPTB disease.

Due to the lack of an accurate diagnostic testing procedure for CPTB, doctors often make treatment decisions under great uncertainty. Reflecting this uncertainty, a definite CPTB diagnosis based on a clinical CPTB case definition as defined in the original study protocol could not be made for 48 per cent of CPTB suspected children in the current study. Taking into account this uncertainty of true CPTB status, we estimated using our latent class model that in our cohort the proportion receiving CPTB treatment for true CPTB negative cases is about 46 percent, while not receiving CPTB treatment among true CPTB positive was estimated at 7 percent (with a wide credible interval). This points to the possibility of a substantial amount of overtreatment and limited (low) undertreatment, reflecting the use a low implicit threshold probability for a decision to treat hospitalized children for CPTB by the treating doctors in a high HIV prevalence country. We stress that this low threshold and consequent high level of overtreatment was likely clinically appropriate in the study cohort due to the high uncertainty

of true CPTB, the high prevalence of TB in this geographical area and the high morbidity and mortality of untreated CPTB.

Diagnostic test evaluation in the absence of an accurate reference standard remains a challenging problem. In recognition of this, the US National Institutes of Health convened an expert panel to propose a uniform clinical case definition for PTB, which recently issued revised definitions [74]. The proposed PTB case definition leads to a classification into one of 3 different classes (confirmed TB, unconfirmed TB, unlikely TB) based on a set of clinical, radiological and microbiological criteria. While this case definition is clearly an important step forward, the middle category (unconfirmed TB) prohibits unambiguous evaluation of diagnostic tests in terms of diagnostic test accuracy and estimation of PTB prevalence and degrees of possible under- and overtreatment. In future work we intend to compare the 3 group classification obtained based on this expert guideline to that obtained using our model. We also intend to evaluate our model in children presenting with suspected PTB with less severe disease in an ambulatory setting.

Although our latent class analyses have been carefully designed, we acknowledge that our results depend on the assumptions we have made and that it may be difficult to appreciate the validity of such analyses. However, we have made our assumptions explicit in this paper, presented a variety of sensitivity analyses and we consider this preferable over making assumptions that are both left implicit and known to be untenable, e.g. assuming culture is a perfect test. In the absence of a gold standard test (i.e., a test with perfect accuracy) for CPTB, making essentially unverifiable assumptions is inevitable to quantify the accuracy of the diagnostic tests for CPTB.

Appendix A: Full specification of latent class models

Let $T_i\text{Cu}, ..., T_i\text{TS}$ denote the observed outcomes on the 5 binary tests (Cu = Culture, Xp = Xpert, Mi = Microscopy, Ra = Radiography, TS = TST) for subject $i, i = 1, ..., 749$. It is assumed that the true CPTB status of each subject is a Bernoulli latent variable, d_i, where $d_i = 0(1)$ denotes true absence (presence) of CPTB. The joint distribution of test observations is modeled by a two-class latent class model,

$$\mathbb{P}(T_i\text{Cu}, ..., T_i\text{TS}) = \sum_{d_i=0}^{1} \mathbb{P}(d_i)\mathbb{P}(T_i\text{Cu}, ..., T_i\text{TS}|d_i).$$

Conditional dependence latent class models

To relax the conditional independence assumption (M1) we evaluated three conditional dependence latent class models (M2 to M4). We let r denote the random effect representing bacillary load, $r \sim N(0, 1)$. Further, we let $Z_i\text{HhTB}$, $Z_i\text{HIV}$, $Z_i\text{Age}$ and $Z_i\text{Mal}$ denote subject i's observed values on the binary covariates: Household contact with a TB patient, HIV infection status, Age (dichotomized at 24 months) and Malnutrition, respectively.

For M4, the most elaborate latent class model considered here, the latent disease status d_i is regressed on the binary covariates via the logistic function, $\ln\{\mathbb{P}(d_i = 1)/\mathbb{P}(d_i = 0)\} = \alpha_0 + \alpha_1 Z_i\text{HhTB} + \alpha_2 Z_i\text{HIV} + \alpha_3 Z_i\text{Age} + \alpha_4 Z_i\text{Mal}$. For each test j, the marginal conditional probability distribution $\mathbb{P}(T_{ij}|d_i)$ is assumed independent Bernoulli with probability $\phi(\eta_{j|d_i})$ where ϕ denotes the cumulative distribution function (c.d.f) of a standard normal distribution $(N(0, 1))$. The specific forms of $\phi(\eta_{j|d_i})$ for M4 are detailed in Table 4.6.

M1, M2 and M3 are nested within model M4. M1 (conditional independence model) is the special case of M4, where all covariate (including the covariate effects on d_i) and random effects are assumed to be null. In M2, all the covariate effects are assumed to be null and TST outcomes do not depend on the random effect (i.e., $\sigma_4 = \sigma_5 = 0$). In M3, the covariate effects for culture, Xpert, microscopy and radiography are assumed to be null.

Prior distributions

For M1 the prior distributions were as follows,

$$\alpha_0, \beta_{10}, ..., \beta_{50}, \gamma_{30}, ..., \gamma_{50} \sim N(0, 1),$$
$$\gamma_{10} \sim N(3.3023, 0.126), \gamma_{10} \sim N(2.886, 0.126).$$

For M2 the same prior distributions as M1 are used, and we add random effect parameters with corresponding prior distributions,

$$\sigma_1 \sim \text{Uniform}(0,5), \sigma_1 = \sigma_2 = \sigma_3.$$

For M3 we additionally define,

$$\sigma_4 \sim \text{Uniform}(-5,5), \sigma_5 \sim \text{Uniform}(0,5),$$
$$\alpha_1, ..., \alpha_4, \beta_{41}, \beta_{42} \sim N(0,10).$$

Finally,for M4 we add,

$$\gamma_{41}, \beta_{11}, ..., \beta_{42} \sim \text{DoubleExponential}(0,\lambda), \lambda \sim U(0,10).$$

Overtreatment and undertreatment

We used M3 to gain insight in potential overtreatment and undertreatment in the cohort. Let anti-TB treatment received by subject i be denoted by δ_i, taking on the values 0 for no treatment received and 1 for treatment received. We estimate the probability of being a true CPTB case for each individual, $\hat{\mathbb{P}}(d_i = 1)$, by averaging d_i over MCMC samples. The proportion of CPTB positives within the strata of treatment status $\hat{\mathbb{P}}(d = 1|\delta = 0)$ and $\hat{\mathbb{P}}(d = 1|\delta = 1)$ are estimated by the average $\hat{\mathbb{P}}(d = 1)$ within each of the strata. By applying Bayes theorem, we can then estimate the probability of receiving treatment for CPTB positive children $\hat{\mathbb{P}}(d = 1|\delta = 0)$ and probability of not receiving treatment for CPTB positive children $\hat{\mathbb{P}}(d = 0|\delta = 1)$.

Sensitivity analyses

To evaluate the robustness of the random effects structure a 3-class condition independence latent class model (M2b) is developed that is defined by,

$$\mathbb{P}(T_i\text{Cu}, ..., T_i\text{TS}) = \sum_{l_i=0}^{2} \mathbb{P}(l_i) \prod_{j=1}^{J} \mathbb{P}(T_{ij}|l_i), \quad \text{where } \mathbb{P}(T_{ij}|l_i) = \pi_{jl}.$$

Class $l_i = 0$ represents truly CPTB negative children; $l_i = 1$ CPTB disease with TB detectable in respiratory secretions and $l_i = 2$ CPTB disease with TB not detectable in respiratory secretions. The assumptions are reflected in informative prior distributions for culture, Xpert

Table 4.6: Specification of marginal conditional probability distributions of M4.

Test	$\eta_{j\|d_i=0}$	$\eta_{j\|d_i=0}$
Culture	$\eta_{\text{Cu}\|d_i=0} = \gamma_{10}$	$\eta_{\text{Cu}\|d_i=1} = \beta_{10} + \beta_{11} Z_i\text{HIV} + \beta_{12} Z_i\text{Age} + \sigma_1 r_i$
Xpert	$\eta_{\text{Xp}\|d_i=0} = \gamma_{20}$	$\eta_{\text{Xp}\|d_i=1} = \beta_{20} + \beta_{21} Z_i\text{HIV} + \beta_{22} Z_i\text{Age} + \sigma_2 r_i$
Microscopy	$\eta_{\text{Mi}\|d_i=0} = \gamma_{30}$	$\eta_{\text{Mi}\|d_i=1} = \beta_{30} + \beta_{31} Z_i\text{HIV} + \beta_{32} Z_i\text{Age} + \sigma_3 r_i$
Radiography	$\eta_{\text{Ra}\|d_i=0} = \gamma_{40} + \gamma_{41} Z_i\text{HIV}$	$\eta_{\text{Ra}\|d_i=1} = \beta_{40} + \beta_{41} Z_i\text{HIV} + \beta_{42} Z_i\text{Mal}$
TST	$\eta_{\text{Ts}\|d_i=0} = \gamma_{50}$	$\eta_{\text{Ts}\|d_i=1} = \beta_5 0 + \beta_{51} Z_i\text{HIV} + \beta_{52} Z_i\text{Mal} + \sigma_4 r_i - \sigma_4^2 r_i$

and microscopy,

$$\pi_{\text{Cu}_1} \sim \text{Beta}(1, 400), \pi_{\text{Xp}_1} \sim \text{Beta}(1, 198), \pi_{\text{Mi}_1} \sim \text{Beta}(1, 1),$$
$$\pi_{\text{Cu}_2} = 0, \pi_{\text{Xp}_1} = 0, \pi_{\text{Mi}_1} = 0.$$

It is also assumed that for TST, $\pi_{\text{Ts}_1} = \pi_{\text{Ts}_2}$. The remaining π-parameters have Beta(1,1) prior distributions. The prior distribution for $\mathbb{P}(l_i)$ is Dirichlet(1,1,1).

Finally, we perform a sensitivity analysis for M3 by adjusting its prior distributions. M3b, $\sigma_2, \sigma_3 \sim \text{Uniform}(0, 5)$, M3c, $\gamma_{10} \sim N(0, 1)$, M3d, $\gamma_{20} \sim N(0, 1)$.

Appendix B: Sensitivity analysis for model M2 and M3. Posterior median estimates (95% Credible Interval) of marginalized sensitivity, specificity and PTB prevalence for M2b: 3-class latent class model (defined in appendix A), M3a (Random effect Culture, XPERT, Microscopy freely estimated), M3b (non-informative priors for Culture) and M3c (non-informative priors for Xpert).

Test	Parameters	M2	M2b	M3	M3a	M3b	M3c
	Prevalence	28.7	29.7	26.7	26.6	25.6	25.9
		(22.2; 36.3)	(20.2; 49.7)	(20.8; 35.2)	(20.2; 35.0)	(20.0; 33.8)	(20.6; 33.4)
Culture	Sensitivity	57.2	54.0	60.0	60.2	61.6	61.7
		(44.8; 73.5)	(32.0; 77.8)	(45.7; 75.5)	(46.1; 78.0)	(46.9; 77.5)	(48.4; 76.4)
	Specificity	99.9	99.8	99.6	99.6	98.9	99.6
		(99.3; 100.0)	(99.1; 100.0)	(98.7;100.0)	(98.6; 100.0)	(97.6; 99.7)	(98.7; 100.0)
Xpert	Sensitivity	46.7	42.2	49.4	49.5	51.2	50.0
		(37.1; 59.1)	(25.6; 61.1)	(37.7; 62.2)	(37.9; 64.0)	(39.1; 64.8)	(39.1; 62.4)
	Specificity	98.9	98.3	98.6	98.6	98.6	97.8
		(97.3; 99.9)	(96.9; 99.4)	(97.3; 99.5)	(97.2; 99.5)	(97.2; 99.5)	(96.2; 99.0)
Microscopy	Sensitivity	20.4	18.5	22.3	22.3	23.2	23.1
		(14.6; 27.9)	(10.5; 29.0)	(15.6; 30.3)	(15.5; 31.4)	(16.3; 31.6)	(16.5; 30.8)
	Specificity	99.7	99.7	99.7	99.7	99.7	99.7
		(99.0; 100.0)	(99.0; 100.0)	(99.0;100.0)	(99.0; 100.0)	(99.0; 100.0)	(99.0;100.0)
Radiography	Sensitivity	64.7	66.9	64.2	64.3	64.9	65.0
		(56.0; 73.0)	(47.3; 79.5)	(54.9; 72.8)	(55.2; 73.0)	(55.6; 73.5)	(56.0; 73.6)
	Specificity	79.4	80.8	78.0	78.0	77.7	77.9
		(74.2; 84.9)	(74.1; 88.3)	(73.4; 83.4)	(73.3; 83.5)	(73.1; 82.8)	(73.4; 82.9)
TST	Sensitivity	69.3	68.6	75.2	75.3	75.8	75.2
		(61.1; 76.8)	(59.9; 76.4)	(61.2; 83.8)	(61.7; 83.8)	(62.9; 83.8)	(61.2; 83.5)
	Specificity	67.8	68.1	69.3	69.2	68.8	68.8
		(62.6; 73.4)	(62.2; 81.6)	(63.2; 75.9)	(63.1; 75.8)	(63.1; 75.3)	(62.8; 75.1)

Appendix C: Sensitivity and specificity estimates within sub-groups defined by covariates estimated by M4.

Test	Parameter	Sub-group	Posterior Median (95% Credible Interval)	Sub-group	Posterior Median (95% Credible Interval)
Culture	Sensitivity	HIV positive	64.0 (46.9,96.3)	HIV negative	57.9 (44.1,72.5)
	Sensitivity	Age > 24 months	56.2 (40.0,72.4)	Age ≤ 24 months	61.3 (46.5,76.0)
Xpert	Sensitivity	HIV positive	49.9 (36.1,75.9)	HIV negative	48.4 (36.8,60.9)
	Sensitivity	Age > 24 months	47.3 (33.6,60.8)	Age ≤24 months	49.9 (37.8,62.7)
Microscopy	Sensitivity	HIV positive	24.8 (16.1,47.9)	HIV negative	21.2 (14.1,29.6)
	Sensitivity	Age > 24 months	21.4 (13.7,30.3)	Age ≤ 24 months	22.3 (15.1,31.0)
Radiography	Sensitivity	HIV positive	73.4 (59.0,87.7)	HIV negative	63.6 (54.6,72.0)
	Sensitivity	Age > 24 months	52.5 (39.4,66.0)	Age ≤ 24 months	75.0 (62.6,85.8)
	Sensitivity	Malnourished	64.6 (53.4,75.9)	Not malnourished	65.4 (55.5,74.1)
	Specificity	HIV positive	77.7 (70.1,85.0)	HIV negative	78.3 (73.3,84.3)
TST	Sensitivity	HIV positive	66.9 (39.2,84.0)	HIV negative	78.5 (65.8,86.6)
	Sensitivity	Malnourished	71.8 (53.2,83.7)	Not malnourished	78.9 (65.1,87.1)

Chapter 5

Revisiting the cautionary note on conditional dependence latent class models

Abstract

In this chapter we revisit the evidence presented in the influential paper by Albert and Dodd (2004) that cautioned against the use of latent class analysis in the absence of a gold standard. We identify problems with the evidence from simulations that have important consequences for their interpretation. Later studies that build on the paper suffer from the same problems. New research showing if and when latent class analysis yields valid results for drawing inferences about diagnostic test accuracy and disease prevalence is needed.

In the context of estimating diagnostic test accuracy and disease prevalence in the absence of a gold standard, the seminal paper by Albert and Dodd (2004, herafter A&D) cautioned on the robustness of latent class analysis. The authors concluded after performing extensive simulations, that: (i) significant bias in estimators of test accuracy can occur if the dependence structure among diagnostic test outcomes is incorrectly specified, and (ii) in typical diagnostic research settings where the number of tests carried out is lower than 10 and the sample size is small or moderate (in the order of a few hundred), the observed data is generally insufficient to distinguish between alternative dependence structures. More than 10 years after publication of the manuscript, these findings still have much relevance to the practicing diagnostic researcher and important consequences for the further development of the latent class methodology (e.g., [41, 136, 168]).

In this chapter we revisit the findings presented in the orginal paper by A&D. We found that many of the presented simulation results have flaws and therefore do not adequate support their strong cautions against the use of latent class analysis. Notably, a substantial number of these results concern the ability to distinguish between saturated latent class models with alternative dependence structures (relating to conclusion ii of the original paper), without being identified as such. It is well known that the evaluation of fit of saturated models is never meaningful [115]. In addition, we will argue that the majority of the simulation settings have little relevance to the practicing diagnostic researcher. These results relate to a special case of a latent class model that has found little application in practice. Based on the results of a new simulation study we conclude that a different view on the value of latent class analysis is warranted: new research into both merits and pitfalls of latent class analysis is needed. We end with a discussion relating our findings to more recent papers suffering partly from the same problems [5, 7].

Latent class models used by A&D

Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ})'$ denote the vector of binary diagnostic test outcomes on test $j, j = 1, \cdots, J$, for subject $i, i = 1, \cdots, N$, taking on the value $x_{ij} = 1$ when positive and $x_{ij} = 0$ otherwise. Assuming a 2-class latent class model, the true disease status for subject i is a Bernoulli random variable d_i. We let the probability for the true disease be denoted by $\mathbb{P}(d_i = 1) = \pi_1$. A&D focus on the case where diagnostic test errors are correlated (i.e., subjects' test outcomes are not independent Bernoulli conditional on the true disease status), due to an unobserved factor r, hereafter referred to as the random effect variable.

For their analyses, A&D used a Gaussian Random Effect (GRE) model, assuming $r \sim \mathrm{N}(0,1)$, where, $\mathbb{P}(x_{ij} = 1|d_i, r_i) = \Phi(\mu_{jd_i} + \sigma_{d_i} r_i)$ are independent Bernoulli, $\Phi(\cdot)$ is the cumulative distribution function (c.d.f) of a standard normal variable. The probability that the i^{th} subject has test pattern $f(\boldsymbol{x}_i)$ is

$$f(\boldsymbol{x}_i) = \sum_{d_i=0}^{1} \pi_{d_i} \int \prod_{j=1}^{J} Pr(x_{ij}|d_i, r)\psi(r)dr, \tag{5.1}$$

where $\psi(r)$ is a standard normal density. Under the GRE model the vector of unknown parameters is denoted by $\theta_{GRE} = \{\pi_1, \sigma_0, \sigma_1, \mu_{01}, \cdots, \mu_{1J}\}'$. The GRE is a special case of the Gaussian random effects model proposed by Qu et al. [140], constraining the σ-parameters to be equal for all tests. Xu et al. [180] noticed that this constraint amounts to assuming equal within class correlation between all pairs of tests.

As an alternative for the random effects distribution, A&D used a Finite Mixture (FM) model [8], assuming r_i is a Bernoulli random variable, $\mathbb{P}(x_{ij} = 1|d_i, r_i) = \gamma_{jd_ir_i}$,

$$f(\boldsymbol{x}_i) = \sum_{d_i=0}^{1} \pi_{d_i} \sum_{r_i=0}^{1} \omega_{d_ir_i} \prod_{j=1}^{J} \gamma_{jd_ir_i}^{x_{ij}} (1 - \gamma_{jd_ir_i})^{1-x_{ij}}, \tag{5.2}$$

where $\omega_{d_ir_i}$ is a class specific mixture weight ($\omega_{d_i1} = 1 - \omega_{d_i0}$). It is further assumed that at one level of the random effect variable ($r_i = 1$) the probabilities of false positive and false negative test outcomes are structurally zero, i.e. $\gamma_{j01} = 1 - \gamma_{j11} = 0$. The parameter vector is given by $\theta_{FM} = \{\pi_1, \omega_{01}, \omega_{11}, \gamma_{100}, \cdots, \gamma_{J10}\}'$.

Estimates of the sensitivity $\hat{\tau}_{j1}$ and specificity $1 - \hat{\tau}_{j0}$ for j^{th} diagnostic tests are the averages over r for the estimated values of $\mathbb{P}(x_j = 1|d = 1, r)$ and $\mathbb{P}(x_j = 0|d = 0, r)$, respectively. For the GRE model, $\hat{\tau}_{j1} = \Phi(\hat{\mu}_{j1}/(1 + \hat{\sigma}_1^2)^{1/2})$ and $1 - \hat{\tau}_{j0} = 1 - \Phi(\hat{\mu}_{j0}/(1 + \hat{\sigma}_0^2)^{1/2})$. Under the FM model, $\hat{\tau}_{j1} = \hat{\omega_{11}} + \hat{\omega_{10}}\hat{\gamma_{j10}}$ and $1 - \hat{\tau}_{j0} = \hat{\omega_{01}} + \hat{\omega_{00}}(1 - \hat{\gamma_{j00}})$.

The two random effects models reviewed here may be viewed as two extreme cases of conditional dependence structures. Under the GRE model, for most subjects $\mathbb{P}(x_j = 1|d_i, r_i)$ corresponds to values close to the mean value of the random effect $r_i = 0$. Under the FM model, $\mathbb{P}(x_j = 1|d_i, r_i)$ across subjects is concentrated at one of two extreme values, corresponding to $r_i = 0$ or $r_i = 1$.

Latent class models for the number of positive test results

Provided that diagnostic tests have identical (or 'common') true sensitivity and true specificity, i.e. $\mathbb{P}(x_{i1}|d_i) = \mathbb{P}(x_{i2}|d_i) = \cdots = \mathbb{P}(x_{iJ}|d_i)$, the sum of binary test outcomes variable, $s_i = \sum_{j=1}^{J} \boldsymbol{x}_{ij}$, $s_i \in 0, \cdots, J$, is a sufficient statistic for a Mixture Binomial (MB) model [64, 162]. Three MB models are presented by A&D. First, they discuss a Mixture Binomial Gausian Random Effect (MB-GRE) model,

$$f(s_i) = \sum_{d_i=0}^{1} \pi_{d_i} \binom{J}{s} \int \Phi(\mu_{d_i} + \sigma_{d_i} r)^{s_i} (1 - \Phi(\mu_{d_i} + \sigma_{d_i} r))^{J-s_i} \psi(r) dr. \tag{5.3}$$

Estimates of the common sensitivity and specificity can be calculated from $\hat{\tau}_1 = \Phi(\hat{\mu}_1/(1 + \hat{\sigma}_1^2)^{1/2})$ and $1 - \hat{\tau}_0 = 1 - \Phi(\hat{\mu}_0/(1 + \hat{\sigma}_0^2)^{1/2})$. Secondly, they discuss a Mixture Binomial Finite Mixture model (MB-FM),

$$f(s_i) = \sum_{d_i=0}^{1} \pi_{d_i} \sum_{r_i=0}^{1} \omega_{d_i r_i} \binom{J}{s} \gamma_{d_i r_i}^{s_i} (1 - \gamma_{d_i r_i})^{J-s_i}, \tag{5.4}$$

$\gamma_{01} = 1 - \gamma_{11} = 0$, $\hat{\tau}_1 = \hat{\omega_{11}} + \hat{\omega_{10}}\hat{\gamma_{10}}$ and $1 - \hat{\tau}_1 = \hat{\omega_{01}} + \hat{\omega_{00}}(1 - \hat{\gamma_{00}})$.

The parameter vectors for the MB-GRE and MB-FM are given by $\theta_{MB-GRE} = \{\pi_1, \sigma_0, \sigma_1, \mu_0, \mu_1\}'$ and $\theta_{MB-FM} = \{\pi_1, \omega_{01}, \omega_{11}, \gamma_{00}, \gamma_{J10}\}'$, respectively. Not detailed here, a third alternative proposed by A&D is based on a Beta (MB-B) model with parameters $\theta_{MB-B} = \{\pi_1, \alpha_0, \alpha_1, \beta_0, \beta_1\}'$.

Estimation and degrees of freedom

Details about estimation and identifiability of the models discussed in the preceding section can be found in the papers in which the respective models were originally proposed [8, 140]. In short, the parameters can be estimated by maximum likelihood through, for example, an expectation-maximization (EM) algorithm [51]. Estimation is based on maximizing $\mathcal{L} = \prod_{i=1}^{N} f(\boldsymbol{x}_i)$ for binary test data or $\mathcal{L} = \prod_{i=1}^{N} f(s_i)$ for mixture binomial models. Only for the GRE and MB-GRE models estimation requires integration. These integrals will be approximated numerically by Gauss-Hermite quadrature [6, 140] with 50 quadrature nodes. The minimal requirement for identifiability for the GRE and FM models is $J \geq 4$ and for MB-GRE, MB-FM and MB-B $J \geq 5$.

When fitting these models to observed data, one often reports the residual degrees of freedom (df_{res}, the number of degrees of freedom (df) minus the number of parameters (P), [2, 3]).

First, it provides information on both the number of (additional) parameters that can be estimated from the available data (minimal requirement $df_{res} \geq 0$). Second, goodness-of-fit testing is often performed using χ^2-statistics fit statistics (we use the likelihood ratio statistic G^2, [63]) which under perfect fit have asymptotic χ^2 distributions with degrees of freedom equal to the residual degrees of freedom. For the FM and GRE models, $df_{res} = 2^J - 2J - 4$ and for the MB-FM, MB-GRE and MB-B models $df_{res} = J - 5$. For binary test data, this follows naturally from the observations that i) $P = 2J + 3$ corresponds to the size of the parameters vectors θ_{GRE} and θ_{FM} and ii) $df = 2^J - 1$, since the sample space consists of 2^J possible test outcome patterns. Similarly, for the number of positive test outcomes data, i) from θ_{MB-GRE}, θ_{MB-FM} and θ_{MB-B} it can be seen that $P = 5$, ii) s_i can take on the values between all test negative ($s = 0$) and all test positives ($s = J$), hence, the sample space of s_i consists of $J + 1$ possible sum scores, providing J degrees of freedom (e.g., see [162]).

Revisiting the analyses

In this section we will re-analyze the motivating example presented by A&D and point out some problems with the earlier inference. Further, we briefly discuss the evidence based on their examination of asymptotic properties (section 4, A&D) and Monte Carlo experiments (section 5, A&D).

Dentistry data revisited

A&D presented an analysis on Handelman's dentistry data [59, 62, 82] that consists of the classifications of $N = 3869$ teeth as sound (negative) or carious (positive) based on X-rays by $J = 5$ dentists. In Table 5.1 we present a re-analysis assuming FM and GRE conditional dependence structures. We used Latent Gold 5.0 software. Focusing first on the binary test data, as pointed out in A&D, there are substantial differences in parameter estimates between the FM and GRE models. The differences are most pronounced in the estimated sensitivity of dentist 1 and 3 and the prevalence of caries. However, based on a likelihood ratio test, we find that both the FM and GRE model exhibit large degrees of misfit on the dentistry data ($G^2 = 53.1$ and $G^2 = 42.7$, $df = 18$).

When estimating the MB-GRE, MB-FM and MB-B models (we ignore the unrealistic 'conditional independence model'), A&D reported fit to the dentistry data for each of these models is close to perfect (goodness-of fit χ^2-statistics are estimated close to zero) and next to equivalent (minimal differences in log-likelihood values), making it difficult to distinguish between these

Table 5.1: Dentistry data revisited (N = 3869). Point estimates and standard errors[a] from latent class models

	Models for binary data		Models for binomial data	
	FM	GRE	MB-FM	MB-GRE
Prevalence	.17 (.016)	.11 (.076)	.17 (.017)	.31 (.093)
Average Se	.	.	.65 (.027)	.44 (.097)
Average Sp	.	.	.89 (.006)	.91 (.006)
Se dentist 1	.45 (.038)	.54 (.126)	.	.
Sp dentist 1	.99 (.003)	.97 (.014)	.	.
Se dentist 2	.74 (.033)	.77 (.107)	.	.
Sp dentist 2	.88 (.009)	.85 (.027)	.	.
Se dentist 3	.65 (.041)	.81 (.195)	.	.
Sp dentist 3	.98 (.006)	.96 (.023)	.	.
Se dentist 4	.51 (.025)	.50 (.071)	.	.
Sp dentist 4	.96 (.007)	.93 (.022)	.	.
Se dentist 5	.92 (.018)	.93 (.077)	.	.
Sp dentist 5	.68 (.012)	.64 (.034)	.	.
P	13	13	5	5
residual *df*	18	18	0	0
log likelihood	-7427.00	-7421.70	-5226.22	-5226.37
G^2	53.08	42.74	NA[b]	NA[b]

[a] non-parametric bootstrap procedure with 10,000 bootstrap samples. Point estimates and fit statistics obtained by averaging over bootstrap samples.

[b] model is saturated: fit not determined.

Prevalence: ($\hat{\pi}_1$); Average Se: common sensitivity ($\hat{\tau}_1$), Average Sp: common specificity ($1 - \hat{\tau}_0$), Se: Sensitivity dentist j ($\hat{\tau}_{j1}$); Sp: Specificity dentist j ($1-\hat{\tau}_{j0}$); P: Number of estimated parameters; G^2: Likelihood ratio statistic;

models based on the observed data. However, the reader can verify using subsection 5 of the current manuscript that for the MB-GRE, MB-FM and MB-B models on the dentistry data: df_{res} = 0. Hence, these MB-models are saturated. Consequently, the fit of these models to the data is perfect and equivalent besides possible rounding errors [115], irrespectively of the structure of the dentistry data. Notice this is general: when J = 5, the fit to the observed data of the mixture binomial models defined in subsection 5 cannot meaningfully be evaluated, unless (additional) constraints to the parameters of these models are imposed. A&D (Table 1, pp. 429) erroneously report: df_{res} = 1 for the MB-GRE, MB-FM and MB-B models.

We also observed a large spread in observed proportions of positive diagnoses by each of the dentists: 8.8%, 22.2%, 12.8%, 12.1% and 42.5%. This clearly contraindicates the assumption of identical dentists' sensitivity and specificity in diagnosing caries. Hence, the use of the mixture binomial models for these data is not appropriate. Again, this issue is not specific to

the dentistry data. [162] details specific conditions in which identical sensitivity and specificity can hold. Due to these rather strong assumptions, mixture binomial models are rarely found in diagnostic research literature as the diagnostic tests under study often may have different accuracy.

Revisiting the simulations in A&D sections 3 and 4

The simulation results presented by A&D are tabulated in their sections three and four (Tables 3 to 6). A&D Table 3 and Table 5 reports closeness of goodness-of-fit for $J < 10$ of correctly and misspecified MB random effect models, both asymptotically (by expected subjects' contributions to the likelihood) and by Monte Carlo simulation. However, in both sections 50% of these results are flawed as these concern the case $J = 5$, for which we have earlier noted that the MB models are saturated. In Table 4 and 6 A&D report in total 5 simulation scenarios based on binary latent class models and that only for the minimum of $J = 4$ tests that are needed for identifiability.

Simulation study

We conducted a few simulations to investigate the robustness of the simulation results of A&D. We repeated the simulation scenario reported in Table 6 of their manuscript. Hereafter we refer to this scenario as data generating mechanism 1 (dgm 1). Further we investigated whether small changes to the simulation scenarios may lead to different conclusions about the ability to distinguish between latent class models based on their fit on binary data. Simulations are performed in Latent Gold 5.0 software.

For dgm 1, $R = 5000$ datasets of $N = 1000$ with $J = 4$ were generated under the FM model. The simulation parameters for this simulation scenario are: $\gamma_{j10} = (.6, .7, .8, .9), 1 - \gamma_{j00} = (.9, .8, .7, .6), \pi_k = .5, \omega_{01} = \omega_{11} = .5$. This corresponds to a setting where 50% of subjects are correctly classified by all tests, sensitivities of tests x_1 to x_4 are 0.95, 0.9, 0.85 and 0.8, respectively, while specificities are 0.8, 0.85, 0.9 and 0.95. On each simulated dataset, the GRE (i.e. the misspecified model) and FM (i.e. the correctly specified model) are fitted; for each model the 11 parameters are estimated. The simulation set-up implies within-class pairwise correlations between tests ranges from 0.07 to 0.21. We also considered two different simulation settings (dgm 2 and 3). Dgm 2 is equivalent to dgm 1, except that $\gamma_{j10} = (.6, .6, .6, .6), 1 - \gamma_{j00} = (.6, .6, .6, .6)$. This implies the sensitivity and specificity of all 4 tests is equal to 0.8. Dgm 3 is an extension of dgm 2 to 5 tests. The within-class pairwise

correlations between tests is .11 under both dgm2 and dgm3.

Similar to the original A&D publication we find that for dgm 1, the probability of estimating a log likelihood value for the FM model higher than the value for the constrained GRE model is unsatisfactory (success rate: 62%). The small adjustments to the dgm 1 scenario, in dgm 2 and 3 increases success rates to satisfactory levels: .88 and >.99 respectively. Similarly, the power (approximated by simulation) to reject the GRE model based on the likelihood ratio test (G^2) is very low: 6% at $\alpha = .05$ in dgm 1, increasing to 59% and 88% for dgm 2 and 3 respectively.

Discussion

The frequently cited paper by Albert and Dodd [6] has warned readers against the use of latent class analysis in the absence of a gold standard. In this paper we have shown that the evidence reported in this paper has important flaws that need reconsideration. First, the authors fail to report that a substantial number of their simulation scenarios involve saturated models. Their observation that the fit of these models to the observed data is near equivalent is therefore both unsurprising and uninformative. In addition, much of the evidence presented in their paper is based on examinations under Mixture Binomial models which involve the assumption of equivalent sensitivity and specificity across tests. These models are rarely applied in practice. Lastly, while the A&D paper suggest it may be possible to distinguish between alternative dependence structures only with very large number of tests ($J = 10$), we found it was possible to find a scenario where $J = 5$ was sufficient to distinguishing between alternative dependence structures with high probability.

In our view, a more nuanced view on the value and problems of latent class analysis is warranted. We also recognize that latent class analysis can yield biased results when the dependence structure is misspecified and it may not always be possible to distinguish between alternative dependence structures in practice. As the conditions under which this occurs are yet unclear, more research in this area is needed. Albert and Dodd [7] and Albert [5] suggested potentially more robust methods for estimating accuracy and disease prevalence by bringing in information in latent class analysis from partial or imperfect verification using a reference standard. However, these papers contain similar scenarios as found in the paper reviewed here, where model fit is evaluated for latent class models that are saturated. Therefore, caution should be exercised as well when interpreting the later studies building on the original A&D manuscript.

Chapter 6

Value of composite reference standards in the absence of a gold standard

A common challenge in diagnostic studies is to obtain a correct final diagnosis in all participants. Ideally, a single error-free reference test, known as a gold standard, is used to determine the final diagnosis [145] and estimate the accuracy of the test or diagnostic model under evaluation. If the reference standard does not perfectly correspond to true target disease status, estimates of the accuracy of the test or model under study (index test), such as sensitivity, specificity, predictive values, or area under the curve, can be biased [174]. This is known as imperfect reference standard bias. One method to reduce this bias is to use a fixed rule to combine results of several imperfect tests into a composite reference standard [9]. When the combination of several component tests provides a better perspective on disease than any of the individual tests alone, accuracy estimates of the test under evaluation (the index test) will be less biased than if only one imperfect test is used as the reference standard. Comparing the index test against each component test separately and then averaging the accuracy estimates is not recommended; it is better to insightfully combine component tests together into a CRS.

The hallmark of CRSs is that each combination of test results leads to a particular final diagnosis; in its simplest form, disease present or absent. For example, in a study on the accuracy of a rapid antigen test for detecting trichomoniasis, researchers decided against using the traditional gold standard of culture because it probably misses some cases [87]. As they believed that microscopy picks up additional true cases, they instead considered patients as diseased if either microscopy or culture results were abnormal. Table 6.1 gives further examples. Although the choice of component tests and the rules used to combine them affects the estimates of accuracy of the test under study [79], little guidance exists on how to develop and define a composite reference standard. Additionally, there is a lack of consensus in the way the term composite reference standard is used and reporting of results is generally poor. To address these problems, we provide an explanation of the methods for composite reference standards and make recommendations for development and reporting.

What is a composite reference standard?

A composite reference standard is a fixed rule used to make a final diagnosis based on the results of two or more tests, referred to as component tests. For each possible pattern of component test results (test profiles), a decision is made about whether it reflects presence or absence of the target disease.

Composite reference standards are appealing because of their similarity to clinical practice; they strongly resemble diagnostic rules that exist for several conditions, such as rheumatic fever

Table 6.1: Examples of composite reference standards.

Condition	Example	Rule for combination
Trichomoniasis [87]	'Samples were labeled as positive if the results of either mount microscopy or culture were positive ... samples were labeled negative if both mount preparations and culture were negative'	Any positive rule
Typhoid fever [150]	'A composite reference standard of blood culture and polymerase chain reaction was used'	Any positive rule
Adherence to isoniazid preventive therapy for latent tuberculosis [129]	Adherence defined as ≥ 3 points when tests receive the following weights: 2 points for a positive urine isoniazid test result 1 point for patient observed taking tablets 1 for hospital records 1 point for patient self reporting	Heavier weights given to more accurate tests

and depression. Their main advantage is reproducibility of results, which is made possible by the transparency and consistency in the way that the final diagnosis is reached across participants. However, they also have disadvantages, the most glaring being the subjectivity introduced in the development of the rule.

The term composite reference standard is often loosely used as a catch-all term to describe any situation in which multiple reference tests are used to evaluate the accuracy of the index test. It is sometimes mistakenly used to describe differential verification, when different reference standards are used for different groups of participants (Table 6.2) [49, 125]. It has also been used to describe discrepant analysis, a method in which the reference standard is re-run or re-evaluated, or a different reference standard is used, when the first one does not agree with the index test [77]. Both these approaches can lead to seriously biased estimates of accuracy and should be avoided whenever possible.

In the example in Table 6.2 of a study on deep venous thrombosis in which differential verification from was mislabelled as a composite reference standard, the reference standard for participants with a negative index test result was clinical follow-up while those with a positive result received the preferred reference standard, computed tomography [69]. If minor thromboembolisms that would have been picked up by computed tomography were missed during follow-up, the number of false negatives will be underestimated and the number of true negatives overestimated, thus biasing the accuracy estimates. Ethical or practical difficulties sometimes make it impossible to implement the same reference standard in all participants, but it is important that the term differential verification is used to describe such situations.

Table 6.2: Examples of misuse of the term composite reference standard.

Disease	Example	Explanation of misuse
Congenital heart defect [60]	Pulse oximetry was performed prior to discharge and the results of this index test were compared with a composite reference standard (echocardiography, clinical follow-up and follow-up through interrogation of clinical databases).	This is differential verification because some patients received an intensive clinical work-up while others were followed-up in clinical databases
Deep venous thrombosis [69]	All patients were diagnosed according to local protocols. Pulmonary embolism was confirmed or refuted on the basis of a composite reference standard, including spiral computed tomography and three months follow-up.	This is differential verification because high risk patients had computed tomography whereas other patients were followed- up
Coronary artery stenosis [102]	Diagnosis stenosis using composite findings from both [the index and the reference] tests as an enhanced reference standard ... If a stenosis ≥ 50% had been seen on one [imaging test] but not on the other test, the observers closely re-evaluated the respective coronary artery segment showing discordant findings in order to confirm or revise their initial interpretation.	This is an example of discrepant analysis in which the index test influences the reference standard result

Table 6.2 also gives an example of discrepant analysis from an imaging study for coronary artery stenosis in which the reference standard results were re-evaluated when they did not agree with the index test results [102]. Such re-evaluation can only lead to increased agreement between index test and the reference standard, which in turn can only lead to overestimates of accuracy. Although discrepant analysis his highly discouraged, situations in which the reference standard is repeated or a different reference standard is applied in those patients where the index test and first reference standard disagree, should be termed discrepant analysis.

To avoid confusion we recommend using the term composite reference standard exclusively for situations in which, by design, all patients are intended to receive the same component tests and these component tests are interpreted and combined in a fixed way for all patients.

Developing a composite reference standard

As the choice of component tests and the rule for combining them strongly influences the accuracy of composite reference standards [116], careful attention is required when developing the decision rule. Ideally, the combination of test results and the corresponding final diagnosis should be specified before the study to prevent data driven decisions. However, if there is uncertainty about the best composite reference standard, a sensitivity analysis could be

Table 6.3: Effect of using different rules to produce composite reference standard on estimates of accuracy using example inspired by a study on the accuracy of rapid antigen detection test for trichomoniasis [87].

Result of component reference tests		Diagnosis with composite reference standard		Index test (rapid antigen detection test, n = 100)	
Culture	Microscopy	Any positive rule[a]	All positive rule[b]	No. positive result	No. negative result
+	+	+	+	25	1
+	-	+	-	10	3
-	+	+	-	4	1
-	-	-	-	1	55

[a] Accuracy estimate using the any positive rule: sensitivity = $(25 + 10 + 4)/((25 + 10 + 4) + (1 + 3 + 1)) = 0.89$; specificity= $55/(55 + 1) = 0.98$.

[b] Accuracy estimate using the all positive rule: sensitivity = $25/(25 + 1) = 0.96$; specificity = $(3 + 1 + 55)/((3 + 1 + 55) + (10 + 4 + 1)) = 0.8$.

planned to see how sensitive the results are to the particular choice of tests or combination rule. It is also important that the composite reference standard is clinically relevant. In other words, it should detect cases that will benefit from clinical intervention rather than simply the presence of disease [114]. For clinical situations when the true disease status cannot be defined the composite reference standard should reflect the provisional working definition. Keeping diagnostic guidelines in mind and seeking advice from experts in the field will help ensure that the chosen standard is clinically relevant and interpretable.

Defining rules to combine component tests

Two rules exist for combining component tests into a composite reference standard. In the simplest scenario of two dichotomous component tests, participants could be considered to have the disease if either test is indicative of disease (any positive rule, also known as the or rule). The alternative is that participants are considered to have the disease only if both tests detect disease (all positive or and rule). If there are more than two component tests a combination of these two rules can be used. Increasing the number of component tests will increase the number of participants categorised as diseased. If the any positive rule is used, this will increase the sensitivity of the composite reference standard (more diseased subjects will be classified as diseased) but decrease its specificity (more non-diseased subjects will be classified as having the disease). The reverse is true for the all positive rule; sensitivity of the composite reference standard decreases while specificity increases. Table 6.3 gives an example of how the choice of combination rule affects the accuracy of the composite reference standard, which in turn affects the accuracy estimates of the test under study [174].

There is almost always a trade-off between sensitivity and specificity when considering alternative ways to combine component tests [116]. The exception is when a component test in an any positive rule has perfect sensitivity, which makes a composite reference standard with perfect sensitivity, or when a component test in an all positive rule has perfect specificity, which makes a composite standard with perfect specificity [9]. Near perfect sensitivity or specificity of a component test is often the reasoning provided for the rule chosen.

Selection of component tests

Although it may be tempting to include numerous component tests, the gain in sensitivity or specificity of the resulting composite reference standard decreases (and the clinical interpretability may diminish) as more tests are added. This is because additional tests may fail to provide new information. In the trichomoniasis example, if another test such as polymerase chain reaction amplification is added, new true cases may be detected [87]. However, if yet another test is added, fewer additional true cases will be detected because fewer remained undetected. Eventually, all true cases are detected and additional tests will only result in false positive results, thus decreasing the specificity of the composite reference standard.

Multiple tests will be useful only if the component tests catch each others mistakes. For example, in a group of patients who truly have trichomoniasis, if microscopy identifies disease in the same participants as culture does, microscopy does not add any information and therefore the sensitivity of the composite reference standard will not be higher than that of culture alone [174]. When component tests make the same classifications in truly diseased or non-diseased patients more or less often than is expected by chance alone, this is referred to as conditional dependence.

In some cases, conditional dependence can be avoided or reduced by choosing component tests that look at different biological aspects of the disease [66]. To avoid causing the tests to make the same mistakes, you should consider blinding the observer of each component test to the results of the other component tests if knowledge of these other test results can influence interpretation.

Extensions to the basic composite reference standard

The basic composite reference standard categorises patients simply as diseased or non-diseased. However, multiple disease categories can also be defined, such as subtypes, stages, or degree

Table 6.4: Use of a composite reference standard to determine different categories of diagnosis for turberculosis [165].

Final diagnosis	Individual tests				
	Acid fast bacilli smear	Culture	Radiology	Histology	Follow-up
Confirmed	+/-	+	+/-	+/-	+
Probable	+/-	-	+	+	+
	+/-	-	+	-	+
	+/-	-	-	+	+
Possible	+/-	-	-	-	+
Not tuberculosis	-	-	-	-	-

of certainty of disease. An example is a study on tuberculosis in which people were categorised into one of four levels of disease certainty (Table 6.4) [165]. The basic composite reference standard gives equal weight to all tests, but in clinical practice tests carry different weights. The relative importance of the component tests can be incorporated by assigning weights. For example, in the assessment of adherence to isoniazid treatment for latent tuberculosis in Table 6.1, the most reliable test was given twice the weight of the other tests [129].

Missing values on component tests

As with all diagnostic accuracy studies, results may be biased when not all participants receive the intended reference standard [49]. Careful attention needs to be paid to missing values in component tests. For example, if the any positive rule is used and the result of component test 1 is positive, we can conclude that a patient is diseased without knowing the result of component test 2. For efficiency, researchers might consider skipping the second test in participants whose first test result is positive [9, 91]. However, if component test 1 is negative, component test 2 becomes necessary for determining the diagnosis.

When a result is missing from a component test that must be present under the combination rules, the composite reference standard is also missing. This may affect the accuracy estimates of the index test and mathematical methods should be used to tentatively correct for this bias [48].

Table 6.5: Template for reporting results when using a composite reference standard.

Composite reference standard				Index test[a]	
Test 1	Test 2	Test 3	Final diagnosis	No. positive result	No. negative result
+	+	+	+	p1	n1
+	+	-	+ or -	p2	n2
+	-	+	+ or -	p3	n3
+	-	-	+ or -	p4	n4
-	+	+	+ or -	p5	n5
-	+	-	+ or -	p6	n6
-	-	+	+ or -	p7	n7
-	-	-	-	p8	n8

[a] When the results of the test being studied are not dichotomous or it is a diagnostic model, results from the optimal or most common cut-off point should be presented in the index test column. This template can be extended for situations with more than 3 composite tests.

Reporting guidelines

Complete and accurate reporting of the reference standard procedure is critical to allow readers to judge the potential risk of bias in accuracy estimates. This is especially important for systematic reviews of diagnostic tests. The validity of comparing accuracy estimates between studies and pooling of estimates across studies is challenged when studies use different reference standards or when reference standards are poorly defined or reported [111, 177]. We therefore recommend that in addition to using current reporting guidelines [26], authors of diagnostic accuracy studies should include the following details about studies with composite reference standards:

- The rationale behind the selection of component tests and the combination rule
- The corresponding final diagnosis for each combination of test results
- Whether component test results were missing and whether this resulted in a missing composite reference standard
- The number of participants with each combination of test results. For continuous tests, this information should at least be provided for the optimal or most common cut-off point.

Table 6.5 gives a template for reporting. The availability of all of the above information will allow studies using composite reference standards to be compared with those using only one of the component tests as the reference standard.

Conclusions and recommendations

Combining multiple tests to define a target disease status rather than using a single imperfect test is a transparent and reproducible method for dealing with the common problem of imperfect reference standard bias. Although composite reference standards may reduce the amount of such bias, they cannot completely eliminate it because it is unlikely that a combination of imperfect tests will produce a composite standard with perfect sensitivity and specificity.

Other methods for dealing with bias resulting from imperfect reference standards are panel diagnosis and latent class analysis [9,145]. In panel diagnosis, multiple experts review relevant patient characteristics, test results, and sometimes follow-up information before coming to a consensus about the final diagnosis in each patient. Latent class analysis estimates accuracy by assuming that true disease status is unobservable and relating the results of multiple tests to it in a statistical model [9, 136]. The choice of method to deal with imperfect reference standard bias will probably depend on the type, number, and accuracy of the pieces of diagnostic information available in a particular study. Results from all three methods could be presented to strengthen their face validity. Researchers who use a composite reference standard can improve the transparency and reproducibility of their results by following our recommendations on reporting.

Chapter 7

Bias due to composite reference standards in the absence of a gold standard

Abstract

Composite reference standards (CRSs) have been advocated in diagnostic accuracy studies in the absence of a perfect reference standard. The rationale is that combining results of multiple imperfect tests leads to a more accurate reference than any one test in isolation. Focusing on a CRS that classifies subjects as disease positive if at least one component test is positive, we derive algebraic expressions for sensitivity and specificity of this CRS, sensitivity and specificity of a new (index) test compared to this CRS, as well as the CRS-based prevalence. We use as a motivating example the problem of evaluating a new test for *Chlamydia trachomatis*, an asymptomatic disease for which no gold-standard test exists. As the number of component tests increases, sensitivity of this CRS increases at the expense specificity, unless all tests have perfect specificity. Therefore, such a CRS can lead to significantly biased accuracy estimates of the index test. The bias depends on disease prevalence and accuracy of the CRS. Further, conditional dependence between the CRS and index test can lead to over-estimation of index test accuracy estimates. This commonly-used CRS combines results from multiple imperfect tests in a way that ignores information, and therefore is not guaranteed to improve over a single imperfect reference unless each component test has perfect specificity and the CRS is conditionally independent of the index test. When these conditions are not met, as in the case of *C. trachomatis* testing, more realistic statistical models should be researched instead of relying on such CRSs.

For many diseases there is no gold standard diagnostic test having both perfect sensitivity and specificity. Obtaining a definitive diagnosis for each subject in a diagnostic test evaluation study or disease prevalence study therefore becomes challenging [143]. Disease classification based on any imperfect single reference test will lead to biased inferences [29, 154, 176].

Consider the problem of evaluating a new test for *Chlamydia trachomatis* (*C. trachomatis*), an asymptomatic, infectious disease with harmful medical consequences if missed, and potentially serious social consequences if falsely diagnosed. The reference test of choice was previously cell culture, which is thought to have near-perfect specificity but only moderate sensitivity [21]. More recently developed nucleic acid amplification tests (NAATs) are considered more sensitive than culture, but not as specific.

To reduce the misclassification of *C. trachomatis* status, a number of studies have used a Composite Reference Standard (CRS) for evaluating a new test (or index test) [9, 15, 22, 37, 99, 149]. For example, Alonzo et al. [9] defined a CRS based on two imperfect tests - cell culture and polymerase chain reaction (PCR). Subjects were classified as *C. trachomatis*-positive if positive on at least one of the two component tests and as negative otherwise. This CRS is based on what we will refer to as an OR decision rule (i.e., logical disjunction or 'any positive' rule) according to which a positive result on one or more component tests is defined as having the disease of interest. Another common way to define a CRS is the AND decision rule (i.e., logical conjunction or 'all positive' rule) according to which a positive result on all component tests is required to classify a subject as having the disease. Besides applications in *C. trachomatis*, CRSs are widely used, e.g. for extra-pulmonary tuberculosis, community-acquired pneumonia and pertussis [55, 113, 151].

The CRS is appealing because it provides a simple rule to assign a final 'diagnosis' to each study subject. It is also described as being 'unaffected by' or 'independent of' the test under evaluation because the test under evaluation is not used in defining the final diagnosis [9, 22]. However, though the CRS aims to reduce mis-classifications of the disease status, it will not eliminate them [124]. Nor are the component tests necessarily independent of the index test [79]. For *C. trachomatis*, the aim is to reduce the number of false negatives by increasing the number of component tests in combination with the OR-rule. The prospect of reducing misclassification by increasing the number of component tests has led to a diversity of CRSs encountered in practice - some having as few as two component tests [9, 23], others as many as 9 [149], while the FDA draft guidance requires 4 [163].

So far, little attention has been paid to studying the accuracy of the CRS itself, and to how its accuracy is affected by its composition [54, 79]. Another aspect that has not received much attention is the impact of dependence between the errors of the component and index tests. Alternative statistical methods, particularly latent class analysis, have been the subject of much scrutiny with regards to the impact of conditional dependence and studies have shown that ignoring conditional dependence can result in over-estimation of accuracy of the index test [135, 136, 160, 164]. Dependence of errors may occur, for example, if there exists a spectrum of disease severity with both component and index tests being more likely to detect more severe cases and to miss less severe cases [31, 174] or when component and index tests are based on the same technology. In the approach currently used by the FDA to evaluate new NAATs for *C. trachomatis* infection, a new test is typically compared against a CRS made up of other NAATs based on the same biological mechanism [38, 163]. It is highly likely that there is a conditional dependence between the CRS and the test under evaluation in this setting.

In this article, we aim to bridge these gaps by elucidating the workings of the CRS. We will focus on the OR definition of the CRS as this applies to most *C. trachomatis* settings, where the specificity of the component tests are considered to be relatively high and there is an interest in improving the overall sensitivity of the CRS beyond that of any single component test. We first describe a motivating example of evaluating a new test for *C. trachomatis*. We derive algebraic expressions for the accuracy of the CRS, as well as expressions for the index test's accuracy based on the CRS. Further, we examine the behaviour of these parameters in several simulated scenarios. We also briefly investigate CRSs defined by other decision rules. We conclude with a Discussion.

Estimating the accuracy of a test for *C. trachomatis* infection

To illustrate application of a CRS in practice we use data from a diagnostic accuracy study of *C. trachomatis* tests [22]. Data are available on 743 asymptomatic women (see Supplementary Material) on 3 types of tests: A) 4 NAATs: polymerase chain reaction (PCR) and ligase chain reaction (LCR) tests each carried out on both cervical (PCRC, LCRC) and urine specimens (PCRU, LCRU), B) 2 culture tests one each on cervical (CULC) and urethral specimens (CULU), and C) a DNA hybridization test (DNAP).

Though based on different mechanisms, all tests are designed to have high specificity [138]. NAATs are designed to detect *C. trachomatis* DNA at a low organism load and are therefore considered most sensitive [21]. However, these tests cannot distinguish between DNA from

Table 7.1: Impact of increasing number of tests in the CRS on estimated sensitivity ($\hat{S}^*_{PCRU}$) and estimated specificity ($\hat{C}^*_{PCRU}$) of the PCRU test. (For each CRS component tests used in OR decision rule appear in brackets.)

Reference standard	$\hat{S}^*_{PCRU}$	$\hat{C}^*_{PCRU}$
	Estimate (95% Confidence Interval)	
CULC+	0.833 (0.752 – 0.892)	0.943 (0.923 – 0.959)
CRS1 (CULC+ or CULU+)	0.825 (0.747 – 0.883)	0.957 (0.938 – 0.970)
CRS2 (CRS1+ or DNAP+)	0.810 (0.732 – 0.869)	0.961 (0.943 – 0.974)
CRS3 (CRS2+ or LCRC+)	0.804 (0.705 – 0.838)	0.983 (0.969 – 0.991)
CRS4 (CRS3+ or LCRU+)	0.776 (0.704 – 0.834)	0.992 (0.980 – 0.996)

viable and non-viable organisms contributing to less than perfect specificity. In comparison, cell culture and DNAP tests have lower sensitivity [21]. The cervical culture test is believed to have near perfect specificity [22]. We use one NAAT(PCRU), which was new at the time of the data collection, as the index test.

To evaluate the robustness of CRS-based accuracy estimates we considered different CRSs that could reasonably be defined with the available data. To examine the impact of increasing the number of component tests we defined a sequence of reference standards starting with a single test (CULC) followed by CRS1 = CULC+ OR CULU+, CRS2= CRS1+ OR DNAP+, CRS3=CRS2+ OR LCRC+ and CRS4=CRS3+ OR LCRU+.

When the reference standard was CULC, estimated sensitivity of the index test PCRU was 0.83 $(95\%CI, 0.75 – 0.89)$ but it decreased to 0.78 $(95\%CI, 0.70 – 0.83)$ using a CRS with 5 component tests (see Table 7.1). Simultaneously, the estimated specificity of PCRU ranged from 0.94 $(95\%CI, 0.92 – 0.96)$ based on CULC to 0.99 $(95\%CI, 0.98 – 0.99)$ based on 5 component tests. Our example shows that estimates of sensitivity and specificity of the index test are clearly sensitive to composition of the reference standard. This has also been noted by others, e.g. in a systematic review of tests for *C trachomatis*, Cook et al [43] found that those studies that had used a CRS with more component tests reported lower sensitivity and higher specificity. To understand how the composition of the CRS affects its accuracy and the estimated accuracy of the index test, we derive analytical expressions for the properties of the CRS as a function of its component tests in the following section.

Accuracy of CRS and accuracy of index test with respect to CRS

Let $T_i = (T_{i1}, \ldots, T_{iP})$ denote the vector of observed diagnostic test results for the i^{th} subject. T_{ij} takes values 1 if positive and 0 if negative on test j, and $Pr(T_i)$ denotes the probability function of observing T_i. The first $(P-1)$ tests are used to define the CRS. The P^{th} test is an index test to be evaluated. The CRS based on the OR-rule is formally defined as

$$CRS_i = I(T_{i1}, \cdots, T_{iP-1}) = \begin{cases} 1 & \text{if } max(T_{i1}, \ldots, T_{iP-1}) = 1 \\ 0 & \text{if } max(T_{i1}, \ldots, T_{iP-1}) = 0, \end{cases} \tag{7.1}$$

where I is an indicator function and CRS_i=1 implies the i^{th} subject is classified as having the disease. For brevity we drop the subscript i from the remainder of the presentation.

In the development below, we examine two scenarios: i) the P tests (i.e. both component and index) are independent conditional on the true (unobserved) disease status, ii) the P tests are dependent conditional on the true disease status.

Case when tests are conditionally independent

Assume that all P tests are stochastically independent conditional on the target disease status. The target disease status, denoted by D, takes the value $d = 0$ if the target disease is absent, and $d = 1$ if it is present. The probability function of each combination of test results can be written as,

$$Pr(T) = \sum_{d=0}^{1} Pr(T_1, \ldots, T_P | D = d) Pr(D = d) = \sum_{d=0}^{1} Pr(D = d) \prod_{j=1}^{P} Pr(T_j | D = d). \tag{7.2}$$

The sensitivity of the CRS based on the OR-rule is given by,

$$\begin{aligned} S_{CRS} &= Pr(CRS = 1 | D = 1) = Pr(max(T_1, \ldots, T_{P-1}) = 1 | D = 1) \\ &= 1 - \prod_{j=1}^{P-1} (1 - Pr(T_j = 1 | D = 1)) = 1 - \prod_{j=1}^{P-1} (1 - S_j), \end{aligned} \tag{7.3}$$

where S_j denotes the true sensitivity of the j^{th} component test. The specificity of the CRS is

given by,

$$\begin{aligned} C_{CRS} &= Pr(CRS = 0|D = 0) = Pr(max(T_1, \ldots, T_{P-1}) = 0|D = 0) \\ &= \prod_{j=1}^{P-1} Pr(T_j = 0|D = 0) = \prod_{j=1}^{P-1} C_j, \end{aligned} \tag{7.4}$$

where C_j denotes the true specificity of the j^{th} component test.

A perfect CRS would have zero probability of misclassification, i.e.: $Pr(CRS = 1|D = 1) = Pr(CRS = 0|D = 0) = 1$. From equations (7.3) and (7.4) it follows that for a CRS based on a finite number of component tests, under the assumption of conditional independence of all component tests, the OR-rule leads to a perfect CRS if and only if the following conditions are satisfied:

I) $C_1 = C_2 = \cdots = C_{(P-1)} = 1$,

II) $S_j = 1$ for at least one $j = 1, \cdots, P-1$.

Put in words, the conditions given above imply that for a CRS to be perfect it should include at least one component test with perfect sensitivity and specificity (i.e., is a gold standard which would then make the use of the CRS unnecessary) and all other component tests should at least have perfect specificity. Clearly, these conditions will not be met in practice. It is therefore of interest to see how the use of the CRS influences the estimated accuracy of the index tests.

Let the sensitivity of index test T_P with respect to the CRS be denoted by S_P^*. It can be shown that,

$$\begin{aligned} S_P^* &= Pr(T_P = 1|CRS = 1) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 1|D = d)\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|D = d))\}Pr(D = d)}{\sum_{d=0}^{1}\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|D = d))\}Pr(D = d)}, \end{aligned} \tag{7.5}$$

illustrating that CRS based sensitivity of index test is a function of the sensitivity and specificity of the component tests and the true disease prevalence. From equation (7.5), it can be seen that the sensitivity of the index test with respect to the CRS will be exactly equal to the true sensitivity of the index test ($S_P^* = S_P$) if any of the following conditions are satisfied:

I) $Pr(D = 1) = 1$

II) $C_1 = C_2 = \cdots = C_{(P-1)} = 1$

III) $S_P = 1 - C_P$.

The specificity of index test P with respect to the CRS is given by,

$$\begin{aligned} C_P^* &= Pr(T_P = 0|CRS = 0) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 0|D = d)Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 0|D = d)}{\sum_{d=0}^{1} Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 0|D = d)}, \end{aligned} \tag{7.6}$$

and thus also relies on the accuracy of the component tests and the true disease prevalence. It follows directly from equation (7.6) that under conditional independence of all P tests, the OR-rule leads to $C_P^* = C_P$ if any of the following conditions are satisfied:

I) $Pr(D = 0) = 1$

II) $S_1 = S_2 = \cdots = S_{(P-1)} = 1$

III) $C_P = 1 - S_P$.

Considered in the context of our example on *C. trachomatis*, the conditions I and III above for unbiased sensitivity or unbiased specificity are not relevant. A diagnostic study is unlikely to be designed in a population where the disease is either present or absent in 100% of individuals, nor is it realistic that the index test has no discriminatory value (i.e. $S_P + C_P = 1$). Under conditional independence, an unbiased estimate of the sensitivity (specificity respectively) of an index test would be obtained only if all specificities (sensitivities respectively) of $P - 1$ component tests equal 1. This means that even in the ideal scenario where all component tests have perfect specificity leading to S_P^* being unbiased, C_P^* will be biased. For the case of *C. trachomatis*, while it is true that cell culture is believed to have near perfect specificity, the condition that all component tests have perfect specificity will not be met. We will study the consequences of having high, though not perfect, specificities on S_P^* and C_P^* in Section 4.

The estimated disease prevalence with respect to the CRS is defined by:

$$Pr(CRS = 1) = Pr(D = 1)(1 - \prod_{j=1}^{P-1}(1 - S_j)) + Pr(D = 0)(1 - \prod_{j=1}^{P-1} C_j). \tag{7.7}$$

Case when tests are conditionally dependent

Conditional dependence between the P observed tests would arise due to their association with a variable besides the disease status D. Here we will study the situation where conditional

dependence is due to two dichotomous random variables Z_0 and Z_1 - the variable Z_1 causing dependence among subjects where $D = 1$ and the variable Z_0 causing dependence among subjects where $D = 0$. Conditional dependence between component and index tests among true disease positives would imply they both have imperfect sensitivity. Likewise for specificity.

In the case of our motivating example on *C. trachomatis* tests, Z_1 could be the severity of the disease status (or organism load) among those with the infection, while Z_0 could represent the presence of *Chlamydia* DNA among subjects who had a recent infection but are currently disease negative. The three dichotomous variables Z_1, Z_0 and D define between them 4 possible classes of subjects. We will denote these classes by: $L_1 = (Z_1 = 1, D = 1)$, $L_2 = (Z_1 = 0, D = 1)$, $L_3 = (Z_0 = 1, D = 0)$ and $L_4 = (Z_0 = 0, D = 0)$. The joint probability function of the observed test results can be defined as follows,

$$\begin{aligned} Pr(T) &= \sum_{Z_1=0}^{1} Pr(T_1, \ldots, T_P | Z_1, D = 1) Pr(Z_1, D = 1) \\ &\qquad + \sum_{Z_0=0}^{1} Pr(T_1, \ldots, T_P | Z_0, D = 0) Pr(Z_0, D = 0), \\ &= \sum_{k=1}^{4} Pr(T_1, \ldots, T_P | L_k) Pr(L_k) = \sum_{k=1}^{4} Pr(L_k) \prod_{p=1}^{P} Pr(T_p | L_k), \end{aligned} \tag{7.8}$$

assuming conditional independence between all P tests within the four classes. Notice that when the P tests are independent conditional on the target disease status $Pr(T_p|L_1) = Pr(T_p|L_2)$ and $Pr(T_p|L_3) = Pr(T_p|L_4)$.

The sensitivity and specificity of the j^{th} test with respect to the true disease status are given by,

$$S_j = Pr(T_j = 1|D = 1) = \frac{Pr(T_j = 1|L_1)Pr(L_1) + Pr(T_j = 1|L_2)Pr(L_2)}{Pr(L_1) + Pr(L_2)}, \text{and}$$

$$C_j = Pr(T_j = 0|D = 0) = \frac{Pr(T_j = 0|L_3)Pr(L_3) + Pr(T_j = 0|L_4)Pr(L_4)}{Pr(L_3) + Pr(L_4)}. \tag{7.9}$$

The sensitivity and specificity of the CRS are given by,

$$\begin{aligned} S_{CRS} &= \frac{\sum_{k=1}^{2} Pr(L_k)(1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|L_k)))}{\sum_{k=1}^{2} Pr(L_k)}, \text{and} \\ C_{CRS} &= \frac{\sum_{k=3}^{4} Pr(L_k) \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|L_k))}{\sum_{k=3}^{4} Pr(L_k)}. \end{aligned} \tag{7.10}$$

The sensitivity and specificity of index test P with respect to the CRS are given by,

$$S_P^* = \frac{\sum_{k=1}^{4} Pr(T_P = 1|L_k)Pr(L_k)\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|L_k))\}}{\sum_{k=1}^{4}\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|L_k))\}Pr(L_k)}, \quad (7.11)$$

$$C_P^* = \frac{\sum_{k=1}^{4} Pr(T_P = 0|L_k)Pr(L_k)\prod_{j=1}^{P-1} Pr(T_j = 0|L_k)}{\sum_{k=1}^{4} Pr(L_k)\prod_{j=1}^{P-1} Pr(T_j = 0|L_k)}. \quad (7.12)$$

Examining the sensitivity and specificity of the CRS

We study the effect of various factors on S_{CRS} and C_{CRS}. We consider the true sensitivity, specificity and disease prevalence values in the following ranges: $S_j \in (0.30, 0.90)$, $C_j \in (0.90, 0.99)$ and $Pr(D = 1) \in (0.05, 0.30)$. These values are motivated by those expected for testing of *C. trachomatis* and other infectious diseases, such as tuberculosis or pneumonia, where sensitivities can range from low to high depending on the type of test, specificities are generally high and disease prevalence is low in the general population but can be higher in high-risk sub-groups. For ease of illustration, we assume that all $P-1$ component tests used to define the CRS have identical sensitivity and specificity, i.e., $S_j = S, C_j = C, j = 1, \cdots, P-1$. The number of component tests are varied from 2 to 10.

We also examine the case where the observations are generated under the model specified by equation (7.8), where the P tests are conditionally dependent. The prevalence of the four classes are set at $Pr(L_1) = 0.05$, $Pr(L_2) = 0.05$, $Pr(L_3) = 0.05$, and $Pr(L_4) = 0.85$. The following constraints are applied without loss of generality: $Pr(T_j = 1|L_1) > Pr(T_j = 1|L_2)$ and $Pr(T_j = 1|L_3) > Pr(T_j = 1|L_4)$, $j = 1, \cdots P-1$.

Accuracy of CRS under conditional independence

From equation (7.3) it follows that S_{CRS} depends only on the sensitivities (S) of the component tests. As the number of component tests increases the sensitivity of the CRS itself increases. A simple intuitive explanation for this relation is that as the number of component tests increases the number of patients classified with a 'positive' diagnosis increases and therefore the probability of correctly classifying at least one target disease positive subject increases. Table 7.2 shows that combining 2 component tests both with sensitivity $S = 0.6$ will lead to a composite reference standard having sensitivity 0.84, i.e. the probability that at least one of the 2 tests correctly diagnose a disease positive subject is 0.84. By simply adding a third component test with sensitivity 0.6, the CRS's sensitivity reaches 0.94. In theory if we let the

Table 7.2: Relation between number of component tests and the accuracy of a CRS under settings of conditional independence and conditional dependence ((a) strong, (b) weak). True sensitivity and specificity of each component test are $S = 0.60$ and $C = 0.95$, respectively.

	Conditional independence		Conditional dependence Setting (a)		Conditional dependence Setting (b)	
$P-1$	S_{CRS}	C_{CRS}	S_{CRS}	C_{CRS}	S_{CRS}	C_{CRS}
2	0.840	0.903	0.800	0.930	0.840	0.903
3	0.936	0.857	0.888	0.917	0.936	0.857
4	0.974	0.815	0.934	0.908	0.974	0.815
5	0.98976	0.77378	0.96096	0.89824	0.98970	0.77383
10	0.99990	0.59874	0.99700	0.85414	0.99989	0.59891

number of component tests tend to infinity, the sensitivity would approach 1 (see Equation 7.3).

Similarly, from equation (7.4) it can be seen that the specificity of the CRS is completely determined by the specificities (C) of the component tests. The specificity of the CRS is a decreasing function of the number of tests. Again, the intuitive explanation is that the probability of all component tests having a negative outcome for a given target disease negative subject becomes lower as the number of tests increases. As shown in Table 7.2, two tests with $C = 0.95$ specificity would together form a CRS with specificity 0.90. Adding a third test with the same properties ($S = 0.6, C = 0.95$) would drive the CRS's specificity down to 0.86. In theory, adding an infinite number of tests would reduce the specificity of the CRS toward 0 (see Equation 7.4). Thus we observe that one cannot expect to improve both CRS sensitivity and specificity simultaneously, unless $C = 1$ for all component tests.

Figure 7.1 shows that for a given sensitivity of the component tests (x-axis), S_{CRS} increases as the number of component tests increases from $P - 1 = 2$ to $P - 1 = 10$. On the other hand, for a given specificity of the component tests, C_{CRS} decreases as $P - 1$ increases from 2 to 10.

Accuracy of CRS under conditional dependence

To allow for comparison with the results under conditional independence, two particular settings are considered with greater and lesser conditional dependence. Conditional

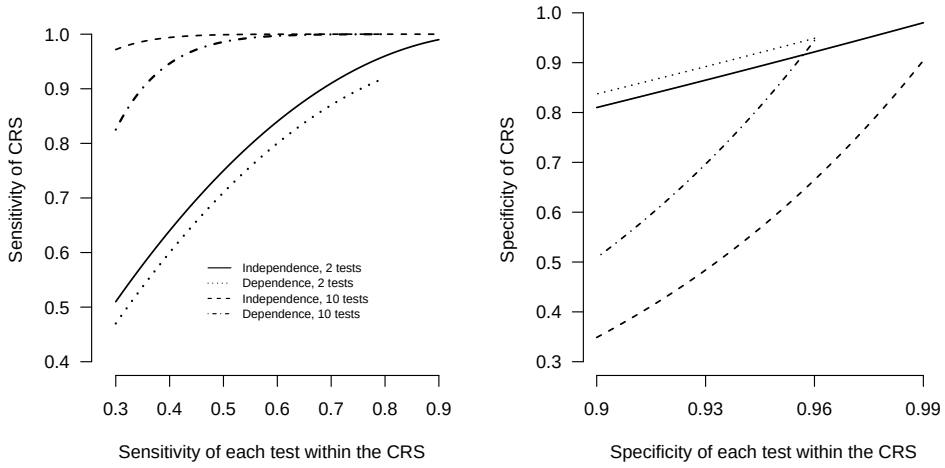

Figure 7.1: Sensitivity and specificity of the CRS vs. sensitivity and specificity, respectively, of conditionally independent and conditionally dependent component tests.

dependence within disease positive is created by setting $Pr(T_j = 1|L_1) \neq Pr(T_j = 1|L_2)$, and conditional dependence within disease negative by setting $Pr(T_j = 1|L_3) \neq Pr(T_j = 1|L_4)$. In setting (a), we consider a case where the magnitude of conditional dependence is greater. We set $Pr(T_j = 1|L_1) = 0.80, Pr(T_j = 1|L_2) = 0.40, Pr(T_j = 1|L_3) = 0.73$ and $Pr(T_j = 1|L_4) = 0.01$, $j = 1, \cdots, P-1$, so that the probability of a false negative outcome is substantially larger in L_2 than L_1 and the probability of a false positive outcome much larger in L_3 than L_4. In setting (b) $Pr(T_j = 1|L_1) = 0.61, Pr(T_j = 1|L_2) = 0.59, Pr(T_j = 1|L_3) = 0.06$ and $Pr(T_j = 1|L_4) = 0.0494$, corresponding to a situation where conditional dependence is weak. In both cases the probabilities stated above ensured $S = 0.60$ and $C = 0.95$.

From Table 7.1 and Figure 7.1 it can be seen that in setting (a) S_{CRS} does not increase as quickly when $P - 1$ increases from 2 to 10, while C_{CRS} also does not decline as quickly compared to the case when the tests are conditionally independent. From Table 7.1, under setting (b), changes in S_{CRS} and C_{CRS} resemble those observed under conditional independence. When the CRS is based on all conditionally dependent tests it will never achieve a specificity of 1. For example, under the conditional dependence setting (a), the component tests have a maximum possible specificity of 0.96 and C_{CRS} reaches a maximum value of 0.948 (Figure 7.1).

Bias in CRS-based accuracy estimates

We now investigate the behavior of S_P^* and C_P^* in relation to the composition of the CRS. Of particular interest is the impact of the sensitivity and specificity of the component tests and the true disease prevalence on the bias in CRS-based sensitivity and specificity ($S_P^* - S_P$, $C_P^* - C_P$). We consider the same ranges for accuracy of the component tests and true disease prevalence as in the previous section. The true values of the accuracy of the index test are set to $S_P = 0.90$ and $C_P = 0.90$. Under conditional dependence these values are obtained by setting $Pr(T_P = 1|L_1) = 0.98$ and $Pr(T_P = 1|L_2) = 0.82$, resulting in $S_P = \frac{0.05\times0.98+0.05\times0.82}{0.05+0.05} = 0.9$. The probabilities contributing to the index test specificity are set to $Pr(T_P = 1|L_3) = 0.95$ and $Pr(T_P = 1|L_4) = 0.05$, resulting in $C_P = \frac{0.05\times0.05+0.85\times0.95}{0.05+0.85} = 0.9$.

Trends in CRS-based sensitivity and specificity with changes in accuracy of component tests and disease prevalence

Figure 7.2 presents CRS-based sensitivity (S_P^*) and specificity (C_P^*) for varying accuracy of the component tests and disease prevalence assuming all tests are conditionally independent. The lines in each plot correspond to a different number of component tests ranging from 2 to 10. It should be noted that the values plotted are the expected values S_P^* and C_P^* (derived using equations (7.5) and (7.6)).

From the three panels on the left of Figure 7.2 it can be seen that the CRS-based sensitivity S_P^* is much lower than the true value of $S_P = 0.90$ in all settings. We find that the bias in S_P^* is determined primarily by worsening specificity of the component tests and decreasing true disease prevalence (see sharper slopes in Figures 7.2(c) and 7.2(e) vs. Figure 7.2(a)). On the other hand, the bias in C_P^* is primarily due to lower sensitivity of the component tests and increasing prevalence (Figures 7.2(b) and 7.2(f) vs. Figure 7.2(d)). The bias illustrated in these figures corresponds to the particular case where the component tests share the same properties. From equation (7.5) we can calculate precisely the bias from more general cases, such as when each successive component test in the CRS has better accuracy. For example, assume $Pr(D = 1) = 0.10$, and that the second test in the CRS improves over the first one in terms of specificity such that $S_1 = S_2 = 0.6$, $C_1 = 0.94$ and $C_2 = 0.96$. From equation (7.5) we can calculate that $S_3^* = 0.49$.

Figures 7.3 and 7.4 summarize the trends in S_P^* and C_P^*, respectively, when there is conditional dependence between the component tests of the CRS and the index test. The true disease prevalence is set at 0.10 in all scenarios. Greater conditional dependence within disease

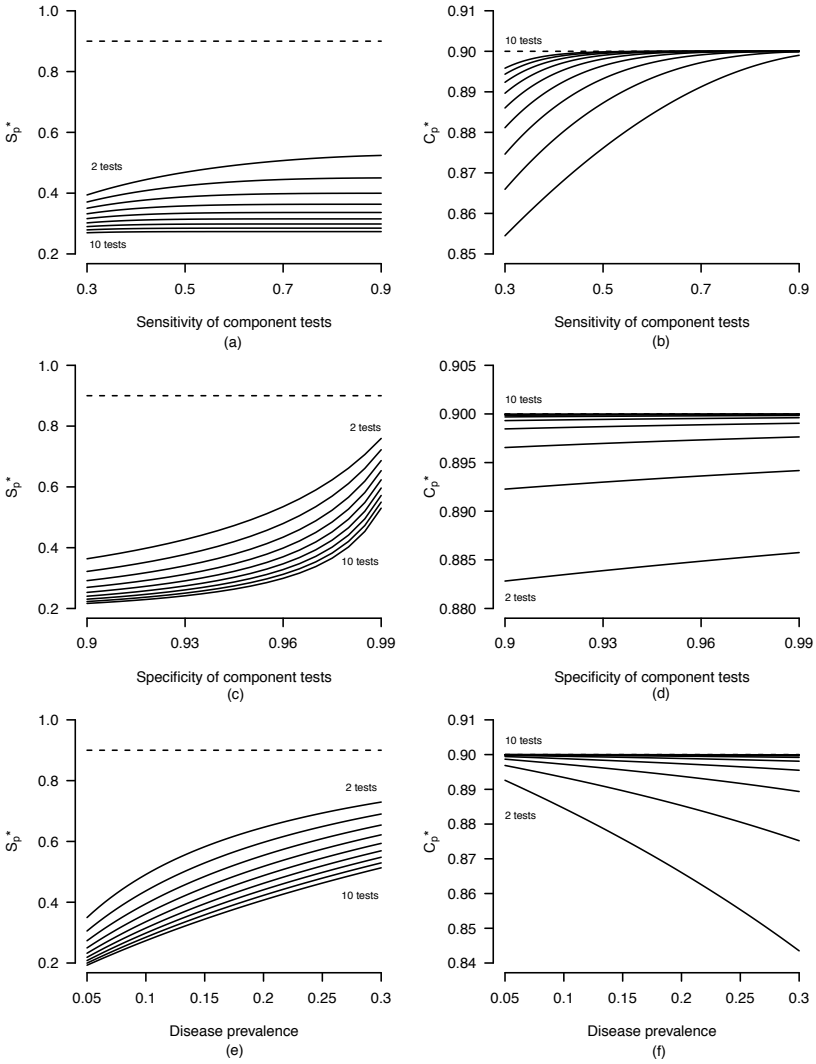

Figure 7.2: CRS-based sensitivity (S_P^*) and specificity (C_P^*) against accuracy of the component tests and disease prevalence while all tests are conditionally independent. (Upper panel (a and b): change in $S = (0.30, 0.90)$, while $Pr(D = 1) = 0.10$, C $= 0.95$. Middle panel (c and d): change in $C = (0.90, 0.99)$, while $Pr(D = 1) = 0.10$, S $= 0.60$. Lower panel (e and f): change in $Pr(D = 1) = (0.05, 0.30)$, while C $= 0.95$, S $= 0.60$. Each curve corresponds to a CRS with a different number of component tests $2 \leq (P - 1) \leq 10$)

negative subjects tends to result in decreasing the amount of bias in S_P^* (see Figure 7.3(a) vs 7.3(b)). This can be explained by the observations of Table 7.2 where we found that the specificity of the CRS decreases more slowly with each additional test in the presence of conditional dependence. It should be noted that in other settings, it is also possible S_P^* will be

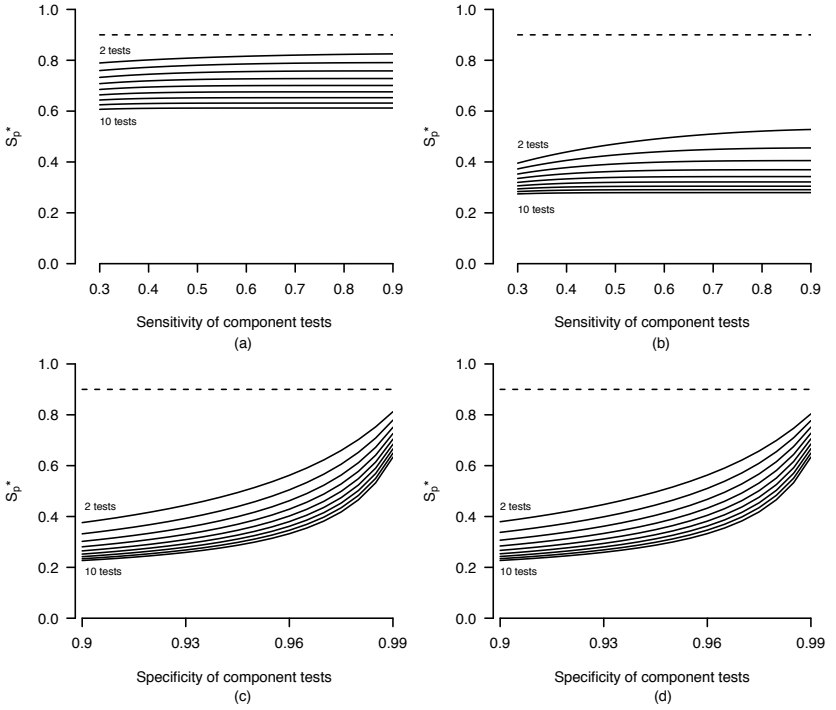

Figure 7.3: CRS-based sensitivity (S_P^*) against sensitivity of component tests (upper panels) and specificity of component tests (lower panels) while assuming conditional dependence between all tests. Each curve corresponds to a CRS with a different number of component tests $2 \leq (P-1) \leq 10$. $Pr(T_j = 1|L_1) - Pr(T_j = 1|L_2) = 0.1$ in upper panel. $Pr(T_j = 1|L_3) - Pr(T_j = 1|L_4) = 0.053$ in lower panel. Details: a. $Pr(D = 1) = 0.10$, $C = 0.95$ ($Pr(T_{ij} = 1|L_3) = 0.73$ and $Pr(T_{ij} = 1|L_4) = 0.01$); b. $Pr(D = 1) = 0.10$, $C = 0.95$ ($Pr(T_{ij} = 1|L_3) = 0.06$ and $Pr(T_{ij} = 1|L_4) = 0.049411764$; c. Change in $C = (0.90, 0.99)$, while $Pr(D = 1) = 0.10$, $S = 0.60$ ($Pr(T_{ij} = 1|L_1) = 0.61$ and $Pr(T_{ij} = 1|L_2) = 0.59$); d. Change in $C = (0.90, 0.99)$, while $Pr(D = 1) = 0.10$, $S = 0.60$ ($Pr(T_{ij} = 1|L_1) = 0.61$ and $Pr(T_{ij} = 1|L_2) = 0.59$).

overestimated [54]. For instance, changing only $Pr(T_j = 1|L_3) = 0.90$ and $Pr(T_j = 1|L_4) = 0$, it can be shown from Equation (7.11) that S_P^* will overestimated, tending to the limit of 0.916 with increasing number of component tests.

Figure 7.4 shows that greater conditional dependence within disease negative will also impact C_P^*, resulting in an overestimate compared to C_P (Figure 7.4(a) vs7.4(b)). The reason for the over-estimation can be understood by examining equation (7.12). We set $Pr(T_j = 1|L_4) < Pr(T_j = 1|L_k)$ $k = 1, 2, 3$, implying that as the sensitivity of the component tests increases $Pr(T_j = 0|L_4) >> Pr(T_j = 0|L_k)$, $k = 1, 2$. Therefore, $P(T_{iP} = 0|CRS = 0)$ is over-estimated due to the increasing influence of $Pr(T_j = 0|L_4)$ with higher sensitivities. Similarly, greater conditional dependence within disease positive can also contribute to overestimation in C_P^*

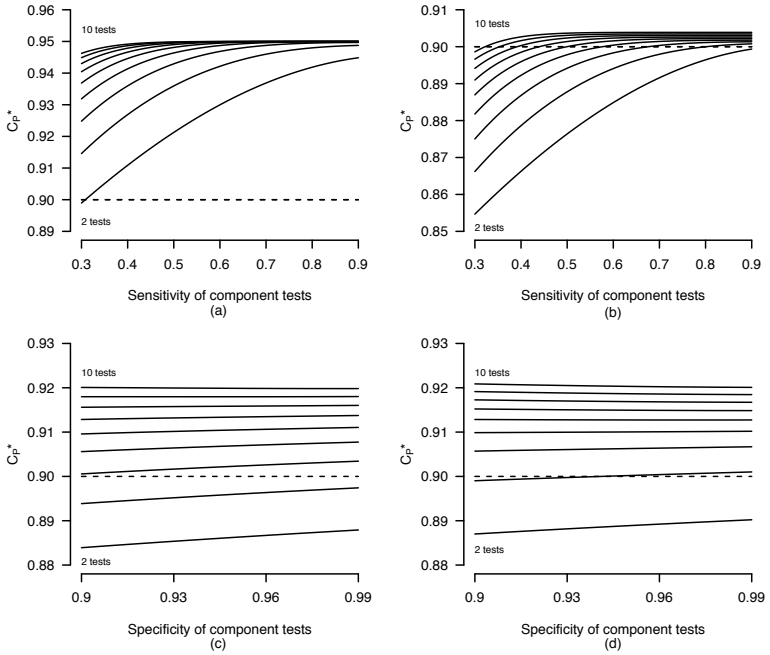

Figure 7.4: CRS-based specificity (C_P^*) against sensitivity of component tests (upper panels) and specificity of componente test (lower panels) while assuming conditional dependence between all tests. Each curve corresponds to a CRS with a different number of component tests $2 \leq (P-1) \leq 10$. $Pr(T_j = 1|L_1) - Pr(T_j = 1|L_2) = 0.1$ in upper panel. $Pr(T_j = 1|L_3) - Pr(T_j = 1|L_4) = 0.053$ in lower panel. Details: a. $Pr(D = 1) = 0.10$, $C = 0.95$ ($Pr(T_{ij} = 1|L_3) = 0.73$ and $Pr(T_{ij} = 1|L_4) = 0.01$); b. $Pr(D = 1) = 0.10$, $C = 0.95$ ($Pr(T_{ij} = 1|L_3) = 0.06$ and $Pr(T_{ij} = 1|L_4) = 0.049411764$); c. $Pr(D = 1) = 0.10$, $S = 0.60$ ($Pr(T_{ij} = 1|L_1) = 0.80$ and $Pr(T_{ij} = 1|L_2) = 0.40$); d. $Pr(D = 1) = 0.10$, $S = 0.60$ ($Pr(T_{ij} = 1|L_1) = 0.61$ and $Pr(T_{ij} = 1|L_2) = 0.59$).

(Figure 7.4(c) and 7.4(d)).

For example, suppose that when the CRS comprises 3 conditionally independent tests having sensitivity $S = 0.90$ and specificity $C = 0.95$, and true disease prevalence is 0.10, $S_P^* = 0.45$ is considerably lower than the true value of S_P (Figure 7.2(a) but $C_P^* = 0.8998 \approx C_P$ (7.2(b)). If the tests were conditionally dependent, the bias in S_P^* may be decreased (Figure 7.3), however C_P^* will be an overestimate (Figure 7.4). Though the percentage bias in C_P^* is seemingly small, it could translate into a large underestimate of the false-positive percentage attributable to the index test in a low prevalence setting.

Impact of increasing number of tests in the CRS

For the scenarios considered, as the number of component tests in the CRS increases, S_P^* tends to be increasingly underestimated (Figures 7.2 and 7.3). On the other hand, with increasing component tests C_P^* may either be estimated better provided the tests are conditionally independent or tends to become overestimated if the tests are conditionally dependent (Figures 7.2 and 7.4).

As the number of component tests increases, more subjects will be classified by the CRS as having the disease. Though in practice the number of component tests $(P-1)$ in the CRS will be finite and relatively small, it is instructive to see what happens when the number of component tests increases infinitely $(P-1 \rightarrow \infty)$. Provided the set of conditions for achieving unbiased S_P^* and C_P^* listed under equations 7.5 and 7.6 are not met, we will observe the asymptotic results described below.

From equation (7.7) it can be seen that as $lim_{P-1\rightarrow\infty}\ Pr(CRS=1)=1$. Accordingly, under the assumption of conditional independence, it can also be shown that $lim_{P-1\rightarrow\infty} Pr(T_P = 1|CRS=1) = Pr(T_P=1) = 0.18$, which is substantially smaller than $S_P = 0.90$. Paradoxically, $lim_{P-1\rightarrow\infty} Pr(T_P=0|CRS=0) = Pr(T_P=0|D=0) = C_P$, meaning the CRS-based specificity of the index test converges to an unbiased estimate in a situation where in fact it is not possible to estimate specificity, as the number of subjects classified as disease negative converges to zero.

A re-examination of equation (7.5) helps understand the limiting behavior of S_P^*. It can be shown that $lim_{P-1\rightarrow\infty}\{1-\prod_{j=1}^{P-1}(1-Pr(T_j=1|D=d))\}=1$. Therefore, for a sufficiently large number of component tests, $S_P^* = Pr(T_P=1|CRS=1) \approx \sum_{d=0}^{1} Pr(T_P=1|D=d)Pr(D=d) = Pr(T_P=1)$, implying that as the number of component tests increases their accuracy has decreasing influence on the estimate of index test sensitivity. As we have chosen $C_j > 1-S_j$, then $\prod_{j=1}^{P-1} Pr(T_j=0|D=0) = \prod_{j=1}^{P-1} C_j = C^{P-1}$ decreases toward 0 much more slowly than the term $\prod_{j=1}^{P-1} Pr(T_j=0|D=1) = \prod_{j=1}^{P-1}(1-S_j) = (1-S_j)^{P-1}$. Therefore, all terms of Equation (7.6) except $Pr(T_P=0|D=0)$ become negligible and C_P^* tends toward the true specificity C_P as $P-1$ increases (even though it cannot be estimated).

When there is strong conditional dependence in the disease negative group, as the number of component tests in the CRS increases, $lim_{P-1\rightarrow\infty}\{\prod_{j=1}^{P-1}(1-Pr(T_j=1|L_k))\}=0,\ k=1,2,3,4$. However, the limit converges much faster for $k=1,2,3$ than for $k=4$, hence the index test specificity C_P^* converges to $Pr(T_P=0|CRS=0) \approx Pr(T_P=0|L_4) = 0.95$, which is an over-

estimate.

Returning to evaluation of a test for C. trachomatis infection

In the light of the preceding sections, we can see that the decline in $\hat{S}^*_{PCRU}$ and increase in $\hat{C}^*_{PCRU}$ seen in Table 7.1 is to be expected with increasing number of component tests in the CRS. Given CULC is supposed to have near perfect specificity but only moderate sensitivity, CULC positive patients are probably a subset of disease positive patients with a higher organism load. Since the sensitivity of the PCR test is also affected by the organism load, the sensitivities of the two tests are likely to be conditionally dependent. Thus the CULC-based sensitivity estimate could be an over-estimate (if the conditional dependence was sufficiently high) or an underestimate. The CULC-based specificity estimate is probably an underestimate due to the imperfect sensitivity of CULC, and may thus serve as a lower bound of the true PCR specificity.

The CRSs in Table 7.1 may be constructed in practice with the expectation they improve over CULC and will add insight, particularly by narrowing the range of uncertainty around the accuracy of the specificity of PCR. However, as we can see from our simulations, with each additional component test CRS sensitivity will improve but with a loss of specificity, given the component tests in the example in Table 1 do not have perfect specificity. Therefore we can expect that increasing $P-1$ will probably result in increasing underestimation of $\hat{S}^*_{PCRU}$. Further, with the addition of NAATs to the CRS, we can expect conditional dependence within disease negative of the CRS and the PCR test. Therefore, the $\hat{C}^*_{PCRU}$ values in Table 1 could be either underestimates or overestimates though it is unknown at what value of $P-1$ the direction of bias changes. Thus, though the definition of a CRS may be considered 'transparent', the resulting estimates are highly likely to be biased and are not easily interpretable even as lower or upper bounds of the new test's accuracy.

Other decision rules for defining a CRS

CRS based on the AND decision rule

While our focus so far has been on the OR rule motivated by the case of *Chlamydia trachomatis* testing, other alternative compositions of the CRS exist. For other applications, the OR-rule may be seen as being too liberal requiring only a single positive component test to classify a patient as true positive. At the other extreme, a CRS based on the AND rule (denoted CRS_a) would be very conservative requiring all tests to be positive as follows

$$CRS_a = I(T_1, \cdots, T_{P-1}) = \begin{cases} 1 & \text{if } min(T_1, \ldots, T_{P-1}) = 1 \\ 0 & \text{if } min(T_1, \ldots, T_{P-1}) = 0, \end{cases}$$

As we will show below, the expressions for the accuracy of this CRS and for the estimated accuracy of an index test with respect to this CRS are symmetric to the expressions defined previously in Section 3. By replacing C_P by S_P, S_P by C_P, C_j by S_j, S_j by C_j, and $1 - Pr(D = 1)$ by $Pr(D = 1)$ in equations (7.3), (7.4), (7.5) and (7.6), we can obtain the corresponding expressions for the case when all tests are conditionally independent. The sensitivity of CRS_a is given by

$$\begin{aligned} S_{CRS_a} &= Pr(CRS_a = 1|D = 1) = Pr(min(T_1, \ldots, T_{P-1}) = 1|D = 1) \\ &= Pr(T_1 = 1, \ldots, T_{P-1} = 1|D = 1) \\ &= \prod_{j=1}^{P-1} Pr(T_j = 1|D = 1) = \prod_{j=1}^{P-1} S_j, \end{aligned} \tag{7.13}$$

the specificity of CRS$_a$ is given by

$$\begin{aligned} C_{CRS_a} &= Pr(CRS_a = 0|D = 0) = Pr(min(T_1, \ldots, T_{P-1}) = 0|D = 0) \\ &= 1 - Pr(min(T_1, \ldots, T_{P-1}) = 1|D = 0) = 1 - Pr(T_1 = 1, \ldots, T_{P-1} = 1|D = 1) \\ &= 1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 0|D = 0)) = 1 - \prod_{j=1}^{P-1}(1 - C_j), \end{aligned} \tag{7.14}$$

the sensitivity of index test P with respect to CRS$_a$ is given by

$$\begin{aligned} S^*_{Pa} &= Pr(T_P = 1|CRS_a = 1) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 1|D = d) Pr(D = d) \prod_{j=1}^{P-1} Pr(T_j = 1|D = d)}{\sum_{d=0}^{1} Pr(D = d) \prod_{j=1}^{P-1} Pr(T_j = 1|D = d)}, \end{aligned} \tag{7.15}$$

and the specificity of index test P with respect to CRS$_a$ is given by,

$$\begin{aligned} C^*_{Pa} &= Pr(T_P = 0|CRS_a = 0) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 0|D = d)\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 0|D = d))\} Pr(D = d)}{\sum_{d=0}^{1}\{1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 0|D = d))\} Pr(D = d)}. \end{aligned} \tag{7.16}$$

Due to the symmetry, S_{CRS_a} decreases and C_{CRS_a} increases as the number of component tests becomes large. If we were to use an AND decision rule for the situation in Table 7.2 (i.e. each component tests having $S = 0.6$ and $C = 0.9$), S_{CRS_a} would drop from 0.36 with 2 component tests to 0.0060 with 10 component tests. Correspondingly, C_{CRS_a} would increase from 0.9975 to 1. The expressions in sections 7 can similarly be converted into equivalent expressions for CRS_a.

CRS based on the 'at least two positive tests' decision rule

The OR and AND rules can be seen as the boundaries of a family of decision rules which are defined by a positivity criterion of the form "at least m positive component tests" as follows

$$CRS_m = I(T_1, \cdots, T_{P-1}) = \begin{cases} 1 & \text{if } \sum_{j=1}^{P-1} T_j \geq m \\ 0 & \text{if } \sum_{j=1}^{P-1} T_j < m, \end{cases}$$

for $1 < m < P - 1$. We can see that if $m = 1$ we have the OR-decision rule and if $m = P - 1$ the AND-decision rule. We will examine in some detail the rule that requires at least two positive tests to classify a patient as positive, i.e. when $m = 2$. CRS_2 is defined as follows

$$CRS_2 = I(T_1, \cdots, T_{P-1}) = \begin{cases} 1 & \text{if } \sum_{j=1}^{P-1} T_j \geq 2 \\ 0 & \text{if } \sum_{j=1}^{P-1} T_j < 2, \end{cases}$$

Assuming conditional independence, the sensitivity of CRS_2 is given by,

$$\begin{aligned} S_{CRS_2} &= Pr(CRS_2 = 1|D = 1) = Pr(\sum_{j=1}^{P-1} T_j \geq 2|D = 1) \\ &= 1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|D = 1)) - \sum_{l=1}^{P-1} Pr(T_l = 1|D = 1) \prod_{l \neq j}(1 - Pr(T_j = 1|D = 1)) \\ &= 1 - \prod_{j=1}^{P-1}(1 - S_j) - \sum_{l=1}^{P-1} S_l \prod_{l \neq j}(1 - S_j), \end{aligned} \tag{7.17}$$

and the specificity of the CRS_2 is given by,

$$\begin{aligned} C_{CRS_2} &= Pr(CRS_2 = 0|D = 0) = Pr(\sum_{j=1}^{P-1} T_j < 2|D = 0) \\ &= \prod_{j=1}^{P-1} Pr(T_j = 0|D = 0) + \sum_{l=1}^{P-1} (1 - Pr(T_l = 0|D = 0)) \prod_{l \neq j} Pr(T_j = 0|D = 0) \\ &= \prod_{j=1}^{P-1} C_j + \sum_{l=1}^{P-1} (1 - C_l) \prod_{l \neq j} C_j. \end{aligned} \tag{7.18}$$

The sensitivity of index test P with respect to the CRS_2 is given by,

$$\begin{aligned} S^*_{P2} &= Pr(T_P = 1|CRS_2 = 1) \\ &= \frac{S_P Pr(D = 1) S_{CRS_2} + (1 - C_P)(1 - Pr(D = 1))(1 - C_{CRS_2})}{Pr(D = 1) S_{CRS_2} + (1 - Pr(D = 1))(1 - C_{CRS_2})}, \end{aligned} \tag{7.19}$$

and the specificity of index test P with respect to the CRS_2 is given by,

$$\begin{aligned} C^*_{P2} &= Pr(T_P = 0|CRS_2 = 0) \\ &= \frac{(1 - S_P) Pr(D = 1)(1 - S_{CRS_2}) + C_P(1 - Pr(D = 1)) C_{CRS_2}}{Pr(D = 1)(1 - S_{CRS_2}) + (1 - Pr(D = 1)) C_{CRS_2}}. \end{aligned} \tag{7.20}$$

As with the other decision rules defined previously, the sensitivity of CRS_2 depends only on the component tests' sensitivities, while the specificity of CRS_2 depends only on the component tests' specificities. Similar to the OR rule, as we increase the number of component tests, CRS_2's sensitivity will increase while its specificity will decline, though the change is at a slower pace. For the settings in Table 7.2 , i.e. each component test having $S = 0.60$ and $C = 0.95$, we have $S_{CRS2} = 0.36$ and $C_{CRS2} = 0.9975$. Note that this happens to be identical to the AND rule for the particular case of two component tests. If we increase the number of component tests to 10, then we would have $S_{CRS2} = 0.9983$ and $C_{CRS2} = 0.9139$ respectively.

Comparison of index test accuracy estimates based on different CRS decision rules

Figure 7.5 compares the CRS-based estimates of sensitivity (S^*_P) and specificity (C^*_P) based on three different CRS decision rules: the OR rule (black lines), the AND rule (red lines) and the 'at least two positive tests' rule (blue lines). In each colour, the solid and dashed lines correspond to a CRS based on 3 and 10 component tests, respectively. The true values of the index test accuracy are $S_P = 0.9$ and $C_P = 0.9$. In all plots we see that the OR and AND rules

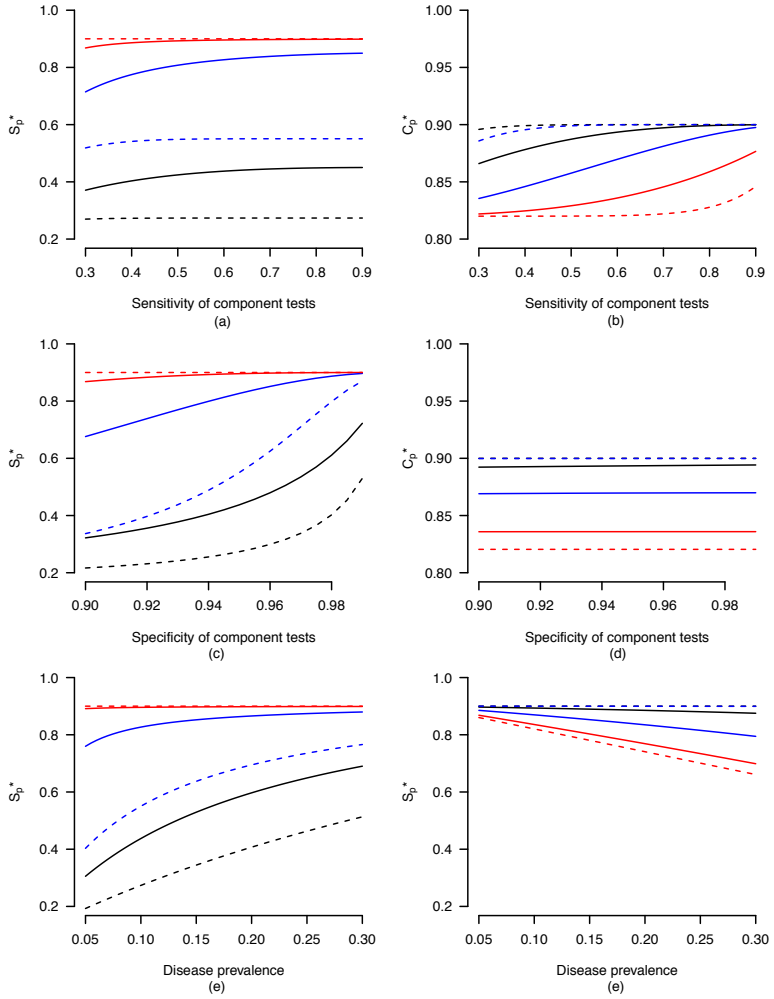

Figure 7.5: A comparison of estimated sensitivity (S_P^*) and estimated specificity (C_P^*) based on 3 different decision rules vs accuracy of the component tests and disease prevalence while all tests are conditionally independent (true parameter values are $S_P = 0.9$ and $C_P = 0.9$ in all cases). Black lines = OR decision rule, blue line = at least 2 positive tests decision rule, red = AND decision rule. Solid line = CRS with 3 component tests, dashed line = CRS with 10 component tests.Upper panel (a and b): $Pr(D = 1) = 0.10$, C $= 0.95$. Middle panel (c and d): $Pr(D = 1) = 0.10$, S $= 0.60$. Lower panel (e and f): C $= 0.95$, S $= 0.60$.

define boundary estimates, and the 'at least two positive tests' rule is intermediate between them. The most bias in S_P^* (left panel) is observed under the OR decision rule (black curve) while the least bias can be seen under the AND rule (red curve), irrespective of the number of component tests used. By symmetry, the AND rule will create the most bias in C_P^* (right panel) while the OR decision rule will create the least bias, among the family of rules defined in the preceding subsection.

Discussion

We studied in detail the performance of a composite reference standard (CRS) based on an OR decision rule. This type of CRS has been used in diagnostic research studies to improve the sensitivity in identifying the disease of interest over any single imperfect reference test. We showed that even if all component tests have excellent performance, e.g. with 0.90 sensitivity and 0.95 specificity, the resulting estimates of index test accuracy may be highly biased. In practice, the magnitude and direction of the biases will be difficult to quantify precisely as they depend on the unknown accuracy of the component tests, the disease prevalence and the degree of conditional dependence between the tests.

The definition of the CRS does not involve the test under evaluation. Therefore, it is perceived as being 'independent' of the test under evaluation [9, 22]. However, as we have shown, conditional dependence between component and index tests could arise due to their common dependence on a variable besides the true disease status. Intuitively, one can imagine that if the new test and the CRS systematically make the same errors (i.e. are conditionally dependent), the accuracy of the new test will be over-estimated. Further, we showed that CRS-based estimates of sensitivity and specificity are dependent on the true unknown disease prevalence. Therefore the same CRS applied in different prevalence settings to evaluate the same index test would give different estimates of sensitivity and specificity, unlike the true sensitivity and specificity of the index test which are mathematically independent of disease prevalence.

Our simulation settings were inspired by those encountered in *C. trachomatis* testing, but the results may be generalized to other settings. It should be noted that other settings may result in different biases, e.g. using the same sensitivities and specificities for the component tests as we have but increasing the prevalence to > 0.50 could result in a greater magnitude of bias in S_P^* compared to C_P^*, and greater impact of conditional dependence within disease positive.

The sensitivity of a CRS based on the OR rule will increase with an increase in the number of component tests, though this will be at the cost of a loss of specificity unless all tests have perfect specificity. In the limiting case when the number of component tests tends to infinity we found the paradoxical result that the estimated prevalence tends to 100% while the CRS-based specificity tends to the true value of the specificity. Further, the estimated sensitivity tends to the probability of a positive test, suggesting the properties of the component tests in the CRS play no role in its estimation. Thus as more information becomes available due to results of multiple tests being gathered, the performance of the CRS may worsen. This is in contrast to the expected performance of a well-defined statistical method. The apparent paradox can be

explained by the fact that the CRS is overly simplistic in its construction. It ignores information on inter-relations between the component tests and reduces their joint results to a dichotomous result.

It has been argued that the CRS is clinically meaningful because it represents a clinical diagnosis and not the true disease status, which is impossible to determine in the absence of a gold-standard [135]. In a clinical setting, a physician faced with results of multiple imperfect tests may use them in a composite decision rule after weighing the risks of missed diagnosis vs. overdiagnosis. But it is questionable whether the same composite rule should be applied to the evaluation of a new test or estimation of disease prevalence with no attempt to correct for the false-positive or false-negative errors in the CRS. It is rather like requiring that the new test replicate the errors of the CRS. Further, clinical diagnosis is more complex than a simple composite decision rule, taking into consideration additional variables, e.g. disease history, and the particular combination of positive and negative results.

Based on our findings, we recommend that a CRS based on combining results of NAATs via an OR or AND decision rule (or other variations) should not be used for estimating the accuracy of new *C. tracomatis* tests as: i) the component tests are not guaranteed to have perfect sensitivity or specificity, ii) this approach would ignore the conditional dependence between the tests, iii) this approach would ignore the inter-relation between the different component tests and the index test.

The problem of evaluating a new test or estimating disease prevalence in the absence of a gold-standard reference remains a challenging one. CRSs have been promoted as a better approach than alternatives like latent class models (LCMs) [9, 135]. Yet, as we have shown, CRSs based on an any-positive (or all positive rule) suffer from a number problems that have not been acknowledged previously. Other CRSs based on more complex any-positive rules have also been found to result in bias [79]. We conclude that future research in this area should be directed towards approaches based on realistic statistical modeling of the observed data. Such models should take into account: i) the inter-relations between all component tests and the index test, ii) model conditional dependence between all tests (component and index), iii) model disease prevalence, and iv) incorporate external information if available, thus making complete use of the collected data while acknowledging all the different parameters (prevalence and individual test accuracies) that may come into play [52, 53, 101]. While there are acknowledged challenges with estimating more complex statistical models [52, 168], improving our understanding of them is necessary to make optimal use of the data gathered.

Chapter 8

No rationale for 1 variable per 10 events criterion when considering sample size for binary logistic regression analysis

Abstract

The number of events per variable (EPV) is considered a key factor in the performance of a binary logistic regression analysis. Ten EPV is a widely advocated minimal criterion for sample size considerations in logistic regression analysis. However, of three previous simulation studies that examined this minimal EPV criterion, only one supports and recommends the use of a minimum of 10 EPV. In this paper, we examine the potential reasons for the large heterogeneity in results between these extensive simulation studies that studied the minimal EPV criterion for binary logistic regression. We show that, besides EPV, the problems associated with low EPV (bias, non-nominal coverage of confidence intervals), depend on other factors such as the total sample size. We also demonstrate that simulation results can be dominated by even a few simulated data sets for which the prediction of the outcome by the covariates is perfect. This issue is known as 'separation'. We reveal that different approaches for identifying and handling this separation problem leads to substantially different simulation results. We further show that a simple correction method can be used to improve the accuracy of regression coefficients and alleviate the problems associated with separation. We conclude that the current evidence supporting EPV rules for binary logistic regression is weak and inconsistent. Given our findings, there is an urgent need for new research to provide guidance for supporting sample size considerations when using logistic regression techniques.

The number of subjects in the smallest of two outcome groups ('number of events') relative to the number of regression coefficients estimated (excluding intercept) has been identified as a key factor driving the performance of binary logistic regression models [84, 85, 155]. This ratio is known as Events Per Variable (EPV). In small data sets, where EPV is low, estimated associations between covariates and the outcome are often imprecise and biased in the direction of more extreme values [68, 98, 127]. Similarly, prediction models built using logistic regression in small data sets will lead to poor predictions that are too extreme and uncertain [10, 85, 155, 169]. A value of 10 EPV is a widely adopted minimal criterion for sample size considerations for binary logistic regression analysis [121, 122, 133].

Despite the wide acceptance of the minimal 10 EPV rule, the results of three well-known simulation studies examining the minimal EPV criterion for binary logistic regression models are highly heterogeneous [45, 134, 171]. These heterogeneous simulation results have in turn led to conflicting minimal EPV recommendations in these papers. Only the study of Peduzzi et al. [134] supports the 10 EPV rule, after concluding that 'no major problem occurred' if EPV exceeds 10. In contrast, Vittinghoff and McCulloch [171] have argued that an EPV of 10 as a minimal guideline criterion is too conservative, showing that severe problems mainly occur in the EPV = 2 to EPV = 4 range. Conversely, Courvoisier et al. [45] showed that substantial problems may still occur 'even if the number of EPV exceeds 10'. They showed that the performance of the logistic model may depend on various factors other than EPV, including the true strength of associations between covariates and outcome and the correlation between covariates.

We aim to explain the heterogeneity between minimal EPV recommendations from previous simulation studies [45, 134, 171]. In line with to these earlier studies, we focus on the accuracy of logistic regression coefficients (i.e., logit coefficients) in low EPV settings. Two issues are known to complicate the interpretation of logit coefficients in this setting. First, the estimation of logit coefficients that are estimated by maximum likelihood can be inaccurate when EPV is low. Second, 'separation' is likely to occur in low EPV settings. When separation occurs, the logit coefficients that are estimated by maximum likelihood may not be estimable at all. We first briefly discuss each of these two issues.

Accuracy of logit coefficients in small samples

In a typical binary logistic regression analysis, the strength of associations between covariates and outcome are quantified by the logit coefficients, which are estimated by the method of

maximum likelihood. While these estimators of the (adjusted) log-odds ratio have attractive asymptotic properties (e.g., unbiasedness and consistency), these properties have been shown not to apply in small samples. For example, the logit coefficients suffer from small sample bias [68,98], leading to systematically overestimated associations. Also, asymptotic confidence intervals often do not have nominal coverage rates in studies with small data sets [33, 134]. Both problems are expected to decrease with increasing sample size and increasing EPV.

The inaccuracies in the coefficients and corresponding confidence intervals lead to inaccurate inferences about the true covariate-outcome associations. Hereafter we refer to these problems as 'inaccuracy in logit coefficients'.

Separation

A second source of difficulty may occur when a single covariate or a linear combination of multiple covariates perfectly separates all events from all non-events [4,89]. This phenomenon is referred to as 'separation' or 'monotone likelihood' (illustrated in Figure 8.1). Estimating a logistic regression model by maximum likelihood on a 'separated data set' leads to non-unique point estimates and standard errors of coefficients near the extremes of parameter space [88]. On separated data, convergence of the iterative maximum likelihood estimation procedure may not be achieved as the upper bound on the number of iterations is reached ('non-convergence'). Or, the solution may converge to a point that is not the maximum likelihood [4]. Because convergence criteria will often differ between software programs, estimates can vary considerably between software programs when fitting a logistic model on separated data.

The probability of separation occurring increases with decreasing sample size and increasing number of covariates. Hence, separation is likely to occur in low EPV data sets. In simulation studies, including those that examined the minimal EPV criterion for binary logistic regression, the occurrence of separated data sets has typically been treated as a nuisance. Researchers remove the simulation data set when separation is detected. Doing so, however, a non-random subset of simulated data sets will be missing when analyzing the simulation results: particularly those data sets with strong associations between the covariates and the outcome [156]. The approaches to identify and handle separated data may therefore strongly affect the results and inferences of simulation studies.

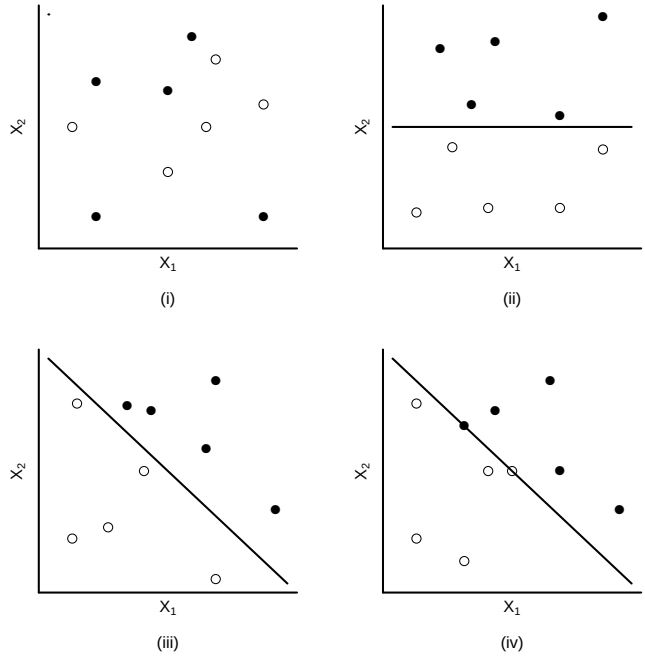

Figure 8.1: Graphical representation of separation (complete and quasi-complete) adapted from Albert and Anderson [4]. Sample points for two variables X_1 and X_2 by outcome (Y): open and filled circles represent different levels of the outcome ($Y = 0$ or 1). (i) No separation; (ii) complete separation by variable X_2; (iii) complete separation by variables X_1 and X_2; (iv) quasi-complete separation by variable X_1 and X_2.

Outline of the paper

In simulation studies involving small samples and low EPV, some degree of inaccuracy in logit coefficients and separation is likely to coexist. Simulation results will therefore reflect the net effect of inaccurate estimation and handling of separated data sets. To gain insight into both problems separately, we will first investigate the factors driving the accuracy of logit coefficients by examining scenarios where drawing separated data sets is highly unlikely (part I). Next, we examine a range of scenarios where the probability of drawing a separated data set is substantially larger than zero (part II). In part II, we monitor the variations in simulation results due to different approaches of detecting and handling separated data sets. In both parts we will evaluate whether a simple-to-apply penalized estimation procedure suggested by Firth [61, 89] in combination with profile likelihood based confidence intervals can effectively improve the accuracy of logit coefficients in small samples. In the discussion, we will return to the differences in results of the previous minimal EPV simulation studies [45, 134, 171] using the findings from our simulations.

Methods

General methods

For each simulated data set, N covariate vectors $X_1, ..., X_P$ were drawn from either an independent multivariate normal distribution (in part I and part II) or an independent Bernoulli distribution (in part II). The outcome variable (Y) for each covariate vector was generated from a Bernoulli distribution with a covariate vector specific probability derived by applying the logistic function using the true values of the data generating model on the simulated covariate data. The data generating models only included first order covariate (main) effects, thus were of the form: $\text{logit}(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_P X_P$.

On each generated data set we fitted the logistic regression model by maximum likelihood that had the same form as the data generating model (i.e., fitting the correctly specified logistic regression model). We additionally applied the modified score equations procedure suggested by Firth [61] that removes a portion of the small sample bias that can be anticipated in the maximum likelihood estimates, by introducing a penalty on the likelihood. The penalty function is a Jeffries invariant prior [61]. An additional advantage of Firth's correction is that its coefficients, $\hat{\beta}_1^F, ..., \hat{\beta}_P^F$, are finite even when estimated on a data set that is separated.

We examined the accuracy of estimating a primary logit coefficient, arbitrarily taking the coefficient for the first covariate, $\hat{\beta}_1$, as the primary one. Based on guidance by Burton et al. [36], we calculated the following quantities: i) bias in the primary coefficient, defined by: $\bar{\hat{\beta}}_1 - \beta_1$, where $\bar{\hat{\beta}}$ is the arithmetic mean of $\hat{\beta}_1^{ML}$ or $\hat{\beta}_1^F$ over all simulated data sets; ii) relative bias in the primary coefficient, defined by $(\bar{\hat{\beta}}_1 - \beta_1)/\beta_1$, iii) coverage of the 90% confidence interval by calculating for each data set the Wald confidence interval by $\hat{\beta}_1^{ML} \pm 1.645 \times \text{SE}(\hat{\beta}_1^{ML})$, where $\text{SE}(\hat{\beta}_1^{ML})$ is the estimated (ML) standard error for $\hat{\beta}_1^{ML}$. For $\hat{\beta}_1^F$ we estimated the profile likelihood 90% confidence interval [88]; iv) average 90% confidence interval width, defined by average of the difference between the upper and lower bounds of the 90% confidence intervals; v) mean square error (MSE): $(\bar{\hat{\beta}}_1 - \beta_1)^2 + (\text{SD}(\hat{\beta}_1))^2$, where $\text{SD}(\hat{\beta}_1)$ is the standard deviation of $\hat{\beta}_1^{ML}$ or $\hat{\beta}_1^F$ over the simulation data sets.

Simulation procedures

In total, 465 different simulation scenarios were examined. For each of these scenarios, $10,000$ data sets were generated using `R` software version 3.1.1 [141]. For each data set, sampling was continued until the prespecified criteria for sample size and the number of events were met, keeping the first events and non-events generated up to the required number of each. This

procedure ensured a fixed sample size (N) and number of events (EPV) in each data set. This approach, which is equivalent to the approach used by Vittinghoff and McCulloch [171], takes advantage of the properties of the logistic model where only the intercept is affected by this sampling procedure.

The logistic regression models fitted by maximum likelihood and Firth's correction were implemented using the `glm` function in the `stats` library (version: 3.1.1) and the `logistf` function in the `logistf` library (version: 1.21), respectively. To identify separation of simulation data sets the maximum likelihood standard errors of parameters were monitored through a re-estimation process [109]. This procedure is explained in detail in the Appendix. Unless otherwise specified: the default software criteria for convergence were used, calculation of the regression coefficient accuracy measures were based only on converged simulation results and maximum likelihood estimates for data sets that exhibited separation were excluded from the calculation of simulation results.

Methods Part I: accuracy of logit coefficients

A series of scenarios were set-up to identify which factors are driving the accuracy of the logit coefficient. In this first part we limited ourselves to scenarios where the probability of drawing a separated data set was close to zero (maximum percentage separated data sets in a single simulation scenario of 0.3%; zero separated data sets in 98% of scenarios). To keep the probability of drawing a separated data set low, in part I, covariate data were sampled only from continuous (multivariate normal) distributions. Part I was further subdivided into four small-scale factorial simulation studies (Ia to Id). In study Ia, the role of EPV and the true value of β_1 on accuracy of logit coefficients was studied for the case of a single continuous covariate. The role of the number of covariates (P) was evaluated in study Ib. In study Ic, the role of the sample size was examined, reflecting the effect of increasing the number in the largest group (non-events). The role of covariate correlations was studied in study Id. Details of these four studies are summarized in Table 8.1.

Methods Part II: Detection and handling of separated data sets

In part II we evaluated the impact of different approaches for the detection and handling of separated data sets on simulation results and inferences. Two different simulation studies were conducted, which are explained below.

Table 8.1: Design factorial simulation studies Ia to Id.

	Study			
Factors	Ia	Ib	Ic	Id
Sample size				
EPV (with steps of)	15 to 150 (5)	15 to 150 (5)	6 to 30 (2)	6 to 30 (2)
Outcome prevalence	1/2	1/2	1/2,1/3,1/4,1/5,1/10	1/4
Effect size				
Value of e^{β_1}	1/4, 1/2, 1, 2, 4	2, 4	2	2
Value of $e^{\beta_j}, j > 1$	Not applicable	$\beta_1 = ... = \beta_P$	2	2
Covariates				
Number (P)	1	2, 3, 4	2	2
Distribution	(Multivariate) standard normal			
Correlation	Not applicable	0	0	.1, .15, .2, .25

IIa. Binary single covariate

In study IIa, we investigated the extent to which simulation results differ between using all simulated data sets (a naive approach, using the software output regardless of convergence status) versus removing all separated data sets for quantifying the accuracy of logit coefficients. We also explored how the simulation results in terms of bias are affected by replacing the results of separated data sets by the highest estimated coefficient on non-separated data (an ad-hoc approach). Data were sampled for a single binary covariate with probability of sampling either observation of .5. The manipulated factors were: EPV and the true value of β_1. We considered EPV values between 6 and 30, at incremental steps of size 2 and the values of the primary coefficient (β_1) were chosen as log(1), log(2) and log(4).

IIb. Single simulation scenario, continuous covariate

In study IIb, we evaluated the impact of using different methods to detect the presence of separated data sets. In the first approach we used likelihood non-convergence as a criterion for removing simulation data sets, as was done in previous studies [134, 171]. This type of non-convergence occurs when the tolerance convergence criterion is not met while the upper bound on the number of iterations is reached. We compare this convergence criterion to our (computationally intensive) method of separation detection (see Appendix), and to the method used by Courvoisier [45]: a simulation data set is removed if for any parameter $j \neq 0, |\hat{\beta}_j| >$log(50). To evaluate the effect of changing the likelihood criterion, four additional criteria for convergence tolerance (tol) and maximum number of Fisher scoring iterations (max-iter)

are used: tol: 1e-8, max-iter: 25 (`glm` function default), tol: 1e-6, max-iter: 25 (Type I), tol: 1e-10, max-iter: 25 (Type II), tol: 1e-10, max-iter: 50 (Type III). Univariate covariate data were generated from standard normal distribution, the ratio of events to non-events was kept constant at 1:1. EPV was fixed at 4 and $\beta_1 = \log(4)$.

Results Part I: accuracy of logit coefficients

Figure 8.2 shows the simulation results for study Ia. With traditional logistic regression (upper left panel), for true non-zero covariate-outcome associations the primary logit coefficient (β_1^{ML}) was biased towards more extreme values (away from zero). Bias decreased with increasing EPV through a non-linear function (that can be approximated by: $\log(|\text{bias}(\beta_1^{ML})|) = \lambda_0 - \lambda_1 \log(\text{EPV})$, where $\lambda_0 > 0$ and $\lambda_1 > 0$, for which the values depend on the simulation setting). Bias in the logit coefficient did not reduce strictly to zero even for EPV as large as 150. Bias depended on the true effect size of the coefficient with bias increasing in case of stronger associations. The figure further illustrates that bias is symmetric but in opposite directions for the conditions with the same true effect size (the effect of recoding the outcome variable: such that $\beta = \log(2)$ becomes $\beta = \log(1/2)$ and $\beta = \log(4)$ becomes $\beta = \log(1/4)$, or vice versa). Bias in Firth's estimator (β_1^F, upper right panel) was close to zero for all studied EPV values and across all true effect sizes.

The middle left panel in Figure 8.2 shows slight over-coverage of the 90% Wald-confidence interval for EPV $<$30. The profile likelihood confidence interval for Firth's estimator, however, was close to the nominal level for all studied conditions. The mean square error of β_1^{ML} and β_1^F decreased with true effect size and EPV. The mean square error for β_1^F was systematically lower than for β_1^{ML}.

The empirical sampling distributions of $\hat{\beta}_1^{ML}$ and $\hat{\beta}_1^F$ at EPV = 20 (study Ia) are presented in Figure 8.3. The sampling distributions show severe non-normality in case of non-zero covariate-outcome associations. The degree of non-normality increased with true effect size. The effect of Firth's correction is illustrated by comparing the distribution of $\hat{\beta}_1^F$ estimates to the $\hat{\beta}_1^{ML}$ distribution: the $\hat{\beta}_1^F$ estimates were shrunken towards zero; the magnitude of shrinkage was proportional to the estimated effect size. The arithmetic mean of the $\hat{\beta}_1^F$ distribution for a non-zero true association was closer to zero and the long tail (tail in the direction of stronger effect size) was smaller.

Figure 8.4 shows the relative bias under varying number of covariates (study Ib), sample size

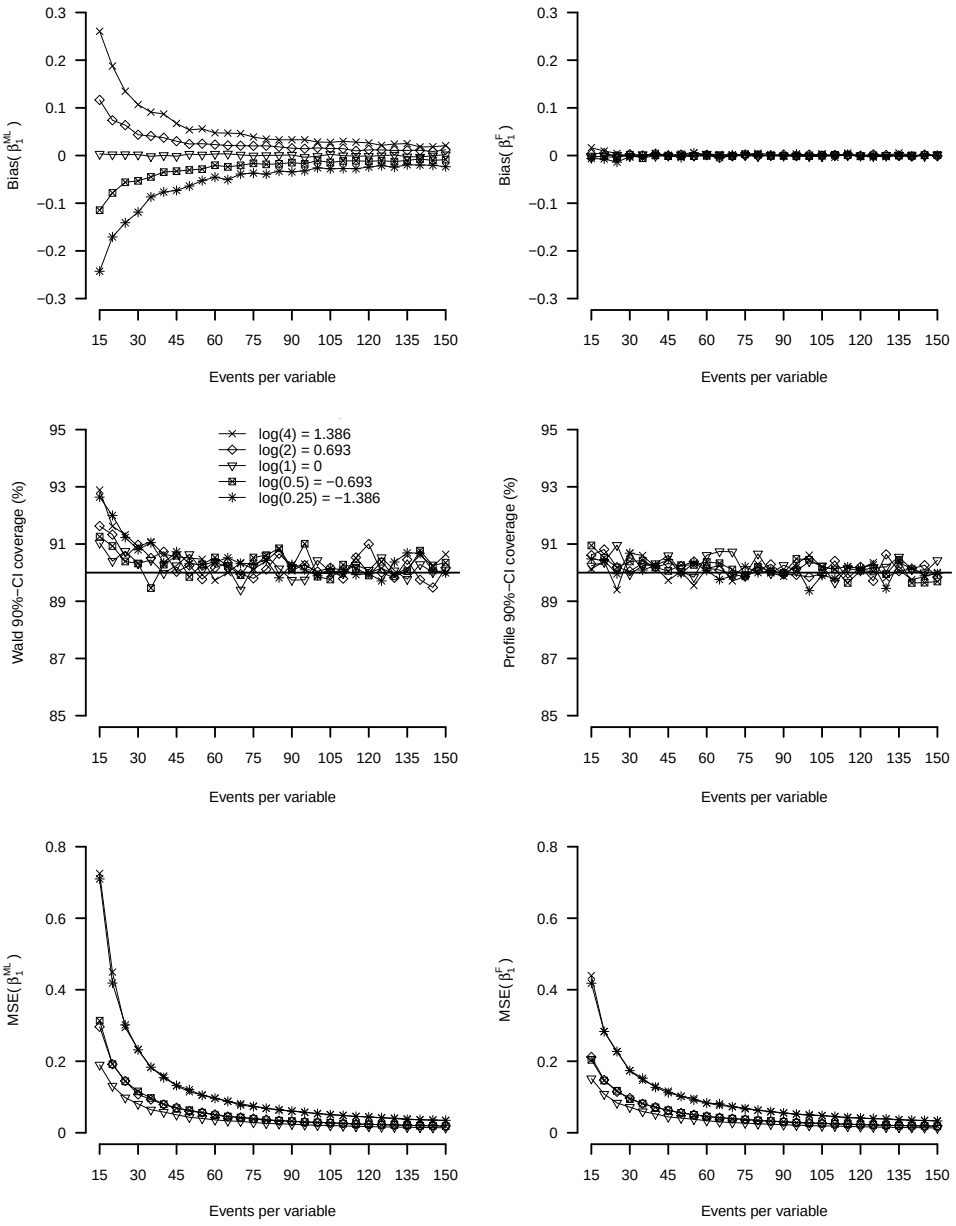

Figure 8.2: Results of simulation study Ia. Accuracy as a function of EPV and true value of the log-odds ratio (β_1). Left panel: maximum likelihood logistic regression, right panel: Firth's correction.

(study Ic) and covariate correlation settings (study Id). The maximum likelihood estimates were always biased away from zero. Bias decreased with the addition of more covariates and

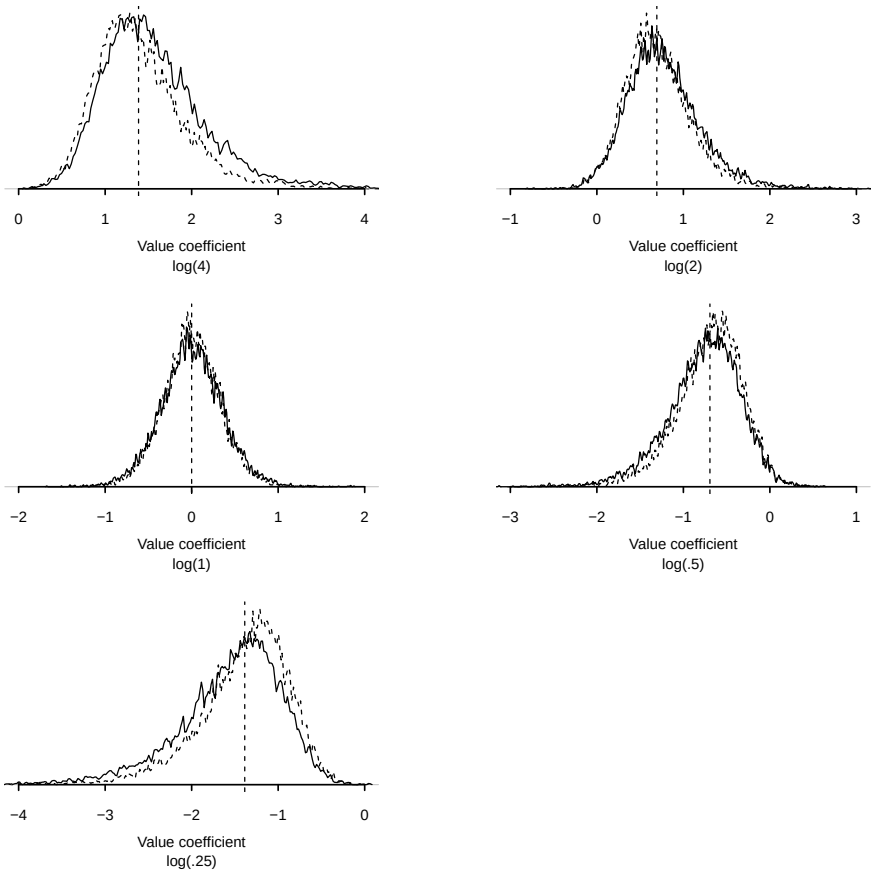

Figure 8.3: Density of estimated coefficients in simulation at EPV = 20 (study Ia) for different true values of the log-odds ratio. Vertical dashed line is true value of the regression coefficient. Solid line: maximum likelihood logistic regression; dashed line: Firth's correction.

was affected by the size of the true effect (Figure 8.4, upper panel) and the total sample size (Figure 8.4, middle panel). There was no apparent effect on bias by varying the amount of correlation between covariates in the model (Figure 8.4, lower panel). In each study and each simulation condition, β_1^F was a close to unbiased estimator.

Table 8.2 summarizes the results for the four factorial simulation studies. Average bias and average mean square error decreased with increasing EPV in case of maximum likelihood estimates. Average coverage for the maximum likelihood Wald confidence interval based and Firth's correction profile likelihood confidence intervals were close to nominal (90%) in most situations, with a small over-coverage in lower EPV settings (though not exceeding 93%). The

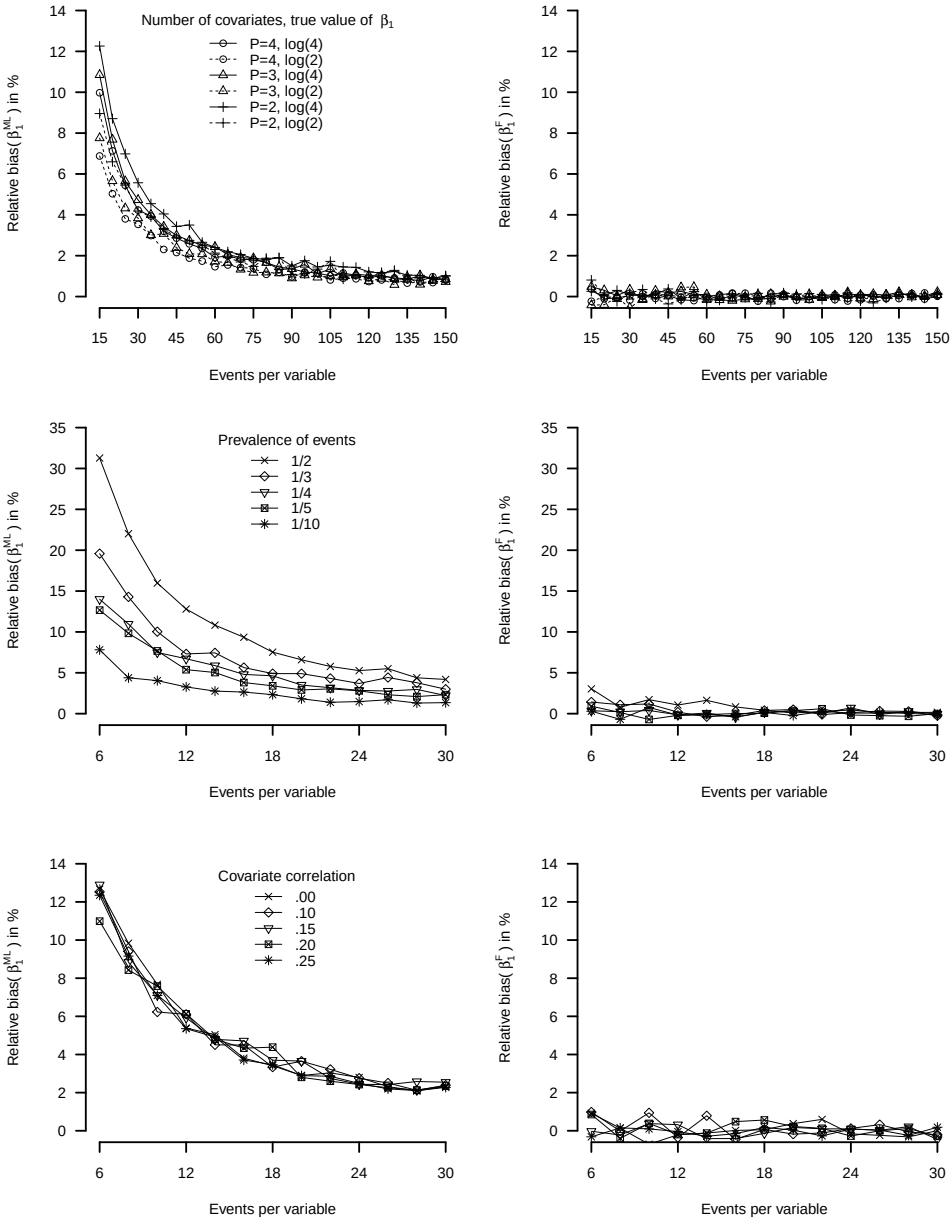

Figure 8.4: Relative bias simulation studies Ib, Ic, and Id. Left panel: maximum likelihood logistic regression, right panel: Firth's correction.

average width of the confidence intervals and mean squared error were systematically smaller after applying Firth's correction.

Table 8.2: Results simulation studies Ia to Id.

Study	Study Ia* and Ib						Study Ic and Id					
EPV	15 to 30		35 to 50		55 to 150		6 to10		12 to 18		20 to 30	
Estimator	β_1^{ML}	β_1^F	β_1^{ML}	β_1^F	β_1^{ML}	β_1^F	β_1^{ML}	β_1^F	β_1^{ML}	β_1^F	β_1^{ML}	β_1^F
Bias												
Average bias	0.084	0.002	0.038	0.001	0.016	0.000	0.069	0.002	0.033	0.000	0.020	0.000
max	0.261	0.016	0.091	0.005	0.056	0.006	0.217	0.021	0.075	0.011	0.046	0.005
min	0.025	-0.004	0.013	-0.002	0.004	-0.005	0.023	-0.005	0.016	-0.003	0.009	-0.003
Average relative bias (%)	7.8	0.1	3.6	0.1	1.5	0.0	8.4	0.4	4.8	0	2.9	0
max	18.8	1.2	6.6	0.5	4.0	0.5	31.2	3.0	10.8	1.6	6.5	0.7
min	3.5	-0.5	1.9	-0.3	0.5	-0.7	3.3	-0.7	2.3	-0.5	1.3	-0.0
$>$+10% relative bias (%)	18.8	0	0	0	0	0	37.5	0	3	0	0	0
Coverage 90% CI												
Average coverage (%)	90.4	90.1	90.2	90.2	90.1	90.0	90.4	90.3	90.2	90.2	90.1	90.2
max	92.9	90.8	91.1	90.7	91.0	90.7	92.1	91.2	90.8	90.6	90.9	90.8
min	89.1	89.4	89.3	89.6	89.4	89.2	89.6	89.6	89.7	89.6	89.3	89.6
$>\pm$ 1% nominal (%)	15.6	0	3.1	0	0.6	0	10	2.5	0	0	0	0
Average width	1.102	1.059	0.752	0.738	0.487	0.483	1.183	1.133	0.828	0.811	0.653	0.646
Mean Square Error												
Average MSE	0.160	0.118	0.063	0.055	0.025	0.024	0.169	0.125	0.070	0.062	0.042	0.039
Separated data sets												
Total (%)	0.006		0		0		0.001		0		0	

* only for $\beta_1 \geq log(1)$

Results part II: Detection and handling of separated data sets

The results for study IIa are given in Table 8.3 and Figure 8.5. In Table 8.3 the simulation results were calculated twice, once by removing the separated data sets from analysis and once by leaving the separated data sets in, using the estimates at the point at which the model had converged. Between these approaches the calculated bias and MSE for EPV values between 4 and 18 were noticeably different. Average coverage in those EPV ranges was not markedly different, while average width of the confidence interval differed strongly depending on the handling of separated data sets. For EPV values between 55 and 150, separation was detected just eight times. In these simulations, only the calculated average width of the confidence interval and, to a lesser extent, mean square error were different between the two approaches of handling the separated data sets.

In the lower panel of Figure 8.5 it can clearly be seen that separation of the simulation data sets was rare for EPV values of 18 or higher. For these scenarios, bias in the maximum likelihood estimates (upper panel) for the non-zero true associations decreased with increasing EPV. For an EPV values of 16 and lower, separation occurred more frequently. The likelihood of drawing separated data sets also increases with true effect size of the coefficient. When removing those data sets from the analysis (upper panel, solid line), for the non-null associations the bias is underestimated, and even becomes negative at EPV values of 6 and 8. When replacing the results for the separated data sets by the highest estimated effect sizes (dashed lines, upper panel), the simulation outcomes are more in line with the patterns we observed in Part I. Finally, using Firth's correction (Figure 8.5, middle panel) all data sets were analyzed and the relative bias was near zero across the whole range of EPV.

The results for study IIb are shown in Table 8.4. In this single scenario study, the prevalence of separated data sets was 5.8% (as detected through the preferred re-estimation process, see Appendix). The differences in the calculated simulation results between the six methods of separation detection and estimation were large. Differences were noticeable especially in the calculated (relative) bias, mean square error and width of confidence intervals. Coverage was not significantly affected across the 6 approaches to detect separation. The success rate of using convergence as a criterion to detect separation depended on the convergence criteria. Relying on the Type III convergence criterion (only .09% non-convergence) makes the simulation results non-interpretable. The use of $|\hat{\beta}_j^{ML}| >$log(50) as a separation criterion in this scenario shows very different results compared to our preferred re-estimation method to detect separation.

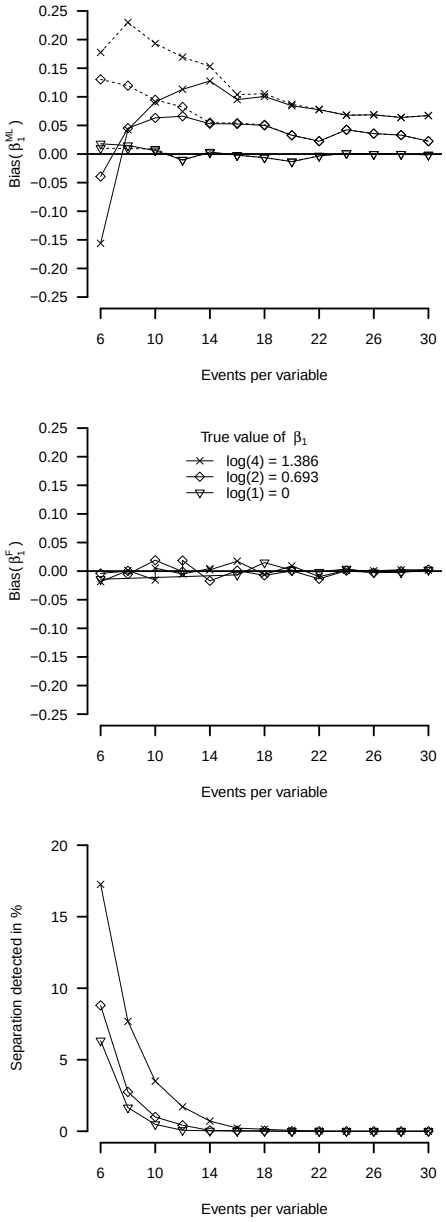

Figure 8.5: Simulation study IIb results. Upper panel, solid line: data sets removed from analysis; Upper panel, dashed line: data sets replaced by maximum non-separated effect size. Middle panel: Firth correction. Lower panel: percentage of separated data sets by true effect size.

Table 8.3: Results simulation study Ib.

EPV	15 to 30		35 to 50		55 to 150	
Separated data removed	Yes	No	Yes	No	Yes	No
Bias						
Average bias	-0.097	2.255	0.083	0.161	0.051	0.053
max	0.091	7.074	0.127	0.439	0.084	0.096
min	-0.556	0.234	0.050	0.056	0.048	0.022
Average relative bias (%)	-0.087	2.110	0.079	0.145	0.048	0.049
max	0.091	5.103	0.095	0.317	0.061	0.069
min	-0.401	0.338	0.069	0.081	0.032	0.032
$>+10\%$ relative bias (%)	0	100	0	37.5	0	0
Coverage 90% CI						
Average coverage (%)	92.7	93.4	89.1	89.1	90.4	90.4
max	98.3	98.8	90.6	90.6	91.8	91.8
min	89.7	89.8	87.9	87.9	89.2	89.2
$>\pm$ 1% nominal (%)	75	75	50	37.5	25	25
Average width	4087	4437.2	2656	49.2	2005	2645
Mean Square Error						
Average MSE	1251	64571	0.709	2243	0.397	0.422
Separated data sets						
Total (%)	13.2		4.2		0.006	

Discussion

Heterogeneity in findings between three previous simulation studies [45, 134, 171] that have examined the minimal EPV criterion for binary logistic regression analysis was the motivation of our study. EPV, which is thought to be an important driver behind the performance of logistic regression models, is frequently used in sample size considerations and as a methodological quality item for critically appraising published studies [121, 122, 133]. To explain the large heterogeneity in minimal EPV recommendations, we distinguished between two small sample issues that often coexist in small sample simulation studies in the context of logistic regression, including minimal EPV studies: biased estimation of logit coefficients and the problem of separation. While biased estimation of coefficients is often of primary interest, separated data sets are an important nuisance. The approach to detect and handle to dealing with separated data can have a strong impact on the results from simulation studies. We now discuss separately: i) the drivers of the accuracy of logit coefficients; ii) the influence of separated data sets on simulation results; iii) reasons for heterogeneity between the earlier minimal EPV simulation studies.

Table 8.4: Results simulation study IIb.

Estimator	β_1^F	β_1^{ML}	β_1^{ML}	β_1^{ML}	β_1^{ML}	β_1^{ML}	β_1^{ML}
Separation detection	NA	Tracing[b]	Estimate[c]	None	None	None	None
Convergence criterion[a]	Default	Default	Default	Default	Type I	Type II	Type III
Data sets removed (%)	0	8.06	16.64	5.12	0.34	6.29	0.09
Bias	0.012	0.569	0.186	1.672	17.5	0.856	41.3
Coverage 90% CI	0.919	0.949	0.937	0.944	0.947	0.944	0.947
Mean width 90% CI	4.32	4.50	3.64	5018	13620	6.03	1135784
MSE	1.080	2.681	0.904	71.563	11532	319	173726

[a] default: tol: 1e-8, max-iter: 25, Type I: tol: 1e-6, max-iter: 25, Type II: tol: 1e-10, max-iter:25, Type III: tol: 1e-10, max-iter:50.
[b] criterion: re-estimation process, variance of scaled standard errors >20 (see appendix).
[c] criterion: if for any parameter $j \neq 0, |\hat{\beta}_j| >$log(50).

Drivers of the accuracy of logit coefficients

Our results illustrate that the estimated coefficients in binary logistic regression analysis are typically overoptimistic estimates of the true associations in small to moderated-sized data sets. This over-optimism is commonly referred to as finite sample bias [103], and is well described in statistics literature [10,155]. The bias can to a large extent be attributed to skewed sampling distributions of the estimator in small data. Our simulations show that the finite sample bias is larger for data sets with small EPV, and may not reduce strictly to zero even for an EPV of 150.

In simulations where by design separation of data sets occurred only rarely, we found that bias depends on various factors besides EPV, notably, the true (multivariable) effect size of the regression coefficient. This latter finding is to be expected, based on the analytical work of Cordeiro and McCullagh [44]. Further, we showed that bias can be reduced by increasing the total sample size while keeping EPV constant (i.e., increasing the number of non-events). Bias at a fixed value of EPV also decreases with the number of covariates included. For a few conditions, we found that the Wald confidence interval showed slight over-coverage at smaller values of the EPV, i.e., for EPV <30 in the case of a single covariate. We could find no evidence to support that the amount of correlation between covariates in the model affected the accuracy of the coefficients as previously suggested [45].

Our study further suggests that Firth's correction [61] can reduce finite sample bias close to zero and reduce mean square error. Profile likelihood confidence intervals for the Firth's corrected estimates showed close to nominal behavior, and on average have smaller width than the traditional Wald confidence interval for the maximum likelihood estimates. Firth's correction is one among several methods that have been demonstrated to meaningfully

improve small sample performance of a logistic regression analysis [33, 133]. In particular, these alternatives seem beneficial for analyzing data sets with sample sizes in the order of a few hundreds. Procedures implementing Firth's correction for logistic regression (and Cox regression) are available in many statistical software packages (such as `SAS`, `Stata` and `R`), yet it is still are rarely used.

The impact of separated data sets on simulation results

The traditional (maximum likelihood) logistic regression analysis of a dataset in which the included covariates perfectly separate the binary outcome variable cannot be trusted. In such cases, typically, very low or very high parameter estimates with large maximum likelihood standard errors are returned by the statistical software program. The estimated values, however, are rather arbitrary and depending on software settings such as likelihood convergence criteria. In the context of simulation studies these extreme values can have a large influence on simulation results.

Methods to detect separation in simulation studies can be computationally intensive [40, 109] and likely therefore not routinely applied in most simulation studies. We also showed that convergence as a criterion for separation detection often fails. Separated data sets may therefore often remain undetected.

If separation is detected, the common approach is to remove the results based on separated data sets from the analysis. Steyerberg et al. [156] recognized that this causes informative missingness of simulation results. Our simulations confirm that even when the proportion of separated data sets is relatively small (~5%), removing separated data sets from analysis has a large impact on (apparent) bias, mean square error and width of the confidence intervals. Alternatively, replacing these results, for example by the largest non-separated simulated effects, may be a more realistic approach. It must be recognized that the choice of the replacing value (mechanism) is again rather arbitrary and may heavily influence the simulation results.

Separation of the outcome by covariates not only occurs in the setting of the binary logistic model. For example, separation can also occur with logistic regression for more than two outcomes and Cox's proportional hazards regression [34, 35]. Reporting on the proportion of separated simulation data sets' is, however, highly uncommon in simulation studies.

By applying Firth's correction, the problems associated with separation can be avoided altogether.

Reasons for heterogeneity between EPV simulation studies

We identified two major reasons for the heterogeneity between the preceding simulation studies [45, 134, 171]. First, differences in the design of the simulation studies may have contributed to variations in simulation outcomes at the same level of EPV. The preceding studies [45, 134, 171] differ, for example, in their range of simulated true effect sizes of the regression coefficient, total sample size and the number of included covariates. Second, none of these studies have sufficiently addressed the issue of separated simulation data sets. We illustrated that separated data sets can lead to misleading simulation outcomes. As separated data sets occur most frequently in low EPV settings, these settings are likely most affected.

The probability of drawing separated data in simulations depends on a multitude of factors, including the total sample size and the true effect sizes of the coefficients [89]. Developing simulation scenarios in realistic contexts where this probability is close to zero is difficult. For example, it was difficult to design small sample simulation settings with binary predictor variables while avoiding separation. Hence, in the setting of small EPV simulation studies, developing realistic full factorial simulation designs (i.e., a simulation design where all possible combinations of simulation factors are evaluated) in which the probability of drawing separated data sets in each condition is close to zero does not appear to be possible.

Steyerberg et al. [156] suggested the use of Firth's correction as a method to perform minimal EPV simulation studies and we have shown that this solves the problem of separated data sets. However, due to the impact of Firth's correction on the estimated coefficients even in the absence of separation, only little is learned about the behavior of traditional logistic regression analysis that is commonly used and is based on the generally well-trusted principles of maximum likelihood.

Conclusion

We conclude that the current evidence underlying the EPV = 10 rule as a minimal sample size criterion for binary logistic regression analysis is weak. Due to the lack of solid evidence, there is an urgent need for new guidance on determining the necessary sample sizes in view of the number of evaluated covariates, for logistic regression analysis. Guidance on appropriate sample size calculation for the development of diagnostic and prognostic prediction models using binary logistic regression is also lacking. Much of the attention has been directed toward performance of estimating the relations between covariates and outcome. However, logistic regression is often used to develop multivariable models for prediction purposes and

the impact of small samples in relation to number of covariates with respect to the model's ability to correctly predict outcomes (e.g. model calibration and discrimination) has received little attention so far. New research in that area, building upon the results of our study, is urgently needed. These new studies into minimal EPV criteria may also focus on alternative logistic regression approaches, such as by applying Firth's correction for which we showed significantly improved performance of logistic regression analysis in small samples.

Appendix

To detect separation in a data sets it is sufficient to monitor the maximum likelihood standard errors of parameters during the estimation process [109]. The logistic regression model is re-fitted on each simulation data set with $1, 2, ..., 30$ Fisher scoring iterations. The maximum likelihood standard errors for each of the 30 refits are collected. This approach to identification of separation is similar to the default method for separation detection in the `brglm` package (Version 0.5-9) for `R` by Ioannis Kosmidis. Separation for a parameter is said to occur if the variance of scaled standard errors (such that standard errors on first iteration equal 1) over refits was larger than 20. This cut-off value was chosen based on a small pilot study. Results not shown.

Chapter 9

Guidance to construct and use a nomogram for multinomial logistic regression models

Abstract

The use of multinomial logistic regression models is advocated when modeling the assocations of covariates with three or more mutually exclusive outcome categories. As compared to a binary logistic regression analysis, the simultaneous modeling of multiple outcome categories within a multinomial model often better resembles the clinical setting, where a physician typically must distinguish between more than two possible diagnoses or outcome events for an individual patient. A disadvantage of the multinomial logistic model is that the interpretation of its results is often complex. In particular, the calculation of predicted probabilities for the various outcomes requires a series of careful calculations. Nomograms are widely used in studies reporting binary logistic regression models to facilitate the interpretation of the results and allow the calculation of the predicted probability for individuals. In this paper we outline a general approach for deriving a scoring chart and a nomogram for multinomial logistic regression models, irrespective of the number of outcome categories that are present. We illustrate the use of the nomogram and its interpretation using a clinical example.

Introduction

The use of multinomial logistic regression modeling has been encouraged to study multiple outcomes categories (and possibly their combination) simultaneously [17, 20]. The multinomial logistic model can be considered to be an extension of the popular binary logistic regression model, which is often used in the presence of two mutually exclusive outcome categories. More specifically, multinomial logistic regression analysis can be viewed as a series of binary logistic regression analyses where one of the outcome categories is the reference category in each binary sub-model.

Currently, researchers often collapse multiple outcome categories into a single binary (composite) outcome and use binary logistic regression models to analyze their data. By collapsing potentially important information about the different outcome categories is lost. It can also be argued that a binary outcome often does not accurately reflect clinical practice, where physicians commonly have to make decisions while considering more than two relevant choices. For instance, physicians often consider the presence of differential diagnoses (and prognoses) for an individual patient simultaneously [20].

One of the key reasons why researchers might refrain from multinomial logistic regression analysis is that the results from these models are more complex to interpret and more elaborate than results from a binary logistic regression analysis. In the multinomial context, regression coefficients are estimated for each binary sub-model reflecting the relation of covariates to one outcome category relative to the reference category. The number of estimated parameters quickly increases with additional outcome categories considered. The large amount of information from multinomial models can easily overwhelm researchers and readers. In addition, when using the multinomial logistic model for estimating probabilities for individual patients, the computation involved for the various outcomes requires a series of careful calculations.

A nomogram is often being used in the reporting of binary logistic regression models (for recent examples, see [13, 131]). A nomogram can not only improve insights of readers into the results of a logistic model, it can also be used to arrive at a predicted probability of outcome(s) of interest that is (are) tailored to the profile of an individual patient in a graphical manner. Nomograms can thus facilitate clinical decision making during clinical encounters. So far, however, nomograms have been used primarily for improving the reporting of models with only two outcomes categories. We are aware of one recent paper that has reported the construction of a nomogram for multinomial models, but this applications is limited to the results of a

specific dataset and a limit number of outcome categories [11].

In this manuscript, we present how to construct, interpret, and use scoring charts and nomograms for a multinomial logistic regression model. We will first specify the multinomial regression model and then present a general approach for deriving the nomogram and scoring chart for such models irrespective of the number of outcome categories that are present. We will illustrate the use of the nomogram and its interpretation using a clinical example on the risk of operative delivery [147].

Multinomial logistic model

Let y_i denote the single observed outcome category of individual $i, i = 1, ..., N$. Assuming that this outcome is in one of K categories (e.g., a disease among K possible diseases), we may assume Y_i to be a multinomial random variable with probabilities $\pi_{i1}, ..., \pi_{iK}$. Conditional on J observed covariate values in vector $\boldsymbol{x}_i$, $\boldsymbol{x}_i = \{x_{i1}, ..., x_{ij}, ..., x_{iJ}\}$, the probability of observing category k is denoted by $\pi_k(\boldsymbol{x}_i)$. The multinomial logistic model where category K is treated as the reference category, can then be defined as

$$\ln \frac{\pi_k(\boldsymbol{x}_i)}{\pi_K(\boldsymbol{x}_i)} = \text{lp}_k(\boldsymbol{x}_i) = \alpha_k + \boldsymbol{\beta}_k' \boldsymbol{x}_i, \quad k = 1, ..., K-1, \tag{9.1}$$

where α_k is an intercept term, $\boldsymbol{\beta}_k = \{\beta_{k1}, ..., \beta_{kJ}\}$ is a vector of regression coefficients and $\text{lp}_k(\boldsymbol{x}_i)$ is one of $K-1$ linear predictors for individual i. We can now define the probability of each possible category k by

$$\pi_k(\boldsymbol{x}_i) = \begin{cases} \exp\{\text{lp}_k(\boldsymbol{x}_i)\} / \left[1 + \sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}\right] & \text{if } k = 1, ..., K-1 \\ 1 / \left[1 + \sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}\right] & \text{if } k = K. \end{cases} \tag{9.2}$$

Constructing the nomogram

The suggested nomogram is a special case of a group of nomograms that are formally known as parallel scale nomograms. Doerfler [57] outlined the parallel scale nomogram that can be constructed if a particular value can be calculated from the sum of two functions. To use this approach for multinomial logistic models we make use of a natural logarithm transformation

applied to the elements of equation 9.2, such that,

$$\ln \pi_k(\boldsymbol{x}_i) = \begin{cases} \text{lp}_k(\boldsymbol{x}_i) - \ln\left[1 + \sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}\right] & \text{if } k = 1, ..., K-1 \\ -\ln\left[1 + \sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}\right] & \text{if } k = K. \end{cases} \tag{9.3}$$

To simplify notation we let, $o_{ik} = \ln \pi_k(\boldsymbol{x}_i)$, $l_{ik} = \text{lp}_k(\boldsymbol{x}_i)$ and $s_i = -\ln\left[1 + \sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}\right]$.

The parallel scale nomogram makes use of the relation: $o_{ik} = l_{ik} + s_i$. Each of these three elements in this relation corresponds to one of the three vertical axes of the nomogram. The axes are denoted by L (left axis), O (middle axis) and S (right axis). Axis L is a scaled function of linear predictor $k, m_1 l_{ik}$, where m_1 is the scaling factor. Axis O corresponds to the probability of observing category $k, m_2 o_{ik}$. Lastly, axis S is a scaled function of the sum of exponentiated linear predictors, $m_3 s_i$.

The nomogram is depicted in Figure 9.1. Below we detail the 4 step procedure to arrive at this nomogram. For further details about the construction of the parallel scale nomogram, we refer to Doerfler [57].

Step 1: Placing the outer axes (L and S)

To obtain an adequately sized nomogram, determine the desired common height (h) for the outer two axes (L and S) and the horizontal distance between them (d). The two parallel axes are placed in the vertical direction. The values for h and d are assigned at the discretion of the researcher in a common metric (e.g., centimeters or inches). Larger values for h and d will allow for more precise reading of values.

Step 2: Determine the scaling factors $\boldsymbol{m_1, m_2, m_3}$

For determining the scaling factors for the outer axes (m_1 and m_3) the relevant ranges of l_{ik} and s_i need to be considered. The limits of these ranges ($l^{low}, l^{up}, s^{low}, s^{up}$) may be determined by the range observed in the data set where the model was developed, e.g., $l^{low} = \min(\hat{l}_{ik})$ and $l^{up} = \max(\hat{l}_{ik})$. These limits define the corresponding limits of the axes. Once these ranges are chosen, the scaling factors m_1 and m_3 are computed by $m_1 = h/(l^{up} - l^{low})$ and $m_3 = h/(s^{up} - s^{low})$. The remaining scaling factor is given by $m_2 = m_1 m_3/(m_1 + m_3)$.

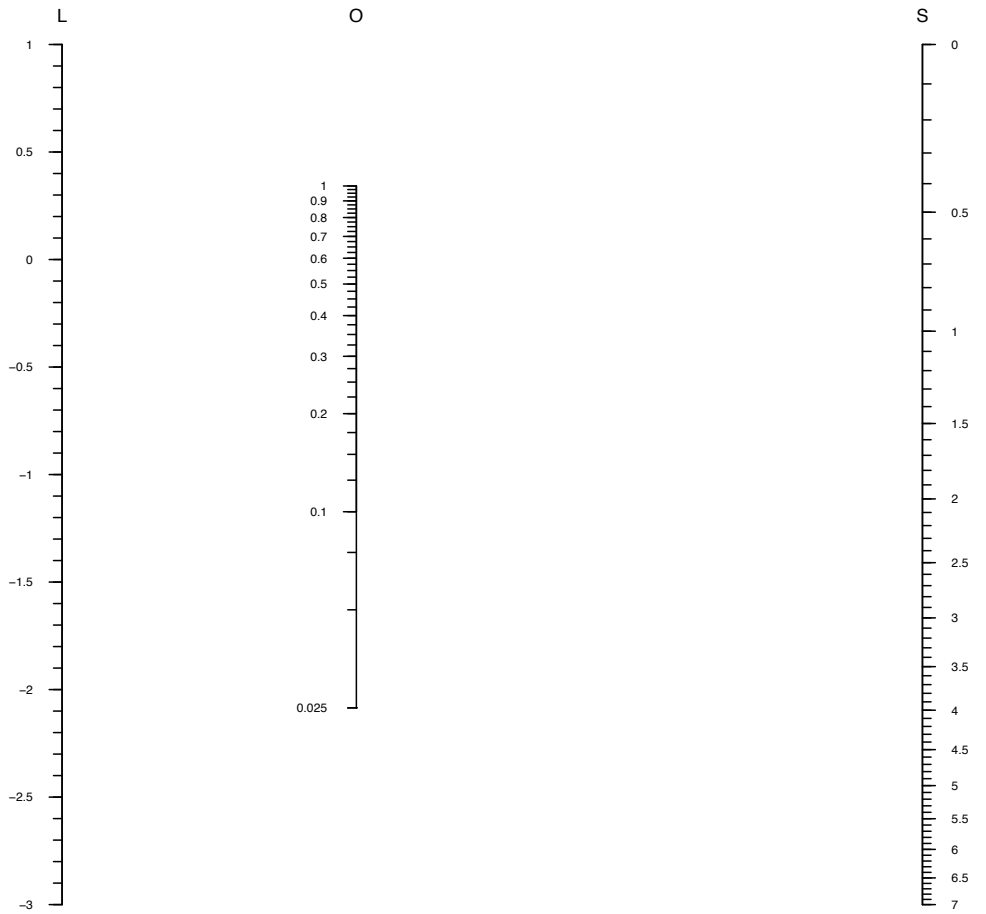

Figure 9.1: Nomogram for reporting multinomial logistic regression analysis. Axis L: $\text{lp}_k(\text{x}_i)$, Axis O: probability of outcome k, Axis S: $\sum_p \exp\{\text{lp}_p(\text{x}_i)\}$

Step 3: Placing the middle axis (O)

The O axis is placed parallel to the outer axes. The horizontal distance between the axes L and O is given by $d_{LO} = d - d/(m1/m3 + 1)$.

Step 4: Placing tickmarks and labels

For the outer axes L and S two sequences of values for the tickmarks and corresponding labels are defined: $l_T = (l^{low}, ..., l_t, ..., l^{up})$ and $s_T = (s^{low}, ..., s^t, ..., s^{up})$. Tickmark t on the axis

should be placed relative to the lower end of that axis at a distance of: $m_1 \times (l_t - l^{low})$ for axis L and $m_3 \times (s_t - l^{low})$ for axis S. For axis O, we define: $o^{low} = \exp\{l^{low} + s^{low}\}$. Because the axis O represents a log transformed probability scale, first a sequence of arithmetic probabilities o_T^* is defined with values between $\exp\{o^{low}\}$ and 1. Then, tickmark t for this sequence may be placed at $m_2 \times (\ln(o_t^*) - f(o^{low}))$ labeled by o_t^*. Axis S is labeled by $\sum_{p=1}^{K-1} \exp\{\text{lp}_p(\boldsymbol{x}_i)\}$.

Constructing the scoring chart

The use of the nomogram by health professionals can be improved by additionally presenting a scoring chart. This scoring chart provides a graphical approach to arriving at the values for the two outer axes L and S of the nomogram for any relevant combination of values on the covariates $(\boldsymbol{x}_i)$. For brevity, in this section we only consider the case of a multinomial logistic regression model with first order main effects. The scoring chart (and nomogram) can be extended to accommodate situations where higher order and interaction effects are present.

To make the scoring chart user-friendly, the individual effects of covariate j, $(\hat{\beta}_{jk} x_{ij})$, that make up the linear predictor k are rescaled to a 'standardized' score. The sum over these individual effects together with a baseline-score make up a 'standardized' total score. This total score is a linear transformation of l_{ik}. To facilitate the applicability of this standardized total score approach to the nomogram, the scaling of axis L should be adjusted accordingly. Below we detail the 3 step procedure to arrive at the scoring system that makes up the scoring chart.

Step 5: Standardized covariate effects: Points

The estimated multinomial logistic regression coefficients, $\hat{\beta}_{jk}$, are rescaled relative to the largest (conditional) covariate effect on a scale that has a minimum of 0 and a maximum of 100. First, the relevant ranges for each of the covariate variables are considered. Let the boundaries of these relevant ranges be denoted: x_j^{low} and x_j^{up}. The rescaling factor and rescaled coefficients are then computed by: $r = 100/\max_{j,k}(|\hat{\beta}_{jk} x_j^{up} - \hat{\beta}_{jk} x_j^{low}|)$ and $\hat{\beta}_{jk}^* = r \times \hat{\beta}_{jk}$. The covariate effects are 'standardized' by, $\text{Points}_{jk}(x_{ij}) = \hat{\beta}_{jk}^* x_{ij} - \min(\hat{\beta}_{jk}^* x_j^{up}, \hat{\beta}_{jk}^* x_j^{low})$.

Step 6: Standardized total effect: Total points

A baseline score for each category (except the reference category) is defined that that takes into account the standardization that has been performed at step 5. The baseline score is computed by: $bl_k = r \times \hat{\alpha}_k + \sum_j \min(\hat{\beta}_{jk}^* x_j^{up}, \hat{\beta}_{jk}^* x_j^{low})$. To also obtain a 'standardized' baseline score such that the minimum rescaled baseline score is zero, we subtract the minimum baseline score,

$bl_k^* = bl_k - \min_k(bl_k)$. The standardized total effect for category k given covariate values is then given by, $\text{Total}_k = bl_k^* + \sum_j \text{Points}_{jk}(x_{ij})$. Notice that, $l_{ik} = \text{lp}_k(\boldsymbol{x}_i) = (\text{Total}_k + \min_k(bl_k))/r$.

Step 7: Connecting the standardized total effects to the S-axis

A horizontal axes representing Total_k is placed near the lower end of the scoring chart. Another parallel horizontal axis is placed: the values on this axis are related to the former axis by $\exp\{(\text{Total}_k + \min_k(bl_k))/r\}$. Taking the sum over the values that can be read off from the axis for all categories (except the reference category) is all the information that is necessary for determining the value on the S axis.

Empirical Example: Predicting the risk of operative delivery

To illustrate the suggested scoring chart and nomogram to report multinomial logistic models, we use a previously published model on predicting the risk of operative delivery [147]. This model has been developed using data from a randomized clinical trial conducted in the Netherlands [175].

In short, the model was developed in 5667 laboring women with high-risk vertex (i.e. babies in a normal position in the uterus) singleton pregnancies beyond 36 weeks of gestation that met the inclusion criteria of the randomized clinical trial. Based on the combination of the intervention (i.e. instrumental vaginal delivery (IVD) or caesarean section (CS)) and the indication for the intervention (i.e. fetal distress (FD) or failure to progress (FTP)) women were assigned to one of five distinctive outcome categories: spontaneous vaginal delivery (reference category); instrumental vaginal delivery due to suspected fetal distress (IVD-FD); caesarean section due to suspected fetal distress (CS-FD); instrumental vaginal delivery due to failure to progress (IVD-FTP); or caesarean section due to failure to progress (CS-FTP).

The model included the antepartum variables: maternal age, parity, gestational age, maternal diabetes mellitus, previous caesarean delivery, fetal gender, maternal hypertensive disorder, suspected intrauterine growth restriction and antepartum estimated fetal weight. An antepartum prediction model was developed using this set of variables (i.e. model 1 in Schuit et al. 2012; see Table 9.1). For more details on the various outcome categories and candidate predictors we refer to the original publication [147].

Table 9.1: Multivariable associations for multinomial antepartum prediction model, predicting the risk of operative delivery.

	IVD-FD vs spont.		CS-FD vs spont.		IVD-FTP vs spont.		CS-FTP vs spont.	
	$\hat{\beta}_{jk}$	OR(95% CI)	$\hat{\beta}_{jk}$	OR(95% CI)	$\hat{\beta}_{jk}$	OR(95% CI)	$\hat{\beta}_{jk}$	OR(95% CI)
Intercept	-13.1		-15.6		-11.1		-15.4	
Maternal age, years	0.029	1.03 (1.01, 1.05)	0.052	1.05 (1.02, 1.09)	0.054	1.06 (1.03, 1.08)	0.056	1.06 (1.04, 1.08)
Gestational age, weeks	0.26	1.29 (1.18, 1.41)	0.32	1.38 (1.22, 1.56)	0.038	1.04 (0.95, 1.13)	0.13	1.14 (1.05, 1.24)
Nulliparous	2.05	7.79 (5.26, 11.5)	1.13	3.09 (2.094.55)	3.39	29.7 (17.251.1)	2.65	14.1 (9.7820.3)
Previous caesarean delivery	1.77	5.87 (3.70, 9.32)	1.06	2.88 (1.74, 4.76)	2.39	10.9 (5.92, 20.1)	2.23	9.34 (6.17, 1.41)
Neonatal female gender	-0.19	0.83 (0.67, 1.03)	-0.5	0.61 (0.45, 0.83)	-0.25	0.78 (0.63, 0.96)	0.013	0.99 (0.81, 1.20)
Birthweight, 100-g increments	-0.059	0.94 (0.92, 0.97)	-0.079	0.92 (0.89, 0.96)	0.083	1.09 (1.06, 1.11)	0.12	1.12 (1.10, 1.15)
Maternal diabetes mellitus	0.32	1.37 (0.65, 2.91)	0.99	2.69 (1.29, 5.60)	-0.24	0.79 (0.35, 1.76)	0.87	2.38 (1.44, 3.95)

We used the reported multivariable associations to construct a scoring chart (Figure 9.2) that presents the prediction model in a more insightful way. The different lines in the scoring chart in between the ”Points” axis (upper end scoring chart) and the ”Total points” axis (lower end scoring chart) can be interpreted as the different effect sizes of the model coefficients relative to the relevant range of the covariates, with longer lines representing stronger effects.

Consider a non-diabetic patient with a maternal age of 32 years, a gestational age of 40.2 weeks, expecting her second baby boy with an estimated birth weight of 3540 grams. With her firstborn she did not have a previous caesarean delivery. Using the scoring chart, these patient characteristics can easily be converted into scores for the different outcome categories. For example, to assess the score of this patient for risk of instrumental vaginal delivery due to suspected fetal distress one first determines the position of 32 years on the IVD-FD line for maternal age. When you draw a vertical line to the upper line indicated by ”Points” one sees that a maternal age of 32 years corresponds to 10 points. The same can be done for all other predictors for this outcome. Adding all points of the separate predictors to the baseline points corresponding to the chosen outcome (i.e. IVD-FD) gives the total points for this particular outcome, which can be marked at the line indicated by ”Total Points”. These steps can be repeated to calculate the total scores for the other 3 outcome categories (i.e. CS-FD, IVD-FTP and CS-FTP).

Finally, when a vertical line is drawn from the calculated total points per category marked on the ”Total Points” axis to the parallel axis right under the former axis, you end up with all information needed to draw the lines within the actual nomogram (Figure 3 9.3) to estimate (read-off) the predicted probability of each outcome category for this particular patient.

To get the predicted probability for each of the four outcome categories (i.e. IVD-FD, CS-FD, IVD-FTP and CS-FTP) you draw four lines from the point on the right axis (i.e. called ”S”) to the respective total points this particular patient received for the respective outcome categories. On the middle ”Predictive Probability” axis you now find the predictive probabilities of the four outcomes for this particular patient. Finally, subtracting these 4 predictive probabilities from 1 gives the probability of the reference category, in this case the probability of a ”spontaneous delivery”.

Conclusion

We expect that our general approach to construct and report score charts and nomograms for multinomial logistic regression models will facilitate the interpretation and use of such models to a wider audience. Our approach is flexible and generalizable and can be used irrespective of the number of outcome categories and types of covariates present.

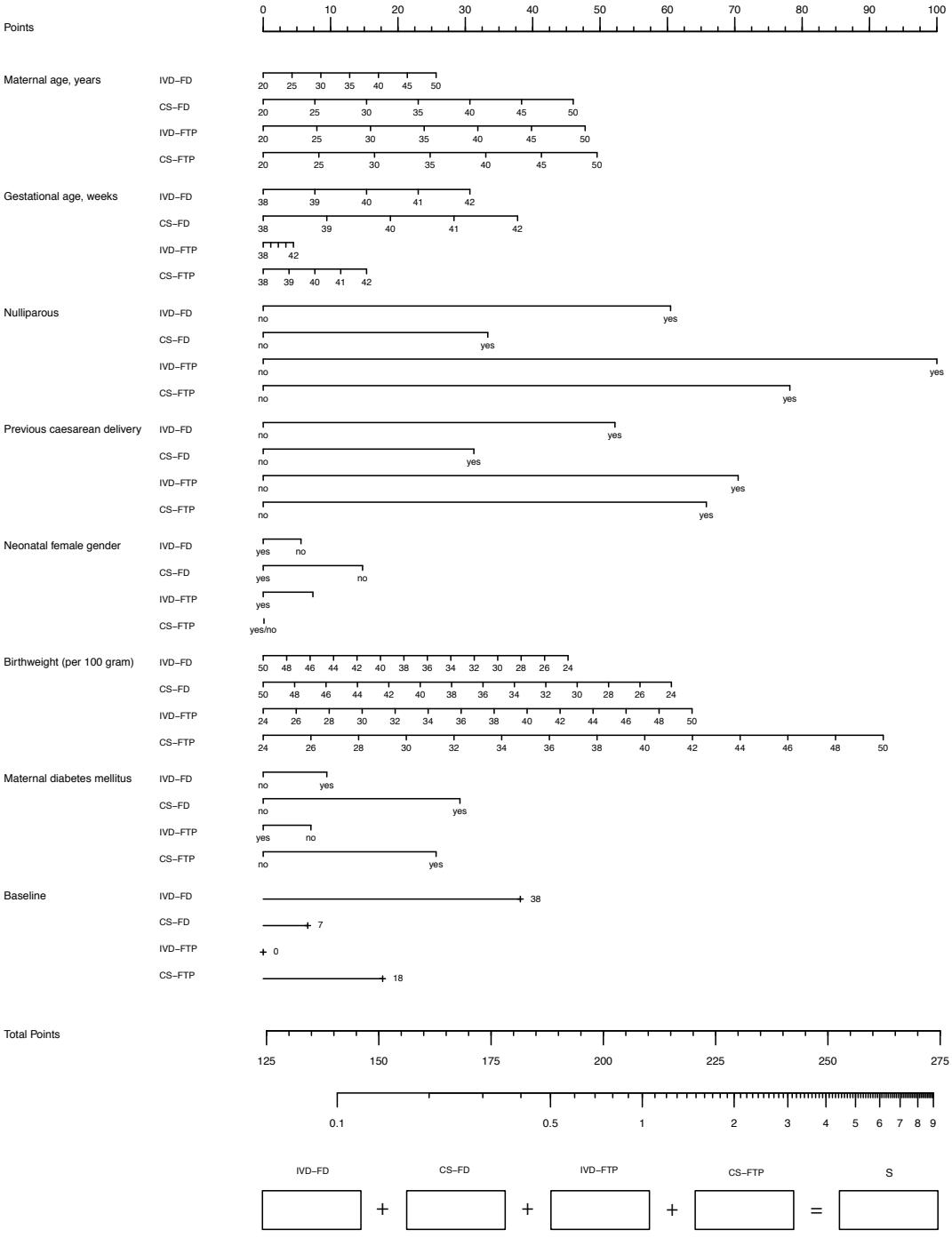

Figure 9.2: Score chart - case study Schuit et al.

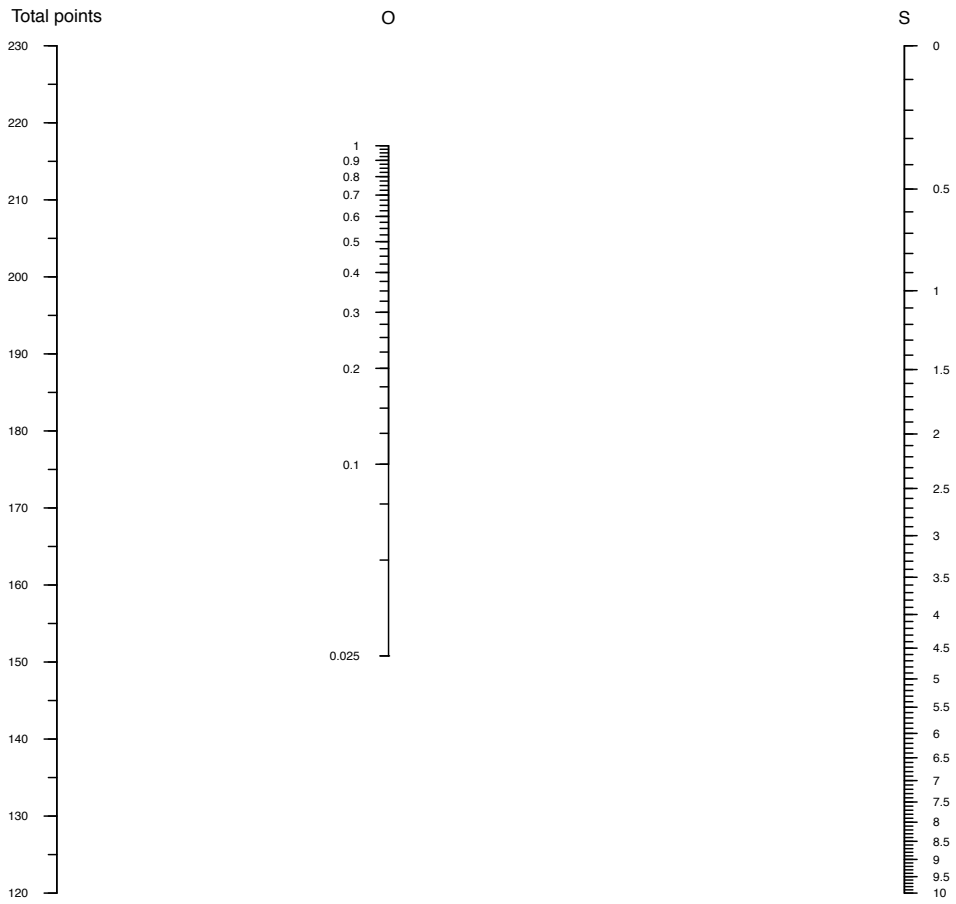

Figure 9.3: Nomogram - case study Schuit et al.

Chapter 10

General Discussion

A key challenge in a diagnostic accuracy study is to discriminate between subjects who have the target condition of interest and those who do not. This process of determining each subject's target condition status is known as verification [16, 18]. In most diagnostic accuracy studies, where the interest is in studying the accuracy of a diagnostic index test(s), verification proceeds by applying a single diagnostic reference test to the study subjects [104, 145]. The reference test results are then compared with those of the diagnostic index test(s) to calculate measures of accuracy for the diagnostic index test(s), such as sensitivity, specificity, predictive values, area under the ROC curve and diagnostic odds ratios [104].

This classical diagnostic accuracy paradigm [145] relies on the important assumption that the single reference test that is used for verification has perfect accuracy for detecting the target condition. A reference test that has perfect accuracy is referred to as a 'gold standard' [28, 104, 143]. Unfortunately, due to the absence of a gold standard for many diseases [135, 145, 166], researchers often face situations where the true target condition status of studied subjects cannot be verified with sufficient accuracy. The classifications of subjects to their target condition status are then prone to misclassification error. When ignored, these errors can in turn lead to severely biased inferences about the accuracy of the index test(s) [29, 112, 176]. This type of bias is often referred to as 'imperfect verification bias' [28], which can also affect the evaluation of markers and diagnostic models.

Alternative approaches to verification have been proposed that aim at improving the evaluation of diagnostic accuracy of index tests in the absence of a gold standard. In the studies presented in this thesis, we critically reviewed two of these approaches, namely: latent class analysis (LCA) and composite reference standards (CRS). Both of these approaches have found wide application in the existing literature (see Chapters 2 and 6). At the same time, however, both the use of LCA and CRS have been critiqued for being imperfect solutions for the gold standard problems in diagnostic accuracy research [54, 136].

This chapter is structured as follows. First, we summarize the principles, strengths and limitations of both LCA and CRS in the context of our findings and current literature. Further, we will focus on the problem of studying diagnostic accuracy in the absence of a gold standard with only few diagnostic tests. An overview is given of potential approaches to improve LCA in the presence of few diagnostic tests. Finally, this chapter will end with some concluding remarks.

Studying diagnostic accuracy using LCA or CRS

LCA and CRS both rely on combining the results of multiple diagnostic tests. By definition, in the absence of a gold standard, none of these diagnostic tests is assumed to have perfect accuracy for detecting the target condition. The rationale for using multiple imperfect diagnostic tests rather than a single reference test is that combining the results of multiple imperfect tests leads to a more accurate verification of the target condition than by any available single reference test in isolation. Consequently, the improved verification will reduce - and ideally eliminates - imperfect verification bias and thus leads to more valid estimates of accuracy of the index test(s).

Despite their similarities, LCA and CRS differ substantially in the way the results of multiple imperfect tests are combined. The CRS can be viewed as an inherently deterministic approach to verification: each study subject is classified into a definitive target condition status. This classification is based on an a priori defined and fixed classification rule [9]. In contrast, LCA is an inherently probabilistic approach, where the uncertainty of the true target condition for each subject is reflected in a target condition status probability derived for each individual subject. The target condition is a latent variable that is defined by a statistical model using the association between diagnostic test results in the data [94].

The CRS is often considered the most intuitive and clinically appealing approach to alleviate the problem of the absence of a gold standard. Indeed, when accurately reported, this method for target condition verification is transparent and reproducible. In Chapters 6 and 7 of this thesis, however, we clearly document that there are serious problems with the use of a CRS. Firstly, a CRS does not eliminate imperfect verification bias; in some cases target condition verification using a CRS will in fact be worse than using one of the 'component' tests used in this CRS as a reference test. Secondly, the direction of this verification bias is hard to predict or to adjust for. Hence, these properties make the validity of the CRS approach for the general use in diagnostic accuracy studies that lack a single acceptable reference standard for the target condition under study questionable. The application of LCA in such instances has also received much critique in literature. The key arguments against the use of LCA and possible nuances to these arguments are summarized in Table 10.1.

One issue that can affect both the applications of LCA and the CRS approach is the fact that, very often, the diagnostic tests that can be applied to study subjects is limited to only a few diagnostic tests. Both approaches rely on combining the results of multiple tests, hence, can be affected by limitations in the number of diagnostic tests available. In the next section we will

Table 10.1: Classical arguments against latent class analysis and possible nuances.

Arguments	Nuances
For clinicians, the results of a latent class analysis are output of a 'black box' [9]	The 'black box' character can be reduced by: i) explicitly mentioning and checking assumptions (Chapter 4); ii) explicating rationale for modeling choices and performing sensitivity analyses (Chapter 4); iii) reporting additional information about that supports face validity (Chapter 2)
Latent class analysis is sensitive to correlated test errors (i.e., violation of the conditional independence assumption) [136]	Latent class analysis comprises a broad family of models including those that explicitly account for correlated test errors (Chapter 2 to 5); other approaches that use multiple imperfect tests to define the true disease status are also affected by correlated errors between tests (Chapter 7)
Latent class analysis requires assumptions that are essentially unverifiable (i.e., cannot be fully tested [136])	This is true for any method that deals with missing data or measurement error; when adequate degrees of freedom are available, the critical assumptions of the latent class model (e.g. dependence structure) can in fact be tested by comparing model fit to the observed data (Chapter 3 and 5)
The target condition is not explicitly defined in a latent class analysis: it is an implicit mathematically defined entity [135]	When the object(s) of study cannot be observed directly it is often replaced by something comparable that can be observed (a proxy) or a mathematical model. Using mathematical models for making inferences about such unobserved objects or relations is also common in other areas of research, for example, confounding adjustment in non-randomized intervention studies and (multiple) imputation for missing data

focus specifically on this issue.

Diagnostic research in the absence of a gold standard with only few diagnostic tests

Unnecessary diagnostic testing for the purpose of gathering data for diagnostic studies can be unethical, particularly when these tests are invasive or after a treatment decision for a subject can already be made. Diagnostic testing can also be expensive and time consuming, e.g., in the case of imaging tests. Diagnostic test accuracy research is therefore often performed in situations where the number of diagnostic tests that is applied to study subjects is limited to only a few tests. For example, in Chapter 2 we reviewed the literature and found that latent class analysis is often performed is situations where only 3 or 4 diagnostic tests were available.

Performing a diagnostic accuracy study using the CRS approach requires data on a minimum of three diagnostic tests: one index test and two so-called component tests that are used in the composite rule. However, data obtained from only two component tests may not always be deemed sufficient for target condition verification. Applications of CRS that require as many as nine component tests can be found in literature [149].

Similar to a CRS, a typical LCA that involves applying the standard 2-class latent class model (see Chapters 2 and 3) also requires data on at least three diagnostic tests [136]. However, such an analysis relies on the assumption that these three diagnostic tests have uncorrelated errors. More precisely, for this analysis it is assumed that the diagnostic test results are locally independent (i.e., independent given the latent target condition status). In practice, this assumption is often not reasonable [31, 174]. In addition, with only three diagnostic tests the standard 2-class latent class model is 'saturated', meaning that the fit of this model to the data cannot be determined (see Chapter 5). LCA is thus particularly vulnerable when data on only a few correlated diagnostic tests are observed. To avoid making unrealistic simplifying assumptions and to account for correlated diagnostic test errors in a LCA with few test (three or even less), the diagnostic test data can be augmented. The advantages of such data augmenting in latent class analysis has so far received only limited attention. In the next section we provide an overview of approaches that can be used to augment the diagnostic test data to improve a LCA.

Overview of methods to augment the diagnostic test data for LCA

The commonly applied 2-class local independence LCA (2-class LI latent class model, for details see Chapter 2 and 3) is said to be 'nonidentifiable' when data from only 2 diagnostic tests are available. A model is nonidentifiable if there exist at least two choices of parameters for which the distributions of observed data are the same [75]. A model violates a minimal requirement for identifiability when more parameters are estimated than there are degrees of freedom available in the data. For the case of the 2-class LI latent class model and two binary diagnostic tests: 5 parameters are estimated while 3 degrees of freedom are available in the data. When the diagnostic tests have correlated errors (i.e., violate the local independence assumption), latent class models that account for this dependence need to be specified (see Chapter 2, 3 and 5). These local dependence latent class models require additional parameters to be estimated and therefore need a minimum of 4 diagnostic tests to be identifiable (the exact number of diagnostic tests needed depends on the specification of the dependence structure and the number of latent classes).

Table 10.2: Approaches to augment imperfect diagnostic test data and alleviate identification constraints for improving latent class analysis.

Approach	Explanation
Add diagnostic test(s)	Augment existing study data by adding the results of additional diagnostic tests
Add covariate(s) / grouping variable(s)	Covariates that modify the accuracy of a test and prevalence of the (latent) target condition can provide additional information to estimate parameters of the latent class model
Fix parameter(s) to a 'known value'	When a parameter of the latent class model[a] can be assumed known based on substantive knowledge (e.g., the sensitivity or specificity of a certain diagnostic test can be assumed to be 100%), the corresponding parameter can be fixed a priori to this 'known value' while the remaining parameters are estimated freely
Use partial gold standard verification	When a gold standard exists and it can be applied to only a subset of patients, the verification information of this subset of patients can be used in a latent class analysis
Use informative prior distribution(s)	In a Bayesian context, informative prior distribution(s) can be employed for the latent class model parameters[a] for which there is reliable prior information available

[a] In the simplest case where there are 2 latent classes within which all tests are conditionally independent, the latent class model is parameterized by the sensitivity and specificity of the tests and the prevalence of the target condition; in more complex latent class models the parameters may need to be transformed to be expressed in terms of sensitivity, specificity and target condition prevalence.

Several data augmenting approaches can be used to alleviate identification constraints when the number of diagnostic tests is small. In addition, some of these approaches increase precision of estimates, account for existing local dependencies between diagnostic tests and improve robustness of inferences to misspecification of the latent class model. An overview of these approaches is given in Table 10.2. In the following we briefly discuss each of these approaches.

Adding diagnostic test(s)

The most obvious approach to augment the diagnostic tests data is by obtaining data on additional diagnostic tests. By increasing the number of diagnostic tests, the degrees of freedom in the data increases (degrees of freedom are calculated $2^J - 1$, where J is the number of diagnostic tests). Besides degrees of freedom, also insight into the accuracy of these added tests is gained. In contrast, adding diagnostic tests may also complicate modeling when an added diagnostic test exhibits local dependencies with the other diagnostic tests.

Adding covariate(s) / grouping variable(s)

When data are obtained from more than one different groups (i.e., populations): each of these groups contributes $2^J - 1$ to the degrees of freedom. This principle underlies the well-known Hui-Walter latent class model [94]. In their seminal paper Hui and Walter described the case of 2 groups and $J = 2$ locally independent diagnostic tests for which the model is identifiable, provided that the prevalences of the target condition in the two groups are different. Generalizations of the Hui-Walter model to latent class models for diagnostic tests results that are assumed conditionally dependent have also been described (e.g., see [30, 58]). In a similar manner, the availability of a covariate that modifies diagnostic test accuracy can also alleviate the problem of non-identifiability. Furthermore, local dependencies can be adjusted for by letting the accuracy parameters vary according to the covariate [80]. It should be noted that despite the apparent increase in degrees of freedom, especially when the covariates are weak, latent class models with covariates or grouping variables are not always identifiable [75, 100].

Fix parameter(s) to a known value

When one or more parameters of the latent class model can be fixed to a 'known' value, the number of parameters that need to be estimated is reduced [72]. This approach may be appropriate when, for example, the sensitivity or specificity of one of the diagnostic tests is widely acknowledged to be perfect and the parameters of the latent class model are fixed accordingly. Albert [5] described an application of a latent class model where for one diagnostic test both sensitivity and specificity were assumed to be known. By performing simulation, he further showed that this approach increased the robustness of inferences from the LCA.

Use gold standard verification in a subset of patients

Albert and Dodd [7] have discussed the value of incomplete but perfect verification with a gold standard on a (potentially non-random) subset of subjects. This approach is suitable when a gold standard exists for the target condition of interest but is prohibitively expensive or invasive, and thus cannot be carried out in all subjects. This approach has similarities to the two-part likelihood functions that have been described in the diagnostic literature on partial verification bias [18].

Use informative prior distribution(s)

In a Bayesian context, the elicitation of informative prior distributions based on expert opinion may be possible for some of the parameters in the latent class model. While fixed value constraints require a very high level of certainty, in practice it might be more realistic to define a small range of possible values for one or more parameters. Provided that the prior information is proper, informative prior distributions can be used to obtain estimates for models that are otherwise non-identifiable [53, 71, 101].

Concluding remarks

The absence of a gold standard is an ubiquitous problem in diagnostic research. Of special interest is the increasing number of research situations where the index test may outperform the existing reference standard. The strength of LCA and its various extensions is that it allows researchers to estimate various diagnostic performance measures of interest using all available information. The credibility of LCA is increased when modelling assumptions and choices are made explicit and the impact of alternative modelling choices are investigated. Investigating these model assumptions and fit is however more problematic if the number of available tests that can be used for LCA is limited. We have discussed various approaches to improve LCA in such low-information situations. Future research initiatives around LCA should be directed towards the effectiveness and applicability of these approaches.

Bibliography

[1] A. Agresti. Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, 1(2):201–218, 1992.

[2] A. Agresti. *Categorical Data Analysis*, volume 2. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

[3] A. Agresti and J. B. Lang. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, 49(1):131–139, 1993.

[4] A. Albert and J. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[5] P. S. Albert. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine*, 28(5):780–797, 2009.

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

[7] P. S. Albert and L. E. Dodd. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481):61–73, 2008.

[8] P. S. Albert, L. M. McShane, and J. H. Shih. 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, 2001.

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

[10] D. G. Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19(4):453–473, 2000.

[11] I. Ardoino, M. Lanzoni, G. Marano, P. Boracchi, E. Sagrini, A. Gianstefani, F. Piscaglia, and E. M. Biganzoli. Widen NomoGram for multinomial logistic regression: an application to staging liver fibrosis in chronic hepatitis C patients. *Statistical Methods in Medical Research*, 2014.

[12] L. Audige, J. Hunter, A. M. Weinberg, J. Magidson, and T. Slongo. Development and evaluation process of a pediatric long-bone fracture classification proposal. *European Journal of Trauma*, 30(4):248–254, 2004.

[13] G. Aviram, H. Shmueli, S. Z. Adam, A. Bendet, T. Ziv-Baran, A. Steinvil, A. S. Berliner, N. Nesher, Y. Ben-Gal, and Y. Topilsky. Pulmonary Hypertension: A nomogram based on CT pulmonary angiographic data for prediction in patients without pulmonary embolism. *Radiology*, 277(1):236–246, 2015.

[14] J. H. Battershill. Cutaneous testing in the elderly patient with tuberculosis. *CHEST Journal*, 77(2):188, 1980.

[15] A. L. Baughman, K. M. Bisgard, M. M. Cortese, W. W. Thompson, G. N. Sanden, and P. M. Strebel. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clinical and Vaccine Immunology*, 15(1):106–114, 2008.

[16] C. B. Begg. Biases in the assessment of diagnostic tests. *Statistics in Medicine*, 6(4):411–423, 1987.

[17] C. B. Begg and R. Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1):11, 1984.

[18] C. B. Begg and R. A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 207–215, 1983.

[19] L. C. M. Bertens, B. D. L. Broekhuizen, C. A. Naaktgeboren, F. H. Rutten, A. W. Hoes, Y. van Mourik, K. G. M. Moons, and J. B. Reitsma. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Medicine*, 10(10):e1001531, 2013.

[20] C. J. Biesheuvel, Y. Vergouwe, E. W. Steyerberg, D. E. Grobbee, and K. G. M. Moons. Polytomous logistic regression analysis could be applied more often in diagnostic research. *Journal of Clinical Epidemiology*, 61(2):125–34, 2008.

[21] C. M. Black. Current methods of laboratory diagnosis of Chlamydia trachomatis infections. *Clinical Microbiology Reviews*, 10(1):160–84, 1997.

[22] C. M. Black, J. Marrazzo, R. E. Johnson, E. W. Hook, R. B. Jones, T. A. Green, J. Schachter, W. E. Stamm, G. Bolan, M. E. St Louis, and D. H. Martin. Head-to-head multicenter comparison of DNA probe and nucleic acid amplification tests for Chlamydia trachomatis infection in women performed with an improved reference standard. *Journal of Clinical Microbiology*, 40(10):3757–3763, 2002.

[23] M. A. Black and B. A. Craig. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine*, 21(18):2653–2669, 2002.

[24] M. Boelaert, S. el Safi, E. Goetghebeur, S. Gomes-Pereira, D. Le Ray, and P. Van der Stuyft. Latent class analysis permits unbiased estimates of the validity of DAT for the diagnosis of visceral leishmaniasis. *Tropical Medicine & International Health*, 4(5):395–401, 1999.

[25] M. Boelaert, S. Rijal, S. Regmi, R. Singh, B. Karki, D. Jacquet, F. Chappuis, L. Campino, P. Desjeux, D. Le Ray, S. Koirala, and P. Van Der Stuyft. A comparative study of the effectiveness of diagnostic tests for visceral leishmaniasis. *American Journal of Tropical Medicine and Hygiene*, 70(1):72–77, 2004.

[26] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, and STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Family Practice*, 21(1):4–10, 2004.

[27] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, D. Moher, D. Rennie, H. C. de Vet, and J. G. Lijmer. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry*, 49(1):7–18, 2003.

[28] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, D. Moher, D. Rennie, H. C. W. de Vet, and J. G. Lijmer. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*, 138(1):W1–12, 2003.

[29] E. J. Boyko, B. W. Alderman, and A. E. Baron. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *Journal of General Internal Medicine*, 3(5):476–481, 1988.

[30] A. J. Branscum, I. A. Gardner, and W. O. Johnson. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*, 68(2-4):145–163, 2005.

[31] H. Brenner. How independent are multiple 'independent' diagnostic classifications? *Statistics in Medicine*, 15(13):1377–1386, 1996.

[32] I. Bross. Misclassification in 2 X 2 tables. *Biometrics*, 10(4):478–486, 1954.

[33] S. B. Bull, C. M. T. Greenwood, and W. W. Hauck. Jacknife bias reduction for polychotomous logistic regression. *Statistics in Medicine*, 16(5):545–560, 1997.

[34] S. B. Bull, J. P. Lewinger, and S. S. F. Lee. Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine*, 26(4):903–918, feb 2007.

[35] S. B. Bull, C. Mak, and C. M. T. Greenwood. A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis*, 39:57–74, 2002.

[36] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

[37] Centers for Disease Control and prevention. Screening tests to detect Chlamydia trachomatis and Neisseria gonorrhoeae infections. *MMWR*, 51(RR-15):1–384, 2002.

[38] Centers for Disease Control and prevention. Recommendations for the laboratory-based detection of Chlamydia trachomatis and Neisseria gonorrhoeae. *Morbidity and mortality weekly report*, 63:1–19, 2014.

[39] H. Chu, Y. Zhou, S. R. Cole, and J. G. Ibrahim. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. *Statistics in Medicine*, 29(11):1206–1218, 2010.

[40] D. B. Clarkson and R. I. Jennrich. Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society. Series B*, 53(2):417–426, 1991.

[41] J. Collins and M. Huynh. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine*, 33(24):4141–4169, 2014.

[42] L. M. Collins, P. L. Fidler, S. E. Wugalter, and J. D. Long. Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28(3):375–389, 1993.

[43] R. L. Cook, S. L. Hutchison, L. Ostergaard, R. S. Braithwaite, and R. B. Ness. Systematic review: noninvasive testing for Chlamydia trachomatis and Neisseria gonorrhoeae. *Annals of Internal Medicine*, 142(11):914–925, 2005.

[44] G. Cordeiro and P. McCullagh. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B.*, 53(3):629–643, 1991.

[45] D. S. Courvoisier, C. Combescure, T. Agoritsas, A. Gayet-Ageron, and T. V. Perneger. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 64(9):993–1000, 2011.

[46] L. E. Cuevas, R. Browning, P. Bossuyt, M. Casenghi, M. F. Cotton, A. T. Cruz, L. E. Dodd, F. Drobniewski, M. Gale, S. M. Graham, M. Grzemska, N. Heinrich, A. C. Hesseling, R. Huebner, P. Jean-Philippe, S. K. Kabra, B. Kampmann, D. Lewinsohn, M. Li, C. Lienhardt, A. M. Mandalakas, B. J. Marais, H. J. Menzies, G. Montepiedra, C. Mwansambo, R. Oberhelman, P. Palumbo, E. Russek-Cohen, D. E. Shapiro, B. Smith, G. Soto-Castellares, J. R. Starke, S. Swaminathan, C. Wingfield, and C. Worrell. Evaluation of Tuberculosis Diagnostics in Children: 2. Methodological issues for conducting and reporting research evaluations of tuberculosis diagnostics for intrathoracic tuberculosis in children. consensus from an expert panel. *Journal of Infectious Diseases*, 205(suppl 2):S209–S215, 2012.

[47] G. De Bock, J. Houwing-Duistermaat, M. Springer, J. Kievit, and J. van Houwelingen. Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. *Journal of Clinical Epidemiology*, 47(12):1343–1352, 1994.

[48] J. A. H. de Groot, K. J. M. Janssen, A. H. Zwinderman, P. M. Bossuyt, J. B. Reitsma, and K. G. M. Moons. Correcting for partial verification bias: a comparison of methods. *Annals of Epidemiology*, 21(2):139–148, 2011.

[49] J. A. H. de Groot, P. M. M. Bossuyt, J. B. Reitsma, A. W. S. Rutjes, N. Dendukuri, K. J. M. Janssen, and K. G. M. Moons. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ (Clinical research ed.)*, 343:d4770, 2011.

[50] M. de Onis, A. W. Onyango, E. Borghi, A. Siyam, C. Nishida, and J. Siekman. Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, 85(09):660–667, 2007.

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

[52] N. Dendukuri, A. Hadgu, and L. Wang. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in Medicine*, 28(3):441–61, 2009.

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

[54] N. Dendukuri, L. Wang, and A. Hadgu. Evaluating diagnostic tests for Chlamydia trachomatis in the absence of a gold standard: a comparison of three statistical methods. *Statistics in Biopharmaceutical Research*, 3(2):385–397, 2011.

[55] C. M. Denkinger, S. G. Schumacher, C. C. Boehme, N. Dendukuri, M. Pai, and K. R. Steingart. Xpert MTB/RIF assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal*, 44(2):435–446, 2014.

[56] A. K. Detjen, A. R. DiNardo, J. Leyden, K. R. Steingart, D. Menzies, I. Schiller, N. Dendukuri, and A. M. Mandalakas. Xpert MTB/RIF assay for the diagnosis of pulmonary tuberculosis in children: a systematic review and meta-analysis. *The Lancet Respiratory Medicine*, 3(6):451–461, 2015.

[57] R. Doerfler. On Jargon-The Lost Art of Nomography. *The UMAP Journal*, 30(4):457, 2009.

[58] C. Enoe, M. P. Georgiadis, and W. O. Johnson. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine*, 45(1):61–81, 2000.

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

[60] A. K. Ewer, A. T. Furmston, L. J. Middleton, J. J. Deeks, J. P. Daniels, H. M. Pattison, R. Powell, T. E. Roberts, P. Barton, P. Auguste, A. Bhoyar, S. Thangaratinam, A. M. Tonks, P. Satodia, S. Deshpande, B. Kumararatne, S. Sivakumar, R. Mupanemunda, and K. S. Khan. Pulse oximetry as a screening test for congenital heart defects in newborn infants: a test accuracy study with evaluation of acceptability and cost-effectiveness. *Health Technology Assessment*, 16(2):v–xiii, 1–184, 2012.

[61] D. Firth. bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[62] A. K. Formann. Measurement errors in caries diagnosis: some further latent class models. *Biometrics*, 50(3):865–871, 1994.

[63] A. K. Formann. Latent class model diagnostics - a review and some proposals. *Computational Statistics & Data Analysis*, 41(3-4):549–559, 2003.

[64] A. K. Formann and T. Kohlmann. Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5(2):179–211, 1996.

[65] R. Gagnon, B. Charlin, M. Coletti, E. Sauvé, and C. Van Der Vleuten. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education*, 39(3):284–291, 2005.

[66] I. A. Gardner, H. Stryhn, P. Lind, and M. T. Collins. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive Veterinary Medicine*, 45(1-2):107–122, 2000.

[67] E. S. Garrett, W. W. Eaton, and S. Zeger. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine*, 21(9):1289–1307, 2002.

[68] J. Gart and J. Zweifel. On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika*, 54(1):181–187, 1967.

[69] G. Geersing, P. M. G. Erkens, W. A. M. Lucassen, H. R. Büller, H. T. Cate, A. W. Hoes, K. G. M. Moons, M. H. Prins, R. Oudega, H. C. P. M. van Weert, and H. E. J. H. Stoffers. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer

testing in primary care: prospective cohort study. *BMJ (Clinical research ed.)*, 345:e6564, 2012.

[70] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapmann & Hall/CRC, London, 2003.

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

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

[73] S. M. Graham, T. Ahmed, F. Amanullah, R. Browning, V. Cardenas, M. Casenghi, L. E. Cuevas, M. Gale, R. P. Gie, M. Grzemska, E. Handelsman, M. Hatherill, A. C. Hesseling, P. Jean-Philippe, B. Kampmann, S. K. Kabra, C. Lienhardt, J. Lighter-Fisher, S. Madhi, M. Makhene, B. J. Marais, D. F. McNeeley, H. Menzies, C. Mitchell, S. Modi, L. Mofenson, P. Musoke, S. Nachman, C. Powell, M. Rigaud, V. Rouzier, J. R. Starke, S. Swaminathan, and C. Wingfield. Evaluation of tuberculosis diagnostics in children: 1. Proposed cliinical case definitions for classification of intrathoracic tuberculosis disease. consensus from an expert panel. *Journal of Infectious Diseases*, 205(suppl 2):S199–S208, 2012.

[74] S. M. Graham, L. E. Cuevas, P. Jean-Philippe, R. Browning, M. Casenghi, A. K. Detjen, D. Gnanashanmugam, A. C. Hesseling, B. Kampmann, A. Mandalakas, B. J. Marais, M. Schito, H. M. L. Spiegel, J. R. Starke, C. Worrell, and H. J. Zar. Clinical case definitions for classification of intrathoracic tuberculosis in children: an update. *Clinical Infectious Diseases*, 61Suppl 3:S179–87, 2015.

[75] P. Gustafson. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140, 2005.

[76] P. Gustafson. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in Medicine*, 24(8):1203–17, 2005.

[77] A. Hadgu. The discrepancy in discrepant analysis. *Lancet*, 348(9027):592–3, 1996.

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

[79] A. Hadgu, N. Dendukuri, and L. Wang. Evaluation of screening tests for detecting Chlamydia trachomatis: bias associated with the patient-infected-status algorithm. *Epidemiology*, 23(1):72–82, 2012.

[80] A. Hadgu and Y. Qu. A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society. Series C*, 47(4):603–616, 1998.

[81] J. A. Hagenaars. *Applied Latent Class Analysis*. Cambridge University Press, Cambridge, 2002.

[82] S. L. Handelman, D. H. Leverett, M. A. Espeland, and J. A. Curzon. Clinical radiographic evaluation of sealed carious and sound tooth surfaces. *Journal of the American Dental Association*, 113(5):751–754, 1986.

[83] D. Harper. Local dependence latent structure models. *Psychometrika*, 37(1):53–59, 1972.

[84] F. E. Harrell. *Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York, 2001.

[85] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, 1984.

[86] M. Hatherill, M. Hanslo, T. Hawkridge, F. Little, L. Workman, H. Mahomed, M. Tameris, S. Moyo, H. Geldenhuys, W. Hanekom, L. Geiter, and G. Hussey. Structured approaches for the screening and diagnosis of childhood tuberculosis in a high prevalence region of South Africa. *Bulletin of the World Health Organization*, 88(4):312–320, 2010.

[87] M. M. Hegazy, N. L. El-Tantawy, M. M. Soliman, E. S. El-Sadeek, and H. S. El-Nagar. Performance of rapid immunochromatographic assay in the diagnosis of Trichomoniasis vaginalis. *Diagnostic Microbiology and Infectious Disease*, 74(1):49–53, 2012.

[88] G. Heinze. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, 25(24):4216–4226, 2006.

[89] G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

[90] A. C. Hesseling, H. S. Schaaf, R. P. Gie, J. R. Starke, and N. Beyers. A critical review of diagnostic approaches used in the diagnosis of childhood tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, 6(12):1038–1045, 2002.

[91] J. Hilden. Boolean algebra, Boolean nodes. In M. Kattan and M. E. Cowen, editors, *Encyclopedia of Medical Decision Making*, 94–98. Sage, 2009.

[92] M. Holden, M. R. Dubin, and P. H. Diamond. Frequency of negative intermediate-strength tuberculin sensitivity in patients with active tuberculosis. *New England Journal of Medicine*, 285(27):1506–1509, 1971.

[93] W. L. Howard. The Loss of Tuberculin Sensitivity in Certain Patients with Active Pulmonary Tuberculosis. *CHEST Journal*, 57(6):530, 1970.

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

[95] S. L. Hui and X. H. Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7(4):354–370, 1998.

[96] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

[97] H. E. Jenkins, A. W. Tolman, C. M. Yuen, J. B. Parr, S. Keshavjee, C. M. Pérez-Vélez, M. Pagano, M. C. Becerra, and T. Cohen. Incidence of multidrug-resistant tuberculosis disease in children: systematic review and global estimates. *The Lancet*, 383(9928):1572–1579, 2014.

[98] N. Jewell. Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics*, 40(2):421–435, 1984.

[99] R. E. Johnson, T. A. Green, J. Schachter, R. B. Jones, E. W. Hook, C. M. Black, D. H. Martin, M. E. St Louis, and W. E. Stamm. Evaluation of nucleic acid amplification tests as reference tests for Chlamydia trachomatis infections in asymptomatic men. *Journal of Clinical Microbiology*, 38(12):4382–4386, 2000.

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

[101] L. Joseph, T. W. Gyorkos, and L. Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–273, 1995.

[102] J. M. Kerl, U. J. Schoepf, P. L. Zwerner, R. W. Bauer, J. A. Abro, C. Thilo, T. J. Vogl, and C. Herzog. Accuracy of coronary artery stenosis detection with CT versus conventional coronary angiography compared with composite findings from both tests as an enhanced reference standard. *European Radiology*, 21(9):1895–1903, 2011.

[103] G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

[104] J. A. Knottnerus, editor. *The Evidence Base of Clinical Diagnosis*. BMJ Books, London, United Kingdom, 2002.

[105] M. Ladouceur, E. Rahme, P. Bélisle, A. N. Scott, K. Schwartzman, and L. Joseph. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. *Statistics in Medicine*, 30(21):2648–2662, 2011.

[106] R. Langeheine, J. Pannekoek, and F. van de Pol. Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24(4):492–516, 1996.

[107] P. F. Lazarsfeld. The logical and mathematical foundation of latent structure analysis. In S. Stouffer, editor, *Measurement and Prediction*, 362–412. Princeton University Press, Princeton, 1950.

[108] P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.

[109] E. Lesaffre and A. Albert. Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society. Series B*, 51(1):109–116, 1989.

[110] F. I. Lewis, G. J. Gunn, I. J. McKendrick, and F. M. Murray. Bayesian inference for within-herd prevalence of Leptospira interrogans serovar Hardjo using bulk milk antibody testing. *Biostatistics*, 10(4):719–728, 2009.

[111] J. G. Lijmer, P. M. M. Bossuyt, and S. H. Heisterkamp. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine*, 21(11):1525–37, 2002.

[112] J. G. Lijmer, B. W. Mol, S. Heisterkamp, G. J. Bonsel, M. H. Prins, J. H. van der Meulen, and P. M. Bossuyt. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 282(11):1061–1066, 1999.

[113] L. Lind-Brandberg, C. Welinder-Olsson, T. Lagergard, J. Taranger, B. Trollfors, and G. Zackrisson. Evaluation of PCR for diagnosis of Bordetella pertussis and Bordetella parapertussis infections. *Journal of Clinical Microbiology*, 36(3):679–683, 1998.

[114] S. J. Lord, L. P. Staub, P. M. M. Bossuyt, and L. M. Irwig. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ (Clinical research ed.)*, 343:d4684, 2011.

[115] T. Luijben. Equivalent models in covariance structure analysis. *Psychometrika*, 56(4):653–665, 1991.

[116] P. Macaskill, S. D. Walter, L. Irwig, and E. L. Franco. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Statistics in Medicine*, 21(17):2527–2546, 2002.

[117] A. Maydeu-Olivares and H. Joe. Limited- and full-information estimation and goodness-of-fit testing in 2 n contingency tables. *Journal of the American Statistical Association*, 100(471):1009–1020, 2005.

[118] A. L. McCutcheon. *Latent Class Analysis*. Sage, Newbury Park, CA, 1987.

[119] K. G. M. Moons, C. J. Biesheuvel, and D. E. Grobbee. Test research versus diagnostic research. *Clinical Chemistry*, 50(3):473–476, 2004.

[120] K. G. M. Moons, G. A. van Es, B. C. Michel, H. R. Buller, J. D. Habbema, and D. E. Grobbee. Redundancy of single diagnostic test evaluation. *Epidemiology*, 10(3):276–281, 1999.

[121] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, 162(1):W1–W73, 2015.

[122] K. G. M. Moons, J. A. H. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J. B. Reitsma, and G. S. Collins. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS Checklist. *PLoS Medicine*, 11(10):e1001744, 2014.

[123] K. G. M. Moons and D. E. Grobbee. When should we remain blind and when should our eyes remain open in diagnostic studies? *Journal of Clinical Epidemiology*, 55(7):633–636, 2002.

[124] C. A. Naaktgeboren, L. C. M. Bertens, M. van Smeden, J. A. H. de Groot, K. G. M. Moons, and J. B. Reitsma. Value of composite reference standards in diagnostic research. *BMJ (Clinical research ed.)*, 347:f5605, 2013.

[125] C. A. Naaktgeboren, J. A. H. de Groot, M. van Smeden, K. G. M. Moons, and J. B. Reitsma. Evaluating diagnostic accuracy in the face of multiple reference standards. *Annals of Internal Medicine*, 159(3):195–202, 2013.

[126] L. J. Nelson and C. D. Wells. Global epidemiology of childhood tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, 8(5):636–647, 2004.

[127] S. Nemes, J. Jonasson, A. Genell, and G. Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1):56, 2009.

[128] M. P. Nicol, L. Workman, W. Isaacs, J. Munro, F. Black, B. Eley, C. C. Boehme, W. Zemanay, and H. J. Zar. Accuracy of the Xpert MTB/RIF test for the diagnosis of

pulmonary tuberculosis in children admitted to hospital in Cape Town, South Africa: a descriptive study. *The Lancet Infectious Diseases*, 11(11):819–824, 2011.

[129] I. Nicolau, L. Tian, D. Menzies, G. Ostiguy, and M. Pai. Point-of-care urine tests for smoking status and isoniazid treatment monitoring in adult patients. *PloS One*, 7(9):e45913, 2012.

[130] D. L. Oberski, G. H. Kollenburg, and J. K. Vermunt. A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3):267–279, 2013.

[131] K. Ogura, T. Fujiwara, H. Yasunaga, H. Matsui, D.G. Jeon, W. H. Cho, H. Hiraga, T. Ishii, T. Yonemoto, H. Kamoda, T. Ozaki, E. Kozawa, Y. Nishida, H. Morioka, T. Hiruma, S. Kakunaga, T. Ueda, Y. Tsuda, H. Kawano, and A. Kawai. Development and external validation of nomograms predicting distant metastases and overall survival after neoadjuvant chemotherapy and surgery for patients with nonmetastatic osteosarcoma: A multi-institutional study. *Cancer*, 121(21):3844–3852, 2015.

[132] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[133] M. Pavlou, G. Ambler, S. R. Seaman, O. Guttmann, P. Elliott, M. King, and R. Z. Omar. How to develop a more accurate risk prediction model when there are few events. *BMJ*, h3868, 2015.

[134] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, 1996.

[135] M. S. Pepe. Study design and hypothesis testing. In *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Chapter 8. Oxford University Press, Oxford, 2003.

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

[137] C. M. Perez-Velez and B. J. Marais. Tuberculosis in children. *New England Journal of Medicine*, 367(4):348–361, 2012.

[138] D. H. Persing, F. C. Tenover, J. Versalovic, T. YiWei, E. R. Unger, D. A. Relman, and T. J. White. *Molecular Microbiology: Diagnostic Principles and Practice*. ASM press, 2004.

[139] M. Plummer. JAGS : A program for analysis of bayesian graphical models using gibbs sampling JAGS : Just Another Gibbs Sampler. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, 2003.

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

[141] R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, 2014.

[142] M. Reiser and Y. Lin. A goodness of fit test for the latent class model when expected frequencies are small. *Sociological Methodology*, 29(1):81–111, 1999.

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

[144] D. Rindskopf and W. Rindskopf. The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5(1):21–27, 1986.

[145] A. W. S. Rutjes, J. B. Reitsma, A. Coomarasamy, K. S. Khan, and P. M. M. Bossuyt. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technology Assessment*, 11(50):ix–51, 2007.

[146] A. W. S. Rutjes, J. B. Reitsma, M. Di Nisio, N. Smidt, J. C. van Rijn, and P. M. M. Bossuyt. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*, 174(4):469–476, 2006.

[147] E. Schuit, A. Kwee, M. Westerhuis, H. Van Dessel, G. Graziosi, J. Van Lith, J. Nijhuis, S. Oei, H. Oosterbaan, N. Schuitemaker, M. Wouters, G. Visser, B. Mol, K. Moons, and R. Groenwold. A clinical prediction model to assess the risk of operative delivery. *BJOG*, 119(8):915–923, 2012.

[148] R. Sepúlveda, J. L. Vicente-Villardón, and M. P. Galindo. The Biplot as a diagnostic tool of local dependence in latent class models. A medical application. *Statistics in Medicine*, 27(11):1855–1869, 2008.

[149] L. A. Shrier, D. Dean, E. Klein, K. Harter, and P. A. Rice. Limitations of screening tests for the detection of Chlamydia trachomatis in asymptomatic adolescent and young adult women. *American Journal of Obstetrics and Gynecology*, 190(3):654–62, 2004.

[150] V. Siba, P. F. Horwood, K. Vanuga, J. Wapling, R. Sehuko, P. M. Siba, and A. R. Greenhill. Evaluation of serological diagnostic tests for typhoid fever in Papua New Guinea using a composite reference standard. *Clinical and Vaccine Immunology*, 19(11):1833–1837, 2012.

[151] A. Sinclair, X. Xie, M. Teltscher, and N. Dendukuri. Systematic review and meta-analysis of a urine-based pneumococcal antigen test for diagnosis of community-acquired

pneumonia caused by Streptococcus pneumoniae. *Journal of Clinical Microbiology*, 51(7):2303–2310, 2013.

[152] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. CRC Press, 2004.

[153] B. D. Spencer. When do latent class models overstate accuracy for diagnostic and other classifiers in the absence of a gold standard? *Biometrics*, 68(2):559–566, 2012.

[154] M. Staquet, M. Rozencweig, Y. J. Lee, and F. M. Muggia. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases*, 34(12):599–610, 1981.

[155] E. W. Steyerberg. *Clinical Prediction Models*. Statistics for Biology and Health. Springer New York, New York, NY, 2009.

[156] E. W. Steyerberg, M. Schemper, and F. E. Harrell. Logistic regression modeling and the number of events per variable: selection bias dominates. *Journal of Clinical Epidemiology*, 64(12):1464–1465, 2011.

[157] A. Subtil, M. R. de Oliveira, and L. Gonçalves. Conditional dependence diagnostic in the latent class model: a simulation study. *Statistics & Probability Letters*, 82(7):1407–1412, 2012.

[158] S. Swaminathan and B. Rekha. Pediatric Tuberculosis: Global Overview and Challenges. *Clinical Infectious Diseases*, 50(s3):S184–S194, 2010.

[159] L. Thibodeau. Evaluating diagnostic tests. *Biometrics*, 37(4):801–804, 1981.

[160] V. L. Torrance-Rynard and S. D. Walter. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16(19):2157–2175, 1997.

[161] L. Tuyisenge, C. P. Ndimubanzi, G. Ndayisaba, N. Muganga, J. Menten, M. Boelaert, and J. Van Den Ende. Evaluation of latent class analysis and decision thresholds to guide the diagnosis of pediatric tuberculosis in a rwandan reference hospital. *Pediatric Infectious Disease Journal*, 29(2):e11–e18, 2010.

[162] J. S. Uebersax. Validity inferences from interobserver agreement. *Psychological Bulletin*, 104(3):405–416, 1988.

[163] US Food and Drug Administration. Draft guidance for industry and food and drug administration staff - establishing the performance characteristics of in vitro diagnostic devices for Chlamydia trachomatis and/or Neisseria gonorrhoea: screening and diagnostic testing. Technical report, 2011.

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

[165] V. Vadwai, C. Boehme, P. Nabeta, A. Shetty, D. Alland, and C. Rodrigues. Xpert MTB/RIF: a new pillar in diagnosis of extrapulmonary tuberculosis? *Journal of Clinical Microbiology*, 49(7):2540–2545, 2011.

[166] P. N. Valenstein. Evaluating diagnostic tests with imperfect standards. *American Journal of Clinical Pathology*, 93(2):252–8, 1990.

[167] G. H. van Kollenburg, J. Mulder, and J. K. Vermunt. Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology*, 11(2):65–79, 2015.

[168] M. van Smeden, C. A. Naaktgeboren, J. B. Reitsma, K. G. M. Moons, and J. A. H. de Groot. Latent class models in diagnostic studies when there is no reference standard-a systematic review. *American Journal of Epidemiology*, 179(4):423–431, 2014.

[169] Y. Vergouwe, E. W. Steyerberg, M. J. Eijkemans, and J. D. F. Habbema. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, 58(5):475–483, 2005.

[170] J. K. Vermunt and J. Magidson. *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont Massachusetts, 2005.

[171] E. Vittinghoff and C. E. McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6):710–718, 2007.

[172] S. D. Walter. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology*, 10(1):67–72, jan 1999.

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

[174] S. D. Walter, P. Macaskill, S. J. Lord, and L. Irwig. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Statistics in Medicine*, 31(11-12):1129–1138, 2012.

[175] M. E. M. H. Westerhuis, G. H. A. Visser, K. G. M. Moons, E. van Beek, M. J. Benders, S. M. Bijvoet, H. J. H. M. van Dessel, A. P. Drogtrop, H. P. van Geijn, G. C. Graziosi, F. Groenendaal, J. M. M. van Lith, J. G. Nijhuis, S. G. Oei, H. P. Oosterbaan, M. M. Porath, R. J. P. Rijnders, N. W. E. Schuitemaker, L. M. Sopacua, I. van der Tweel, L. D. E. Wijnberger, C. Willekes, N. P. A. Zuithoff, B. W. J. Mol, and A. Kwee. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring. *Obstetrics & Gynecology*, 115(6):1173–1180, 2010.

[176] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, and S. Mallett. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology*, 66(10):1093–104, 2013.

[177] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. G. Leeflang, J. A. C. Sterne, P. M. M. Bossuyt, and QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, 155(8):529–536, 2011.

[178] World Health Organization. Global tuberculosis report. Technical report, 2014.

[179] A. Worster and C. Carpenter. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. *CJEM*, 10(2):174–175, 2008.

[180] H. Xu and B. A. Craig. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65(4):1145–1155, 2009.

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

[182] M. A. Young. Evaluating diagnostic criteria: a latent class paradigm. *Journal of Psychiatric Research*, 17(3):285–296, 1982.

[183] H. J. Zar, T. G. Connell, and M. Nicol. Diagnosis of pulmonary tuberculosis in children: new advances. *Expert Review of Anti-infective Therapy*, 8(3):277–288, 2010.

[184] H. J. Zar, L. Workman, W. Isaacs, K. Dheda, W. Zemanay, and M. P. Nicol. Rapid diagnosis of pulmonary tuberculosis in African children in a primary care setting by use of Xpert MTB/RIF on respiratory specimens: a prospective study. *The Lancet. Global Health*, 1(2):e97–104, 2013.

[185] H. J. Zar, L. Workman, W. Isaacs, J. Munro, F. Black, B. Eley, V. Allen, C. C. Boehme, W. Zemanay, and M. P. Nicol. Rapid molecular diagnosis of pulmonary tuberculosis in children using nasopharyngeal specimens. *Clinical Infectious Diseases*, 55(8):1088–1095, 2012.

Summary

Summary

A key problem faced in many diagnostic studies is the absence of a single reference standard that can accurately distinguish between patients with and without the target condition. This problem is commonly referred to as absence of a gold standard. In the absence of a gold standard, the classification of the target condition is prone to error. When these classification errors are ignored, the evaluation of the accuracy of the test(s) under evaluation or the estimation of the prevalence of the target condition can be severely biased.

This thesis examines potential solutions that are aimed at alleviating the problems associated with the absence of a gold standard. In particular, we focus on latent class modeling and composite reference standards: two commonly used methods in diagnostic test evaluation literature.

Chapter 2 provides a systematic review of the use of latent class models in diagnostic accuracy studies where there is no acceptable reference standard. We identify a sharp increase in the use of latent class models in the past decade, notably in the domain of infectious diseases. A critical assumption underlying the majority of the reviewed latent class model applications was that diagnostic test results are stochastically independent conditional on the unobserved (latent) disease status. Testing this local independence assumption is essential as its violation can lead to biased inference about the true value of diagnostic tests and prevalence of disease. Our review shows that 28% of the reviewed studies failed to report any information that enables verification of model assumptions or performance.

Chapter 3 examines the performance of goodness-of-fit testing to detect violation of the local independence assumption underlying the criticized 'standard' 2-class latent class model. Our Monte Carlo simulation study shows that goodness-of-fit tests lack power to detect relevant local independence violations at sample sizes that are typically found in empirical diagnostic studies. The study also shows that a parametric bootstrap procedure improves the evaluation of goodness-of-fit in terms of Type-I error control in the case of sparse diagnostic test data.

Chapter 4 presents a Bayesian latent class analysis on the results of five diagnostic tests for childhood pulmonary tuberculosis obtained from 749 hospitalized South African children. Our study confirms the widely accepted belief that commonly used confirmatory tests for pulmonary tuberculosis lack sensitivity when performed in children. Furthermore, using the latent class analysis we estimate that approximately 46% of true pulmonary tuberculosis negative children in the cohort received anti-tuberculosis treatment, indicating substantial overtreatment.

Chapter 5 revisits the evidence presented in the influential paper by Albert and Dodd, *Biometrics*, 2004, that cautioned against the use of latent class models. We identify problems with the evidence from their simulations that have important consequences for the interpretation. Also, later studies building on the 2004 paper suffer from the same problems. A targeted simulation study shows that the evidence on non-distinguishability between alternative random effect latent class models is not as conclusive as earlier suggested. We argue that new research is needed showing if and when latent class analysis yields valid results for drawing inferences about diagnostic test accuracy and disease prevalence.

Chapter 6 focuses on composite reference standards, pre-defined classification rules combining the results of two or more imperfect component diagnostic tests. The aim of the composite reference standard approach is to improve the accuracy of classification of subjects with respect to their target condition over that of any of the individual component tests in isolation. We provide suggestions for improving the transparency of reporting such composite reference standard analyses.

Chapter 7 studies the potential bias in estimates of diagnostic test accuracy that can be anticipated when using composite reference standards. We show that a composite reference standard can lead to significantly biased inferences about the diagnostic value of test(s) under study. In an individual diagnostic study, the magnitude and direction of this bias may be difficult to predict and adjust for, especially when diagnostic test results are not stochastically independent conditional on the (latent) disease status.

In **Chapter 8** the widely used criterion of a minimum of 10 events per variable for logistic regression analysis is challenged. Reasons for the large heterogeneity in results and recommendations between earlier events per variable simulation studies are identified and explained. It is shown that the issue known as 'separation' can dominate the results of such a simulation study. A simple correction method (known as Firths correction) can however alleviate some of the problems associated with a low number of events per variable in logistic regression.

Chapter 9 outlines a general approach for deriving scoring charts and nomograms for multinomial logistic regression models. We argue that using a scoring chart in combination with a nomogram can improve the reporting of multinomial logistic regression models and facilitate clinical decision making during clinical encounters. A clinical example on the risk of several types of operative delivery is used to illustrate the application.

The absence of a gold standard for many diseases poses a frequently encountered problem in diagnostic research. There is currently no consensus on how to best address this problem. In this thesis we have shown that composite reference standards can cause large biases in inferences about diagnostic test accuracy and target condition disease prevalence. It was also shown that latent class analysis should be used with caution. In **Chapter 10** we suggest various approaches to improve latent class analysis by augmenting the diagnostic test data, for example by using covariate data or by elicitation of informative prior distributions in a Bayesian analysis. Future research should be directed at further elucidating the merits and pitfalls of statistical modeling to account for the absence of a gold standard.

Nederlandse samenvatting

Een veelvoorkomend probleem in diagnostisch onderzoek naar de waarde van diagnostisch testen is het ontbreken van een referentiestandaard waarmee met zekerheid de aanwezigheid van de aandoening onder studie (zoals een specifieke ziekte) kan worden aangetoond of uitgesloten. Dit probleem staat bekend als het ontbreken van een gouden standaard. Het ontbreken van een gouden standaard leidt vaak tot fouten in de vaststelling van de aandoening onder studie. Het negeren van deze fouten leidt vervolgens tot vertekende schattingen van de accuratesse van de diagnostische test(en) onder studie of tot vertekende schattingen van de prevalentie van de aandoening.

In dit proefschrift onderzoeken we potentiële oplossingen voor het probleem van het ontbreken van een gouden standaard. We richten ons specifiek op latente klasse analyse en samengestelde (composite) referentiestandaarden. Beide methoden worden veelvuldig gebruikt in studies naar de accuratesse van diagnostisch testen wanneer een gouden standaard ontbreekt.

Hoofdstuk 2 geeft een systematisch literatuuroverzicht van het gebruik van latente klasse modellen in diagnostische studies waarin een gouden standaard ontbreekt. Het gebruik van latente klasse modellen is de afgelopen tien jaar sterk toegenomen. Deze toename is vooral aanwezig in het domein van infectieziekten. Een cruciale aanname die ten grondslag ligt aan de meerderheid van de onderzochte latente klasse analyses is dat de diagnostische testresultaten onderling onafhankelijk zijn conditioneel op de latente aandoeningsstatus (aan- of afwezigheid van de aandoening). Het testen van deze aanname van lokale onafhankelijkheid is essentieel omdat schending ervan kan leiden tot vertekende schattingen van de accuratesse van diagnostische testen onder studie en de prevalentie van de aandoening. Ons overzicht laat zien dat in 28% van de onderzochte studies geen enkele informatie wordt gerapporteerd waarmee de modelaannames of modelprestaties van de gebruikte latente klasse modellen kunnen worden gecontroleerd.

Hoofdstuk 3 evalueert de prestaties van 'goodness-of-fit' testen om schendingen van de aanname van lokale onafhankelijkheid te detecteren bij het gebruik van een standaard 2-klassen latente klasse model. Onze Monte Carlo simulatiestudie laat zien dat de gebruikelijke goodness-of-fit testen te weinig power hebben om relevante schendingen van de lokale onafhankelijkheidsaanname te detecteren in steekproeven met gangbare grootte. Onze studie laat ook zien dat een parametrische bootstrap procedure de evaluatie van goodness-of-fit verbetert in termen van Type-I fout controle in diagnostische studies waar de data dun verspreid zijn.

Hoofdstuk 4 presenteert een Bayesiaanse latente klasse analyse in een studie met 749 in het ziekenhuis opgenomen Zuid-Afrikaanse kinderen die vijf diagnostische testen voor pulmonaire tuberculose hebben ondergaan. Onze studie bevestigt de algemene overtuiging dat enkele veelgebruikte testen om pulmonaire tuberculose aan te tonen een lage sensitiviteit hebben in kinderen. Verder wordt met behulp van de latente klasse analyse geschat dat circa 46% van kinderen zonder pulmonaire tuberculose toch anti-tuberculose behandeling hebben gekregen.

Hoofdstuk 5 onderzoekt de resultaten van het invloedrijke manuscript van Albert en Dodd, *Biometrics*, 2004, waarin gewaarschuwd wordt voor het gebruik van latente klasse modellen. We identificeren problemen met het gepresenteerde bewijs in het manuscript uit 2004 en in latere studies over dit onderwerp. Deze problemen hebben belangrijke consequenties voor de interpretatie van de studieresultaten. Een nieuwe, doelgerichte simulatiestudie laat zien dat het eerder aangetoonde gebrek aan onderscheid tussen verschillende random effect latente klasse modellen minder algemeen geldend is dan eerder gesuggereerd werd. Nieuw onderzoek is nodig waaruit blijkt in welke situaties latente klasse analyse valide resultaten oplevert en wanneer niet.

In **Hoofdstuk 6** ligt de focus op samengestelde referentiestandaarden. Dit zijn vooraf gedefinieerde classificatieregels waarmee de resultaten van twee of meer imperfecte (component) diagnostische testen worden gecombineerd. Het doel van het gebruik van de samengestelde referentiestandaard is het verbeteren van de vaststelling van de aandoening onder studie ten opzichte van de classificatie op basis van een enkele component test. Suggesties voor het verbeteren van de volledigheid en transparantie van de rapportage van samengestelde referentiestandaard analyses worden gegeven.

Hoofdstuk 7 bestudeert de vertekening in schattingen van de accuratesse van diagnostische testen door het gebruik van samengestelde referentiestandaarden. We laten zien dat een samengestelde referentiestandaard tot sterk vertekende conclusies over de waarde van diagnostische testen kan leiden. De omvang en richting van deze vertekening is moeilijk te voorspellen en te corrigeren, vooral wanneer de diagnostische test resultaten lokaal afhankelijk zijn. Wij geven expliciete formules waarmee de omvang van deze vertekening kan worden bestudeerd.

In **Hoofdstuk 8** wordt het veelgebruikte criterium van 10 events per variabele voor logistische regressie analyse betwist. In dit hoofdstuk tonen wij waarom eerdere simulatiestudies over events per variabele grote verschillen in resultaten en aanbevelingen lieten zien. We laten zien

dat een probleem dat bekend staat als separatie de resultaten van een simulatiestudie kunnen domineren. Een eenvoudige correctiemethode (bekend als de correctie volgens Firth) kan de vertekening in schattingen van de regressie-coëfficiënt aanzienlijk verlagen in logistische regressie met weinig data.

Hoofdstuk 9 beschrijft een algemene aanpak voor het afleiden van scorekaarten en nomogrammen voor multinomiale logistische regressie modellen. We laten zien dat het gebruik van een scorekaart in combinatie met een nomogram de rapportage van de resultaten van een multinomiaal logistisch regressie model kan verbeteren en bovendien klinische beslissingen kunnen faciliteren. Wij illustreren onze aanpak aan de hand van een klinisch voorbeeld over het risico op verschillende vormen van operatieve bevallingen.

Het ontbreken van een gouden standaard voor veel aandoeningen is een veelvoorkomend probleem in diagnostisch wetenschappelijk onderzoek. Er is momenteel geen consensus over hoe dit probleem het best kan worden aangepakt. In dit proefschrift hebben we laten zien dat het gebruik van samengestelde referentiestandaarden kan leiden tot sterk vertekende conclusies over de accuratesse van diagnostische testen en de prevalentie van de aandoening onder studie. Ook het gebruik van latente klasse analyse is niet zonder risicos. In **Hoofdstuk 10** worden methoden aangedragen waarmee latente klasse analyse kan worden verbeterd door extra data toe te voegen aan de resultaten van de diagnostische testen, bijvoorbeeld door het gebruik van covariabelen en informatieve priorverdelingen in een Bayesiaanse analyse. Toekomstig onderzoek moet gericht zijn op het verder aanscherpen van de mogelijkheden en grenzen van statistische oplossingen bij afwezigheid van een gouden standaard.

List of affiliations

List of affiliations

Douglas G Altman	Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, United Kingdom
Loes CM Bertens	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Gary S Collins	Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, United Kingdom
Nandini Dendukuri	Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
Marinus JC Eijkemans	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Joris AH de Groot	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Alula Hadgu	Division of STD Prevention, Centers for Disease Control, Atlanta, U.S.A.
Lawrence Joseph	Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
Michael Libman	Division of Infectious Diseases, McGill University Health Centre, Montreal, Canada
Karel GM Moons	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Christiana A Naaktgeboren	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Mark P Nicol	Division of Medical Microbiology and Institute for Infectious Diseases and Molecular Medicine, University of Cape Town, National Health Laboratory Service, Groote Schuur Hospital, Cape Town, South Africa
Stavros Nikolakopoulos	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Daniel L Oberski	Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands
Madhukar Pai	Department of Epidemiology, Biostatistics and Occupational Health, McGill University, and McGill International TB Centre, Montreal, Canada
Johannes B Reitsma	Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands
Ian Schiller	Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
Samuel G Schumacher	Department of Epidemiology, Biostatistics and Occupational Health, McGill University, and McGill International TB Centre, Montreal, Canada
Jeroen K Vermunt	Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands
Heather Zar	Department of Paediatrics and Child Health, and MRC Unit on Child & Adolescent Health, University of Cape Town and Red Cross War Memorial Childrens Hospital, Cape Town, South Africa

Dankwoord

Prof. dr. K.G.M. Moons, geachte promotor, beste Carl, promovendus zijn onder jouw supervisie is enorm motiverend: jouw enthousiasme voor relevante methodologische onderwerpen is aanstekelijk. Het meest bijzonder vind ik het geduld dat je toonde en het vertrouwen dat je me gaf wanneer een onderzoek weer eens een heel andere kant op ging dan oorspronkelijk gepland. Ik ben trots dat ik mijn promotieonderzoek onder jouw supervisie heb mogen doen en ik hoop in de toekomst onze samenwerking te kunnen blijven voortzetten.

Dr. J.B. Reitsma, geachte co-promotor, beste Hans, je brede kennis en interesse zijn bewonderenswaardig. Tegelijkertijd schuw je ook de diepgang niet: onze discussies over statistisch-technische details duurde vaak tot ver na 17 uur. Bovenal ben je een bijzonder prettige supervisor waarbij ik me altijd welkom voelde om te praten over alle mogelijke onderwerpen: van onderzoekinhoud tot en met de voetbaluitslagen. Ik hoop onze samenwerking op het gebied van methodologie- en statistiekonderzoek ook in de toekomst te kunnen blijven voortzetten.

Dr. J.A.H. de Groot, geachte co-promotor, beste Joris, om je àltijd positieve houding kun je niet heen. Onze wekelijkse gesprekken waren daardoor altijd gezellig en ongedwongen. Bovendien werden in deze gesprekken de door mij aangedragen problemen vaak snel en behendig door je opgelost. Jouw proefschrift en onze gesprekken zijn heel belangrijk geweest voor de totstandkoming van dit proefschrift. Ook met jou hoop ik de samenwerking in de toekomst te kunnen blijven voortzetten.

Dr N. Dendukuri, dear Nandini, your input and support has truly been invaluable. I am very grateful that I have had the opportunity to visit you at McGill University for little over 6 months. These months have not only been a great learning experience, it has been essential in shaping this thesis. Foremost, I admire your positive attitude and extensive knowledge. I have really enjoyed collaborating on our projects and I look forward to continuing our collaboration in the future.

Dr C.A. Naaktgeboren, beste Christiana, je bent bovenal een zeer aangenaam persoon om mee samen te werken. Met jouw doelgerichtheid ga je het zeker ver schoppen in de wetenschap.

Many co-authors have contributed to the research presented in this thesis. I thank all of the authors for the numerous discussions and commitment to our projects.

De leden van de beoordelingscommissie bestaande uit: Prof dr. T. Leiner (voorzitter), Prof. dr. Ir. M.J.C. Eijkemans, Prof. dr. M. Nielen, Prof. dr. H.J.A. Hoijtink en Prof. dr. A.H. Zwinderman

dank ik voor hun bereidheid dit proefschrift te lezen en te beoordelen.

Het wekelijkse methodologieoverleg is een belangrijke inspiratiebron en klankbord geweest. Dank aan alle betrokken junior en senior onderzoekers voor de open discussies.

Veel dank ben ik verschuldigd aan al mijn collegae van de afdeling Biostatistiek en Research Support. Ik wil enkele van hen in het bijzonder noemen. René, dank voor de ruimte die je me het afgelopen jaar gaf om mijn proefschrift af te ronden. Het docententeam, en dan vooral Cas, Paul en Rebecca, door jullie betrokkenheid in het algemeen en onderwijs in het bijzonder, plus het in mij gestelde vertrouwen heb ik met veel plezier (veel) onderwijservaring kunnen opdoen. Kamergenote Caroline, vooral dank voor je bemoedigende woorden in de laatste weken voor de deadline van dit proefschrift.

Speciale dank aan mijn kamergenoten van Stratenum 6.101: Anoukh, Maaike, Manon, Noor, Sara, Sophie en Willemijn. De goede sfeer in onze kamer is een echte motivatiebron geweest. Nog leuker zijn de herinneringen aan de activiteiten buiten werktijd met ook belangrijke inbreng van: Julien, Floriaan, Stavros en Wouter. Naast kamergenoten zijn er heel veel andere collegas van het Julius Centrum en het UMCU die ik ook wil bedanken voor gezelligheid en goede gesprekken en discussies: Anneke, Ardine, Axel, Bert, Carla, Coby, Douwe, Erik, Eva, Ewoud, Faas, Ganna, Gbenga, Giske, Hanneke, Heidi, Henk, Henri, Henrike, Henok, Indira, Janneke, John, Jorien, Judith, Ingeborg, Karlijn, Katrien, Kim, Linda, Loes, Lotty, Maaike, Marie, Marieke, Marijn, Mart, Martine, Mirjam, Nienke, Paco, Peter, Pushpa, Putri, Rob, Rolf, Romin, Rutger, Sander, Shona, Thomas, Timo, Victor, Welling, WJK, en alle (andere) JOB-borrelaars en Promovenskiërs.

I am very grateful to the Methods and Statistics of Social and Behavioral Science Masters program. Where thanks to the excellent lectures my interest in statistical and methodological research has begun. Foremost, I thank my fellow students, my dear friends: Anja, Rianne, Stavros and Suzette. Many thanks to dr. David Hessen for his excellent guidance and supervision during my Masters thesis project.

Lieve vrienden en familie, dank voor de leuke momenten buiten werktijd die er voor zorgden voor de broodnodige afleiding en mij er aan herinneren dat er meer is dan onderzoek alleen. Many thanks to my Canadian friends Ben Rich and Stephan Trudel for making Montreal feel like home.

Beste Kees, ik wil je bedanken voor de wiskunde bijscholing die je me hebt gegeven. Zonder

jouw ondersteuning op dat moment was dit proefschrift er waarschijnlijk nooit gekomen.

Beste Mark, met veel plezier denk ik terug aan ons SenS-project, daar waar mijn interesse in onderzoek is begonnen. Ik weet zeker dat je mooie jaren tegemoet gaat!

Paranimfen, Gerard en Peter, ik ben trots dat jullie deze dag naast mij willen staan.

Schoonfamilie, dank dat ik altijd welkom ben geweest in jullie gezinnen. Frouwke en Peter, ik ben jullie erg dankbaar voor al jullie ondersteuning en betrokkenheid voor en tijdens mijn promotietraject.

Lieve ouders en Bauke, dank voor jullie nooit aflatende steun en interesse. Bauke, met trots kijk ik naar wat je al bereikt hebt. Pa en ma, ik denk met veel plezier terug aan jullie bezoek in Montreal.

Lieve Martina, liefste Marie, dit proefschrift was er natuurlijk nooit gekomen zonder jouw steun. Ik bewonder je niet-ophoudende interesse in de onderwerpen van mijn proefschrift, ik geniet van je zorgzaamheid en put inspiratie uit jouw toewijding aan Cacao. Ik zie uit naar meer tijd samen nu het proefschrift is afgerond.

Curriculum vitae

Maarten van Smeden was born on May 16th, 1986 in Delft, the Netherlands. As of 2006 he studied Psychology focussing on Clinical Psychology and Cognitive Neuroscience at the University of Utrecht. In 2009 he graduated as Bachelor of Science in Psychology. In 2011 he graduated as Master of Science in Methods and Statistics of Behavioral and Social Sciences (cum laude) at the University of Utrecht. During his bachelor and masterstudies he was co-owner of onderzoeksbureau SenS. In June 2011 he started his PhD project, resulting in the research presented in this thesis, at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, under supervision of prof dr. K.G.M. Moons, dr. J.B. Reitsma and dr J.A.H. de Groot. During his PhD studies he was actively involved in teaching as assistant, lecturer and coordinator in several statistics courses on bachelor, master and PhD level. In 2014 he graduated as Master of Science in Epidemiology at the University of Utrecht. Between August 2013 and March 2014 he worked as a visiting PhD candidate at the Department of Epidemiology and Biostatistics at McGill University, Montreal (CA), under supervision of Dr. Nandini Dendukuri. As of Januari 2015, he works as a researcher/lecturer in biostatistics at the Julius Center for Health Science and Primary Care.