INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Assessing Accuracy of a Continuous Medical Diagnostic or Screening Test in the Presence of Verification Bias

Todd Allen Alonzo

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2000

Program Authorized to Offer Degree: Biostatistics

UMI Number: 9975947

Copyright 2000 by
Alonzo, Todd Allen

All rights reserved.

UMI Microform 9975947
Copyright 2000 by Bell & Howell Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© Copyright 2000

Todd Allen Alonzo

In presenting this dissertation in partial fulfillment of the requirements for the Doctorial degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholary purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Todd A. Alonzo

Date 5/25/00

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Todd Allen Alonzo

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Margaret Sullivan Pepe

Margaret Sullivan Pepe

Reading Committee:

Patrick J. Heagerty

Margaret Sullivan Pepe

Mary Lou Thompson

Date: 5-25-00

University of Washington

Abstract

Assessing Accuracy of a Continuous Medical Diagnostic or Screening Test in the Presence of Verification Bias

by Todd Allen Alonzo

Chair of Supervisory Committee

Professor Margaret Sullivan Pepe
Biostatistics

In studies to assess accuracy of a medical diagnostic or screening test, often definitive disease assessment is too invasive or expensive to be ascertained in all subjects. It makes practical and ethical sense to select a higher percentage of subjects with a positive screening test for disease assessment than patients with a negative screening test. However, analyses using only the results for this subset of subjects can cause biased accuracy estimates. This is known as verification bias or work-up bias.

Many screening tests, such as tests based on serum concentrations or biomarkers, yield continuous results; however, available bias-correction methods can only accommodate tests with discrete responses. We propose new verification bias-correction methods to assess accuracy of a continuous test based on (1) an extension of the work of Begg and Greenes (Biometrics, 1983) for discrete tests and (2) mean score, (3) inverse probability weighting, and (4) semi-parametric efficient methods. Asymptotic distribution theory is developed for estimators of disease prevalence, true positive rate, and false positive rate. The different methods are compared using theoretical and simulation results and applied to data from a newborn hearing screening study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I gratefully acknowledge the support provided by a Predoctoral Training Grant from the National Institute of Health (NIH). I would like to thank Susan Norton for permitting access to the audiology data analyzed in this dissertation and Kristin Fletcher for supplying the data.

I would also like to thank the members of my supervisory committee, Tom Fleming, Patrick Heagerty, and Mary Lou Thompson, and Thomas Lumley for their invaluable feedback. I am especially grateful to my dissertation advisor, Margaret Sullivan Pepe, for her guidance, encouragement, and friendship. I look forward to collaborating with her in the future and to exchanging soccer stories.

Many thanks to my classmates and faculty here at UW for making my stay in Seattle such a positive experience. I especially would like to thank my dear friends Tom Braun and Jen Nelson for showing me the ropes and Katherine Guthrie and Julie Stoner for facing it with me. Their support and friendship mean a lot to me.

To my family who continually asked "are you finished with that paper yet?", I am finally done! Thank you for all your love and encouragement.

To my son Peyton, thank you for your contagious laughter and for helping me stay focused on what is important in life. You have taught me so much these past 6 months.

And finally, to my wife, Kim, words cannot express how grateful I am for

your love, encouragement, support, and understanding of the life of a grad student. This is dedicated to you.

Chapter 1

INTRODUCTION

1.1 Aims

The goal of this dissertation is to develop and compare methods for assessing the accuracy of a continuous medical diagnostic or screening test when selection for true disease status ascertainment depends on the results of the test under investigation. Methods for assessing accuracy of continuous tests in the ideal setting when there are complete data are reviewed in Chapter 2. Existing methods for assessing the accuracy of binary and ordinal tests when verification bias exists are reviewed in Chapter 3. In Chapter 4, new approaches for assessing accuracy of continuous tests in the presence of verification bias are proposed. Inferential methods are derived for these methods in Chapter 5. Chapter 6 compares the different approaches through the use of simulation studies. Application of the estimators in the analysis of data from the Neonatal Hearing Screening Project is given in Chapter 7. Conclusions along with plans for future work are summarized in Chapter 8.

1.2 Motivation

Disease screening tests and medical diagnostic tests play an important role in health care. These tests are used in all areas of medicine to detect and diagnose diseases, infections, and medical conditions. Results of these tests are used by physicians and medical care providers to determine appropriate care for patients. The number of

diagnostic tests available for use is increasing dramatically as technologic advances are made. For example, the invention of Polymerase Chain Reaction (PCR) has resulted in numerous screening tests for sexually transmitted diseases. Since diagnostic tests play such an important role in health care and they are being developed at an increasing rate, it is imperative that the accuracy, how well the diagnostic test results correspond with the true state of disease, of these tests be assessed before they are used in practice.

Ideally a diagnostic test is perfect in its diagnosis. However, that is rarely the case. There are several consequences of inaccurate tests. First, failing to treat diseased subjects who are incorrectly diagnosed can have grave ramifications. Next, treating subjects for disease they do not have can be emotionally painful and, in some situations, can be detrimental to the health of the person. Finally, inaccurate tests can accrue unnecessary costs due to the further work-ups ordered after miss-diagnosis. Therefore, before a new test is used in practice, it is imperative to understand its accuracy for detecting disease.

The Federal Food Drug & Cosmetic (FD&C) Act authorizes the U.S. Food and Drug Administration (FDA) to regulate medical devices to assure their safety and effectiveness. According to section 201(h) of the FD&C Act a medical device is "'an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or animals." Therefore, many new screening tests and diagnostic tests are considered medical devices and are, therefore, regulated by the FDA. The FDA recognizes the importance of studying the performance of tests before they are used in practice. In fact, the FD&C Act gives the FDA the authority to regulate the design, clinical evaluation, manufacturing, packaging, labeling, and post market surveliance of medical devices. In particular, the FDA requires that the accuracy be quantified and included in a package insert that

accompanies many new diagnostic and screening tests. These inserts are often used by clinicians to guide their selection of a test.

1.3 Verification Bias

Not only is it important to assess accuracy before tests are used in practice, it is also critical that proper statistical methods exist and are used to evaluate the accuracy. The accuracy of a new test is ideally evaluated by comparison with a perfect gold standard (GS) test which assesses disease status with certainty. In practice, however, a gold standard may be too expensive, such as a behavioral test to diagnose hearing impairment in infants, or too invasive, such as biopsy for diagnosing prostate cancer, for regular use. As a result, it is not feasible for the true disease status to be obtained for all subjects. Although it may be more cost-effective and ethical to ascertain the true disease status with a higher frequency in subjects where the new test suggests disease, estimates of accuracy can be biased in studies with such designs. This bias is known as verification bias (Begg & Greenes, 1983) or work-up bias (Ransohoff & Feinstein, 1978).

Verification bias occurs when subjects selected for disease verification are not a simple random sample of those tested with the new test. Often test results from the test in which accuracy is being assessed is used to identify subjects at higher risk and determine which subjects will receive disease verification. In studies with sampling schemes like this, estimates of the accuracy of a test will suffer from verification bias. The magnitude of the bias depends on the inherent accuracy of the new test and the verification mechanism (Begg & Greenes, 1983). If the bias is ignored, incorrect conclusions regarding the accuracy of new tests may be drawn.

To illustrate how verification bias affects estimates of test accuracy, consider a hypothetical example in the prostate cancer screening setting where the goal is to assess the accuracy of a binary test (e.g. prostate specific antigen (PSA) $>$ 4.0 ng/ml)

Table 1.1: Hypothetical example of verification bias.

	Truth		Observe	
	Diseased	Non-diseased	Diseased	Non-diseased
New Test +	80	90	80	90
New Test −	20	810	2	81
	100	900	82	171

and the gold standard is biopsy. Assume 1000 men receive the new test, prevalence of prostate cancer is 10%, and true sensitivity and specificity of the new test are 0.80 and 0.90, respectively. Further assume that only 10% of subjects with a negative test result are verified using biopsy while all subjects who test positive are biopsied. Table 1.1 summarizes the true data and the data we observe by only verifying a subset of the men. Using the observed data, the sensitivity of the new test is 80/82=0.976 and specificity is 81/171=0.474. Clearly, the estimate of sensitivity is biased upwards and the estimate of specificity is biased downwards. Thus, the diagnostic test appears to be more sensitive and less specific than it actually is.

Even though it is known that verification bias can distort the assessment of diagnostic tests, many studies fail to recognize verification bias. For example, in a review by Reid et al. (1995) 54% of the 112 studies identified through a MEDLINE search of the medical literature from 1978 to 1993 had verification bias; Greenes & Begg (1985) found that at least 26% of 145 studies of diagnostic tests found in a MEDLINE search between 1976 and 1980 had the potential for verification bias and no way of correcting

for the bias; Bates et al. (1993) reviewed 54 studies of pediatric diagnostic tests and found that 36% of the studies were subject to verification bias; Fahey et al. (1995) found that 82% of 62 studies of accuracy of Pap tests had potential for verification bias; and Philbrick et al. (1980) reviewed 33 studies of exercise tests for diagnosing coronary heart disease and found that 94% may have had verification bias.

In practice it is common for the results of the test being evaluated to influence the selection of subjects to receive further work-up to determine disease status. Table 1.2, reproduced from Begg (1987), provides examples of studies with clearly unequal disease verification percentages. For example, Marshall et al. (1984) studied the accuracy of diaphanography for detecting breast cancer. In their study, 833 subjects were tested for breast cancer using diaphanography resulting in 67 positive results and 766 negative results. Biopsy was considered the gold standard test. Fifty-five percent of the test positives were biopsied while only 7% of those with negative and indeterminate test results received disease verification. Clearly, those subjects who were biopsied do not represent a simple random sample, and thus, this study may suffer from verification bias.

Study designs of all types, including prospective, retrospective, observational, and randomized, are subject to verification bias. However, the verification mechanism, which is an integral part of some of the bias-correction methods discussed in Chapters 3 and 4, is more likely to be known in prospective studies than in retrospective studies. For example, in clinical trials verification can be defined by protocol. Conversely, in observational studies often the mechanism is not known, although one suspects that there is an association between the missingness and perceived risk.

1.4 Auxiliary Information

When subjects are referred for a diagnostic test they may have signs, symptoms, results from other tests, or a medical history that are suggestive of disease. This

Table 1.2: Examples of studies where test results appear to influence selection for disease verification.

Author	Test	Result	% Verified
Marshall et al., 1984	Diaphanography	+	55%
	for breast cancer	other	7%
Drum & Christacopoulos, 1972	Hepatic scintigraphy	+	61%
	for liver disease	−	37%
McNeil et al., 1981	Computed tomography	++	77%
	tomography	+	44%
	for fever	0	50%
		−	39%
		− −	27%
Barr & Schumaker, 1984	Serum theophylline	>30 mg/ml	74%
	as toxicity predictor	25−30	42%
		20−25	38%
		15−20	35%
		10−15	22%
		<10	13%
Diamond et al., 1986	Exercise	+	31%
	ventriculography	−	14%
	for CAD		

information will be referred to as auxiliary data and may be any information, other than results from the test for which accuracy is being assessed, that may be informative about disease status. In practice, clinicians often use auxiliary data to identify subjects at higher risk for whom the definitive GS diagnosis is advisable. Therefore, selection for verification may depend on the auxiliary data as well as the new test. Not accounting for this dependence may also result in verification bias. Furthermore, it is possible that including information about auxiliary data in the analysis may increase efficiency even when the sampling does not depend on auxiliary data.

1.5 Examples

There are many examples of studies to assess accuracy of a medical diagnostic or screening test where definitive disease assessment is too invasive or expensive to be ascertained in all subjects. In these studies it makes practical and ethical sense to select a higher percentage of subjects with a positive test result or other auxiliary data suggesting the presence of disease for disease verification.

1.5.1 Audiology

Undetected hearing loss in infants can lead to severe problems with speech, social, and emotional development, but early diagnosis and intervention can minimize the dehabilitating effects (NIH Consensus Statement, 1993). The gold standard for determining true hearing status in infants is the visual reinforcement audiometry (VRA) test. Specifically, VRA is a behavioral test in which clinicians observe whether infants respond to a noise. The drawback to VRA, however, is that it cannot be performed until infants are between 8 and 12 months of age because it is a behavioral test that requires cooperation of the infants.

One way to decrease the time to diagnosis is to consider tests that require much less cooperation from the infants. The Neonatal Hearing Screening Study (NHSS),

a multi-center cohort study funded by the National Institute on Deafness and Other Communication Disorders (NIDCD), enrolled 4800 infants in neonatal intensive care nurseries who were at high risk for hearing loss to assess the accuracy of two new passive electronic devices for detecting hearing impairment. Specifically, this study considered the distortion product otoaccoustic emissions (DPOAE) test and the transient evoked otoaccoustic emissions (TEOAE) test. These tests are non-invasive measures of the cochlear status and do not require active participation of the infants. The study protocol required that all infants be tested with DPOAE and TEOAE in each ear before discharge from the hospital or soon thereafter. The order of ear and tests was chosen at random. All infants were followed after discharge from the hospital so that the VRA test could be given between 8 and 12 months of age.

Tracking all infants from the time the DPOAE and TEOAE tests were performed until VRA was conducted is expensive. Based on the methods developed in Chapter 4 a study could be designed that does not require all infants to be tracked and tested with VRA. This could reduce the cost of a future study. For example, a cost-efficient study could be designed so that the true hearing status, as determined by VRA, is obtained on all infants with positive DPOAE and TEOAE tests and a random subset of infants with negative tests. Data from the Neonatal Hearing Screening Project will be analyzed in Chapter 7 to determine how the methods proposed in Chapter 4 could perform in such a study.

1.5.2 Prostate Cancer

The Prostate Cancer Prevention Trial (PCPT) is a chemoprevention trial funded by the National Cancer Institute in which 18,000 healthy men age 55 and older were randomized to either daily finasteride or placebo tablets for 7 years (Feigl et al., 1995). The primary objective of the study is to assess whether subjects assigned to finasteride have a different prevalence of biopsy identified prostate cancer than subjects randomized to placebo. Another objective of the study is to assess the

sensitivity and specificity of total PSA and digital rectal exam (DRE). All subjects in the study will be scheduled for prostate biopsy following the 7 years of study unless diagnosed with prostate cancer during the course of the study.

The methods developed in Chapter 4 would allow one to assess the accuracy of PSA and DRE without requiring all men to have prostate biopsy. For example, a study could be designed so that all men with high PSA (e.g. PSA $>$ 4.0 ng/ml) or DRE indicative of disease receive biopsy while only a random fraction of those with normal PSA and DRE receive biopsy.

PSA occurs in two forms, free and complex (bound). Recently it has been hypothesized that the ratio of free to total PSA may be a better screening test (lower values suggesting disease) than existing tests. The methods developed in Chapter 4 can also be used to assess the accuracy of this new test without requiring all men in the study to have prostate biopsy.

1.6 Summary

Screening and diagnostic tests play an important role in health care and it is imperative that accuracy of tests is assessed before used in practice. Often when assessing the accuracy of a test it is more cost-effective and ethical to oversample subjects who are more likely to be diseased. However, this biased sampling can lead to biased estimates of accuracy. Existing bias-correction methods focus on studies of diagnostic tests with binary and ordinal responses. These methods are summarized in Chapter 3. As pointed out in a recent review of existing verification bias methods (Zhou, 1998) "bias-correction methods need to be developed if the response of a diagnostic test is continuous". The goal of this dissertation is to develop verification bias correction methods that allow one to assess accuracy of continuous tests.

Chapter 2

ASSESSING ACCURACY WITH COMPLETE DATA

In this chapter we discuss methods for assessing accuracy in the ideal setting where disease status is known for all study subjects. In subsequent chapters we will introduce methods that are appropriate in the presence of verification bias.

2.1 Accuracy

The diagnostic accuracy of a test is the test's ability to discriminate among alternative states of health, e.g. hearing impaired vs. not hearing impaired or cancer vs. cancer-free. Accuracy is the first characteristic of a test to consider because a test that cannot distinguish between different states of health is of no use in practice (Zweig & Campbell, 1993). If a test is shown to be accurate, then the next step is to determine the practical usefulness of the test in managing patients. Since usefulness is not an issue unless the test has first been shown to be accurate, this dissertation focuses on assessing the accuracy of new tests.

2.2 Types of Test Results

Tests may have binary, ordinal, or continuous results. Examples of binary diagnostic tests include home pregnancy tests and the dichotomization of PSA so that values greater than 4.0 ng/ml are considered positive and the rest are considered negative. A radiologist's interpretations of images to quantify the suspicion of cancer or an anethesiologist's assessment of a patient's fitness to determine which patients receive surgery are examples of tests with ordinal rating scales. Examples of continuous

diagnostic tests include tumor-marker concentrations such as PSA and otoacoustic emissions tests for hearing impairment.

2.3 Notation

Let T denote the results of a continuous medical diagnostic or screening test. Let D be the binary disease status variable where $D = 1$ represents diseased subjects and $D = 0$ disease-free subjects. Furthermore, n_D and $n_{\bar{D}}$ are the number of diseased and non-diseased subjects, respectively. Consider a cohort study where (T, D) are known for all $n = n_D + n_{\bar{D}}$ study subjects.

2.4 Assumptions/Convention

D is assumed to be measured without error using a definitive "gold standard" (GS) test. There is an extensive literature that deals with the topic of an imperfect or missing gold standard (e.g. Walter & Irwig, 1988, Hui & Zhou, 1998, Alonzo & Pepe, 1999). The convention is taken that larger values of the diagnostic test, T, are more indicative of disease.

2.5 Assessing Accuracy

Receiver operating characteristic (ROC) curves are a well accepted measure of accuracy for continuous tests (Campbell, 1994, Hanley, 1989). By choosing a cutpoint c on the continuous scale, a binary test may be defined such that a test result with $T \geq c$ is considered positive and if $T < c$ the test is considered negative. An ROC curve is a plot of the true positive rates (TP) versus false positive rates (FP) associated with such binary tests as the cutpoint c is varied from $-\infty$ to $+\infty$. TP(c) and FP(c) are defined as follows:

$$\text{TP}(c) = P(T \geq c | D = 1)$$

$$\mathrm{FP}(c) = P(T \geq c | D = 0).$$

For a particular cutpoint, TP(c) and FP(c) are equivalent to the sensitivity and 1−specificity of the binary test, respectively.

ROC curves have several appealing attributes. First, an ROC curve displays the trade-offs between sensitivity and specificity of the test as the definition of a positive result is varied. Second, an ROC curve can be interpreted as a measure of the amount of separation of the distribution of test results in the diseased population from the distribution of test results in the non-diseased population. The more separated the distributions, the closer the ROC curve is to the upper left-hand corner and, consequently, farther the curve from the forty-five degree line indicating a useless test. Thus, an ROC curve can be used to visually compare the accuracy of different tests. An ROC comparison is particularly useful when tests are measured either in different units or on completely different scales.

ROC curves can be estimated non-parametrically using empirical values of $TP(c)$ and $FP(c)$ for all cutpoints as follows:

$$\widehat{\mathrm{TP}}(c) = \frac{\sum_i I[T_i \geq c]\ D_i}{\sum_i D_i} \tag{2.1}$$

and

$$\widehat{\mathrm{FP}}(c) = \frac{\sum_i I[T_i \geq c]\ (1 - D_i)}{\sum_i (1 - D_i)}. \tag{2.2}$$

The empirical ROC curve is a plot of $\widehat{TP}(c)$ and $\widehat{FP}(c)$ for each cutpoint. Extrapolation of the empirical ROC curve to all possible cutpoints can be made by using a step function or by connecting observed data points linearly. For data with no ties, adjacent points can be connected with horizontal and vertical lines in a unique manner resulting in a step function. As the threshold changes, inclusion of a true positive result produces a vertical line and inclusion of a false positive result produces a horizontal line. When there are ties in the data, both the true positive and false

positive rates change simultaneously, resulting in a point displaced both horizontally and vertically from the last point.

Various summary measures of performance, or accuracy indices, for diagnostic and screening tests are based on the ROC curve. Greenhouse & Mantel (1950) proposed calculating the point on the ROC curve associated with a fixed FP. That is, the sensitivity of the test at a fixed specificity. A standard way to summarize the accuracy of a test is to calculate the area under the ROC curve (AUC). The AUC corresponds to the probability that the test result for a randomly chosen diseased subject exceeds that for a randomly chosen non-diseased subject. Values of the AUC range from 0.5, suggesting that the test is no better than chance alone, to 1.0, indicating a perfect test. The area under the empirical ROC curve is an estimate of the AUC. This non-parametric AUC has been shown to be equal to the Mann-Whitney U-statistic for comparing the distributions of test results in the diseased and non-diseased populations (Bamber, 1975). Bamber (1975) provides a variance estimator.

Another way of summarizing an ROC curve is to calculate the partial AUC (pAUC). The idea here is to restrict the area under the curve to a certain region of the ROC curve that may be of particular interest, such as a region corresponding to low false positive rates. Methods have been developed for estimating the pAUC both non-parametrically (Wieand et al., 1989) and parametrically (Thompson & Zucchini, 1989, McClish, 1989).

2.6 Summary

ROC curves are a standard way to assess the accuracy of continuous tests. They are often summarized using the AUC or pAUC. However, standard methods for estimating ROC curves and summary measures require that all subjects receive disease verification. In practice, however, often subjects more likely to be diseased receive work-up to determine disease status at a higher rate than subjects less likely to be

diseased (i.e. verification bias). In the next chapter we review existing methods that are appropriate for assessing accuracy of a binary or ordinal test in the presence of verification bias. We propose methods for assessing accuracy of continuous tests in the presence of verification bias in Chapter 4.

Chapter 3

EXISTING VERIFICATION BIAS CORRECTION METHODS

In this chapter we review existing methods for assessing accuracy of diagnostic or screening tests in the presence of verification bias. These methods are categorized into existing bias correction methods that focus on tests with binary (Begg & Greenes, 1983, Zhou, 1993) and ordinal (Gray et al., 1984, Hunink et al., 1990, Zhou, 1996, Zhou & Rodenberg, 1998) responses.

3.1 Notation

Consider a cohort study with n subjects on which the test result T and auxiliary data A are measured. Then D is ascertained if $V = 1$ where V is the binary variable indicating whether the definitive disease status was obtained. Therefore, the n_V subjects with disease verification have data $(D, T, A, V = 1)$ and will be referred to as the verification group or verification sample. The other $n_{\bar{V}} = n - n_V$ subjects have data $(T, A, V = 0)$. Auxiliary data A may be a vector of measurements but for convenience is considered to be univariate.

3.2 Assumptions

The methodology developed in this dissertation (Chapter 4) and in most of the existing bias correction methods reviewed in this chapter also assume that conditional on the test result and auxiliary data, disease status is independent of the verification

status. That is,

$$P(D|V,T,A) = P(D|T,A) \tag{3.1}$$

or equivalently

$$P(V|D,T,A) = P(V|T,A). \tag{3.2}$$

These conditions (3.1) and (3.2) are the standard missing at random (MAR) assumptions discussed in the missing data literature (Little & Rubin, 1987). These conditions assume that although the disease process affects both T and A, it only affects selection for disease verification through its influence on T and A. In practice, the MAR assumption will often be satisfied because disease verification is usually only a function of visible factors, test result and auxiliary information. Retrospective studies are more likely to violate this assumption than prospective studies because this assumption is dependent on T and A representing an exhaustive list of the factors that can influence selection for disease verification.

3.3 Binary test

As was illustrated in Section 1.3 verification bias can cause biased estimates of sensitivity and specificity of a binary test. Two methods have been proposed in the literature to correct for this bias. The observed data can be summarized as in Table 3.1.

Begg & Greenes (1983) developed a bias correction method for sensitivity and specificity by using Bayes' Rule and assuming disease status is MAR. First, consider estimating the sensitivity of a test. Bayes' Rule can be used to re-write sensitivity as

$$\begin{aligned} P(T=1|D=1) &= \frac{P(T=1,D=1)}{P(D=1)} \\ &= \frac{P(D=1|T=1)P(T=1)}{P(D=1|T=1)P(T=1)+P(D=1|T=0)P(T=0)}. \end{aligned} \tag{3.3}$$

Table 3.1: Observed data for the verification bias problem when T is binary.

V	D	T=1	T=0
1	1	s_1	s_0
1	0	r_1	r_0
0	Missing	u_1	u_0
	Total	n_1	n_0

Each quantity on the right-hand-side (RHS) of (3.3) can be directly estimated from the observed data using empirical estimates. In particular, $P(T)$ can be estimated using data from all subjects, and $P(D|T)$ can be estimated using the verification group since by the MAR assumption $P(D|T) = P(D|T, V = 1)$. Substituting empirical estimates of the probabilities in (3.3) results in the following unbiased estimate of sensitivity

$$\widehat{P}(T = 1|D = 1) = \frac{\frac{s_1}{s_1+r_1}\frac{n_1}{n}}{\frac{s_1}{s_1+r_1}\frac{n_1}{n} + \frac{s_1}{s_0+r_0}\frac{n_0}{n}} \tag{3.4}$$

where $n = n_1 + n_0$.

A bias-corrected estimate of specificity can be calculated in a similar fashion. It can be shown that these estimators of sensitivity and specificity are, in fact, maximum likelihood (ML) estimators. The delta method is used to develop variance estimators for sensitivity and specificity.

Begg & Greenes (1983) also allow for discrete auxiliary data that may affect the verification process. For example, the corrected estimator for sensitivity that includes auxiliary data is

$$\widehat{P}(T = 1|D = 1) = \frac{\sum_A \widehat{P}(D = 1|T = 1, A)\widehat{P}(T = 1, A)}{\sum_A \widehat{P}(D = 1|T = 1, A)\widehat{P}(T = 1, A) + \widehat{P}(D = 1|T = 0, A)\widehat{P}(T = 0, A)}.$$

With sparse data, Begg and Greenes suggest using a logistic model to estimate $P(D|T, A)$.

Zhou (1993) extended Begg and Greenes' method to allow a more general model for the verification process and derived the maximum likelihood estimators for the sensitivity and specificity of diagnostic test and their corresponding variances. He does not assume D is MAR, but assumes $\lambda_1 = P(V = 1|D = 1, T = 1)/P(V = 1|D = 0, T = 1)$ and $\lambda_0 = P(V = 0|D = 1, T = 1)/P(V = 0|D = 0, T = 1)$ are known. In other words, he assumes the ratio of the probability of selecting for verification a diseased patient with a given test result to that of selecting for verification a non-diseased patient with the same test result is known. However, in practice λ_1 and λ_0 are not usually known and may be tough to estimate. If $\lambda_1 = \lambda_0$, then Zhou's estimators reduce to those of Begg and Greenes.

3.4 Ordinal test

Since verification bias can cause biased estimates of sensitivity and specificity of a binary test, verification bias can also cause biased estimates of $TP(c)$ and $FP(c)$ for any particular cutpoint of an ordinal test. Biased estimates of $TP(c)$ and $FP(c)$ result in a shift in the operating points on the corresponding ROC curve (Hunink et al., 1990, Hunink et al., 1993). Since an operating point on the ROC curve is often chosen to define the test to be used in practice, verification bias can cause less than optimal operating points to be selected. Even though verification bias causes operating points to be biased, the corresponding ROC curve and AUC may or may not be biased (Hunink et al., 1993, Schouw et al., 1994).

Next we consider methods that correct for verification bias when assessing accuracy of ordinal tests. These methods can be categorized into two groups: those based on an empirical ROC curve and those based on a parametric ROC curve. Consider the ordinal test T that has $c = 1, \cdots, C + 1$ categories. The observed data are presented

Table 3.2: Observed data for the verification bias problem when T is ordinal.

		T		
V	D	1	$\cdots$	$C+1$
1	1	s_1	$\cdots$	s_{C+1}
1	0	r_1	$\cdots$	r_{C+1}
0	Missing	u_1	$\cdots$	u_{C+1}
	Total	n_1	$\cdots$	n_{C+1}

in Table 3.2.

3.4.1 Empirical ROC Curve

Gray et al. (1984) extended the work of Begg & Greenes (1983) to ordinal tests. Specifically, they obtained a bias-corrected empirical ROC curve by plotting bias-corrected TP(c) and FP(c) (equivalently, sensitivity and 1-specificity) for each of the empirical operating points. By assuming D is MAR, they use the following as an estimator of TP(c)

$$\widehat{P}(T \geq c|D=1) = \frac{\sum_{T \geq c} \widehat{P}(D=1|T)\widehat{P}(T)}{\sum_{T} \widehat{P}(D=1|T)\widehat{P}(T)}. \tag{3.5}$$

Zhou (1996) developed a non-parametric estimator of the AUC that has been corrected for verification bias. Bamber (1975) showed that based on the trapezoidal rule the AUC was equal to

$$\sum_{i=1}^{C} \sum_{j=i+1}^{C+1} P(T=i|D=0)P(T=j|D=1) + (1/2) \sum_{c=1}^{C+1} P(T=c|D=0)P(T=c|D=1). \tag{3.6}$$

Using Bayes' Rule, Zhou re-writes (3.6) as

$$\frac{\sum_{i=1}^{C} \sum_{j=i+1}^{C+1} (1-\delta_i)\phi_i \delta_j \phi_j + (1/2) \sum_{i=1}^{C+1} (1-\delta_i)\delta_i \phi_i^2}{\sum_{i=1}^{C+1} (1-\delta_i)\phi_i \sum_{j=i+1}^{C+1} \delta_j \phi_j} \tag{3.7}$$

where $\phi_i = P(T = i)$ and $\delta_i = P(D = 1|T = i)$. Under the MAR assumption $P(V|T, D) = P(V|T)$ so the log-likelihood function based on the observed data is

$$\sum_{c=1}^{C+1} n_c \log(\phi_c) + \sum_{c=1}^{C+1} (s_c \log(\delta_c) + r_c \log(1 - \delta_c)). \tag{3.8}$$

This corresponds to the log-likelihood for multinomial data so the ML estimators for ϕ_c and δ_c are

$$\widehat{\phi}_c = \frac{n_c}{n}, \; c = 1, \cdots, C \quad \text{and} \quad \widehat{\delta}_c = \frac{s_c}{s_c + r_c}, \; c = 1, \cdots, C. \tag{3.9}$$

Substituting (3.9) into (3.7) provides an estimator of AUC. Zhou also provides variance estimators.

3.4.2 Parametric ROC Curve

Gray et al. (1984) also developed a parametric estimator of the ROC curve that was corrected for verification bias. In this approach it is assumed that there is an underlying continuous latent variable L, representing the degree of suspicion of disease, and the observed ordinal values result from classifying L into one of $C + 1$ intervals. That is, there are C cutpoints $\{\nu_c; c = 1, \cdots, C\}$ and $T = c \;\Leftrightarrow\; \nu_c < L < \nu_{c+1}$ with $\nu_0 = -\infty$ and $\nu_{C+1} = +\infty$.

Parametric models for the ROC curve of T are derived by assuming a relationship between the observed T and unobserved L and a parametric distribution for L. Gray et al. (1984) assume the binormal model (Dorfman & Alf, 1969). The binormal model assumes L has a normal distribution with mean μ_1 and variance σ_1^2 for diseased subjects and is distributed $N(\mu_0, \sigma_0^2)$ for non-diseased subjects. Let $\nu'_c = (\nu_c - \mu_0)/\sigma_0, \; c = 1, \cdots, C + 1$. Under these assumptions, the ROC curve for L is a plot of

$$P(T \geq c|D = 0) = 1 - \Phi(\nu'_c)$$

versus

$$P(T \geq c|D = 1) = 1 - \Phi(b\, \nu'_{c-1} - a)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $a = (\mu_1 - \mu_0)/\sigma_1$, and $b = \sigma_0/\sigma_1$.

The log-likelihood for the observed data is

$$\sum_{c=1}^{C+1} s_c \log(\theta\ \pi_{c1}) + r_c \log((1-\theta)\ \pi_{c0}) + u_c \log(\theta\ \pi_{c1} + (1-\theta)\ \pi_{c0}) \qquad (3.10)$$

where $\theta \equiv P(D=1)$, $P(T=c|D=1) \equiv \pi_{c1} = \Phi(b\nu'_{c-1} - a) - \Phi(b\nu'_c - a)$ and $P(T=c|D=0) \equiv \pi_{c0} = \Phi(\nu'_{c-1}) - \Phi(\nu'_c)$.

Gray et al. (1984) employed a modified scoring algorithm to maximize this likelihood with respect to a, b, and ν'_c. The resulting ML estimates can be used to obtain a bias-corrected ROC curve. Furthermore, the area under this bias-corrected ROC can be calculated as $\Phi(a/\sqrt{1+b^2})$.

Hunink et al. (1990) present another parametric estimator for the ROC curve that is corrected for verification bias. They correct for verification bias by dividing the observed cell counts ($s_i, r_i;\ i = c, \cdots, C+1$ in Table 3.2) by the probability of verification, $P(V=1|T,A)$, where this probability is modelled using a logistic model. These adjusted cell counts are then modelled using an ordinal regression method, such as PLUM (McCullagh, 1980), to obtain an estimate of the ROC curve.

Zhou & Rodenberg (1998) use the EM-algorithm to adapt ordinal regression methods to estimate a bias-corrected ROC curve but under a non-MAR model for verification rather than a MAR model. They make the same assumption about the verification mechanism as Zhou (1993) did for binary tests. Namely that λ_1 and λ_0 are known.

3.5 Summary

In this chapter we reviewed existing verification bias correction methods. Begg & Greenes (1983) and Zhou (1993) have proposed methods for binary tests while Gray et al. (1984), Zhou (1996), Hunink et al. (1990), and Zhou & Rodenberg (1998) have

proposed methods for ordinal tests. None of these methods can accommodate the large number of continuous diagnostic and screening tests that are being developed. Therefore, in the next chapter we propose bias correction methods for continuous tests.

Chapter 4

NEW BIAS CORRECTION METHODS FOR CONTINUOUS TESTS

Assessing the accuracy of continuous screening or diagnostic tests is the primary aim of many studies conducted today. Standard ROC methods (Chapter 2) require that disease status is obtained on all subjects, which is often not possible in practice. Furthermore, all of the existing statistical methods for data collected with differential disease verification is applicable for tests with discrete response categories (Chapter 3). In this chapter we propose several new methods for assessing accuracy of continuous tests in the presence of verification biased sampling.

4.1 Data

Consider a study where test result T_i and auxiliary data A_i are available for all $i = 1, \cdots, n$ subjects. Furthermore, disease status D_i is only available for those subjects in the verification sample (i.e. $V_i = 1$). In some settings, observations (T_i, A_i, D_i, V_i) will be clustered. This occurs, for example, when individuals are tested several times (e.g. two ears from the subject in the audiology example). The methods developed in this chapter can easily be extended to the clustered data setting.

4.2 Complete Case Approach

When disease status is not ascertained for all study subjects, standard evaluation of a continuous test only includes data from subjects in the verification sample. This is known as a complete case (CC) analysis because only data from those subjects with

complete data are used.

CC estimators of TP(c) and FP(c) are similar to the empirical estimators given in (2.1) and (2.2) for full data except that only data for the subjects in the verification sample are used. That is,

$$\widehat{TP}_{CC}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\ V_i\ D_i}{\sum_{i=1}^{n} V_i\ D_i} \tag{4.1}$$

and

$$\widehat{FP}_{CC}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\ V_i\ (1 - D_i)}{\sum_{i=1}^{n} V_i\ (1 - D_i)}. \tag{4.2}$$

A complete case approach is valid if the verification sample is a simple random sample of subjects in the study, but often this is not the case.

4.3 Basic Approach

The basic approach of the methods we develop in this chapter is to first develop bias-corrected estimators of TP(c) and FP(c) for each observed cutpoint c. Then an empirical bias-corrected ROC curve can be obtained by plotting the bias-corrected TP(c) and FP(c) for all c. This corrected ROC curve can then be used to assess the accuracy of the test either by visual inspection or by computing summary indices such as AUC and pAUC.

Several new approaches for obtaining bias-corrected estimators of TP(c) are developed in this chapter. Estimators of FP(c) can be constructed in a similar fashion. The first approach we consider is an extension of the approach of Begg & Greenes (1983) discussed in Section 3.3 for binary tests. The other approaches we propose are based on the fact that a study with verification bias can be thought of as a study with a two-phase or double sampling design (Neyman, 1938, Tenenbein, 1970). In the first phase, the diagnostic or screening test results and auxiliary data are collected on all subjects in the study. Disease status is then verified for a subset of subjects in the second phase of the study with the selection of the subjects for phase 2 possibly

dependent on the measurements made in phase 1. Thus, the probability of selection for phase 2 may be dependent on data from the first phase.

4.4 *Extension of Begg & Greenes Approach*

Begg & Greenes (1983) considered methods for assessing the accuracy of a binary test when verification bias exists. Gray et al. (1984) extended their approach to ordinal tests. In Chapter 3 we saw that they used Bayes' Rule to re-write TP(c) as

$$\begin{aligned} TP(c) &= P(T \geq c | D = 1) \\ &= \frac{P(T \geq c, D = 1)}{P(D = 1)} \\ &= \frac{\sum_{T \geq c} \sum_A P(D = 1 | T = t, A = a)\ P(T = t, A = a)}{\sum_T \sum_A P(D = 1 | T = t, A = a)\ P(T = t, A = a)}. \end{aligned} \tag{4.3}$$

The MAR assumption (3.1) implies all quantities are estimable with the observed data.

Extending this approach to the continuous test setting is straightforward. Specifically, with continuous T and A, the sums in (4.3) are replaced by integrals and we see that

$$\begin{aligned} TP(c) &= \frac{\int_c^\infty \int P(D = 1 | T = t, A = a)\ P(T = t, A = a)\ da\ dt}{\int \int P(D = 1 | T = t, A = a)\ P(T = t, A = a)\ da\ dt} \\ &= \frac{\int \int P(D = 1 | T = t, A = a)\ I(t \geq c)\ P(T = t, A = a)\ da\ dt}{\int \int P(D = 1 | T = t, A = a)\ P(T = t, A = a)\ da\ dt}. \end{aligned}$$

Although we considered continuous A, this method can also accommodate discrete auxiliary data.

Specifying a parametric model for $P(T, A)$ is challenging, especially as the dimension of A increases, suggesting the use of a non-parametric estimator. If $P(T, A)$ is estimated empirically by giving mass $1/n$ at $(t, a) = \{(T_1, A_1), \cdots, (T_n, A_n)\}$, then

$$\widehat{P}(T \geq c, D = 1) = \frac{1}{n} \sum_{i=1}^{n} I[T_i \geq c]\ \widehat{P}(D_i = 1 | T_i, A_i)$$

and

$$\widehat{P}(D=1) = \frac{1}{n}\sum_{i=1}^{n} \widehat{P}(D_i = 1|T_i, A_i) \tag{4.4}$$

where $\widehat{P}(D_i = 1|T_i, A_i)$ can be obtained from a parametric model, e.g. logistic regression.

Hence a natural generalization of the Gray et al. (1984) estimator (3.5) is

$$\widehat{TP}_{\mathrm{BG}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\ \widehat{P}(D_i = 1|T_i, A_i)}{\sum_{i=1}^{n} \widehat{P}(D_i = 1|T_i, A_i)}. \tag{4.5}$$

Similarly,

$$\widehat{FP}_{\mathrm{BG}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\widehat{P}(D_i = 0|Ti, A_i)}{\sum_{i=1}^{n} \widehat{P}(D_i = 0|T_i, A_i)}. \tag{4.6}$$

We refer to $\widehat{TP}_{\mathrm{BG}}(c)$ and $\widehat{FP}_{\mathrm{BG}}(c)$ as the Begg and Greenes estimators.

4.5 Two-phase Design Approaches

Next we develop approaches based on the fact that a study with verification bias can be thought of as having a two-phase design. As noted before, the true positive rate can be written as

$$\begin{aligned} TP(c) &= P(T \geq c|D=1) \\ &= \frac{P(T \geq c, D=1)}{P(D=1)}. \end{aligned}$$

This formulation suggests that TP(c) can be thought of as a ratio of two "prevalences". The denominator is the usual disease prevalence while the numerator is the prevalence of disease **and** the test being greater than the cutpoint c. Therefore, both the numerator and denominator of the equation for $TP(c)$ are prevalences and methods for estimating prevalence in a two-phase design can be used to develop a valid estimator of the true positive rate.

Estimating prevalence with two-phase sampling is quite common (Pickles et al., 1995). For example, Clayton et al. (1998) studied the use of mean score and inverse probabilitity weighting methods. These methods are among the methods we will consider for estimating prevalence.

Consider estimating the prevalence of disease, $\theta = P(D = 1)$. If D is missing completely at random (MCAR) or MAR, then likelihood inference that ignores the missingness mechanism is valid (Little & Rubin, 1987). Therefore, a fully parametric analysis parameterizes the joint distribution of T, A, and D as $P_{\theta,\gamma}(T, A, D) = P_\theta(D)P_\gamma(T, A|D)$ and estimates θ and γ simultaneously.

With independent data, maximum likelihood estimation can be achieved using the EM-algorithm. Similar methods that adjust for clustering can be used with dependent data. The EM-algorithm solves

$$\sum_{i=1}^{n} V_i S_\theta(D_i) + (1 - V_i)E_{\theta,\gamma}[S_\theta(D)|T_i, A_i] = 0 \tag{4.7}$$

$$\sum_{i=1}^{n} V_i S_\gamma(T_i, A_i|D_i) + (1 - V_i)E_{\theta,\gamma}[S_\gamma(T_i, A_i|D)|T_i, A_i] = 0 \tag{4.8}$$

where $S_\theta(D)$ and $S_\gamma(T, A|D)$ denote the score functions or partial derivatives of $\log P_\theta(D)$ and $\log P_\gamma(T, A|D)$ with respect to θ and γ respectively. A subject not verified contributes $E_{\theta,\gamma}[S_\theta(D)|T_i, A_i]$ to the estimation of θ where the expectation is taken over the "imputation distribution"

$$P(D|T, A) \propto P_\theta(D)P_\gamma(T, A|D).$$

So that

$$E_{\theta,\gamma}[S_\theta(D)|T_i, A_i] = \int S_\theta(D)P(D|T, A)dD$$

where $P(D|T, A) = \frac{P_\gamma(T,A|D)P_\theta(D)}{\int P_\gamma(T,A|D)P_\theta(D)}$.

Not only is $P_\gamma(T, A|D)$ not of interest, but Pepe (1992) showed that misspecification of this model can lead to inconsistent estimation of θ. Another disadvantage

to a fully parametric approach is that it is usually difficult to specify a model for $P_\gamma(T, A|D)$, especially as the dimension of A increases (Clayton et al., 1998). For the special case when D is MCAR, Pepe (1992) proposed using the verification group to empirically estimate $P(T, A|D)$ and then estimate the likelihood component for each subject not verified. This approach estimates θ by solving

$$\sum_{i=1}^{n} V_i S_\theta(D_i) + (1 - V_i)\widehat{E}[S_\theta(D)|T_i, A_i] = 0 \tag{4.9}$$

where $\widehat{E}[S_\theta(D)|T_i, A_i] = \int S_\theta(D) \frac{\widehat{P}_\gamma(T,A|D)P_\theta(D)}{\int \widehat{P}_\gamma(T,A|D)P_\theta(D)} dD$. However, this approach fails in our setting because the MCAR assumption does not hold, verification can depend on T or A.

Alternatives to using a fully parametric approach are mean score, inverse probability weighting, and semi-parametric efficient approaches. These alternative approaches can be used to obtain estimators of both prevalences so that the ratio of the two prevalence estimators should provide a valid estimator of the true positive rate.

4.6 Mean Score Approach

Pepe et al. (1994) and Reilly & Pepe (1995) noted that $P(D|T, A)$ may be estimated using only those subjects in the verification group by assuming the conditional independence of D and V. They proposed estimating θ by solving

$$\sum_{i=1}^{n} V_i \; S_\theta(D_i) + (1 - V_i) \; \widehat{E}[S_\theta(D)|T_i, A_i] = 0 \tag{4.10}$$

where $S_\theta(D)$ is the binomial score, $\widehat{E}[S_\theta(D)|T_i, A_i] = \int S_\theta(D) \; \widehat{P}(D|T_i, A_i)dD$ and $\widehat{P}(D|T, A)$ is estimated non-parametrically using the verification group. However, this approach breaks down as the joint distribution of T and A becomes sparse, which will be the case when T and A are continuous variables. Clayton et al. (1998) proposed using a parametric model for $P(D|T, A)$, for example a logit model, to accommodate settings where T and A are continuous. In other words, they suggest estimating θ

using the score contributions for those in the verification group and imputing mean scores for those not in the verification group over the distribution $P(D|T, A)$ which is modeled parametrically.

If $S_\theta(D_i) = D_i - \theta$, then solving (4.10) gives the following estimator of disease prevalence:

$$\widehat{P}(D = 1) = \frac{1}{n}\sum_{i=1}^{n}\{V_i\ D_i\ +\ (1 - V_i)\ \widehat{P}[D_i = 1|T_i, A_i]\}. \tag{4.11}$$

$P(T \geq c, D = 1)$ can be estimated analogously yielding

$$\widehat{TP}_{\mathrm{MS}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\{V_i\ D_i\ +\ (1 - V_i)\ \widehat{P}(D_i = 1|T_i, A_i)\}}{\sum_{i=1}^{n} V_i\ D_i\ +\ (1 - V_i)\ \widehat{P}(D_i = 1|T_i, A_i)}. \tag{4.12}$$

Similarly,

$$\widehat{FP}_{\mathrm{MS}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\{V_i\ (1 - D_i)\ +\ (1 - V_i)\ \widehat{P}(D_i = 0|T_i, A_i)\}}{\sum_{i=1}^{n} V_i\ (1 - D_i)\ +\ (1 - V_i)\ \widehat{P}(D_i = 0|T_i, A_i)}. \tag{4.13}$$

Again the MAR assumption implies that data from the verification sample can be used to obtain a valid estimate of $P(D|T, A)$.

4.7 Inverse Probability Weighting Approach

Another approach for estimating prevalence in a two-phase design is to use a Horvitz-Thompson type estimator (Horvitz & Thompson, 1951). This approach weights each observation in the verification group by the inverse of the sampling fraction (i.e. probability the subject was selected for verification). With complete data, the expected value of the score contribution, $S_\theta(D_i)$, is 0 at the true value of θ. Thus, θ can be consistently estimated by solving $\sum_i S_\theta(D_i) = 0$. However, this estimating equation does not yield consistent estimates when only applied to the verification group. That is, $E\{V_i\ S_\theta(D_i)\} \neq 0$. However, if each observation in the verification group is given weight equal to π_i, the inverse of the probability that it was selected for verification,

then

$$\begin{aligned} E\left\{V_i\pi_i^{-1}S_\theta(D_i)\right\} &= E\left\{E\left\{V_i\pi_i^{-1}S_\theta(D_i)\right\}|T_i, A_i, D_i\right\} \\ &= E\left\{\pi_i^{-1}S_\theta(D_i)E\{V_i|T_i, A_i, D_i\}\right\} \\ &= E\left\{\pi_i^{-1}S_\theta(D_i)P(V_i = 1|T_i, A_i, D_i)\right\} \\ &= E\left\{\pi_i^{-1}S_\theta(D_i)P(V_i = 1|T_i, A_i)\right\} \text{ by MAR} \\ &= E\{S_\theta(D_i)\} \\ &= 0. \end{aligned}$$

This suggests estimating θ by solving the following estimating equation proposed by Zhao & Lipsitz (1992):

$$\sum_{i=1}^{n} V_i\pi_i^{-1}S_\theta(D_i) = 0. \tag{4.14}$$

If $S_\theta(D_i) = D_i - \theta$, then solving (4.14) yields the famous Horvitz-Thompson estimator of prevalence for sample surveys

$$\widehat{P}(D = 1) = \left(\sum_{i=1}^{n} \pi_i^{-1}V_i\right)^{-1} \sum_{i=1}^{n} V_i\pi_i^{-1}D_i. \tag{4.15}$$

$\widehat{P}(T \geq c, D = 1)$ can be obtained in a similar fashion. Thus,

$$\widehat{TP}_{\text{IPW}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\ V_i\ D_i\ \pi_i^{-1}}{\sum_{i=1}^{n} V_i\ D_i\ \pi_i^{-1}}. \tag{4.16}$$

π_i is the sampling fraction and may be known or may need to be estimated depending on the design of the study. In studies that have a protocol which dictates which subjects are to be verified, the sampling fractions are known and only those in the verification group contribute to (4.16). Even though in some studies the probabilities of selection for verification were known by the original sampling design, the actual selection probabilities may be unknown and may need to be estimated from the data due to drop-out, refusal, or other reasons. These probabilities always have to be estimated in observational studies. If these sampling fractions have to be estimated,

then those not selected for verification as well as those selected contribute to the estimation of θ. For this reason, estimating π_i even when it is known may increase efficiency (Pepe et al., 1994).

This approach bears a resemblance to the approach of Hunink et al. (1990) introduced in Section 3.4. They also weight the observed data by the inverse of the sampling fraction.

4.8 Semi-parametric Efficient Approach

A large body of literature has considered the idea of semi-parametric efficient estimation (e.g. Robins et al., 1994, Rotnitzky & Robins, 1995, Rotnitzky et al., 1998). What is meant by a semi-parametric efficient estimator? An estimator is semi-parametric in the sense that the estimator is asymptotically normal and consistent whatever the nuisance distributions. In the case of estimating disease prevalence, a nuisance distribution is $P(T, A|D)$, the joint distribution of the test result and auxiliary information conditional on disease status. Furthermore, it is efficient in that any estimator with smaller asymptotic variance must belong to a different class of estimators.

Before a semi-parametric efficient estimator can be developed, one has to specify the relevant class of estimators, as defined by restrictions imposed on the probability models. We will consider two different classes of estimators that both assume the expected value of D is θ_0, true disease prevalence. In addition, one class assumes that $P(D|T, A)$ is known or can be modelled correctly while the other assumes that $P(V|T, A)$ is known or can be modelled correctly.

4.8.1 Restrictions on Verification Probabilities

First we consider the class of estimators that assumes the expected value of D is θ_0 and that the verification probabilities are known or a parametric model for the probability

of being selected for verification given the test results and auxiliary information can be specified. These restrictions are the same as those required by the Inverse Probability Weighting approach (Section 4.7).

A class of estimators that satisfy these restrictions has been proposed for several different settings (Robins et al., 1994, Rotnitzky & Robins, 1995, Rotnitzky et al., 1998). Rotnitzky & Robins (1995) consider the case of interest to us where some responses (rather than covariates) are missing. Specifically, they consider a class of estimators that assumes a model for the mean of D conditional on a design matrix X. They propose a class of estimators based on modifications to the IPW estimating equation (4.14). Since $S_\theta(D_i)$ equals X_i times the residual $\epsilon(\theta)$, (4.14) is equivalent to

$$\sum_{i=1}^{n} \frac{V_i}{\pi_i} X_i \epsilon(\theta) = 0. \tag{4.17}$$

Rotnitzky & Robins (1995) modify (4.17) by replacing X_i by $h(X_i)$ where $h(\cdot)$ is an adaptively estimated function (except for special cases) and by adding an additional term

$$-\frac{V_i - \pi_i}{\pi_i}\phi(X_i, T_i, A_i).$$

Thus, they propose using the following estimating equation for estimating θ

$$\sum_{i=1}^{n} \frac{V_i}{\pi_i} h(X_i)\epsilon(\theta) - \frac{V_i - \pi_i}{\pi_i}\phi(X_i, T_i, A_i) = 0. \tag{4.18}$$

This estimating equation (4.18) is equivalent to the usual IPW estimating equations (4.14) and (4.17) when $h(X) = X$ and $\phi(X, T, A) = 0$. Since the expected value of (4.18) is 0 regardless of the choice of $h(\cdot)$ and $\phi(\cdot)$, $\widehat{\theta}$ obtained by solving (4.18) will always be consistent for θ. Therefore, Rotnitzky & Robins (1995) choose $h(\cdot)$ that will lead to the most efficient estimation of θ. They show that for a fixed $h(\cdot)$ the asymptotic variance of $\widehat{\theta}$ is minimized at $\phi(X_i, T_i, A_i) = E\{h(X_i)\epsilon(\theta)|X_i, T_i, A_i\}$. Furthermore, they argue that the optimal estimator in the class of estimators satisfy-

ing (4.18) is semi-parametric efficient and any regular asymptotically linear estimator of θ is asymptotically equivalent to an estimator in their class.

Usually the approach of Rotnitzky & Robins (1995) is tough to implement since it requires an adaptive selection process for $h(\cdot)$. However, when estimating prevalence without covariates it is straightforward to implement. In this case the design matrix X only contains a vector of 1's corresponding to the intercept, so $h(X)$ is equal to a constant and, without loss of generality, will be assumed to equal 1. Clearly, $h(X)$ no longer needs to be selected adaptively. When there are no covariates $\epsilon(\theta)$ equals $D_i - \theta$. Furthermore, in this setting

$$\begin{aligned}
\phi(X_i, T_i, A_i) &= E\{h(X_i)\epsilon(\theta)|X_i, T_i, A_i\} \\
&= E\{(D_i - \theta_i)|T_i, A_i\} \\
&= E\{D|T_i, A_i\} - \theta \qquad (4.19) \\
&= P(D_i = 1|T_i, A_i) - \theta. \qquad (4.20)
\end{aligned}$$

Using (4.18) and (4.20) we obtain the following prevalence estimator

$$\begin{aligned}
\widehat{P}(D = 1) &= \frac{1}{n}\sum_{i=1}^{n} V_i \frac{D_i}{\pi_i} - \left(\frac{V_i - \pi_i}{\pi_i}\right)\widehat{P}(D_i = 1|T_i, A_i) \qquad (4.21) \\
&= \frac{1}{n}\sum_{i=1}^{n} V_i \frac{D_i - \widehat{P}(D_i = 1|T_i, A_i)}{\pi_i} + \widehat{P}(D_i = 1|T_i, A_i). \qquad (4.22)
\end{aligned}$$

Again the MAR assumption implies that consistent estimates of $P(D_i = 1|T_i, A_i)$ can be obtained using the verification sample.

Similarly we can obtain an estimator of $P(T \geq c, D = 1)$. This results in

$$\widehat{TP}_{\mathrm{SP}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\{V_i D_i/\pi_i - (V_i - \pi_i)\widehat{P}(D_i = 1|T_i, A_i)/\pi_i\}}{\sum_{i=1}^{n} V_i D_i/\pi_i - (V_i - \pi_i)\widehat{P}(D_i = 1|T_i, A_i)/\pi_i}$$

and

$$\widehat{FP}_{\mathrm{SP}}(c) = \frac{\sum_{i=1}^{n} I[T_i \geq c]\{V_i(1 - D_i)/\pi_i - (V_i - \pi_i)\widehat{P}(D_i = 0|T_i, A_i)/\pi_i\}}{\sum_{i=1}^{n} V_i(1 - D_i)/\pi_i - (V_i - \pi_i)\widehat{P}(D_i = 0|T_i, A_i)/\pi_i}.$$

We refer to these estimators as the SP or semi-parametric efficient estimators.

4.8.2 Restrictions on Disease Probabilities

The second class of estimators we will consider again assumes the expected value of D is θ_0 but also assumes that $P(D|T, A)$ is known or can be modelled correctly. Consider the estimating function $S^C = \sum_{i=1}^{n} kD_i + (1-k)\rho_i - \theta$ where $\rho_i = P(D_i = 1|T_i, A_i)$ and k is a constant. When there are no missing data, i.e. D is known for all subjects, S^C is consistent for θ provided $E[D] = \theta_0$ and $\rho_i,\ i = 1, \cdots, n$ are known or known up to a vector of constants. To obtain the corresponding estimating function when there are missing data, we take the expectation of S^C conditional on the observed data, W. For S_i^C, the ith contribution of S^C,

$$E[S_i^C|W] = \begin{cases} kD_i + (1-k)\rho_i - \theta, & \text{when } V_i = 1 \\ k\mu_i + (1-k)\rho_i - \theta, & \text{when } V_i = 0 \end{cases}$$

This implies that

$$E[S^C|W] = \sum_{i=1}^{n} V_i\{kD_i + (1-k)\rho_i - \theta\} + (1-V_i)\{\rho_i - \theta\}. \qquad (4.23)$$

Two particular values of k result in familiar estimating functions. If $k = 0$, then (4.23) equals $\sum_{i=1}^{n}\{\rho_i - \theta\}$, the estimating function for the Begg and Greenes approach (Section 4.4). Conversely, if $k = 1$, then (4.23) equals $\sum_{i=1}^{n}\{V_iD_i + (1-V_i)\rho_i - \theta\}$, the estimating function for the Mean Score approach (Section 4.6). By design (4.23) is consistent for all k. Thus, BG and MS are two of many estimating functions that could be used to estimate θ. To find the most efficient estimator in this class of estimators, we can minimize the variance of $\widehat{\theta}$ with respect to k. In Chapter 5 we will attempt to analytically find the most efficient estimator in this class.

4.9 Qualitative Comparison of Approaches

Several new methods for obtaining a bias-corrected estimator of disease prevalence, TP(c), and FP(c) have been described in this chapter. These approaches will be

referred to as BG for the extension of the work by Begg and Greenes (Section 4.4), MS for the mean score approach (Section 4.6), IPW for the method that uses inverse probability weighting (Section 4.7), and SP for the semi-parametric efficient approach (Section 4.8.1).

We now present a qualitative comparison of the different approaches. A quantitative comparison will be given in Chapter 6. We compare each estimator of disease prevalence since disease prevalence is an essential component of $TP(c)$ and $FP(c)$ and it is easiest to see how the approaches differ by studying estimation of this component. Contributions to disease prevalence estimation for each approach are given in Table 4.1. The BG estimator (4.4) estimates disease status for all subjects in the study as a function of the test results and auxiliary data. In contrast, the MS estimator (4.11) estimates disease status only for those subjects not in the verification sample and uses the observed disease status for those in the verification sample. The IPW estimator (4.15) is similar to the CC estimator ($n_V^{-1} \sum_{i=1}^{n} V_i D_i$) in that it uses the *observed* disease status for the verification sample. Unlike CC, however, it corrects for the biased sampling by weighting the observed values by the probability of selection for verification. The SP estimator (4.22) appears to be a combination of the BG, MS, and IPW estimators. Similar to BG and MS, it estimates the probability of disease for those not verified. For those subjects in the verification group, it weights the observed disease status by the probability of being selected for disease verification just like IPW, but also subtracts the product of the probability of being diseased and the odds of not being verified.

All the approaches we propose for correcting for verification bias when assessing accuracy of a continuous test only require a regression model to be fit for a binary response (D or V). Therefore, all approaches are easy to implement even as the dimension of A increases. Table 4.2 provides a summary of the different approaches. Specifically, BG, MS, and SP require estimates of the probability of being diseased given the test results and auxiliary data while IPW and SP require estimates for the

Table 4.1: Contributions to prevalence estimation. $\rho_i = P(D_i = 1|T_i, A_i)$ and $\pi_i = P(V_i = 1|T_i, A_i)$.

Method	$V_i = 1$	$V_i = 0$
CC	D_i	0
BG	ρ_i	ρ_i
MS	D_i	ρ_i
IPW	$D_i\pi_i^{-1}$	0
SP	$\pi_i^{-1}[D_i - \rho_i(1 - \pi_i)]$	ρ_i

probability of being verified. The CC and IPW estimators of TP(c) and FP(c) only sum over those subjects verified while the corresponding BG, MS, and SP estimators sum over all subjects. Therefore, ROC curves for BG, MS, and SP will have $n_{\bar{V}}$ more jumps than the ROC curves for CC and IPW, and hence, ROC curves for BG, MS, and SP will appear to be more smooth. This is evident in Figure 4.1, ROC curves for hypothetical data (we describe how these data were generated In Chapter 6).

As the cutpoint c is decreased, estimators of TP(c) and FP(c) include more and more data. All the quantities in Table 4.1 are positive except $\pi_i^{-1}[D_i - \rho_i(1 - \pi_i)]$ which is negative so the resulting ROC curves will be monotone for all approaches except SP (see Figure 4.1). ROC curves are monotone by definition so we may need to use isotonic regression (Robertson et al., 1988) to make the SP ROC curve monotone. We will see in Chapters 6 and 7 that in practice SP ROC curves do not appear to deviate dramatically from monotonicity.

4.10 Summary

In this chapter we propose several methods for estimating bias-corrected TP(c) and FP(c) for continuous tests when it is either not cost-effective or unethical to verify

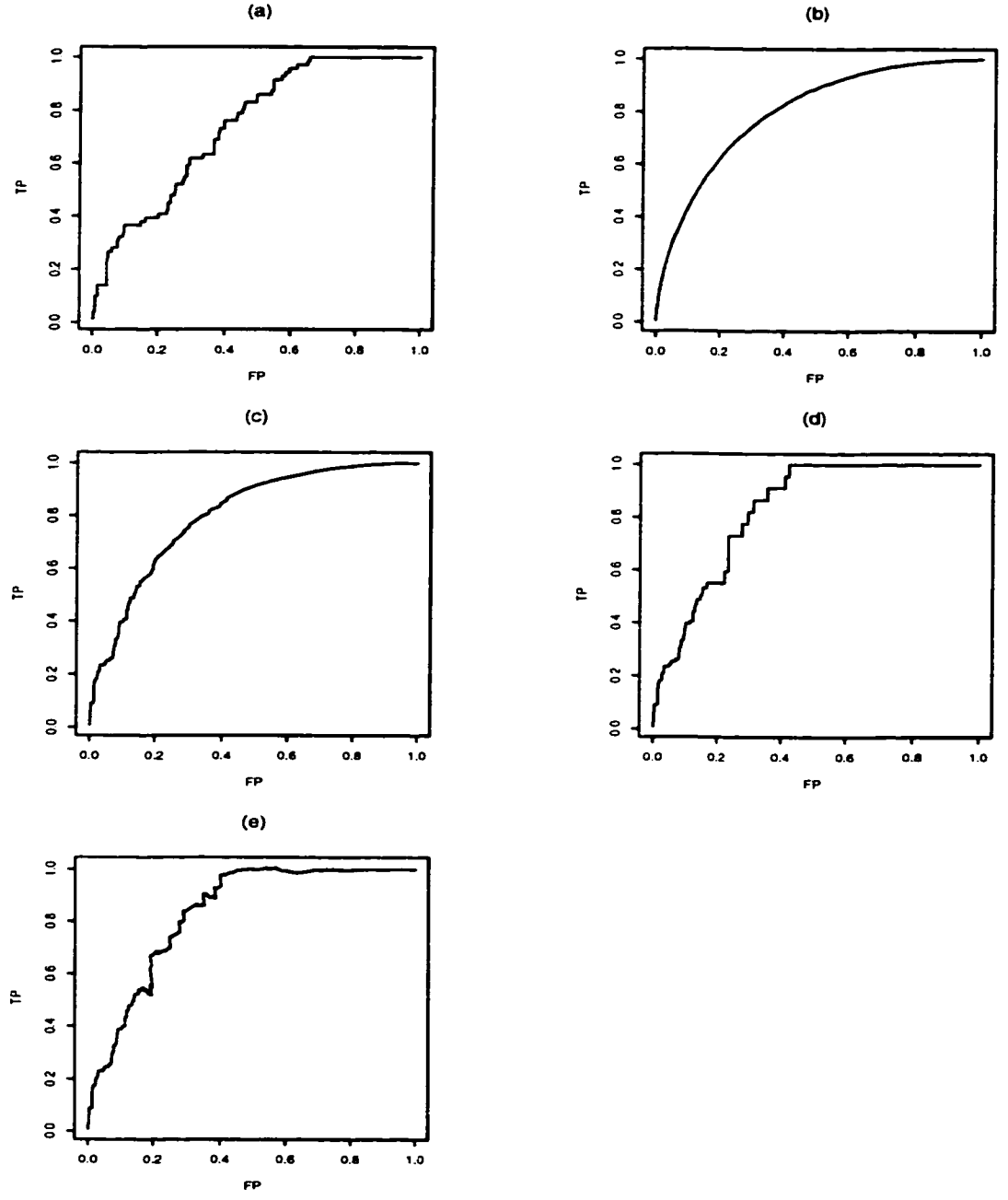

Figure 4.1: ROC curves corresponding to hypothetical data. (a) CC, (b) BG, (c) MS, (d) IPW, and (e) SP.

Table 4.2: Qualitative comparison of the different approaches.

	Require Estimates of		ROC Curve	
Method	$P(D\|T,A)$?	$P(V\|T,A)$?	relatively smooth?	ROC monotone?
CC	No	No	No	Yes
BG	Yes	No	Yes	Yes
MS	Yes	No	Yes	Yes
IPW	No	Yes	No	Yes
SP	Yes	Yes	Yes	No

all study subjects. These methods may not be as efficient as a fully parametric analysis that specifies a model for $P(T,A|D)$. However, it is usually much easier to specify distributions for $P(D|T,A)$ or $P(V|T,A)$, and these distributions require fewer assumptions. In addition, $P(D|T,A)$ and $P(V|T,A)$ can easily handle multivariate auxiliary data A which often arise in practice and can be estimated using only the verification group.

Chapter 5

ASYMPTOTIC DISTRIBUTION THEORY FOR ESTIMATORS OF DISEASE PREVALENCE, TP, AND FP

In the previous chapter we observed that prevalence estimators are the building blocks of $\widehat{TP}(c)$ and $\widehat{FP}(c)$, and the methods proposed differ in how they estimate each prevalence component. Not only is disease prevalence estimation important in its own right, but in Section 5.8 we show the same method used to develop asymptotic distribution theory for disease prevalence estimators can be used to develop asymptotic distribution theory for $TP(c)$ and $FP(c)$ estimators. Therefore, we first consider inference for disease prevalence and then discuss inference for $TP(c)$ and $FP(c)$.

Since it can be shown that the CC, BG, and SP estimators of disease prevalence along with the mean score and inverse probability weighting estimators of disease prevalence considered by Clayton et al. (1998) can be derived as solutions to estimating equations of the same form, we chose to use the corresponding estimating functions when developing theory. By considering these estimating functions along with estimating functions corresponding to the estimation of the nuisance parameters, we are able to account for the uncertainty of estimating nuisance parameters in the estimation of the parameter of interest (Clayton et al., 1998). In our setting, the nuisance parameters arise from modelling the probability of disease conditional on T and A in the verification sample and/or the probability of verification conditional on T and A. Depending on the form of the data, these estimating functions could correspond to, for example, logistic regression.

5.1 Notation

Let $\beta = (\theta, \alpha)^T$ where θ is a scalar corresponding to the parameter of interest and α is a vector of nuisance parameters. Denote β_0 to be the true value of the parameter vector β and let $N_\delta(\beta_0)$ be a δ-neighborhood of β_0. Furthermore, let $\widehat{\beta}$ be the solution to the vector of estimating equations $U_n(\beta) = \sum_{i=1}^n U_i(\beta) = 0$ where $U_i(\beta)$ is the ith subject's contribution to the estimating functions. Let $U^\theta(\beta)$ and $U^\alpha(\beta)$ be the estimation functions corresponding to the estimation of θ and α, respectively.

5.2 Assumptions

The theory developed in this chapter relies on these assumptions:

A1: D is MAR

A2: (D_i, T_i, A_i, V_i) are independent and identically distributed (iid)

A3: (T, A) is bounded

A4: $N_\delta(\beta_0)$ is bounded

A5: $E(\frac{\partial}{\partial\beta}U_i(\beta_0))$ is negative definite

A6: Disease and verification probabilities are bounded away from 0

5.3 Asymptotic Results for Solutions to Estimating Equations

In this section we prove consistency and develop distribution theory for solutions to estimating equations $U_n = \sum_{i=1}^n U_i(\beta) = 0$ where the estimating functions have the following properties:

P1: $U_i(\beta_0)$ iid

P2: Elements of $U_n(\beta)$, $\frac{\partial}{\partial\beta}U_n(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n(\beta)$ exist in $N_\delta(\beta_0)$

P3: $U_i(\beta)$, $\frac{\partial}{\partial\beta}U_i(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_i(\beta)$ are uniformly bounded in $N_\delta(\beta_0)$

P4: $E(U_n(\beta_0)) = 0$.

In Sections 5.4-5.6 we will show that estimating functions $U_n^\theta(\beta)$ corresponding to SP, IPW, BG, MS, and CC prevalence estimators have properties P1-P4 under assumptions A1-A5. We assume P1-P4 are true for the estimating functions corresponding to the estimation of nuisance parameters. Clearly, if generalized linear models, for example logistic regression, are used to estimate the nuisance parameters α_1 and α_2, then the corresponding estimating functions will have properties P1-P4.

5.3.1 Consistency of $\widehat{\beta}$

Theorem 5.1 *(consistency) Given estimating functions that satisfy P1-P4 then a unique solution $\widehat{\beta}$ to $U_n(\beta) = 0$ exists with probability converging to 1 as $n \to \infty$ and $\widehat{\beta} \to_p \beta_0$.*

Theorem 5.1 relies on the following results[1].

Lemma 5.1 *(uniformly bounded averages of $U_i(\beta)$ and its derivatives) If $U_i(\beta)$ and its first and second derivatives are uniformly bounded in $N_\delta(\beta_0)$, then averages of these quantities are also uniformly bounded in $N_\delta(\beta_0)$.*

Lemma 5.2 *(uniformly bounded $\frac{\partial}{\partial\beta}E(\frac{\partial}{\partial\beta}U_i(\beta))$) $\frac{\partial}{\partial\beta}E(\frac{\partial}{\partial\beta}U_i(\beta))$ exists and is uniformly bounded in $N_\delta(\beta_0)$ if $\frac{\partial}{\partial\beta}(\frac{\partial}{\partial\beta}U_i(\beta))$ is uniformly bounded in $N_\delta(\beta_0)$.*

---

[1]Proofs of all lemmas can be found in Appendix A.

5.3.2 Proof of Theorem 5.1 (consistency)

Foutz (1977) proved that the solution to a set of generic estimating equations exists and is unique in $N_\delta(\beta_0)$ and is consistent for β_0 provided the following four conditions are satisfied:

i. Elements of $\frac{\partial}{\partial\beta}U_n(\beta)$ exist and are continuous in $N_\delta(\beta_0)$

ii. $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0)$ is negative definite with probability converging to 1 as $n \to \infty$

iii. $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) \to_p E(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta))$ uniformly for $\beta \in N_\delta(\beta_0)$ as $n \to \infty$

iv. $E(U_n(\beta_0)) = 0$.

Therefore, to prove Theorem 5.1 is true, it suffices to show that estimating functions with properties P1-P4 satisfy conditions (i)-(iv).

First, we consider condition (i). The existence of elements of $\frac{\partial}{\partial\beta}U_n(\beta)$ in $N_\delta(\beta_0)$ is guaranteed to be true by property P2 of the estimating functions. Continuity is also implied by P2 since $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n(\beta)$ exists in $N_\delta(\beta_0)$.

To prove condition (ii) is true, we need to show that $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0)$ is negative definite with probability converging to 1 as $n \to \infty$. We note that

$$\begin{aligned} n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0) &\equiv n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}U_i(\beta_0) \\ &\to_p E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right) \quad \text{by WLLN since } U_i(\beta_0) \text{ iid (P1)} \\ &= E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0)\right) \end{aligned}$$

where WLLN is the Weak Law of Large Numbers. Therefore, $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0)$ is negative definite with probability converging to 1 as $n \to \infty$ since $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0)$ converges in probability to $E(\frac{\partial}{\partial\beta}U_i(\beta_0))$ that is negative definite by assumption A5 (Horn & Johnson, 1991).

Next, we examine condition (iii). To show $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) \rightarrow_p E(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta))$ uniformly for $\beta \in N_\delta(\beta_0)$ as $n \rightarrow \infty$, we need to show that given $\epsilon > 0$ and $\gamma > 0$ there exists N_o such that $\forall n > N_o$

$$P\left(\sup_{\beta\in N_\delta(\beta_0)}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right| > \epsilon\right) < \gamma. \tag{5.1}$$

Given γ, we can choose a finite partition P_k of $N_\delta(\beta_0)$ such that $|\beta_k - \beta| < \psi \ \ \forall\beta \in P_k$. In other words, the distance between any two points in the partition is less than ψ. By the addition and subtraction of $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)$ and $E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)$ and the triangle inequality,

$$\begin{aligned}
&\sup_{\beta\in N_\delta(\beta_0)}\left|n^{-1}\tfrac{\partial}{\partial\beta}U_n(\beta) - E(n^{-1}\tfrac{\partial}{\partial\beta}U_n(\beta))\right| \\
&= \max_k \sup_{\beta\in P_k}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right| \\
&= \max_k \sup_{\beta\in P_k}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) - n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k) + n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right. \\
&\qquad\qquad \left. + E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right| \\
&\leq \max_k \sup_{\beta\in P_k}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) - n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right| \\
&\qquad\qquad + \max_k \sup_{\beta\in P_k}\left|E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right| \\
&\qquad\qquad\qquad + \max_k\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k) - E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right|.
\end{aligned} \tag{5.2}$$

We separately consider each of the three terms on the right-hand-side (RHS) of (5.2). The first term

$$\begin{aligned}
\max_k \sup_{\beta\in P_k}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) - n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right| &= \sup_{\beta\in P_k}\frac{\partial}{\partial\beta}\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta^*)\right)|\beta_j - \beta| \\
&< \frac{\partial}{\partial\beta}\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta^*)\right)\psi \\
&\leq M\psi
\end{aligned}$$

where $\beta^* \in (\beta_j, \beta)$. The mean value theorem yields the equality above while the last inequality is given by P3, the uniform boundedness of $\frac{\partial}{\partial\beta}U_i(\beta)$ for all $\beta \in N_\delta(\beta_0)$, and Lemma 5.1.

Similarly, the mean value theorem and uniform boundedness of $\frac{\partial}{\partial\beta}\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)$, which by Lemma 5.2 implies the uniform boundedness of $\frac{\partial}{\partial\beta}\left(E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right)$, yield that

$$\begin{aligned}\max_k \sup_{\beta\in P_k}\left|E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)-E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right| &= \sup_{\beta\in P_k}\frac{\partial}{\partial\beta}\left(E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta^*)\right)\right)|\beta_j-\beta| \\ &< \frac{\partial}{\partial\beta}\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta^*)\right)\psi \\ &\leq M\psi.\end{aligned}$$

Given $\gamma > 0$ and $\epsilon > 0$, we choose ψ such that $\psi < \frac{\epsilon}{3M}$. Therefore, applying the inequalities we just obtained for the first two terms of (5.2) we get

$$\sup_{\beta\in N_\delta(\beta_0)}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)-E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right| < \frac{2}{3\epsilon}+\max_k\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)-E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right|.$$

Since by the WLLN $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta) \to_p E(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta))$, we can choose N_o such that

$$\max_k\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)-E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_k)\right)\right| < \frac{\epsilon}{3}$$

with probability greater than $1-\gamma$. Hence,

$$P\left(\sup_{\beta\in N_\delta(\beta_0)}\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)-E\left(n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right)\right| > \epsilon\right) < \gamma$$

and we have shown the uniform convergence required in condition (iii).

It is assumed (A5) that we are considering estimating functions that satisfy the final condition, (iv) $E(U_n(\beta_0)) = 0$. Therefore, since we have shown that estimating functions with properties P1-P4 satisfy the four conditions of Foutz, a unique solution $\widehat{\beta}$ to $\sum_{i=1}^n U_i(\beta) = 0$ exists with probability converging to 1 as $n \to \infty$ and $\widehat{\beta} \to_p \beta_0$.

5.3.3 Distribution Theory for $\widehat{\beta}$

Theorem 5.2 *(asymptotic normality) Given estimating functions that satisfy P1-P4 then as n converges to ∞, $n^{\frac{1}{2}}(\widehat{\beta}-\beta_0)$ converges in distribution to a mean 0 normally*

distributed random variable with variance-covariance matrix

$$\sum = \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1} Cov(U_i(\beta_0)) \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1}.$$

Theorem 5.2 relies on the following lemma.

Lemma 5.3 *(*$\left|\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right|\right|$ *bounded away from 0)* $\left|\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right|\right|$ *is bounded in probability away from 0* $\forall\beta \in N_\delta(\beta_0)$ *provided the following conditions are true:*

i. $\frac{\partial}{\partial\beta}\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)$ *is uniformly bounded in* $N_\delta(\beta_0)$

ii. $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0) \to_p E\left(\frac{\partial}{\partial\beta}U_n(\beta_0)\right)$

iii. $E\left(-\frac{\partial}{\partial\beta}U_i(\beta_0)\right)$ *is positive definite*

5.3.4 Proof of Theorem 5.2 (asymptotic normality)

A second-order Taylor series expansion of $n^{-\frac{1}{2}}U_n(\widehat{\beta})$ about β_0 yields:

$$n^{-\frac{1}{2}}U_n(\widehat{\beta}) = n^{-\frac{1}{2}}U_n(\beta_0) + (\widehat{\beta}-\beta_0)n^{-\frac{1}{2}}\frac{\partial}{\partial\beta}U_n(\beta_0) + \frac{1}{2}(\widehat{\beta}-\beta_0)n^{-\frac{1}{2}}\frac{\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^*)(\widehat{\beta}-\beta_0)$$

where β^* is between $\widehat{\beta}$ and β_0. Since $n^{-\frac{1}{2}}U_n(\widehat{\beta}) = 0$,

$$n^{-\frac{1}{2}}U_n(\beta_0) = n^{\frac{1}{2}}(\widehat{\beta}-\beta_0)n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) + n^{\frac{1}{2}}(\widehat{\beta}-\beta_0)n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^*)\frac{(\widehat{\beta}-\beta_0)}{2}. \tag{5.3}$$

Next we aim to show that the second term on the RHS of (5.3) converges to 0 as $n \to \infty$. We first consider $n^{\frac{1}{2}}(\widehat{\beta}-\beta_0)$. A first-order Taylor series expansion of $n^{-\frac{1}{2}}U_n(\widehat{\beta})$ about β_0 yields:

$$n^{-\frac{1}{2}}U_n(\widehat{\beta}) = n^{-\frac{1}{2}}U_n(\beta_0) + n^{-\frac{1}{2}}(\widehat{\beta}-\beta_0)\frac{\partial}{\partial\beta}U_n(\beta^*).$$

This implies

$$n^{-\frac{1}{2}}U_n(\beta_0) = n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^*)$$

since $n^{-\frac{1}{2}}U_n(\widehat{\beta}) = 0$.

$n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) = O_p(1)$, i.e. bounded in probability, if $\left|\left|n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)\right|\right|$ is bounded in probability. Furthermore, $\left|\left|n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)\right|\right|$ is bounded in probability if $\left|\left|n^{-\frac{1}{2}}U_n(\beta_0)\right|\right|$ is bounded in probability and $\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^*)\right|\right|$ is bounded in probability away from 0. By the Central Limit Theorem $\left|\left|n^{-\frac{1}{2}}U_n(\beta_0)\right|\right|$ is bounded in probability and by Lemma 5.3 $\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^*)\right|\right|$ is bounded in probability away from 0. The three conditions required by Lemma 5.3 are true by P3 and Lemma 5.1, P1 and WLLN, and A5 respectively.

Next we consider $n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^*)$. By P3 $\frac{-\partial^2}{\partial\beta\partial\beta^T}U_i(\beta)$ is uniformly bounded in $N_\delta(\beta_0)$. So by Lemma 5.1 $n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta)$ is also uniformly bounded in $N_\delta(\beta_0)$. Thus, $n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^*) = O_p(1)$ since $\beta^* \in N_\delta(\beta_0)$.

The final component of the second term on the RHS of (5.3) to consider is $\frac{(\widehat{\beta}-\beta_0)}{2}$. The consistency of $\widehat{\beta}$ for β_0 (Theorem 5.1) implies that $\frac{(\widehat{\beta}-\beta_0)}{2}$ converges in probability to 0.

The second term on the RHS of (5.3), therefore, converges to 0 in probability since two components are bounded in probability and one component converges to 0 as $n \to \infty$. Hence (5.3) is equal to

$$n^{-\frac{1}{2}}U_n(\beta_0) = n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) + o_p(1). \tag{5.4}$$

The inverse of $n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)$ exists with high probability in large samples since $n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) \to_p E\left(n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right)$ by the WLLN and $E\left(n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right)$ is positive definite by A5 (Horn & Johnson, 1991). This allows us to re-arrange (5.4) as

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) = n^{-\frac{1}{2}}U_n(\beta_0)\left[n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right]^{-1} + o_p(1).$$

Therefore, the asymptotic distribution of $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)$ can be determined by calculating the asymptotic distribution of $n^{-\frac{1}{2}}U_n(\beta_0)\left[n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right]^{-1}$. Since $U_i(\beta_0)$ are

iid, $n^{-\frac{1}{2}}U_n(\beta_0) \rightarrow_d N(0, Cov(U_i(\beta_0)))$ by the CLT. By the WLLN and continuous mapping theorem, we also observe that $\left[n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right]^{-1}$ converges in probability to $\left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1}$.

Hence,

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) \rightarrow_d N\left(0, \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1} Cov(U_i(\beta_0)) \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1}\right) \tag{5.5}$$

by Slutsky's theorem. This completes the proof.

5.3.5 Asymptotic Variance-Covariance Matrix

Theorem 5.2 states that as n converges to ∞, $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)$ converges in distribution to a mean 0 normally distributed random variable with variance-covariance matrix $\not\Sigma = [-E(\frac{\partial}{\partial\beta}U_i(\beta_0))]^{-1}Cov(U_i(\beta_0))[-E(\frac{\partial}{\partial\beta}U_i(\beta_0))]^{-1}$. Recall that $\beta = (\theta, \alpha)^T$ where θ is the parameter of main interest and α are the nuisance parameters. Therefore, $\not\Sigma$ is equivalent to

$$\vartheta(\theta,\alpha)^{-1}\Xi(\theta,\alpha)\vartheta(\theta,\alpha)^{-1}$$

where

$$\vartheta(\theta,\alpha) \equiv \begin{pmatrix} \vartheta_\theta(\theta,\alpha) & \vartheta_{\theta,\alpha}(\theta,\alpha) \\ \vartheta_{\alpha,\theta}(\theta,\alpha) & \vartheta_\alpha(\theta,\alpha) \end{pmatrix} = -E\left(\frac{\partial}{\partial\beta}U_i(\beta)\right)$$

and

$$\Xi(\theta,\alpha) \equiv \begin{pmatrix} \Xi_\theta(\theta,\alpha) & \Xi_{\theta,\alpha}(\theta,\alpha) \\ \Xi_{\alpha,\theta}(\theta,\alpha) & \Xi_\alpha(\theta,\alpha) \end{pmatrix} = Cov(U_i(\beta)).$$

We see that by including estimating functions corresponding to the estimation of nuisance parameters, the resulting asymptotic variance-covariance matrix takes into account the uncertainty of estimating α when determining the variability of $\widehat{\theta}$.

5.3.6 *Distribution Theory for* $\widehat{\theta}$

Our main interest is in $\widehat{\theta}$, the disease prevalence estimator. Theorems 5.1 and 5.2 imply that $\widehat{\theta}$ is consistent for the true disease prevalence and

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \rightarrow_d N\left(0, \sigma_\theta^2\right). \tag{5.6}$$

Next we derive an expression for σ_θ^2. In our setting $U_i^\theta(\theta, \alpha)$, the estimating function used to estimate disease prevalence, depends on the data and both θ and α. Conversely, the vector of estimating functions corresponding to the estimation of the nuisance parameters, $U_i^\alpha(\alpha)$, are only a function of α and the data. Since $U_i^\alpha(\alpha)$ is not a function of θ, $\vartheta_{\theta,\alpha}(\alpha) = 0$. It can then be shown by matrix algebra that

$$\begin{aligned} \sigma_\theta^2 &= \vartheta_\theta^{-1}\Xi_\theta(\vartheta_\theta^{-1})^T + \vartheta_\theta^{-1}\vartheta_{\theta\alpha}Cov(\alpha)\vartheta_{\theta\alpha}^T(\vartheta_\theta^{-1})^T \\ &\quad - \vartheta_\theta^{-1}(\Xi_{\theta\alpha}\vartheta_\alpha^{-1}\vartheta_{\theta\alpha}^T + \vartheta_{\theta\alpha}\vartheta_\alpha^{-1}\Xi_{\theta\alpha}^T)(\vartheta_\theta^{-1})^T \end{aligned} \tag{5.7}$$

where $Cov(\alpha) = \vartheta_\alpha^{-1}\Xi_\alpha(\vartheta_\alpha^{-1})^T$. Furthermore, in our setting $\vartheta_\theta = 1$ and θ is a scalar so (5.7) reduces to

$$\Xi_\theta + \vartheta_{\theta\alpha}Cov(\alpha)\vartheta_{\theta\alpha}^T - 2\Xi_{\theta\alpha}\vartheta_\alpha^{-1}\vartheta_{\theta\alpha}^T. \tag{5.8}$$

We see that σ_θ^2 is comprised of three components. The first component is the variability of $\widehat{\theta}$ if α were known while the last two components take into account the variability in $\widehat{\alpha}$.

5.3.7 *Consistent Variance Estimator for* $\widehat{\beta}$

Theorem 5.3 *(consistent variance estimator) A consistent estimator of* $\not\Sigma$*, the asymptotic variance-covariance matrix for* $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)$*, is*

$$\widehat{\not\Sigma} = n\left[\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}\left[\sum_{i=1}^{n}U_i(\widehat{\beta})U_i(\widehat{\beta})^T\right]\left[\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}$$

5.3.8 Proof of Theorem 5.3 (consistency of variance estimator)

Our goal is to show that each component of $\widehat{\not\Sigma}$ is consistent for the corresponding component of $\not\Sigma$ because then Slutsky's theorem can be used to complete the proof. First, we consider the outside components of $\widehat{\not\Sigma}$. A first-order Taylor series expansion of $n^{-1}\frac{-\partial}{\partial\beta}U_n(\widehat{\beta})$ about β_0 gives:

$$n^{-1}\frac{-\partial}{\partial\beta}U_n(\widehat{\beta}) = n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) + (\widehat{\beta}-\beta_0)n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^\star) \tag{5.9}$$

where $\beta^\star$ is between $\widehat{\beta}$ and β_0. The consistency of $\widehat{\beta}$ (Theorem 5.1) implies that $(\widehat{\beta}-\beta_0)$ converges in probability to 0 as $n\to\infty$. By P3 and Lemma 5.1, $n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta)$ is uniformly bounded in $N_\delta(\beta_0)$. Since $\beta^\star \in N_\delta(\beta_0)$, this implies $n^{-1}\frac{-\partial^2}{\partial\beta\partial\beta^T}U_n(\beta^\star)$ is bounded in probability. Thus, (5.9) can be re-written as

$$n^{-1}\frac{-\partial}{\partial\beta}U_n(\widehat{\beta}) = n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) + o_p(1).$$

By the WLLN, $n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)$ converges in probability to $E\left[\frac{-\partial}{\partial\beta}U_i(\beta_0)\right]$ as $n\to\infty$.

Next, we consider the middle component of $\widehat{\not\Sigma}$. Similarly, a first-order Taylor series expansion of $n^{-1}\sum_{i=1}^n U_i(\widehat{\beta})U_i(\widehat{\beta})^T$ about β_0 gives

$$n^{-1}\sum_{i=1}^n U_i(\widehat{\beta})U_i(\widehat{\beta})^T = n^{-1}\sum_{i=1}^n U_i(\beta_0)U_i(\beta_0)^T + (\widehat{\beta}-\beta_0)n^{-1}\sum_{i=1}^n \frac{\partial}{\partial\beta}\left[U_i(\beta^\star)U_i(\beta^\star)^T\right]. \tag{5.10}$$

Again the second term on the RHS converges in probability to 0 since $\widehat{\beta}$ is consistent for β_0 (Theorem 5.1) and $n^{-1}\sum_{i=1}^n \frac{\partial}{\partial\beta}\left[U_i(\beta^\star)U_i(\beta^\star)^T\right]$ is bounded in probability. The latter follows from the chain rule and property P3 which gives the uniform boundedness of $U_i(\beta)$ and $\frac{\partial}{\partial\beta}U_i(\beta)$ for all $\beta \in N_\delta(\beta_0)$, in particular $\beta^\star$.

So (5.10) can be re-written as

$$n^{-1}\sum_{i=1}^n U_i(\widehat{\beta})U_i(\widehat{\beta})^T = n^{-1}\sum_{i=1}^n U_i(\beta_0)U_i(\beta_0)^T + o_p(1).$$

Furthermore, by the WLLN

$$n^{-1}\sum_{i=1}^{n} U_i(\beta_0)U_i(\beta_0)^T \rightarrow_p E\left[U_i(\beta_0)U_i(\beta_0)^T\right].$$

This implies $n^{-1}\sum_{i=1}^{n} U_i(\widehat{\beta})U_i(\widehat{\beta})^T$ converges in probability to $E\left[U_i(\beta_0)U_i(\beta_0)^T\right]$.

Finally, we combine the convergence results for each component of $\widehat{\not\Sigma}$ using Slutsky's theorem to conclude that

$$\begin{aligned}
\widehat{\not\Sigma} &\equiv n\left[\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}\left[\sum_{i=1}^{n}U_i(\widehat{\beta})U_i(\widehat{\beta})^T\right]\left[\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1} \\
&= \left[n^{-1}\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}\left[n^{-1}\sum_{i=1}^{n}U_i(\widehat{\beta})U_i(\widehat{\beta})^T\right]\left[n^{-1}\sum_{i=1}^{n}\frac{-\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1} \\
&\rightarrow_p \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1} E\left[U_i(\beta_0)U_i(\beta_0)^T\right]\left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1} \\
&= \not\Sigma.
\end{aligned}$$

The last equality follows from the definition of covariance and property P4.

5.4 CC Estimator

Recall that the complete case estimator of disease prevalence is

$$\left(\sum_{i=1}^{n} V_i\right)^{-1}\sum_{i=1}^{n} V_i D_i.$$

This estimator can be obtained by setting the estimating function

$$U_n(\theta) = \sum_{i=1}^{n} U_i^{\theta}(\theta) = \sum_{i=1}^{n} V_i(D_i - \theta) \tag{5.11}$$

equal to 0 and solving for θ. This estimator does not include any nuisance parameters so we only need to show that (5.11) has properties P1-P4 to apply the asymptotic distribution theory developed in Section 5.3.

P1 is satisfied since $U_i^{\theta}(\theta_0) = V_i(D_i - \theta_0)$ and data for different subjects are iid by A2. $\frac{\partial}{\partial\theta}U_i(\theta) = -V_i$ and $\frac{\partial^2}{\partial\theta\partial\theta^T}U_i(\theta) = 0$ exist so averages of these expressions exist as

well and, thus, P2 is true. $U_i^{\theta}(\beta)$, $\frac{\partial}{\partial\beta}U_i^{\theta}(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_i^{\theta}(\beta)$ are uniformly bounded in $N_\delta(\beta_0)$ (P3) because they are only a function of data and θ which are bounded by assumptions A3 and A4. This estimator is only valid when the verification sample is a simple random sample so below we show P4 is true in this case. Otherwise, as we will see in the next chapter P4 can fail.

$$\begin{aligned} E(U_n(\theta_0)) &= \sum_{i=1}^{n} E(U_i(\theta_0)) \\ &= \sum_{i=1}^{n} E(V_i(D_i - \theta_0)) \\ &= \sum_{i=1}^{n} \{E(V_i)E(D_i) - E(V_i)\theta_0\} \\ &= 0. \end{aligned}$$

Since P1-P4 are true, Theorem 5.2 yields

$$n^{\frac{1}{2}}(\widehat{\theta}_{CC} - \theta_0) \rightarrow_d N(0, \theta(1-\theta)/\pi) \tag{5.12}$$

where $\pi = P(V)$. Not surprisingly, the asymptotic variance is the usual binary variance weighted by the probability of verification.

5.5 *Class of Verification Restricted Estimators*

In Section 4.8.1 a class of estimators was developed that required $E(D) = \theta_0$ and verification probabilities to be known or known up to a vector of constants. These estimators simultaneously solve a vector of estimating functions corresponding to the estimation of nuisance parameters, $U_n^{\alpha}(\theta, \alpha)$, and the following estimating function for the parameter of interest

$$U_n^{\theta}(\theta, \alpha) = \sum_{i=1}^{n} g_2(\alpha_2, T_i, A_i)^{-1} V_i(D_i - \theta) - k g_2(\alpha_2, T_i, A_i)^{-1}(V_i - g_2(\alpha_2, T_i, A_i))(g_1(\alpha_1, T_i, A_i) - \theta)$$

where $g_1(\alpha_1, T_i, A_i)$ is an estimate of $P(D_i = 1|T_i, A_i)$ and $g_2(\alpha_2, T_i, A_i)$ is an estimate of $P(V_i = 1|T_i, A_i)$. Thus, for this class of estimators there are two vectors of nuisance

parameters, α_1 and α_2. Note that $g_2(\alpha_2, T_i, A_i)$ was previously denoted π_i and in some studies may not need to be estimated.

Setting these estimating functions equal to 0 and solving for θ results in the IPW estimator when $k = 0$ and the semi-parametric efficient estimator when $k = 1$.

Consistency and asymptotic distribution theory for $\widehat{\theta}$ is ensured by Theorem 5.1 provided the estimating functions have properties P1-P4. Since properties P2 and P3 involve derivatives with respect to all the parameters, P2 and P3 are the only properties that we cannot consider separately for $U^{\theta}(\theta, \alpha)$ and $U^{\alpha}(\alpha)$. Derivatives of $U^{\alpha}(\alpha)$ with respect to θ equal 0 because $U^{\alpha}(\alpha)$ are not a function of θ. Therefore, P2 and P3 are true for $\frac{\partial}{\partial\theta}U^{\alpha}(\alpha)$. We assume properties P1-P4 are satisfied for $U^{\alpha}(\alpha)$. So all that remains to be shown is that $U_n^{\theta}(\theta, \alpha)$ satisfy P1-P4 and $\frac{\partial}{\partial\alpha}U_n^{\theta}(\theta, \alpha)$ satisfy P2-P3. Equivalently we need to show P1 and P4 are true for $U_n^{\theta}(\theta, \alpha)$ and P2 and P3 are true for derivatives of $U_n^{\theta}(\theta, \alpha)$ with respect to β.

$U_i^{\theta}(\theta, \alpha)$ are iid (P1) since each $U_i^{\theta}(\theta, \alpha)$ is only a function of the data which are assumed to be iid (A2). Expressions for $U_n^{\theta}(\beta)$, $\frac{\partial}{\partial\beta}U_n^{\theta}(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n^{\theta}(\beta)$ can be easily derived. Hence existence (P2) is satisfied. Furthermore, these expressions are bounded if all elements of the expressions are bounded. These expressions are only functions of the data, parameters, $g_1(\alpha_2, T_i, A_i)$, and $g_2(\alpha_2, T_i, A_i)$. By assumptions A3 and A4, β and (T, A) are bounded. By assumption A6 $g_1(\alpha_2, T_i, A_i)$ and $g_2(\alpha_2, T_i, A_i)$ are bounded away from 0. Thus, we have shown P3 to be true.

The final property (P4) we must show is true is that the expectation of $U_n^{\theta}(\theta_0, \alpha_0)$ equals 0. We prove that now.

$$
\begin{aligned}
E(U_n^{\theta}(\theta_0,\alpha_0)) &= \sum_{i=1}^{n} E[g_2(\alpha_2,T_i,A_i)^{-1}V_i(D_i-\theta_0) \\
&\quad - kg_2(\alpha_2,T_i,A_i)^{-1}(V_i-g_2(\alpha_2,T_i,A_i))(g_1(\alpha_1,T_i,A_i)-\theta_0)] \\
&= \sum_{i=1}^{n}\{E[E\{g_2(\alpha_2,T_i,A_i)^{-1}V_i(D_i-\theta_0)|T_i,A_i\}] \\
&\quad - kE[E[g_2(\alpha_2,T_i,A_i)^{-1}(V_i-g_2(\alpha_2,T_i,A_i))(g_1(\alpha_1,T_i,A_i)-\theta_0)|T_i,A_i]]\} \\
&= \sum_{i=1}^{n}\{E[g_2(\alpha_2,T_i,A_i)^{-1}E[D_i|T_i,A_i]E[V_i|T_i,A_i]-g_2(\alpha_2,T_i,A_i)^{-1}\theta_0 E[V_i|T_i,A_i]] \\
&\quad - kE[g_2(\alpha_2,T_i,A_i)^{-1}(g_1(\alpha_1,T_i,A_i)-\theta_0)(E[V_i|T_i,A_i]-g_2(\alpha_2,T_i,A_i))]\} \\
&= \sum_{i=1}^{n}\{E[E[D_i|T_i,A_i]-g_2(\alpha_2,T_i,A_i)^{-1}g_2(\alpha_2,T_i,A_i)\theta_0] \\
&\quad - kE[g_2(\alpha_2,T_i,A_i)^{-1}(g_1(\alpha_1,T_i,A_i)-\theta_0)(g_2(\alpha_2,T_i,A_i)-g_2(\alpha_2,T_i,A_i))]\} \\
&= \sum_{i=1}^{n} E[E[D_i|T_i,A_i]-\theta_0]-0 \\
&= \sum_{i=1}^{n}\{E[D_i]-\theta_0\} \\
&= 0
\end{aligned}
$$

where the third equality is a result of the MAR assumption (A1).

Since estimating functions for IPW and SP have properties P1-P4, their prevalence estimators are consistent. In addition, Theorems 5.2 and 5.3 provide asymptotic distribution theory and consistent variance estimators.

5.6 Class of Disease Restricted Estimators

In Section 4.8.2 we developed a class of estimators that is consistent for the true prevalence, θ_0, provided $E(D)=\theta_0$ and $P(D_i=1|T_i,A_i)$ are known or known up to a vector of constants. These estimators simultaneously solve a vector of estimating functions corresponding to the estimation of nuisance parameters, $U_n^{\alpha}(\theta,\alpha)$, and the following estimating function for the parameter of interest

$$
U_n^{\theta}(\theta,\alpha)=V_i(kD_i+(1-k)g(\alpha,T_i,A_i)-\theta)+(1-V_i)(g(\alpha,T_i,A_i)-\theta) \qquad (5.13)
$$

where $g(\alpha, T_i, A_i)$ is an estimate of $P(D_i = 1|T_i, A_i)$. Recall that BG and MS are members of this class when k equals 0 and 1, respectively.

Theorems 5.1 and 5.2 imply consistency of $\widehat{\theta}$ and provide asymptotic distribution theory provided we can show that P1-P4 are true for $U_n(\theta, \alpha)$. Since P1-P4 are assumed to be true for $U_n^{\alpha}(\theta, \alpha)$, similar to the verification restricted class of estimators all that needs to be shown is that $U_n^{\theta}(\theta, \alpha)$ satisfy P1-P4.

Again $U_i^{\theta}(\theta_0, \alpha_0)$ are iid since they are only a function of constants k, θ_0, and α_0 and the data (T_i, A_i, D_i, V_i) which are assumed to be iid by A2. So P1 is satisfied.

Existence of $U_n^{\theta}(\beta)$, $\frac{\partial}{\partial\beta}U_n^{\theta}(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n^{\theta}(\beta)$ (P2) is satisfied since expressions for each of these can easily be derived. Furthermore, these expressions are only functions of the data, parameters, and $g(\alpha, T_i, A_i)$ so they are bounded if each of these are bounded. P3 then follows from assumptions A3, A4, and A6.

The final property to consider is P4.

$$
\begin{aligned}
E(U_n^{\theta}(\theta_0, \alpha_0)) &\equiv \sum_{i=1}^{n} E[V_i(kD_i + (1-k)g(\alpha_0, T_i, A_i)) + (1 - V_i)g(\alpha_0, T_i, A_i) - \theta_0] \\
&= \sum_{i=1}^{n} kE[V_i D_i] + (1-k)E[V_i\, g(\alpha_0, T_i, A_i)] + E[(1 - V_i)g(\alpha_0, T_i, A_i)] - \theta_0 \\
&= \sum_{i=1}^{n} kE\{E[V_i D_i|T_i, A_i]\} + (1-k)E[V_i P(D_i = 1|T_i, A_i)] \\
&\qquad + E\{E[(1 - V_i)g(\alpha_0, T_i, A_i)|T_i, A_i] - \theta_0 \\
&= \sum_{i=1}^{n} kE[V_i]E[D_i] + (1-k)E[V_i]E[D_i] + E[1 - V_i]E[D_i] - \theta_0 \quad \text{by MAR (A1)} \\
&= 0.
\end{aligned}
$$

Since P1-P4 are true for estimating functions corresponding to the class of disease restricted estimators, theoretical results are applicable to all estimators of this class, including BG and MS.

5.6.1 *Most Efficient Estimator*

For the verification restricted class of estimators we know by construction that the most efficient estimator in this class is SP, the semi-parametric efficient estimator. However, the most efficient estimator in the class of disease restricted estimators is not immediately known. By Theorem 5.2,

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \rightarrow_d N\left(0, \sigma_\theta^2\right)$$

where σ_θ^2 is given in (5.8) is a function of k. To find the most efficient estimator in the disease restricted class of estimators, we can minimize σ_θ^2 with respect to k.

First, we consider the case when disease probabilities are known. Denote these known probabilities by ρ_i. In this case σ_θ^2 reduces to

$$\begin{aligned} Cov(U_i^\theta(\theta_0, \rho_i)) &= E[U_i^\theta(\theta_0, \rho_i)^2] \\ &= E[V_i\{kD_i + (1-k)\rho_i - \theta_0\}^2 + (1-V_i)\{\rho_i - \theta_0\}^2 \\ &\quad + V_i(1-V_i)\{kD_i + (1-k)\rho_i - \theta_0\}\{\rho_i - \theta_0\}] \\ &= E[V_i\{kD_i + (1-k)\rho_i - \theta_0\}^2 + (1-V_i)\{\rho_i - \theta_0\}^2]. \end{aligned} \tag{5.14}$$

The last equality holds because $V_i(1-V_i) = 0$. Differentiating (5.14) with respect k we get $2E[k(V_iD_i - V_i\rho_i)(D_i - \rho_i) + (V_i\rho_i - V_i\theta_0)(D_i - \rho_i)]$. Setting this equal to 0 and solving for k we get that

$$\begin{aligned} k &= \frac{-E[(V_i\rho_i - V_i\theta_0)(D_i - \rho_i)]}{E[(V_iD_i - V_i\rho_i)(D_i - \rho_i)]} \\ &= \frac{-\rho_iP(V_iD_i) + \rho_i^2P(V_i) + \theta_0P(V_iD_i) - \theta_0\rho_iP(V_i)}{P(V_iD_i) - 2\rho_iP(V_iD_i) + \rho_i^2P(V_i)} \\ &= \frac{0}{P(V_iD_i)(1-\rho_i)} \\ &= 0. \end{aligned}$$

This implies that if the probabilities of being diseased conditional on test results and auxiliary information are known, then the BG estimator of disease prevalence

($k = 0$) is the most efficient estimator in the disease restricted class of estimators. In other words, if the disease probabilities are known, then it is more efficient to use them instead of the observed D_i.

Next we consider the more realistic setting where a parametric model, for example a logistic model, is correctly specified for $P(D|T, A)$ so that $\rho_i = P(D_i = 1|T_i, A_i)$ are known only up to a vector of constants. In this setting we have an additional vector of estimating functions corresponding to the model $P(D|T, A)$. Although this setting is more complicated, the same procedure for identifying the k that minimizes the asymptotic variance can be used here.

Applying (5.8) to this setting we get that

$$\begin{aligned}\sigma_\theta^2 &= E(U_i^\theta(\theta,\alpha,k)U_i^\theta(\theta,\alpha,k)^T) + E\left(\frac{\partial}{\partial\alpha}U_i^\theta(\theta,\alpha,k)\right) Cov(\alpha) E\left(\frac{\partial}{\partial\alpha}U_i^\theta(\theta,\alpha,k)\right)^T \\ &\quad - 2E(U_i^\theta(\theta,\alpha,k)U_i^\alpha(\alpha)^T)\left[E\left(\frac{\partial}{\partial\alpha}U_i^\alpha(\alpha)\right)\right]^{-1} E(U_i^\theta(\theta,\alpha,k)U_i^\alpha(\alpha)^T)^T. \quad (5.15)\end{aligned}$$

Therefore, the most efficient estimator in this class of estimators corresponds to the value of k that minimizes (5.15). To find k, numerically we can evaluate (5.15) for a range of values of k and find the one that minimizes (5.15). As will be seen in the next chapter, the optimal k depends on the scenario.

5.7 *Alternative Theory Derivation for BG Estimator*

It was shown in Section 5.6 that the asymptotic distribution theory developed in Section 5.3 using an estimating equation approach can be applied to the BG estimator of disease prevalence. Because of the simple form of this estimator, it is also possible to develop asymptotic distribution theory directly. This development is given here.

5.7.1 Notation

We are interested in finding asymptotic distribution theory for $\widehat{\theta} \equiv n^{-1}\sum_{i=1}^n g(\widehat{\alpha}, T_i, A_i)$ where $g(\widehat{\alpha}, T_i, A_i) = \widehat{P}(D_i = 1|T_i, A_i)$ and α is estimated using only data for subjects

in the verification sample. Let $G_n(\alpha) \equiv n^{-1} \sum_{i=1}^{n} g(\alpha, T_i, A_i)$, $G_n'(\alpha) \equiv \frac{\partial}{\partial \alpha} G_n(\alpha)$, and $\theta_0 = E(g(\alpha_0, T_i, A_i))$.

5.7.2 Distribution Theory

The theory developed in this section relies on assumptions A1-A4 and that the following properties are satisfied

P5: $g(\alpha, T_i, A_i)$, $\frac{\partial}{\partial \alpha} g(\alpha, T_i, A_i)$, and $\frac{\partial^2}{\partial \alpha \partial \alpha^T} g(\alpha, T_i, A_i)$ exist and are uniformly bounded in $N_\delta(\alpha_0)$

P6: $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \rightarrow_d N\left(0, \sum\!\!\!\!\!\!\diagup_{\alpha}\right)$ where the convergence in distribution is conditional on $\{(T_i, A_i), \quad i = 1, \cdots, n\}$.

In Appendix B it is shown that when $g(\alpha, T_i, A_i)$ is of logistic form and its first two derivatives exist and are bounded uniformly (P5) provided $N_\delta(\alpha_0)$ is bounded (A4) and the data are bounded (A3). Furthermore, it is also proven in Appendix B that P6, conditional convergence, is true when (T_i, A_i, D_i, V_i) are iid and $\widehat{\alpha}$ corresponds to the maximum likelihood estimator (MLE).

Theorem 5.4 *(asymptotic normality)*

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \rightarrow_d N\left(0, E\left(\frac{\partial}{\partial \alpha} g(\alpha_0, T_i, A_i)\right)^T \sum\!\!\!\!\!\!\diagup_{\alpha} E\left(\frac{\partial}{\partial \alpha} g(\alpha_0, T_i, A_i)\right) + Var(g(\alpha_0, T_i, A_i))\right).$$

Theorem 5.4 relies on Lemma 5.1.

5.7.3 Proof of Theorem 5.4 (asymptotic normality)

By adding and subtracting $n^{\frac{1}{2}} G_n(\alpha_0)$, $n^{\frac{1}{2}} G_n(\widehat{\alpha})$ can be re-written as

$$n^{\frac{1}{2}} G_n(\widehat{\alpha}) = n^{\frac{1}{2}} \{G_n(\widehat{\alpha}) - G_n(\alpha_0) + G_n(\alpha_0)\}. \tag{5.16}$$

Using a second-order Taylor series expansion of $\sqrt{n}G_n(\widehat{\alpha})$ about α_0,

$$n^{\frac{1}{2}}G_n(\widehat{\alpha}) = n^{\frac{1}{2}}G_n(\alpha_0) + n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)G_n'(\alpha_0) + R_G \tag{5.17}$$

where the remainder $R_G = n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha^*, T_i, A_i)\frac{1}{2}(\widehat{\alpha} - \alpha_0)$ and α^* lies between $\widehat{\alpha}$ and α_0. By P6, $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \rightarrow_d N(0, \sum\!\!\!\!\!\!/_{\alpha})$ where the convergence is conditional on $\{(T_i, A_i), \quad i = 1, \cdots, n\}$. This implies $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)$ is bounded in probability. P6 also implies $\widehat{\alpha}$ is consistent for α_0 so $\frac{1}{2}(\widehat{\alpha} - \alpha_0) \rightarrow_p 0$. The last term to consider is $n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha^*, T_i, A_i)$. By Lemma 5.1, $n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha, T_i, A_i)$ is uniformly bounded in $N_\delta(\alpha_0)$. Furthermore, α^* lies between $\widehat{\alpha}$ and α_0 which implies that $\alpha^* \in N_\delta(\alpha_0)$ for large n. Hence $n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha^*, T_i, A_i)$ is bounded in probability. Therefore, $R_G \rightarrow_p 0$ since the first two components of R_G are bounded in probability and the last component converges in probability to 0. Hence

$$n^{\frac{1}{2}}G_n(\widehat{\alpha}) = n^{\frac{1}{2}}G_n(\alpha_0) + n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)G_n'(\alpha_0) + o_p(1). \tag{5.18}$$

To determine the asymptotic distribution of $n^{\frac{1}{2}}G_n(\widehat{\alpha})$, clearly it suffices to find the asymptotic distribution of $n^{\frac{1}{2}}G_n(\alpha_0) + n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)G_n'(\alpha_0)$. The next step, therefore, is to determine the asymptotic joint distribution of $(n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0), n^{\frac{1}{2}}G_n(\alpha_0))$ so that the asymptotic distribution of the sum of the two can be determined. Property P6 states that conditional on $\{(T_i, A_i), \quad i = 1, \cdots, n\}$, $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \rightarrow_d N\left(0, \sum\!\!\!\!\!\!/_{\alpha}\right)$. Since the conditional asymptotic distribution of $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)$ does not depend on the value of (T, A), the asymptotic distribution of $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)$ conditional on any function of (T, A) will also converge in distribution to $N\left(0, \sum\!\!\!\!\!\!/_{\alpha}\right)$. In particular, conditional on $G_n(\alpha_0)$, $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \rightarrow_d N\left(0, \sum\!\!\!\!\!\!/_{\alpha}\right)$. Next we need to determine the asymptotic distribution of $n^{\frac{1}{2}}G_n(\alpha_0)$ where $G_n(\alpha_0) \equiv n^{-1}\sum_{i=1}^{n}g(\alpha_0, T_i, A_i)$.

The central limit theorem can be used to determine the asymptotic distribution of $n^{\frac{1}{2}}G_n(\alpha_0)$ because it is the sum of iid terms and each term is bounded. Applying the CLT yields

$$n^{\frac{1}{2}}(G_n(\alpha_0) - E(g(\alpha_0, T_i, A_i))) \rightarrow_d N(0, Var(g(\alpha_0, T_i, A_i))). \tag{5.19}$$

Since both the conditional distribution of $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)$ and marginal distribution of $n^{\frac{1}{2}}G_n(\alpha_0)$ converge to normal distributions, the joint distribution, $(n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0), n^{\frac{1}{2}}G_n(\alpha_0))$, converges to the product of the asymptotic distributions which is the joint distribution of independent normals with mean $(0, E(g(\alpha_0, T_i, A_i)))$, variance $(\not{\sum}_{\alpha}, Var(g(\alpha_0, T_i, A_i)))$, and covariance equal to 0. That is,

$$\begin{pmatrix} n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \\ n^{\frac{1}{2}}G_n(\alpha_0) \end{pmatrix} \rightarrow_d N\left(\begin{bmatrix} 0 \\ E(g(\alpha_0, T_i, A_i)) \end{bmatrix}, \begin{bmatrix} \not{\sum}_{\alpha} & 0 \\ 0 & Var(g(\alpha_0, T_i, A_i)) \end{bmatrix}\right). \tag{5.20}$$

Now that we have determined the asymptotic joint distribution of $(n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0), n^{\frac{1}{2}}G_n(\alpha_0))$, the final component of (5.18) to consider is $G_n(\alpha_0)' \equiv n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\alpha_0, T_i, A_i)$. By observing that $G_n(\alpha_0)'$ is the sum of iid terms,

$$n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\alpha_0, T_i, A_i) \rightarrow_p E\left(\frac{\partial}{\partial\alpha}g(\alpha_0, T_i, A_i)\right) \tag{5.21}$$

is given by the WLLN and Slutsky's theorem.

Using the multivariate form of Slutsky's theorem along with the results in (5.20) and (5.21) we get that $n^{\frac{1}{2}}(G_n'(\alpha_0))^T(\widehat{\alpha} - \alpha_0) + n^{\frac{1}{2}}G_n(\alpha_0)$ is equal to

$$\begin{pmatrix} G_n'(\alpha_0) \\ 1 \end{pmatrix}^T \begin{pmatrix} n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \\ n^{\frac{1}{2}}G_n(\alpha_0) \end{pmatrix}$$

which converges in distribution to

$$N\left(E(g(\alpha_0, T_i, A_i)), \left[E\left(\frac{\partial}{\partial\alpha}g(\alpha_0, T_i, A_i)\right)^T \not{\sum}_{\alpha} E\left(\frac{\partial}{\partial\alpha}g(\alpha_0, T_i, A_i)\right) + Var(g(\alpha_0, T_i, A_i))\right]\right)$$

where

$$\begin{aligned} E\{g(\alpha_0, T_i, A_i)\} &\equiv E\{P(D_i = 1|T_i, A_i)\} \\ &= E\{E\{I[D_i = 1]|T_i, A_i\}\} \text{ by MAR (A1)} \\ &= E\{I[D_i = 1]\} \\ &= P(D = 1) \\ &\equiv \theta_0. \end{aligned}$$

Hence,

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \rightarrow_d N\left(0, E\left(\frac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right)^T \not{\sum}_{\alpha} E\left(\frac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right) + Var(g(\alpha_0, T_i, A_i))\right)$$

and the proof is complete.

5.7.4 Equality of Asymptotic Distributions

In Section 5.3.6 asymptotic distribution theory obtained using an estimating function approach was applied to $\widehat{\theta}$. We would expect the resulting asymptotic distribution for the BG estimator to be the same as that derived in Section 5.7 using an alternative approach. We now show that indeed the distributions are the same.

Both approaches yield that $n^{\frac{1}{2}}(\widehat{\theta} - \theta_0)$ converges in distribution to a mean 0 normally distributed random variable. All that remains to be shown is that the asymptotic variances are the same. That is, we must show for BG (5.8) equals the asymptotic variance expression given in Theorem 5.4, namely

$$E\left(\frac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right)^T \not{\sum}_{\alpha} E\left(\frac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right) + Var(g(\alpha_0, T_i, A_i)) \qquad (5.22)$$

where $\not{\sum}_{\alpha} = Cov(\alpha_o)$.

It can be shown that (5.8) is equal to

$$\left[E\left(\tfrac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right)\right]^T Cov(\alpha_o) \left[E\left(\tfrac{\partial}{\partial\alpha} g(\alpha_0, T_i, A_i)\right)\right]^T + E(U_i^{\beta}(\theta_0, \alpha_0) U_i^{\beta}(\theta_0, \alpha_0)^T) \qquad (5.23)$$

$$-2E(U_i^{\beta}(\theta_0, \alpha_0) U_i^{\alpha}(\alpha_0)^T) \left[E\left(\tfrac{\partial}{\partial\alpha} U_i^{\alpha}(\alpha_0)\right)\right]^{-1} E\left(\tfrac{\partial}{\partial\alpha} U_i^{\beta}(\theta_0, \alpha_0)\right)^T \qquad (5.24)$$

where $U_i^{\theta}(\theta_0, \alpha_0)$ is the estimating function corresponding to the prevalence estimator and $U_i^{\alpha}(\alpha_0)$ is the vector of estimating functions corresponding to the estimation of $P(D|T, A)$. Therefore, to show equality of the variance expressions, we must show (5.24) is equal to (5.22).

Comparing the expressions we observe that they would be identical if

$$E(U_i^{\theta}(\theta_0,\alpha_0)U_i^{\theta}(\theta_0,\alpha_0)^T) - 2E(U_i^{\theta}(\theta_0,\alpha_0)U_i^{\alpha}(\alpha_0)^T)\left[E\left(\frac{\partial}{\partial\alpha}U_i^{\alpha}(\alpha_0)\right)\right]^{-1} E\left(\frac{\partial}{\partial\alpha}U_i^{\theta}(\theta_0,\alpha_0)\right)^T$$

was equal to $Var(g(\alpha_0, T_i, A_i))$.

First, we consider $E(U_i^{\theta}(\theta_0, \alpha_0)U_i^{\theta}(\theta_0, \alpha_0)^T)$. Since $U_i^{\theta}(\theta_0, \alpha_0) = g(\alpha_0, T_i, A_i) - \theta_0$ for BG,

$$\begin{aligned} E(U_i^{\theta}(\theta_0, \alpha_0)U_i^{\theta}(\theta_0, \alpha_0)^T) &= E\left((g(\alpha_0, T_i, A_i) - \theta_0)^2\right) \\ &= Var(g(\alpha_0, T_i, A_i)). \end{aligned}$$

Next, we consider $E(U_i^{\theta}(\theta_0, \alpha_0)U_i^{\alpha}(\alpha_0)^T)$.

$$\begin{aligned} E(U_i^{\theta}(\theta_0, \alpha_0)U_i^{\alpha}(\alpha_0)^T) &= E(E(U_i^{\theta}(\theta_0, \alpha_0)U_i^{\alpha}(\alpha_0)^T|T_i, A_i)) \\ &= E(U_i^{\theta}(\theta_0, \alpha_0)E(U_i^{\alpha}(\alpha_0)^T|T_i, A_i)) \\ &= 0. \end{aligned}$$

The last equality follows from the fact that $E(U_i^{\alpha}(\alpha_0)^T|T_i, A_i)) = 0$. Hence, we have shown that the variance expressions and consequently, the asymptotic distributions are the same for the two theoretical approaches.

5.7.5 Consistent variance estimator

Theorem 5.5 *(consistent variance estimator) A consistent estimator of σ_{θ}^2, the asymptotic variance for $n^{\frac{1}{2}}(\widehat{\theta} - \theta_0)$, is*

$$\widehat{\sigma_\theta^2} = \left[n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\widehat{\alpha},T_i,A_i)\right]^T \widehat{\not\Sigma}_\alpha \left[n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\widehat{\alpha},T_i,A_i)\right] + (n-1)^{-1}\sum_{i=1}^{n}\left(g(\widehat{\alpha},T_i,A_i) - n^{-1}\sum_{i=1}^{n}g(\widehat{\alpha},T_i,A_i)\right)^2$$

provided a consistent estimator of $\not\Sigma_\alpha$ *exists.*

5.7.6 Proof of Theorem 5.5 (consistent variance estimator)

We aim to show that each component of $\widehat{\sigma_\theta^2}$ is consistent for the corresponding component of σ_θ^2 because then Slutsky's theorem can be used to complete the proof. A first-order Taylor-series expansion of $n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\widehat{\alpha},T_i,A_i)$ about α_0 yields

$$n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\widehat{\alpha},T_i,A_i) = n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\alpha_0,T_i,A_i) + (\widehat{\alpha}-\alpha_0)n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha^*,T_i,A_i) \quad (5.25)$$

where α^* is between $\widehat{\alpha}$ and α_0. By P6, $\widehat{\alpha}$ is consistent for α_0 so $(\widehat{\alpha}-\alpha_0) \to_p 0$. P5 and Lemma 5.1 imply $n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\alpha\partial\alpha^T}g(\alpha^*,T_i,A_i)$ is bounded in probability. Therefore, the second term on the RHS of (5.25) converges in probability to 0. Furthermore, $n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\alpha_0,T_i,A_i)$ converges in probability to $E\left(\frac{\partial}{\partial\alpha}g(\alpha_0,T_i,A_i)\right)$ by the WLLN so $n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}g(\widehat{\alpha},T_i,A_i)$ also converges in probability to $E\left(\frac{\partial}{\partial\alpha}g(\alpha_0,T_i,A_i)\right)$.

The final component to consider is a sample variance estimator for $g(\alpha_0,T_i,A_i)$,

$$(n-1)^{-1}\sum_{i=1}^{n}\left(g(\widehat{\alpha},T_i,A_i) - n^{-1}\sum_{i=1}^{n}g(\widehat{\alpha},T_i,A_i)\right)^2.$$

Expanding this out we get

$$(n-1)^{-1}\sum_{i=1}^{n}g(\widehat{\alpha},T_i,A_i)^2 - \frac{n}{n-1}(n^{-1}\sum_{i=1}^{n}g(\widehat{\alpha},T_i,A_i))^2.$$

First-order Taylor series expansions (similar to the one above) along with the consistency (P6) of $\widehat{\alpha}$ and boundedness (P5) of functions of $g(\alpha,T_i,A_i)$ yield

$$(n-1)^{-1}\sum_{i=1}^{n}g(\widehat{\alpha},T_i,A_i)^2 \to_p E(g(\alpha_0,T_i,A_i)^2) \quad (5.26)$$

and

$$n^{-1}\sum_{i=1}^{n} g(\widehat{\alpha}, T_i, A_i) \rightarrow_p E(g(\alpha_0, T_i, A_i)). \tag{5.27}$$

Applying the continuous mapping theorem to (5.27) implies

$$\left(n^{-1}\sum_{i=1}^{n} g(\widehat{\alpha}, T_i, A_i)\right)^2 \rightarrow_p [E(g(\alpha_0, T_i, A_i))]^2. \tag{5.28}$$

Thus, by Slutsky's theorem, (5.26) and (5.28) imply

$$(n-1)^{-1}\sum_{i=1}^{n}\left(g(\widehat{\alpha}, T_i, A_i) - n^{-1}\sum_{i=1}^{n} g(\widehat{\alpha}, T_i, A_i)\right)^2 \rightarrow_p Var(g(\alpha_0, T_i, A_i).$$

Slutsky's theorem can then be used to combine the convergence results provided above to prove $\widehat{\sigma}_\theta^2 \rightarrow_p \sigma_\theta^2$ provided $\widehat{\not\Sigma}_\alpha \rightarrow_p \not\Sigma_\alpha$.

5.8 *Asymptotic Distribution Theory for Estimators of TP and FP*

$\widehat{TP}(c)$ and $\widehat{FP}(c)$ are ratios of prevalence estimators. The estimating equation framework developed in this chapter can be extended to develop asymptotic distribution theory for $\widehat{TP}(c)$ and $\widehat{FP}(c)$ by ascertaining the joint asymptotic distribution theory for prevalences and applying the delta method.

Let $\beta = (\theta_1, \theta_2, \theta_3, \alpha)$ where $\theta_1 = P(D = 1)$, $\theta_2 = P(T \geq c \text{ and } D = 1)$, $\theta_3 = P(T \geq c \text{ and } D = 0)$, and α is a vector of nuisance parameters corresponding to the estimation of $P(D|T, A)$ and/or $P(V|T, A)$. $\widehat{\beta}$ can be obtained by simultaneously considering the following vector of estimating functions

$$U(\beta) = \begin{pmatrix} U^{\theta_1}(\theta_1, \alpha) \\ U^{\theta_2}(\theta_2, \alpha) \\ U^{\theta_3}(\theta_3, \alpha) \\ U^{\alpha}(\alpha) \end{pmatrix}. \tag{5.29}$$

Theorem 5.2 can be applied to (5.29) to yield $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) \rightarrow_d N\left(0, \not{\sum}\right)$ where

$$\not{\sum} = \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1} Cov(U_i(\beta_0)) \left[-E\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right]^{-1}.$$

Recall that $TP(c) = \theta_2/\theta_1$ and $FP(c) = \theta_3/(1-\theta_1)$. Therefore, the multivariate delta method can then be used to obtain the asymptotic joint distribution of $(\widehat{TP}(c), \widehat{FP}(c))$. Specifically, denote

$$h(\beta) = \begin{pmatrix} h_1(\beta) \\ h_2(\beta) \end{pmatrix} = \begin{pmatrix} \theta_2/\theta_1 \\ \theta_3/(1-\theta_1) \end{pmatrix}.$$

Then the multivariate delta method implies $n^{\frac{1}{2}}(h(\widehat{\beta}) - h(\beta_0)) \rightarrow_d N\left(0, \dot{h}(\beta_0)\not{\sum}\dot{h}(\beta_0)^T\right)$ where

$$\begin{aligned} \dot{h} &= \begin{pmatrix} \frac{\partial}{\partial\beta^T}h_1 \\ \frac{\partial}{\partial\beta^T}h_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial}{\partial\theta_1}h_1 & \frac{\partial}{\partial\theta_2}h_1 & \frac{\partial}{\partial\theta_3}h_1 & \frac{\partial}{\partial\alpha}h_1 \\ \frac{\partial}{\partial\theta_1}h_2 & \frac{\partial}{\partial\theta_2}h_2 & \frac{\partial}{\partial\theta_3}h_2 & \frac{\partial}{\partial\alpha}h_2 \end{pmatrix} \\ &= \begin{pmatrix} -\theta_2\theta_1^{-2} & \theta_1^{-1} & 0 & 0 \\ \theta_3(1-\theta_1)^{-2} & 0 & (1-\theta_1)^{-1} & 0 \end{pmatrix}. \end{aligned}$$

That is,

$$n^{\frac{1}{2}}\left(\begin{pmatrix} \widehat{TP}(c) \\ \widehat{FP}(c) \end{pmatrix} - \begin{pmatrix} TP(c) \\ FP(c) \end{pmatrix}\right) \rightarrow_d N\left(0, \dot{h}(\beta_0)\not{\sum}\dot{h}(\beta_0)^T\right). \tag{5.30}$$

A consistent estimator of $\dot{h}(\beta_0)\not{\sum}\dot{h}(\beta_0)^T$ can be obtained by substituting $\widehat{\theta}$ for θ and replacing population averages by the corresponding sample averages.

Using (5.30) we can calculate pointwise confidence intervals for $\widehat{TP}(c)$ and $\widehat{FP}(c)$ at the cutpoint c. In addition to pointwise confidence intervals, we can calculate joint confidence regions for $(\widehat{TP}(c), \widehat{FP}(c))$.

5.9 Dependent Data

In this chapter we have considered independent data. In practice, however, data may be dependent. For example, in neonatal hearing screening each infant contributes data for each of two ears. Multiple measurements on the same infant will presumably be more similar than measurements on a different infant.

The theoretical results developed in this chapter can be easily extended to dependent data. Consider the setting where $(D_{ij}, T_{ij}, A_{ij}, V_{ij})$ are the available data for the jth observation on the ith subject. Since data for two different subjects are independent, the sum of contributions to the estimating function for each subject $U_i(\beta) = \sum_{j=1}^{m_i} U_{ij}(\beta)$ are iid, where m_i is the number of observations for subject i. As long as m_i, which may vary across subjects, is independent of the data, $U_i(\beta)$ will be iid.

If $U_i(\beta) = \sum_{j=1}^{m_i} U_{ij}(\beta)$ and $U_n = \sum_{i=1}^{n} U_i(\beta)$ are substituted for the corresponding expressions in Theorems 5.1, 5.2, and 5.3, then consistency, asymptotic distribution theory, and consistent variance estimation follow.

5.10 Summary

In this chapter we developed asymptotic distribution theory for disease prevalence estimators using an estimating equation framework that was shown to hold for all estimators we considered. By simultaneously considering estimating functions corresponding to the estimation of the nuisance parameters and the parameter of interest, we were able to account for the uncertainty in estimating nuisance parameters when estimating disease prevalence. An alternative derivation of asymptotic distribution theory for the Begg and Greenes estimator of disease prevalence yielded the same results as that obtained from the estimating function approach.

We then extended the asymptotic distribution theory for disease prevalence estimators to obtain asymptotic distribution theory for estimators of TP(c) and FP(c).

Since we obtained the joint asymptotic distribution of TP(c) and FP(c) estimators, we can calculate joint confidence regions for estimators of TP(c) and FP(c). In addition to confidence intervals for $\widehat{TP}(c)$ and $\widehat{FP}(c)$, we would like to be able to calculate confidence bands for the entire ROC curve and confidence intervals for estimates of the area under the ROC curve. This is an area of future research and will be discussed again in Chapter 8.

In the next chapter we use simulation studies to investigate the performance of the different estimators.

Chapter 6

SIMULATION RESULTS

In this chapter we use simulation studies to evaluate and compare the different methods proposed in the previous chapter.

6.1 Goals

Specifically we are interested in answering the following questions:

1. Is there small sample bias in estimating disease prevalence, TP, FP, or AUC?
2. How efficient are the methods for estimating disease prevalence, TP, FP, and AUC relative to each other in small samples?
3. How efficient are the methods for estimating disease prevalence, TP, and FP relative to each other in large samples?
4. Does asymptotic relative efficiency translate in small samples?
5. Can the variance estimator based on large sample theory be used in small samples for disease prevalence, TP, and FP?
6. Does incorporating auxiliary data improve efficiency?
7. How robust are the methods to model misspecification?

6.2 Simulation Logistics

All data used in the following simulations were created in *Splus*. The seed for the data sets were generated with the *Splus* function **set.seed**. This function generates a unique seed for each data set and uses a single uniform random number generator adapted from Marsaglia (1973). See Kennedy & Gentle (1980) for further information. By recording the seed for each simulation, we can easily reproduce the data.

6.3 Simulation Set-up

Often a disease can be thought of as arising from an underlying continuous disease process which remains subclinical until it reaches some threshold, at which point the disease becomes apparent. Therefore, in this simulation study disease status, D, was generated as a dichotomous variable indicating whether a random variable $Z \sim N(0,1)$ was greater than some threshold h. That is, $D = I[Z > h]$ where the threshold h determines the prevalence of disease.

Frequently there are multiple components to a disease process and diagnostic or screening tests measure different aspects of these components. Thus, we consider Z to be the sum of components Z_1 and Z_2 where each component is distributed $N(0, 0.5)$. Furthermore, continuous test results, T, were constructed as a linear combination of Z_1 and Z_2 plus random normal error, ϵ_1. In particular,

$$T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1, \qquad \epsilon_1 \sim N(0, 0.25).$$

By varying α_1 and β_1 we can consider T that measure different aspects of the disease. Thus the inherent accuracy of T can be altered by changing the values of α_1 and β_1. Table 6.1 summarizes the AUC corresponding to tests generated under various combinations of α_1 and β_1 between 0 and 1. By increasing α_1 and β_1, T becomes a more accurate measure of the true disease state. As expected, $\alpha_1 = \beta_1 = 1$ corresponds to a near perfect test while an $\alpha_1 = \beta_1 = 0$ represents a worthless test.

Table 6.1: Empirical AUC, percentages of the diseased ($D+$) and non-diseased ($D-$) subjects verified when verification depends on the value of the $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Different values of α_1 and β_1 are considered.

α_1	β_1	AUC	% D+ verified	% D- verified
1	1	0.960	92	30
1	0.5	0.919	83	31
0.5	0.5	0.876	76	32
1	0	0.810	66	33
0.5	0	0.721	56	34
0	0	0.500	36	36

Auxiliary data, A, were generated in a similar fashion to T. Specifically,

$$A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2, \qquad \epsilon_2 \sim N(0, 0.25)$$

where ϵ_1 and ϵ_2 are independent. Analogous to T, by varying α_2 and β_2 we can consider auxiliary data which are anywhere from highly informative ($\alpha_2 = \beta_2 = 1$) to non-informative ($\alpha_2 = \beta_2 = 0$). Moreover, α_1, α_2, β_1, and β_2 determine the correlation between T and A.

Since D and T are generated and, thus, known for all subjects, we can calculate the full data prevalence, TP(c) and FP(c) for all cutpoints c, empirical ROC curve, and AUC. To induce verification bias, the binary variable indicating verification status, V, was generated under two different scenarios where verification depended on either the value of (i) the test being studied, T, or (ii) the auxiliary data, A. Sections 6.4 and 6.5 report the results of simulation studies that consider scenarios (i) and (ii), respectively.

The performance of the bias correction methods proposed in Chapter 4 was evaluated by assuming disease status are missing for those subjects with $V = 0$ and then

comparing the estimates to the full data estimates. A summary of the approaches considered in these simulation studies is provided in Table 6.2. BG, MS, and SP require a parametric model for the probability of disease. We chose to use a probit model because it can be shown for the Normal theory assumptions made previously a probit model is the true model. To investigate the effects of incorporating auxiliary data, models for the probability of being diseased were fit both with and without A. We add "$-A$" to methods that included A in the model for disease while "$-\bar{A}$" indicates that the auxiliary data were not included. For example, MS-$\bar{\text{A}}$ refers to the mean score approach with $P(D|T)$ used to estimate disease probability. We consider IPW and SP approaches that use either known or estimated weights corresponding to the verification probabilities. Estimators using these weights are denoted "$-K$" and "$-E$", respectively.

To reflect disease screening studies frequently encountered in practice, we simulate studies with 1000 subjects and a disease prevalence of 0.1. For each realization of data the methods discussed above were used to estimate disease prevalence, TP(c), and FP(c). Estimates of TP(c) and FP(c) were then used to construct an empirical ROC curve from which the AUC was calculated. This process was repeated for 1000 realizations.

6.4 Verification Depends on the Test Under Evaluation

Often results of a diagnostic or screening test are used to determine which subjects receive disease verification. To simulate this dependence of the verification mechanism on the test results, V was generated using a Bernoulli random variable with the probability of verification

$$P(V) = \begin{cases} 1 & T > t^q \\ \delta & \text{else} \end{cases}$$

where t^q is the qth quantile of the distribution of T. That is, all subjects with a test

Table 6.2: Summary of estimators considered in the simulation study. Probit models were used to fit $P(D|T,A)$ and $P(D|T)$.

Estimator	Disease Probability Model	Verification Probabilities
CC	-	-
BG-A	P(D|T,A)	-
BG-$\bar{A}$	P(D|T)	-
MS-A	P(D|T,A)	-
MS-$\bar{A}$	P(D|T)	-
IPW-K	-	Known
IPW-E	-	Estimated
SP-K-A	P(D|T,A)	Known
SP-K-$\bar{A}$	P(D|T)	Known
SP-E-A	P(D|T,A)	Estimated
SP-E-$\bar{A}$	P(D|T)	Estimated

result greater than the threshold t^q were verified while only a random fraction below the threshold received disease verification. In practice, t^q can be based on the results of pilot studies or previously published results. Results presented here consider q and δ equal to 0.8 and 0.2, respectively. These values result in an average of 36% of the subjects receiving disease verification, with 20% having a test result above the threshold and 16% below the threshold. In Section 6.6 we discuss the impact of increasing δ.

In this scenario, empirical estimates of the verification probabilities are 1.0 for those subjects with $T > t^{0.8}$ and the *observed* fraction of subjects receiving verification below the threshold,

$$\frac{\sum_{i=1}^{n} V_i I[T_i \leq t^{0.8}]}{\sum_{i=1}^{n} I[T_i \leq t^{0.8}]},$$

for those subjects with $T \leq t^{0.8}$. This empirical estimate is equivalent to using the predicted probability of verification resulting from a logistic regression with V as the response and $I[T > t^{0.8}]$ as the predictor.

6.4.1 Small Sample Bias

The magnitude of verification bias depends on the inherent accuracy of the test and the verification mechanism (Begg & Greenes, 1983). Therefore, we can assess how well the methods correct for varying amounts of verification bias by varying the parameters that determine T, α_1 and β_1. Table 6.1 suggests that the more accurate the test, the larger the discrepancy in the percentage of diseased and non-diseased subjects verified, and hence, the greater the potential for verification bias. Verification bias does not occur when $\alpha = \beta = 0$ since the same percentage of diseased and non-diseased subjects are verified, i.e. the verification sample is a simple random sample.

Small sample bias results are presented for scenarios in which the inherent accuracy of T is varied from a highly accurate test ($\alpha_1 = \beta_1 = 1$) with great potential for verification bias to a test ($\alpha_1 = \beta_1 = 0$) that is no better than chance alone with

little potential for verification bias. The efficiency gain resulting from incorporating auxiliary data is discussed later so in these simulations A is arbitrarily fixed to be $Z_1 + Z_2 + \epsilon_2$.

Some might argue that a study with 1000 subjects is not a "small sample." However, the effective sample size for the IPW-K disease prevalence estimator is only 360 after taking into account that only 36% of the subjects receive disease verification in this simulation study. The effective sample size is even smaller for the IPW-K estimator of TP(c) and FP(c) which, in addition to only using data on verified subjects, only includes data for subjects with test results greater than or equal to the specified cutpoint. Moreover, in practice screening studies typically contain a sample size substantially greater than 1000. For example, in Chapter 7 we analyze a subset of a neonatal audiology screening study which enrolled over 5000 infants.

Table 6.3 presents mean estimates of disease prevalence across 1000 realizations. Results from the estimation methods are given in the rows while results for the different T considered are provided in the columns. The full data disease prevalence is 0.1 so the amount of bias caused by the dependence of verification on T can be quantified by comparing the estimated value to 0.1.

Clearly all estimators that were shown in Chapter 5 to be asymptotically unbiased yield unbiased estimates of prevalence in the small samples. CC does not adjust for the biased sampling and thus, as expected, yields biased estimates of prevalence in all scenarios with a potential for verification bias ($\alpha_1 > 0$, $\beta_1 > 0$). CC over-estimates prevalence by nearly 0.160, more than twice the full data prevalence, in the most extreme setting. The amount of bias decreases as the accuracy of T and, hence, the discrepancy between the percentage of diseased and non-diseased subjects verified decreases. In the setting when α_1 and β_1 both equal 0, i.e. a simple random sample of subjects is verified, CC has no bias.

To investigate small sample bias in the estimation of operating points on the ROC curve, we considered estimating $TP(c)$ and $FP(c)$ for cutpoints corresponding to full

Table 6.3: Mean disease prevalence of 1000 realizations when verification depends on T. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Biased estimates are in bold face.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
Full Data	0.100	0.100	0.100	0.100
CC	**0.257**	**0.210**	**0.184**	0.100
BG-A	0.100	0.100	0.100	0.099
BG-$\bar{A}$	0.100	0.100	0.100	0.100
MS-A	0.100	0.100	0.100	0.100
MS-$\bar{A}$	0.100	0.100	0.100	0.100
IPW-K	0.100	0.100	0.100	0.100
IPW-E	0.100	0.100	0.100	0.100
SP-K-A	0.100	0.099	0.100	0.100
SP-K-$\bar{A}$	0.100	0.099	0.100	0.100
SP-E-A	0.100	0.100	0.100	0.100
SP-E-$\bar{A}$	0.100	0.099	0.100	0.100

Table 6.4: Mean estimated $TP(c)$, $FP(c)$ where c is such that the full data $FP(c)$=0.20 in 1000 realizations when verification depends on T. A is fixed to be $Z_1+Z_2+\epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1+\beta_1 Z_2+\epsilon_1$. Biased estimates are in bold face.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
Full Data	0.965, 0.200	0.778, 0.200	0.644, 0.200	0.198, 0.200
CC	**0.993, 0.461**	**0.943, 0.493**	**0.893, 0.509**	**0.551, 0.556**
BG-A	0.964, 0.200	0.781, 0.200	0.649 0.200	0.202, 0.200
BG-$\bar{A}$	0.964, 0.200	0.783, 0.200	0.652, 0.200	0.202, 0.200
MS-A	0.964, 0.200	0.781, 0.200	0.649, 0.200	0.201, 0.200
MS-$\bar{A}$	0.964, 0.200	0.784, 0.200	0.651, 0.200	0.203, 0.200
IPW-K	0.968, 0.200	0.787, 0.200	0.653, 0.200	0.204, 0.200
IPW-E	0.968, 0.200	0.787, 0.200	0.653, 0.200	0.204, 0.200
SP-K-A	0.967, 0.200	0.785, 0.200	0.650, 0.200	0.202, 0.200
SP-K-$\bar{A}$	0.968, 0.200	0.787, 0.200	0.653, 0.200	0.204, 0.200
SP-E-A	0.967, 0.200	0.785, 0.200	0.650, 0.200	0.202, 0.200
SP-E-$\bar{A}$	0.968, 0.200	0.787, 0.200	0.653, 0.200	0.204, 0.200

data $FP(c)$ equal to 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Again, very little small sample bias was observed for the bias-correction estimators. Conversely, CC clearly overestimates the $TP(c)$ and $FP(c)$. For example, the mean estimated $TP(c)$ and $FP(c)$ across the 1000 realizations are displayed in Table 6.4 for the scenario when c is fixed so that the full data $FP(c) = 0.2$.

Table 6.5 provides mean estimates of AUC across the 1000 realizations considered previously. Results for estimating AUC are similar to those observed for disease prevalence, TP, and FP. This is not surprising since AUC estimators are a func-

Table 6.5: Mean AUC of 1000 realizations when verification depends on T. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Biased estimates are in bold face.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
Full Data	0.960	0.876	0.810	0.499
CC	**0.913**	**0.826**	**0.774**	0.499
BG-A	0.960	0.876	0.812	0.501
BG-$\bar{A}$	0.960	0.877	0.813	0.502
MS-A	0.960	0.876	0.812	0.501
MS-$\bar{A}$	0.960	0.877	0.813	0.502
IPW-K	0.961	0.879	0.813	0.502
IPW-E	0.961	0.879	0.813	0.502
SP-K-A	0.961	0.878	0.812	0.501
SP-K-$\bar{A}$	0.961	0.879	0.813	0.502
SP-E-A	0.961	0.878	0.812	0.500
SP-E-$\bar{A}$	0.961	0.879	0.813	0.501

tion of $\widehat{TP}(c)$ and $\widehat{FP}(c)$, which are a function of the disease prevalence estimator. Specifically, only CC is biased in the three settings where there is the potential for verification bias and all methods are unbiased when a simple random sample of subjects are verified. As the amount of bias in the sampling decreases, the bias in the CC estimate of AUC decreases. The bias ranges from 0.036 to 0.049.

The bias in the CC approach is also apparent in Figure 6.1, which displays CC and full data ROC curves for one randomly chosen data realization when α_1 and β_1 equal 0.5. Clearly, CC underestimates the full data ROC curve. F1, F2, and F3 in Figure 6.1 correspond to the operating points on the full data ROC curve for FP

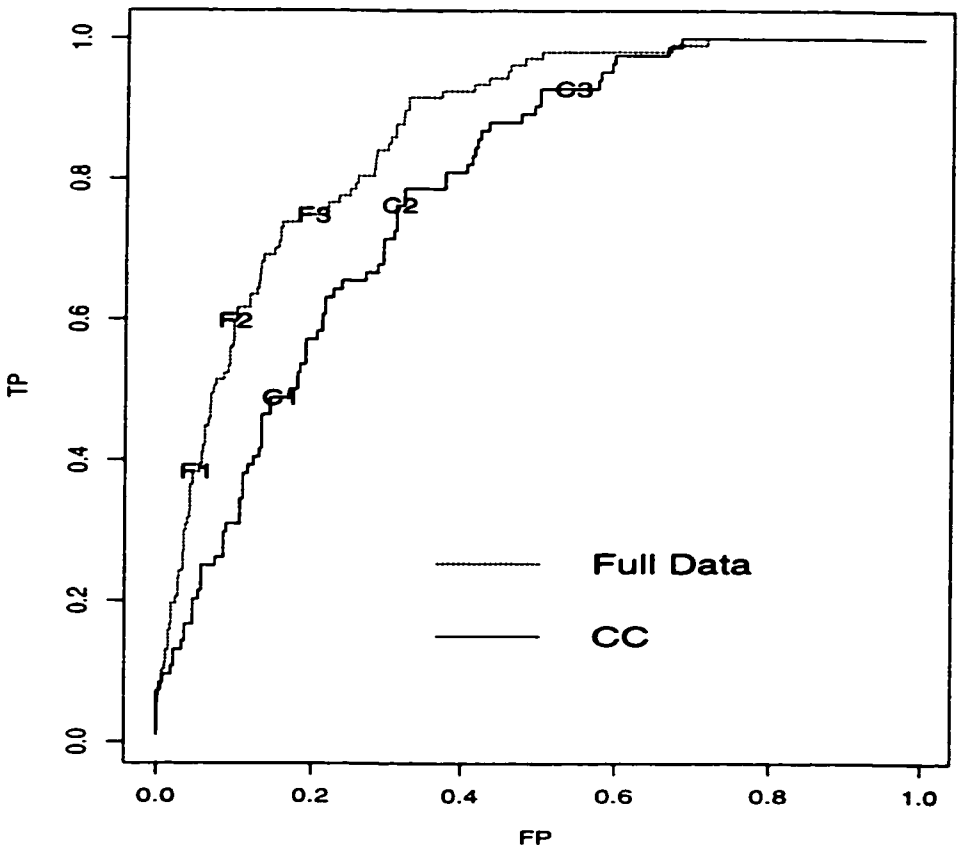

Figure 6.1: Full data and CC ROC curves from a randomly chosen realization of the simulation study when verification depends on the test results, $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$. $A = Z_1 + Z_2 + \epsilon_2$.

rates of 0.05, 0.10, and 0.20, respectively. The cutpoints used to generate F1-F3 were also used to identify the corresponding operating points (C1, C2, and C3) on the CC ROC curve. As noted previously, operating points on the CC ROC curve are biased upwards relative to the full data curve. In practice one must choose the cutpoint used to define a positive test so this bias can have substantial ramifications.

Since BG, MS, IPW, and SP yield unbiased estimates of the AUC in all scenarios,

it is not surprising that their ROC curves closely resemble the full data ROC curve (Figure 6.2). ROC curves for these methods when auxiliary data are not incorporated and/or verification probabilities are estimated are not presented because they are similar to the curves presented. Not surprisingly, methods in the same class of estimators yield similar ROC curves. That is, ROC curves for BG-A and MS-A are similar while curves for IPW-K and SP-K-A are similar.

Although IPW and SP yield unbiased AUC estimates across the 1000 realizations of the data, for any one realization their ROC curves tend not to follow the full data ROC curve as well in the upper part of the ROC curve. This is evident in Figures 6.2 (c) and (d). It makes sense that they perform better in the lower portion of the ROC curve because this part of the curve corresponds to large cutpoints for which a higher percentage of subjects are verified. The non-monotonicity of SP is also evident in Figure 6.2 (d).

6.4.2 Relative Efficiency

Next the efficiency of the different disease prevalence, TP(c), FP(c), and AUC estimators relative to each other is considered. Both small sample relative efficiency (SSRE) and asymptotic relative efficiency (ARE) are calculated. SSRE was calculated as the ratio of the simulation variances for two estimators where simulation variance is the variance of the estimated measure across the 1000 realizations. The asymptotic variance was estimated for disease prevalence, TP(c), and FP(c) by applying the variance estimator (Theorem 5.3 and Section 5.8) to a large sample (n=100000). ARE was then calculated by taking the ratio of these Monte Carlo variance estimates for two different estimators.

SSRE relative to BG-A for disease prevalence estimation is presented in Table 6.6. A SSRE value less than 1 implies the estimator is less efficient than BG-A. Estimators are listed in the rows while the accuracy of T and potential for verification bias decreases moving across the columns. Not only does CC yield biased estimates, but

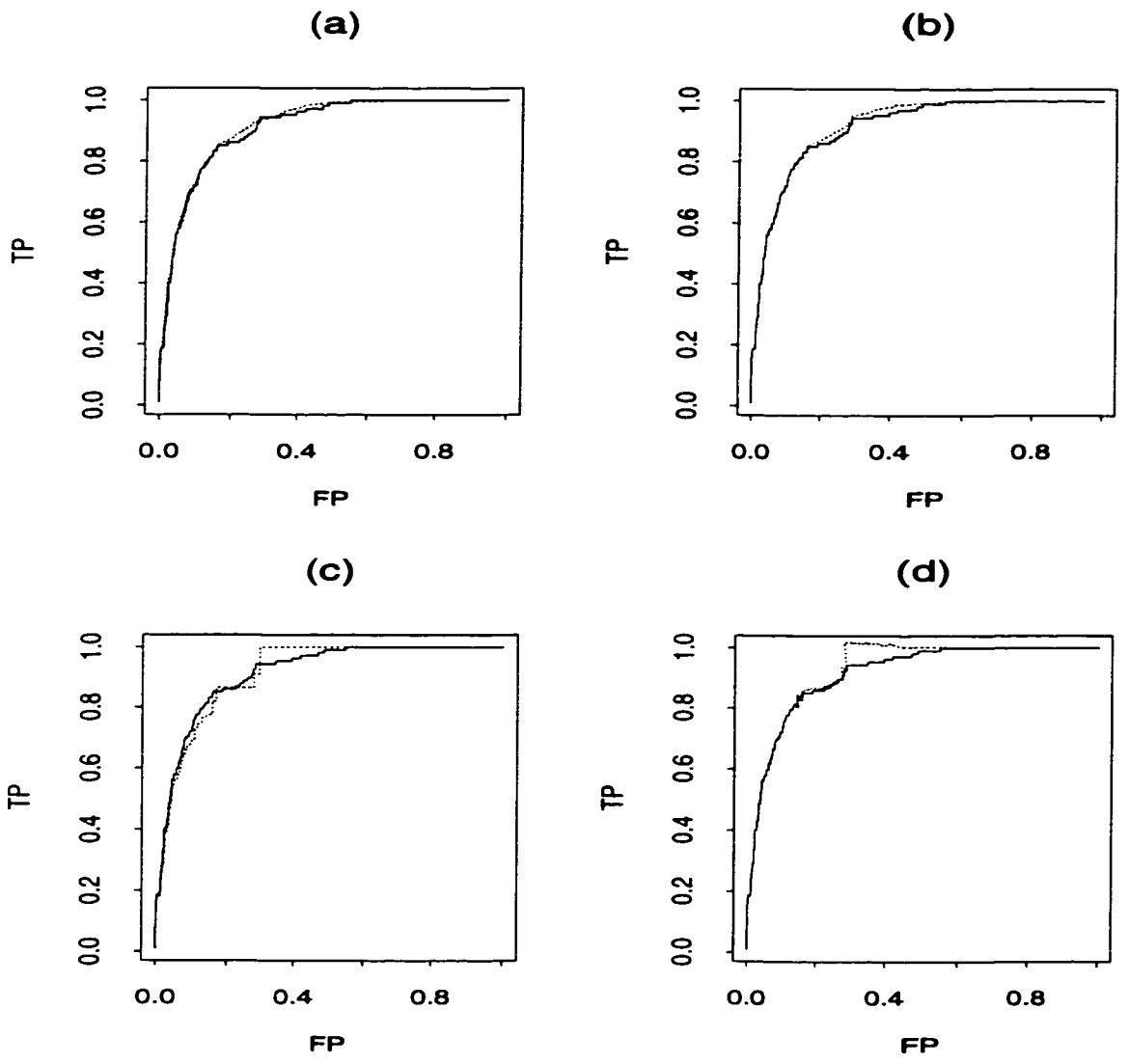

Figure 6.2: Full data (solid line) and (a) BG-A, (b) MS-A, (c) IPW-K, and (d) SP-K-A ROC curves (dashed lines) from a randomly chosen realization of the simulation study when verification depends on the test results, $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$. $A = Z_1 + Z_2 + \epsilon_2$.

CC is also substanially less efficient than the other estimators. MS-A and BG-A have similar efficiency and are the most efficient estimators. MS-$\bar{A}$ and BG-$\bar{A}$ also have similar efficiency. Relative to their counterparts which include the auxiliary data, MS-$\bar{A}$ and BG-$\bar{A}$ range from 3% to 37.5% less efficient. They are increasingly less efficient as the accuracy of T decreases since incorporating A adds more information about D when T is less accurate. Similar trends are observed for SP-K-$\bar{A}$ and SP-E-$\bar{A}$.

Comparing IPW-K and IPW-E we see that IPW with empirically estimated weights is up to 47% more efficient than using known weights. Conversely, using known or correctly estimated weights does not appear to make much of a difference in the efficiency of SP. These observations are consistent with previous literature (Pepe et al., 1994, Rotnitzky & Robins, 1995).

Table 6.6 also suggests that SP-K-A is roughly 9% to 13% less efficient than BG-A and MS-A. However, SP-K-A is 5.5% to 40% more efficient than IPW-K, becoming more efficient as the extent of the biased sampling decreases. Furthermore, a comparison of the efficiency in BG-$\bar{A}$ and MS-$\bar{A}$ to the IPW approaches suggests that even BG-$\bar{A}$ and MS-$\bar{A}$ that do not include auxiliary data are at least 6.8% more efficient than the IPW methods.

It is interesting to note that in the setting where there is no verification bias (i.e. $\alpha_1 = \beta_1 = 0$), using only complete cases is 24.0% more efficient than BG-$\bar{\text{A}}$ and MS-$\bar{\text{A}}$ and 30.8% more efficient than IPW-E.

ARE relative to BG-A is included in parentheses in Table 6.6. ARE is similar to SSRE suggesting that ARE appears to translate in small samples. The performance of the variance estimator is further investigated in Section 6.4.4.

SSRE relative to BG-A when estimating AUC is summarized in Table 6.7. Results are similar to those for disease prevalence with the exception that SSRE for IPW and SP are noticeably smaller. For example, SP-K-A is 23-38% less efficient than BG-A when estimating AUC and only 9-13% less efficient when estimating disease prevalence. This decreased efficiency in SP and IPW may be caused by poor estimation in

Table 6.6: Small sample efficiency relative to BG-A for estimating disease prevalence when verification depends on T. ARE relative to BG-A is provided in parentheses. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
CC	0.200 (0.196)	0.293 (0.299)	0.381 (0.378)	0.865 (0.840)
BG-$\bar{A}$	0.968 (0.955)	0.837 (0.867)	0.815 (0.788)	0.625 (0.623)
MS-A	0.993 (1.000)	0.993 (1.000)	0.996 (1.002)	1.000 (1.002)
MS-$\bar{A}$	0.963 (0.956)	0.834 (0.867)	0.813 (0.789)	0.625 (0.623)
IPW-K	0.666 (0.680)	0.594 (0.644)	0.642 (0.601)	0.558 (0.508)
IPW-E	0.813 (0.823)	0.642 (0.665)	0.671 (0.632)	0.557 (0.563)
SP-K-A	0.867 (0.879)	0.826 (0.839)	0.875 (0.874)	0.911 (0.919)
SP-K-$\bar{A}$	0.829 (0.848)	0.669 (0.690)	0.688 (0.651)	0.554 (0.559)
SP-E-A	0.867 (0.876)	0.825 (0.843)	0.873 (0.877)	0.909 (0.921)
SP-E-$\bar{A}$	0.828 (0.845)	0.669 (0.694)	0.688 (0.654)	0.553 (0.561)

the upper portion of the ROC curve.

Figures 6.3 and 6.4 present asymptotic and small sample efficiency results in the estimation of $TP(c)$ for various cutpoints. Qualitatively the results are similar to those observed for disease prevalence and AUC. Namely, BG-A and MS-A have similar efficiency and are the most efficient estimators followed by BG-$\bar{A}$ and MS-$\bar{A}$, SP-K-A, and IPW-E. We noted in the previous section that although asymptotically SP and IPW estimators of AUC are unbiased, SP and IPW estimated ROC curves tend not to follow the full data ROC curve as well for large FP rates. This increased variability for large FP rates is evident in Figures 6.3 and 6.4.

ARE relative to BG-A is summarized in Figure 6.5 for $FP(c)$. Surprisingly, MS-A, MS-$\bar{A}$, BG-$\bar{A}$, and SP-K-A are more efficient than BG-A for large FP rates. In contrast to the ARE, small sample efficiency results (Figure 6.6) imply similar efficiency for all estimators of $FP(c)$ except IPW.

6.4.3 Most Efficient Estimator of Disease Prevalence in Disease Restricted Class of Estimators

In the previous section we saw that BG and MS estimators of disease prevalence and AUC appear to have similar variability in small and large samples. BG and MS are just two of infinite number of possible estimators in the disease restricted class of estimators (Section 4.8.2) we could consider. To answer the question of whether there is one estimator in this class that is most efficient for all scenarios, we estimated the asymptotic variance of a variety of disease prevalence estimators in many scenarios. Figure 6.7 presents a plot of the Monte Carlo estimate of variance as k, which defines the estimator, is varied. We see that the k that minimizes the asymptotic variance is not the same for the four scenarios. This suggests that there does not appear to one estimator in this class that is the most efficient in all scenarios. Therefore, we focus our attention on the intuitive estimators in this class, namely BG ($k = 0$) and MS ($k = 1$), acknowledging that in some scenarios there may be an estimator in this

Table 6.7: Small sample efficiency relative to BG-A for estimating AUC when verification depends on T. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
CC	0.385	0.828	1.070	0.599
BG-$\bar{A}$	0.949	0.784	0.763	0.670
MS-A	0.994	0.994	0.993	1.000
MS-$\bar{A}$	0.942	0.782	0.757	0.671
IPW-K	0.553	0.522	0.538	0.464
IPW-E	0.560	0.520	0.535	0.462
SP-K-A	0.621	0.688	0.734	0.766
SP-K-$\bar{A}$	0.590	0.557	0.563	0.493
SP-E-A	0.622	0.688	0.732	0.759
SP-E-$\bar{A}$	0.593	0.556	0.554	0.457

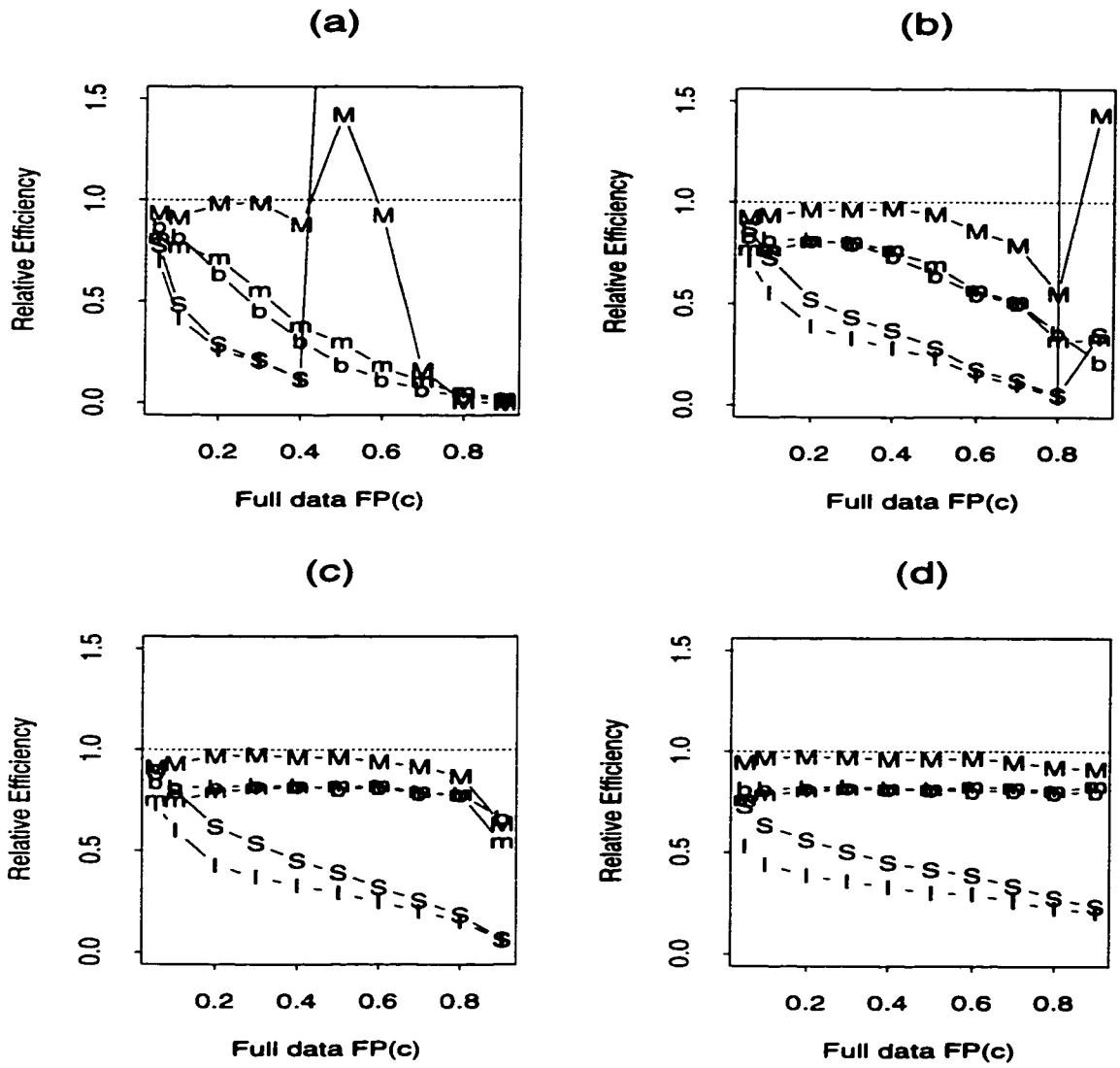

Figure 6.3: ARE relative to BG-A in estimating TP(c) as the cutpoint is varied so that the full data FP(c) $\in (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. M=MS-A, m=MS-$\bar{A}$, b=BG-$\bar{A}$, S=SP-K-A, I=IPW-E. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while verification is a function of (a) $T = Z_1 + Z_2 + \epsilon_1$, (b) $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$, (c) $T = Z_1 + \epsilon_1$, (d) $T = \epsilon_1$.

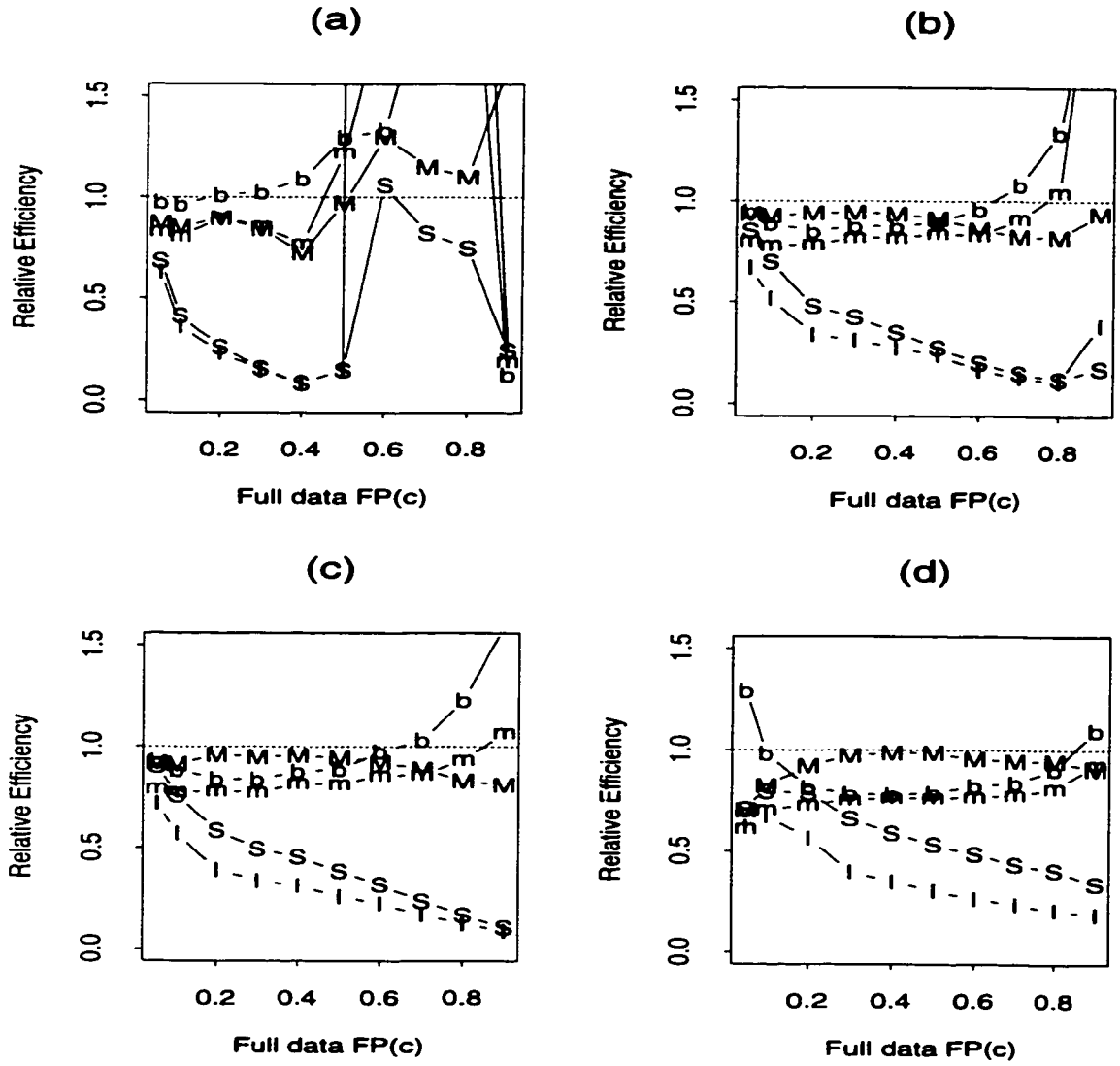

Figure 6.4: SSRE relative to BG-A in estimating TP(c) as the cutpoint is varied so that the full data FP(c) $\in (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. M=MS-A, m=MS-$\bar{A}$, b=BG-$\bar{A}$, S=SP-K-A, I=IPW-E. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while verification is a function of (a) $T = Z_1 + Z_2 + \epsilon_1$, (b) $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$, (c) $T = Z_1 + \epsilon_1$, (d) $T = \epsilon_1$.

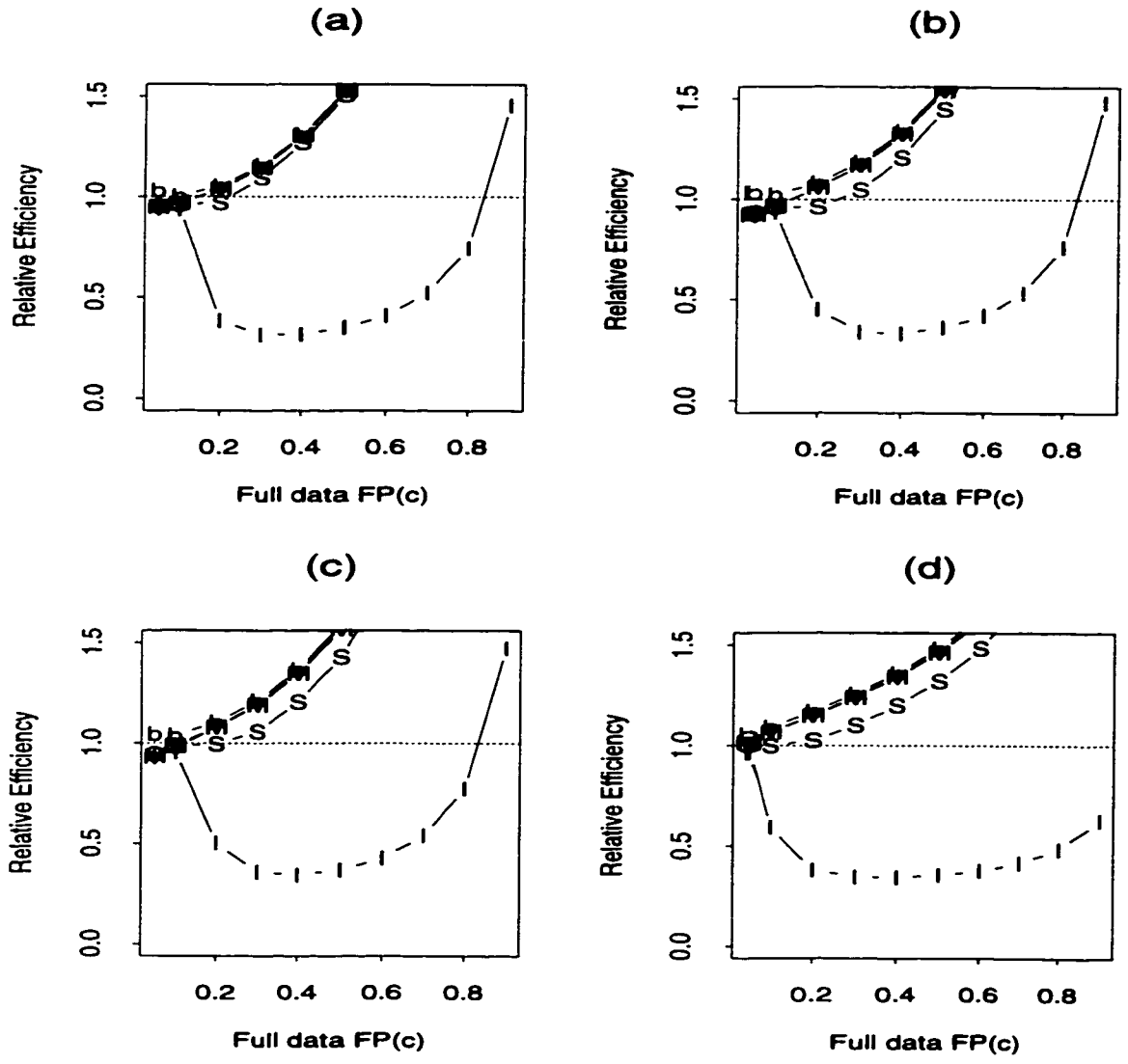

Figure 6.5: ARE relative to BG-A in estimating FP(c) as the cutpoint is varied so that the full data FP(c) $\in (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. M=MS-A, m=MS-$\bar{A}$, b=BG-$\bar{A}$, S=SP-K-A, I=IPW-E. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while verification is a function of (a) $T = Z_1 + Z_2 + \epsilon_1$, (b) $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$, (c) $T = Z_1 + \epsilon_1$, (d) $T = \epsilon_1$.

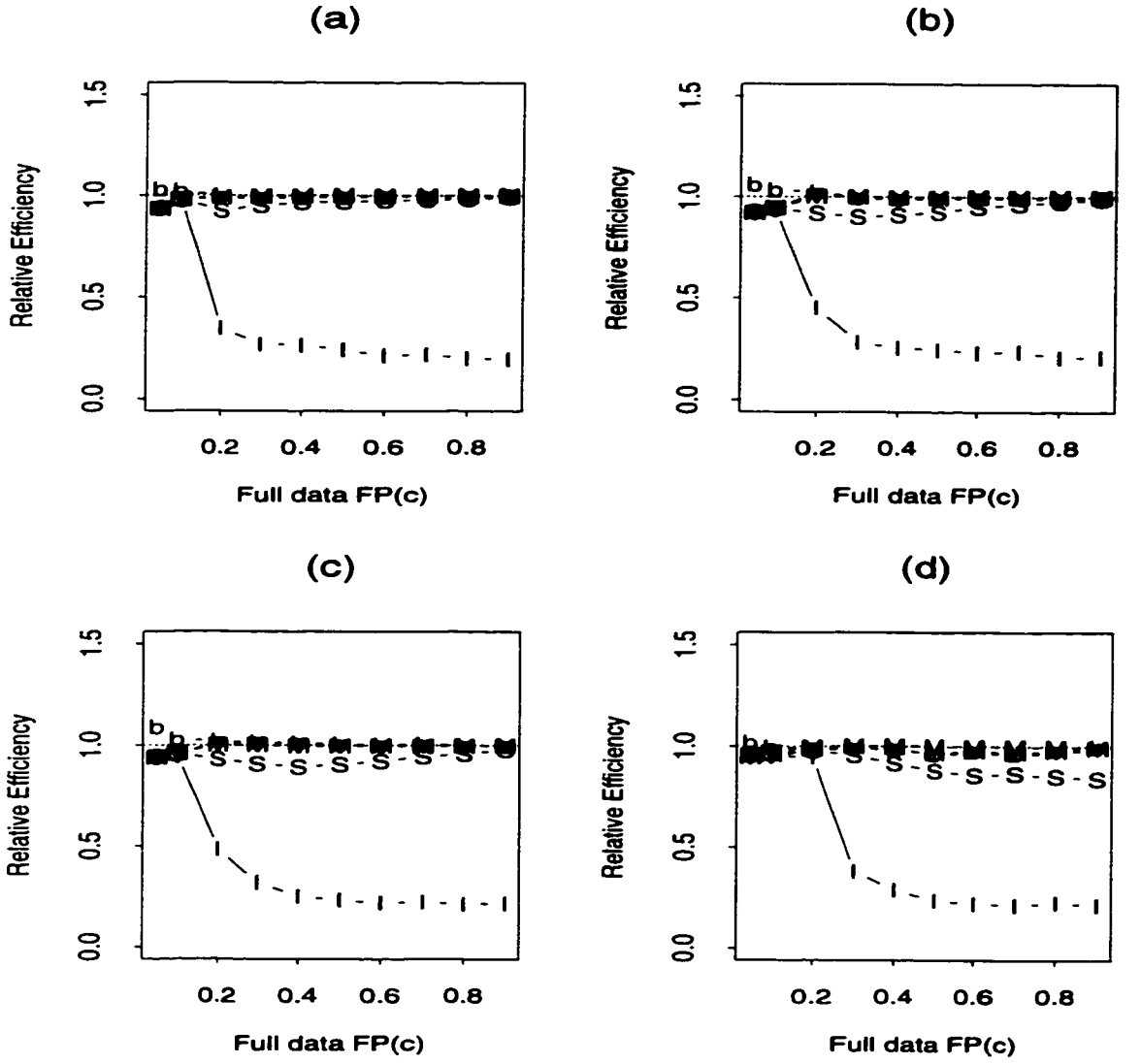

Figure 6.6: SSRE relative to BG-A in estimating FP(c) as the cutpoint is varied so that the full data FP(c) $\in (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. M=MS-A, m=MS-$\bar{A}$, b=BG-$\bar{A}$, S=SP-K-A, I=IPW-E. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while verification is a function of (a) $T = Z_1 + Z_2 + \epsilon_1$, (b) $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$, (c) $T = Z_1 + \epsilon_1$, (d) $T = \epsilon_1$.

class that is more efficient than BG and MS.

6.4.4 Performance of the Variance Estimator

In Section 6.4.2 we saw that asymptotic relative efficiency appeared to be preserved in small samples for disease prevalence and to a lesser extent $TP(c)$ and $FP(c)$. Next we further investigate if the variance estimators for disease prevalence (Theorem 5.3) and $TP(c)$ and $FP(c)$ (Section 5.8) can be used in small samples. The variance of the disease prevalence, $TP(c)$, and $FP(c)$ estimates calculated from the 1000 data realizations (i.e. simulation variance) provides an estimate of the true variability of the estimator. The performance of the variance estimator can be assessed by comparing the average estimate it yields to the this "true variance". In addition, the performance can be assessed by calculating coverage probabilities of the confidence interval corresponding to the variance estimator.

Table 6.8 provides the simulation variance and mean variance estimator for disease prevalence while Table 6.9 summarizes nominal 90% confidence interval (CI) coverage probabilities of disease prevalence for several scenarios. The average disease prevalence variance estimator tends to be slightly larger than the simulation variance. However, nominal 90% coverage probabilities of between 87.8 and 91.8 are achieved for all estimators except CC. CC does not account for biased sampling so it is not surprising that this estimator has poor coverage in the scenarios with potential for verification bias. Since the variance estimate and coverage look good, the disease prevalence variance estimator appears to perform well in small samples and, thus, can be used to generate interval estimates in practice.

The simulation variance and mean variance estimate for $TP(c)$ and $FP(c)$ when c is fixed so that the full data $FP(c)$ equals 0.2 are provided in Tables 6.10 and 6.11. Generally the simulation and mean variance estimate are similar. However, the variance estimator has poor coverage for IPW and SP estimators in several scenarios (Table 6.12). Therefore, we may need to consider using bootstrap techniques to obtain

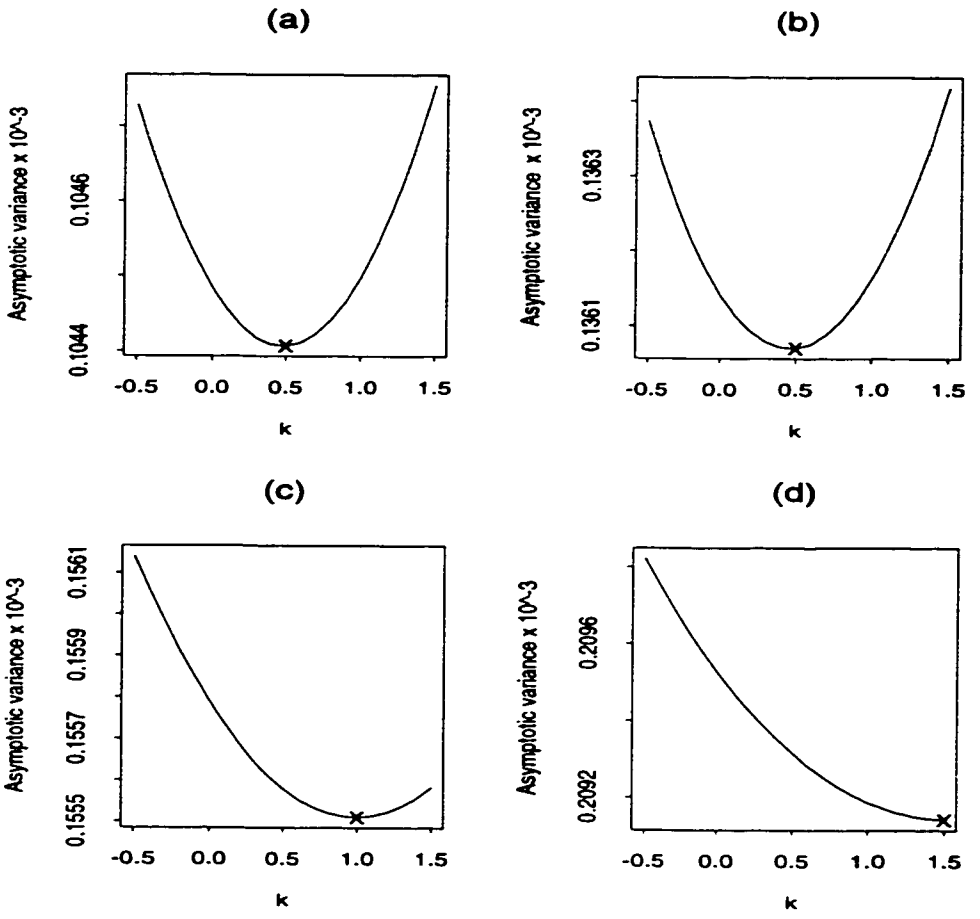

Figure 6.7: Asymptotic variance for a variety of disease prevalence estimators in the disease restricted class of estimators. A is fixed to be $Z_1 + Z_2 + \epsilon_2$. Verification is a function of (a) $T = Z_1 + Z_2 + \epsilon_1$, (b) $T = 0.5Z_1 + 0.5Z_2 + \epsilon_1$, (c) $T = Z_1 + \epsilon_1$, and (d) $T = \epsilon_1$.

Table 6.8: Simulation variance (mean variance estimate) $\times 10^{-4}$ of estimated disease prevalence when verification depends on T. Different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$ and A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	α_1, β_1			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
CC	4.94 (5.29)	4.35 (4.61)	4.07 (4.17)	2.33 (2.49)
BG-A	0.99 (1.07)	1.28 (1.43)	1.55 (1.62)	2.02 (2.24)
BG-$\bar{A}$	1.02 (1.11)	1.52 (1.69)	1.91 (2.04)	3.23 (3.41)
MS-A	0.99 (1.06)	1.29 (1.41)	1.56 (1.61)	2.02 (2.22)
MS-$\bar{A}$	1.02 (1.11)	1.53 (1.68)	1.91 (2.03)	3.23 (3.41)
IPW-K	1.48 (1.52)	2.15 (2.17)	2.42 (2.57)	3.62 (3.79)
IPW-E	1.21 (1.28)	1.99 (2.05)	2.32 (2.51)	3.63 (3.80)
SP-K-A	1.14 (1.18)	1.55 (1.58)	1.78 (1.79)	2.22 (2.28)
SP-K-$\bar{A}$	1.19 (1.24)	1.91 (1.98)	2.26 (2.43)	3.64 (3.81)
SP-E-A	1.14 (1.36)	1.55 (2.10)	1.78 (2.48)	2.22 (3.62)
SP-E-$\bar{A}$	1.19 (1.24)	1.91 (2.12)	2.26 (2.61)	3.65 (4.34)

Table 6.9: 90% confidence interval (CI) coverage probability of disease prevalence variance estimator when verification depends on T. Different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$ and A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	α_1, β_1			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
CC	0.0	0.0	0.3	90.5
BG-A	90.0	90.4	89.4	89.3
BG-$\bar{A}$	88.9	90.9	89.6	90.8
MS-A	89.3	90.1	89.4	89.4
MS-$\bar{A}$	89.1	90.9	89.5	90.8
IPW-K	88.2	88.5	89.3	90.0
IPW-E	88.6	88.9	89.9	89.9
SP-K-A	87.8	89.3	89.5	89.7
SP-K-$\bar{A}$	88.6	88.3	90.2	89.9
SP-E-A	90.1	91.7	91.1	91.8
SP-E-$\bar{A}$	89.1	89.1	89.9	91.5

Table 6.10: Simulation variance (mean variance estimate) of TP(c) $\times 10^{-4}$ when c is fixed so that the full data FP(c) equals 0.2 when verification depends on T. Different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$ and A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	α_1, β_1			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
BG-A	0.34 (0.46)	2.50 (2.86)	3.42 (3.73)	1.74 (1.93)
BG-$\bar{A}$	0.33 (0.66)	2.94 (3.45)	4.07 (4.34)	2.15 (2.18)
MS-A	0.38 (0.44)	2.66 (2.87)	3.56 (3.77)	1.89 (2.00)
MS-$\bar{A}$	0.38 (0.58)	3.19 (3.41)	4.38 (4.45)	2.39 (2.38)
IPW-K	1.43 (1.35)	7.42 (6.65)	8.76 (8.19)	3.18 (3.20)
IPW-E	1.43 (1.35)	7.36 (6.57)	8.70 (8.03)	3.09 (3.06)
SP-K-A	1.29 (1.22)	5.21 (4.85)	5.81 (5.62)	2.22 (2.28)
SP-K-$\bar{A}$	1.39 (1.33)	7.24 (6.39)	8.44 (7.80)	3.09 (3.06)
SP-E-A	1.30 (1.22)	5.23 (4.84)	5.82 (5.61)	2.22 (2.28)
SP-E-$\bar{A}$	1.39 (1.33)	7.25 (6.38)	8.43 (7.79)	3.09 (3.06)

confidence intervals of $TP(c)$ and $FP(c)$ in small samples. Coverage probabilities were close to the nominal 90% when the sample size was increased to 5000 (results not included).

6.4.5 Efficiency Gain by Incorporating Auxiliary Information

In these simulations disease verification is not a function of A, therefore, incorporating auxiliary data does not correct for verification bias but may increase efficiency. To assess gains in efficiency that can be achieved by incorporating auxiliary data, we varied the informativeness of A from highly informative (i.e. $\alpha_2 = \beta_2 = 1$) to non-informative (i.e. $\alpha_2 = \beta_2 = 0$) and varied the potential for verification bias by

Table 6.11: Simulation variance (mean variance estimate) $\times 10^{-3}$ of FP(c) when c is fixed so that the full data FP(c) equals 0.1 when verification depends on T. Different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$ and A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	α_1, β_1			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
BG-A	1.75 (1.88)	1.77 (1.88)	1.67 (1.92)	1.71 (2.13)
BG-$\bar{A}$	1.76 (1.77)	1.73 (1.73)	1.64 (1.74)	1.74 (1.84)
MS-A	1.77 (1.79)	1.76 (1.76)	1.67 (1.77)	1.72 (1.82)
MS-$\bar{A}$	1.78 (1.80)	1.74 (1.76)	1.65 (1.77)	1.75 (1.85)
IPW-K	5.49 (5.28)	4.63 (4.30)	4.25 (3.72)	3.29 (1.94)
IPW-E	5.04 (4.99)	3.90 (4.16)	3.45 (3.64)	1.81 (2.01)
SP-K-A	1.90 (1.91)	1.92 (1.92)	1.79 (1.91)	1.75 (1.85)
SP-K-$\bar{A}$	1.90 (1.95)	1.92 (2.03)	1.79 (2.02)	1.75 (1.93)
SP-E-A	1.90 (1.92)	1.92 (1.93)	1.79 (1.91)	1.75 (1.85)
SP-E-$\bar{A}$	1.93 (1.96)	2.04 (2.04)	1.96 (2.03)	1.81 (1.93)

Table 6.12: 90% confidence interval (CI) coverage probability of TP(c) (FP(c)) given full data FP(c) equals 0.2 when verification depends on T. Different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$ and A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	α_1, β_1			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
BG-A	91.6 (91.6)	90.8 (90.8)	90.3 (92.9)	90.8 (94.1)
BG-$\bar{A}$	92.9 (90.3)	92.4 (89.5)	90.6 (90.6)	90.0 (91.9)
MS-A	90.5 (90.5)	90.4 (89.6)	90.2 (91.4)	89.7 (91.5)
MS-$\bar{A}$	93.9 (90.5)	91.7 (89.7)	91.3 (91.0)	90.2 (91.7)
IPW-K	48.2 (88.4)	86.2 (88.1)	87.5 (87.0)	90.6 (79.6)
IPW-E	54.7 (89.2)	85.4 (90.7)	87.9 (90.3)	89.2 (92.4)
SP-K-A	56.5 (90.3)	85.6 (89.8)	88.7 (91.2)	88.1 (91.5)
SP-K-$\bar{A}$	48.9 (90.6)	85.7 (90.8)	87.8 (91.9)	89.3 (92.2)
SP-E-A	55.8 (90.3)	85.7 (89.5)	88.6 (91.4)	88.2 (91.5)
SP-E-$\bar{A}$	48.1 (90.6)	85.4 (90.9)	87.8 (90.4)	89.3 (91.9)

Table 6.13: % gain in disease prevalence (AUC) small sample efficiency for BG-A relative to BG-$\bar{A}$ when verification depends on T.

α_2	β_2	$\alpha_1 = \beta_1 = 1$	$\alpha_1 = \beta_1 = 0.5$	$\alpha_1 = 1, \beta_1 = 0$	$\alpha_1 = \beta_1 = 0$
1	1	3.2 (5.1)	16.3 (21.6)	18.5 (23.6)	37.5 (33.0)
0.5	1	1.6 (4.0)	9.9 (12.8)	16.7 (21.3)	25.9 (21.8)
0	1	0.5 (2.4)	3.3 (4.1)	13.4 (17.0)	10.8 (8.0)
0.5	0.5	1.0 (1.2)	6.8 (7.6)	7.6 (10.6)	19.0 (16.7)
0	0.5	0.3 (1.3)	1.6 (1.1)	5.6 (7.8)	5.1 (4.1)
0	0	0.2 (0.3)	0.1 (0.0)	-0.2 (0.0)	0.1 (0.0)

considering different values of α_1 and β_1. Table 6.13 quantifies the efficiency gain (100 $\times$ [1-SSRE]) by incorporating auxiliary data relative to BG-$\bar{A}$, which does not incorporate A, when estimating disease prevalence and AUC. Similar results are observed for MS. Moving down a column decreases the informativeness of A and, not surprisingly, decreases the gain in efficiency. It is reassuring to observe that estimators that include non-informative auxiliary data are not less efficient than estimators that incorporate these data because in practice the informativeness of auxiliary data may be hard to quantify. We also observe that efficiency gains generally increase moving across the columns as the accuracy of T decreases. That is, the less accurate T is, the more efficiency that can be gained by incorporating auxiliary data. Table 6.14 provides the gain in efficiency in SP-K-A relative to SP-K-$\bar{A}$. Gains are similar to those observed for BG-A.

6.5 Verification Depends on Auxiliary Data

Next we consider the setting where disease verification is a function of auxiliary data. Again V was generated using a Bernoulli random variable with the probability of

Table 6.14: % gain in disease prevalence (AUC) small sample efficiency for SP-K-A relative to SP-$\bar{A}$ when verification depends on T.

α_2	β_2	$\alpha_1 = \beta_1 = 1$	$\alpha_1 = \beta_1 = 0.5$	$\alpha_1 = 1, \beta_1 = 0$	$\alpha_1 = \beta_1 = 0$
1	1	4.4 (3.1)	19.0 (13.1)	21.5 (17.1)	39.1 (27.3)
0.5	1	2.2 (1.9)	12.2 (9.3)	18.5 (19.4)	27.7 (23.3)
0	1	0.8 (0.5)	4.6 (1.8)	14.2 (13.4)	12.4 (8.5)
0.5	0.5	1.3 (1.0)	8.0 (5.5)	8.3 (7.2)	19.8 (15.6)
0	0.5	0.4 (0.2)	2.1 (1.9)	5.7 (4.5)	5.8 (3.4)
0	0	0.1 (0.2)	0.2 (0.0)	-0.2 (0.0)	0.1 (-0.1)

verification now equal to

$$P(V) = \begin{cases} 1 & A > a^q \\ \delta & \text{else} \end{cases}$$

where a^q is the qth quantile of the distribution of A. Again results are presented for q equal to 0.8 and δ equal to 0.2.

The same estimators considered in Section 6.4 were investigated here.

6.5.1 Small Sample Bias

First consider estimating disease prevalence which is fixed to be 0.1. In these simulations the sampling is biased because of the dependence of verification on the auxiliary data. So it is not surprising that the estimators that properly account for this dependence (BG-A, MS-A, IPW-K, IPW-E, SP-K-A, SP-K-$\bar{A}$, SP-E-A, and SP-E-$\bar{A}$) yield unbiased estimates of disease prevalence (results not shown). Conversely, BG-$\bar{A}$, MS-$\bar{A}$, and CC, which fail to account for this dependence, yield biased results.

Table 6.15 provides the mean estimated disease prevalence across 1000 realizations for BG-$\bar{A}$ and CC. Results for MS-$\bar{A}$ are similar to BG-$\bar{A}$ and, thus, are not presented. The inherent accuracy of T decreases moving across the columns of the table. The

Table 6.15: Mean disease prevalence of 1000 realizations when verification depends on auxiliary data. Different values of α_1, α_2, β_1 and β_2 are considered for $A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2$ and $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Full Data prevalence is 0.100. Results for MS-$\bar{A}$ are similar to BG-$\bar{A}$. Biased estimates are in bold face.

		BG-$\bar{A}$				
α_2	β_2	$\alpha_1 = \beta_1 = 1$	$\alpha_1 = \beta_1 = 0.5$	$\alpha_1 = 1, \beta_1 = 0$	$\alpha_1 = \beta_1 = 0$	CC
1	1	**0.128**	**0.175**	**0.201**	**0.257**	**0.257**
0.5	1	**0.123**	**0.162**	**0.199**	**0.232**	**0.232**
0.5	0.5	**0.119**	**0.151**	**0.170**	**0.210**	**0.210**
0	1	**0.115**	**0.139**	**0.184**	**0.184**	**0.184**
0	0	0.100	0.100	0.100	0.100	0.100

last column of the table corresponds to CC which does not depend on T (since V does not depend on T and the CC prevalence estimator is not a function of T). The informativeness of A and, consequently, potential for verification bias decreases moving down the rows.

Clearly, CC and BG-$\bar{A}$ overestimate disease prevalence in all scenarios except when $\alpha_2 = \beta_2 = 0$, i.e. a simple random sample of subjects is verified. The largest bias is 0.157, more than double the full data prevalence. As expected, the amount of bias decreases moving down the columns. The relationship between T and A also dictates the magnitude of bias observed in BG-$\bar{A}$. As noted earlier, BG-A is unbiased since it includes A when estimating the probability of disease. Therefore, the more that T and A get at similar aspects of disease (i.e. correlated), the less amount of bias in BG-$\bar{A}$ and MS-$\bar{A}$, which only include T when estimating the probability of disease. The correlation between T and A decreases moving across columns.

The bias in BG-$\bar{A}$ is less than that of CC as long as T is a somewhat accurate test (i.e. $\alpha_1 > 0$ and $\beta_1 > 0$). When α_1 and β_1 equal 0, T is a useless test so $\widehat{P}(D|T)$

Table 6.16: Mean TP given FP equals 0.1 of 1000 realizations when verification depends on auxiliary data. Different values of α_1, α_2, β_1 and β_2 are considered for $A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2$ and $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Results for MS-$\bar{A}$ are similar to those for BG-$\bar{A}$. Full data (FD) estimates are provided in parentheses. Biased estimates are in bold face.

		$\alpha_1 = \beta_1 = 1$		$\alpha_1 = \beta_1 = 0.5$		$\alpha_1 = 1, \beta_1 = 0$		$\alpha_1 = \beta_1 = 0$	
		(FD=0.873)		(FD=0.601)		(FD=0.460)		(FD=0.100)	
α_2	β_2	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC
1	1	**0.835**	**0.756**	**0.533**	**0.495**	**0.399**	**0.380**	0.100	0.101
0.5	0.5	**0.857**	**0.804**	**0.561**	**0.533**	**0.420**	**0.406**	0.100	0.101
0	1	**0.865**	**0.822**	**0.575**	**0.549**	**0.440**	**0.434**	0.100	0.101
0	0	0.880	0.874	0.610	0.602	0.464	0.461	0.102	0.101

used by BG-$\bar{A}$ is essentially the observed prevalence which is the CC estimator. When estimating disease prevalence both T and A can be considered surrogates for auxiliary data A. Therefore, it is not surprising that the bias in CC (Tables 6.3 and 6.15) is the same regardless of whether verification depends on T or A.

Similar results are observed for the empirical AUC (Table 6.17) and $TP(c)$ and $FP(c)$ (results not presented). BG-$\bar{A}$, MS-$\bar{A}$, and CC yield biased AUC estimates when $\alpha_2 > 0$ and $\beta_2 > 0$. These biases are apparent in Figures 6.8-6.12 which display ROC curves from a randomly chosen realization of the data where $T = Z_1 + \epsilon_1$ and $A = Z_1 + Z_2 + \epsilon_2$. BG-$\bar{A}$, MS-$\bar{A}$, and CC clearly underestimate the full data ROC curve while ROC curves for the other estimators tend to closely resemble the full data ROC curve. Similar to the observations made in Section 6.4.1, ROC curves for IPW and SP are not as smooth as the others, especially in the upper portion of the ROC curves.

Comparing results for CC in Tables 6.5 and 6.17 we see that there is less bias in the AUC estimator when verification is a function of the auxiliary data than when it

Table 6.17: Mean AUC of 1000 realizations when verification depends on auxiliary data. Different values of α_1, α_2, β_1 and β_2 are considered for $A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2$ and $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Results for MS-$\bar{A}$ are similar to those for BG-$\bar{A}$. Full data (FD) estimates are provided in parentheses. Biased estimates are in bold face.

		$\alpha_1 = \beta_1 = 1$ (FD=0.960)		$\alpha_1 = \beta_1 = 0.5$ (FD=0.875)		$\alpha_1 = 1, \beta_1 = 0$ (FD=0.810)		$\alpha_1 = \beta_1 = 0$ (FD=0.499)	
α_2	β_2	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC	BG-$\bar{A}$	CC
1	1	**0.950**	**0.930**	**0.844**	**0.831**	**0.775**	**0.768**	0.498	0.498
0.5	1	**0.954**	**0.937**	**0.852**	**0.840**	**0.793**	**0.788**	0.498	0.498
0.5	0.5	**0.955**	**0.942**	**0.857**	**0.846**	**0.788**	**0.781**	0.498	0.498
0	1	**0.957**	**0.947**	**0.863**	**0.853**	**0.798**	**0.798**	0.498	0.498
0	0	0.961	0.960	0.877	0.876	0.812	0.811	0.499	0.499

is a function of the test under evaluation. An ROC curve is a plot of FP(c) versus TP(c) for all c and the numerators of TP(c) and FP(c) sum over all subjects with $T \geq c$ so it is not surprising that verification that depends on the distribution of T will have a larger effect.

6.5.2 Relative Efficiency

Table 6.18 presents SSRE and ARE relative to BG-A when estimating disease prevalence and T is fixed to be $Z_1 + Z_2 + \epsilon_1$. The informativeness of A and potential for verification bias decrease moving across the columns. Again BG-A and MS-A have similar efficiency. They are the most efficient estimators for all scenarios except $\alpha_2 = \beta_2 = 0$ when BG-$\bar{\text{A}}$ and MS-$\bar{\text{A}}$ are roughly 20% more efficient. In the latter scenario, T is highly accurate ($\alpha_1 = \beta_1 = 1$) and A is non-informative. It is interesting that incorporating the non-informative A hurts efficiency. This is not the case when estimating AUC (Table 6.19) or in scenarios where the accuracy of T is decreased

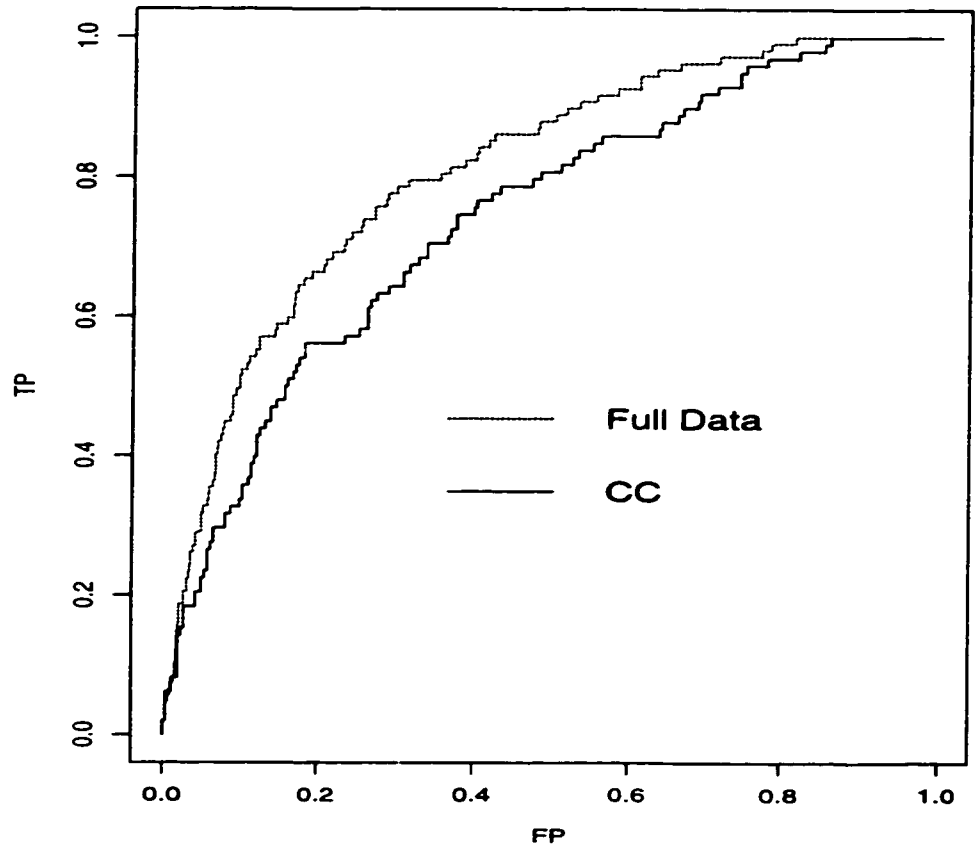

Figure 6.8: Full data and CC ROC curves from a randomly chosen realization of the simulation study when verification depends on the auxiliary data. $A = Z_1 + Z_2 + \epsilon_2$ and $T = Z_1 + \epsilon_1$.

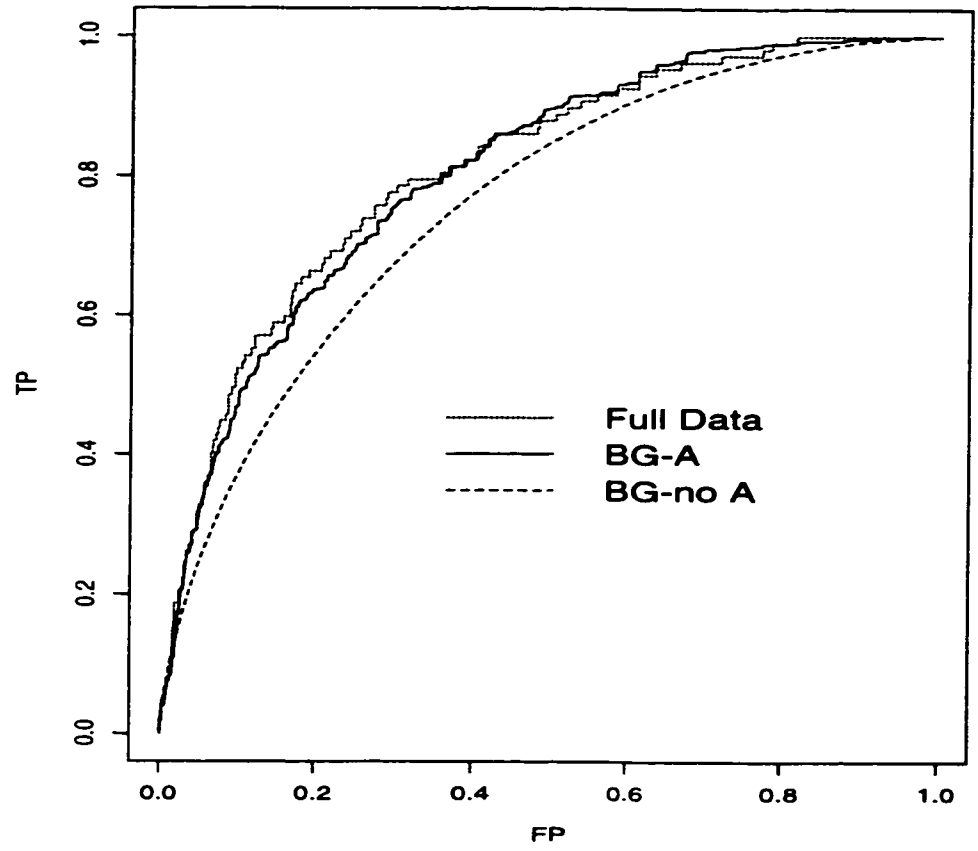

Figure 6.9: Full data, BG-A, and BG-$\bar{A}$ ROC curves from a randomly chosen realization of the simulation study when verification depends on the auxiliary data. $A = Z_1 + Z_2 + \epsilon_2$ and $T = Z_1 + \epsilon_1$.

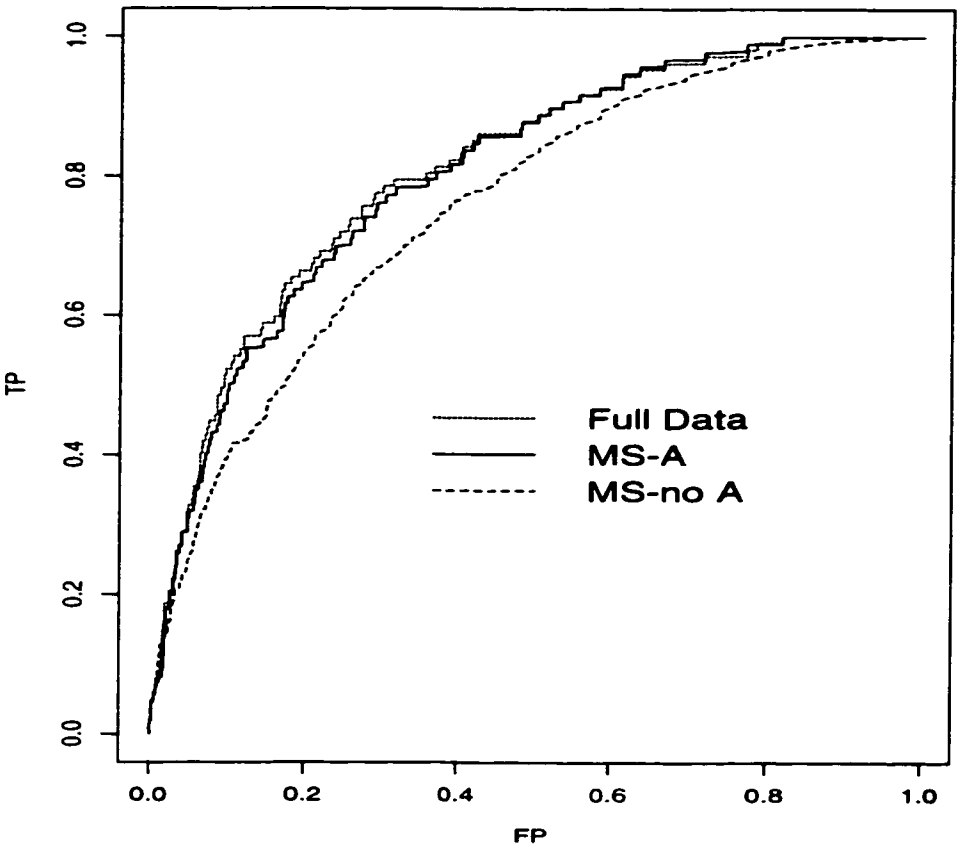

Figure 6.10: Full data, MS-A, and MS-$\bar{A}$ ROC curves from a randomly chosen realization of the simulation study when verification depends on the auxiliary data. $A = Z_1 + Z_2 + \epsilon_2$ and $T = Z_1 + \epsilon_1$.

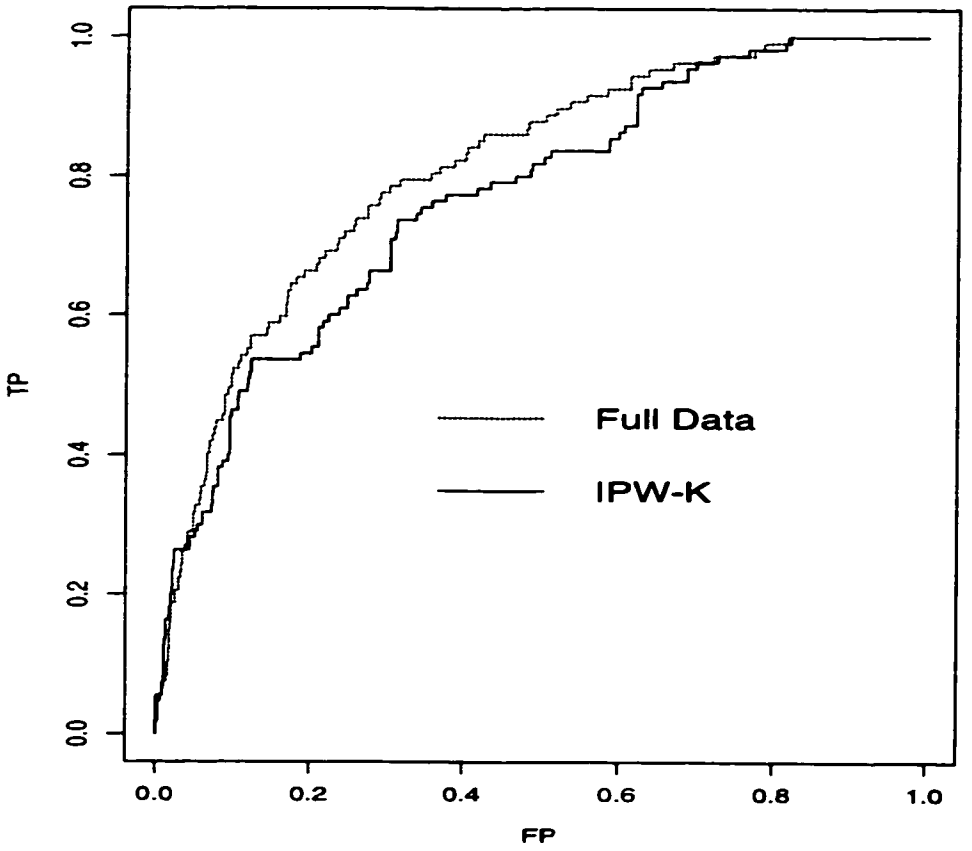

Figure 6.11: Full data and IPW ROC curves from a randomly chosen realization of the simulation study when verification depends on the auxiliary data. $A = Z_1 + Z_2 + \epsilon_2$ and $T = Z_1 + \epsilon_1$.

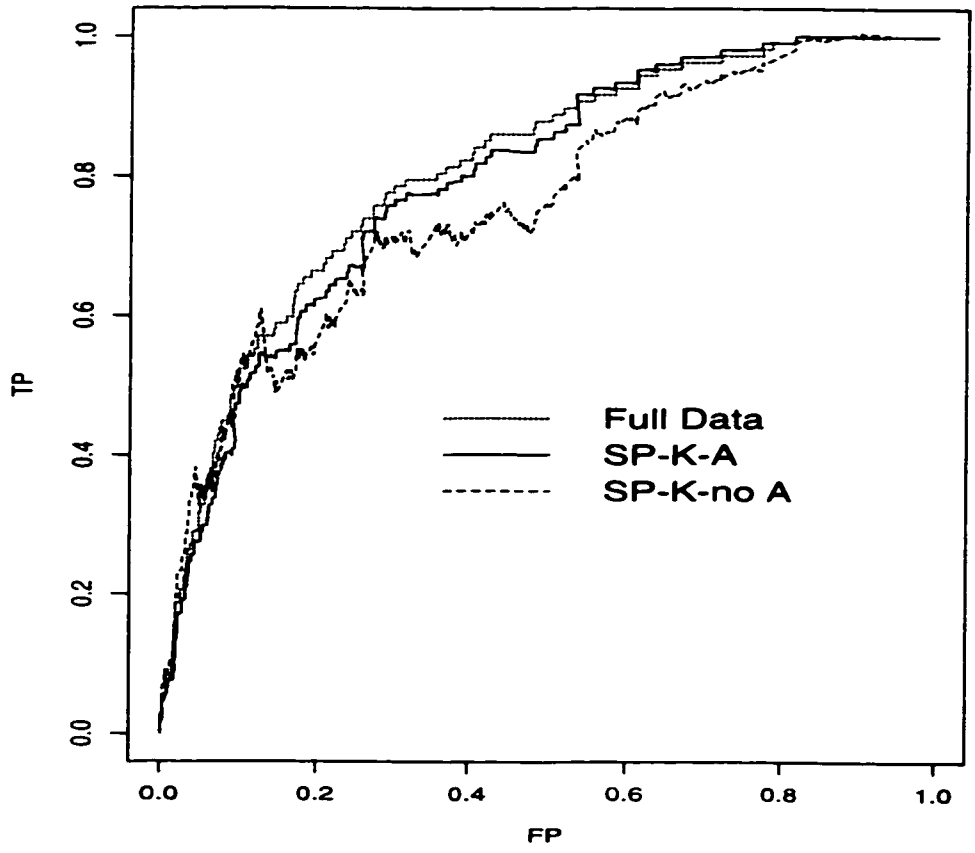

Figure 6.12: Full data, SP-K-A, and SP-K-$\bar{A}$ ROC curves from a randomly chosen realization of the simulation study when verification depends on the auxiliary data. $A = Z_1 + Z_2 + \epsilon_2$ and $T = Z_1 + \epsilon_1$.

Table 6.18: Small sample efficiency relative to BG-A for estimating disease prevalence. ARE in parentheses. 1000 realizations when verification depends on auxiliary data. T is fixed to be $Z_1 + Z_2 + \epsilon_1$ while different values of α_2 and β_2 are considered for $A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2$.

	α_2, β_2			
Method	1, 1	0.5, 0.5	1, 0	0, 0
CC	0.191 (0.194)	0.285 (0.297)	0.357 (0.361)	0.780 (0.824)
BG-$\bar{A}$	0.671 (0.655)	0.897 (0.900)	1.001 (0.991)	1.207 (1.253)
MS-A	0.990 (0.996)	0.996 (1.000)	0.997 (0.997)	0.998 (1.002)
MS-$\bar{A}$	0.669 (0.654)	0.897 (0.897)	0.998 (0.989)	1.204 (1.250)
IPW-K	0.679 (0.784)	0.614 (0.633)	0.592 (0.575)	0.531 (0.495)
IPW-E	0.800 (0.776)	0.652 (0.656)	0.608 (0.605)	0.531 (0.535)
SP-K-A	0.885 (0.851)	0.863 (0.855)	0.857 (0.838)	0.892 (0.920)
SP-K-$\bar{A}$	0.716 (0.671)	0.810 (0.772)	0.825 (0.792)	0.896 (0.910)
SP-E-A	0.884 (0.852)	0.862 (0.819)	0.854 (0.787)	0.894 (0.819)
SP-E-$\bar{A}$	0.745 (0.642)	0.820 (0.700)	0.831 (0.718)	0.897 (0.812)

when estimating prevalence (results not shown).

MS and BG estimators that incorporate the auxiliary data are anywhere from 11% to 28% more efficient than the SP estimators. Furthermore, the SP estimators which incorporate the auxiliary information range from roughly 9% to 41% more efficient than the IPW estimators.

ARE relative to BG-A is included in parentheses in Table 6.18. Comparing SSRE to ARE, we see that they are quite similar. Thus, ARE appears to translate in small samples.

Table 6.19: Small sample efficiency relative to BG-A for estimating AUC when verification depends on auxiliary data. T is fixed to be $Z_1 + Z_2 + \epsilon_1$ while different values of α_2 and β_2 are considered for $A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2$.

	α_2, β_2			
Method	1, 1	0.5, 0.5	1, 0	0, 0
CC	0.335	0.475	0.549	0.859
BG-$\bar{A}$	0.570	0.777	0.854	1.004
MS-A	0.978	0.979	0.969	1.003
MS-$\bar{A}$	0.557	0.763	0.830	1.005
IPW-K	0.552	0.508	0.490	0.550
IPW-E	0.556	0.509	0.491	0.548
SP-K-A	0.701	0.609	0.582	0.668
SP-K-$\bar{A}$	0.605	0.588	0.569	0.666
SP-E-A	0.702	0.606	0.582	0.661
SP-E-$\bar{A}$	0.571	0.578	0.560	0.658

Table 6.20: Simulation variance, mean variance estimate and nominal 90% confidence interval (CI) coverage probability of disease prevalence when verification depends on $A = Z_1 + Z_2 + \epsilon_1$. $T = Z_1 + Z_2 + \epsilon_2$.

	Simulation	Variance Estimator	
	Variance	Mean	90% CI Coverage
	($\times 10^{-4}$)	($\times 10^{-4}$)	of prevalence
CC	5.224	5.281	0.0
BG-A	0.996	1.040	89.7
BG-$\bar{A}$	1.484	1.591	28.5
MS-A	1.006	1.044	89.6
MS-$\bar{A}$	1.489	1.586	28.7
IPW-K	1.255	1.271	90.2
IPW-E	1.245	1.282	90.1
SP-K-A	1.125	1.180	88.7
SP-K-$\bar{A}$	1.391	1.534	89.0
SP-E-A	1.127	1.356	90.4
SP-E-$\bar{A}$	1.337	1.795	90.7

6.5.3 Performace of the Variance Estimator

Tables 6.20 and 6.21 provide the simulation variance, mean variance estimator, and coverage probabilities of a nominal 90% confidence interval for two different scenarios. Again the variance estimator tends to be slightly larger than the simulation variance and coverage probabilities are close to 90% for the unbiased estimators. Poor coverage is to be expected for the biased estimators, CC, BG-$\bar{A}$, and MS-$\bar{A}$. The performance of the variance estimator in these scenarios when verfication is a function of A, therefore, suggest the variance estimator could be used in small samples.

Table 6.21: Simulation variance, mean variance estimate and nominal 90% confidence interval (CI) coverage probability of disease prevalence when verification depends on $A = Z_2 + \epsilon_1$. $T = Z_1 + Z_2 + \epsilon_2$.

	Simulation	Variance Estimator	
	Variance ($\times 10^{-4}$)	Mean ($\times 10^{-4}$)	90% CI Coverage of prevalence
CC	5.224	5.281	0.0
BG-A	1.449	1.607	89.9
BG-$\bar{A}$	1.448	1.534	69.3
MS-A	1.453	1.590	89.9
MS-$\bar{A}$	1.452	1.527	69.5
IPW-K	2.449	2.559	89.7
IPW-E	2.382	2.484	89.7
SP-K-A	1.691	1.769	90.1
SP-K-$\bar{A}$	1.757	1.863	91.0
SP-E-A	1.696	2.466	92.4
SP-E-$\bar{A}$	1.744	2.309	92.6

6.6 Varying Prevalence and % Verified

Simulations in Sections 6.4 and 6.5 considered accuracy studies with 10% disease prevalence and a verification mechanism such that only 36% of the subjects received disease verification. In this section we investigate the effect of varying prevalence and the percentage of subjects below the threshold verified. Results are presented for an accuracy study with 1000 subjects and selection for verification depending on $T = Z_1 + Z_2 + \epsilon_1$.

6.6.1 Varying Prevalence

Consider disease prevalences of 0.10, 0.50, 0.75, and 0.90. Since verification is only a function of T, it is not surprising that there is no small sample bias in estimating prevalence or AUC for any method except CC (results not shown). The bias in CC for estimating prevalence are 0.157, 0.217, 0.110, and 0.040 for prevalences of 0.10, 0.50, 0.75, and 0.90, respectively. Furthermore, CC underestimates the full data AUC (0.960) by 0.047, 0.033, 0.028, and 0.018 for the same disease prevalences.

SSRE relative to BG-A when estimating prevalence and AUC are summarized in Table 6.22 for $T = Z_1 + Z_2 + \epsilon_1$ and $A = Z_1 + Z_2 + \epsilon_2$ (results for other T and A are similar). Again MS-A has similar efficiency to BG-A. In addition, the difference in efficiency between these estimators and SP-K-A, SP-E-A, BG-$\bar{A}$, and MS-$\bar{A}$ decreases as prevalence increases. On the other hand, the difference in efficiency between IPW estimators gets larger for larger prevalence.

Therefore, qualitatively we arrive at similar conclusions regarding the relative merits of the estimators with larger disease prevalence.

6.6.2 Varying % Verified

By varying δ we can investigate the effect of varying the percentage of subjects below the threshold verified. Again the only small sample bias observed when estimating

Table 6.22: SSRE relative to BG-A in estimating disease prevalence (AUC) as disease prevalence is varied. Verification depends on $T = Z_1 + Z_2 + \epsilon_1$. $A = Z_1 + Z_2 + \epsilon_2$.

	Prevalence			
Method	0.10	0.50	0.75	0.90
BG-$\bar{A}$	0.968 (0.949)	0.845 (1.010)	0.847 (0.992)	0.826 (0.859)
MS-A	0.993 (0.994)	1.000 (0.991)	1.000 (0.997)	1.000 (1.000)
MS-$\bar{A}$	0.963 (0.942)	0.845 (0.999)	0.847 (0.989)	0.826 (0.884)
IPW-K	0.812 (0.553)	0.326 (0.524)	0.177 (0.619)	0.071 (0.709)
IPW-E	0.813 (0.558)	0.528 (0.530)	0.500 (0.616)	0.499 (0.711)
SP-K-A	0.883 (0.621)	0.966 (0.695)	0.994 (0.861)	1.000 (0.943)
SP-K-$\bar{A}$	0.829 (0.590)	0.810 (0.663)	0.838 (0.803)	0.825 (0.913)
SP-E-A	0.867 (0.622)	0.966 (0.698)	0.994 (0.862)	1.000 (0.943)
SP-E-$\bar{A}$	0.828 (0.593)	0.810 (0.618)	0.838 (0.749)	0.825 (0.829)

prevalence and AUC was for CC. As expected, bias in CC decreases as the percentage of subjects verified is increased. Specifically, the bias in estimating prevalence is 0.157, 0.059, 0.022, and 0.0 for δ equal to 0.2, 0.5, 0.75, and 1.0. The bias in estimating AUC is 0.047, 0.014, 0.005, and 0.018 for the same δ.

As the percentage of subjects verified is increased, the difference in efficiency between BG-A and MS-A and the rest decreases (Table 6.23). In the extreme case where all subjects are verified ($\delta = 1.0$), all methods are identical to CC (provided models for the probability of disease and for the probability of verification are correct).

6.7 Robustness to Model Misspecification

BG, MS, and SP require a model for the probability of disease while IPW and SP require a model for the verification probability. Here we investigate how robust these

Table 6.23: SSRE relative to BG-A in estimating disease prevalence (AUC) as δ, i.e. the probability a subject below the threshold is selected for verification, is varied. Verification depends on $T = Z_1 + Z_2 + \epsilon_1$. A is fixed to be $Z_1 + Z_2 + \epsilon_2$.

	δ (% Verified)			
Method	0.20 (36)	0.50 (60)	0.75 (80)	1.00 (100)
CC	0.200 (0.385)	0.431 (0.759)	0.683 (0.908)	1.000 (0.999)
BG-$\bar{A}$	0.968 (0.949)	0.855 (0.983)	0.996 (0.882)	1.000 (0.999)
MS-A	0.993 (0.994)	1.000 (1.020)	1.000 (1.010)	1.000 (0.999)
MS-$\bar{A}$	0.963 (0.942)	0.983 (0.925)	0.996 (0.947)	1.000 (0.999)
IPW-K	0.812 (0.553)	0.957 (0.867)	0.990 (0.930)	1.000 (0.999)
IPW-E	0.813 (0.560)	0.961 (0.867)	0.989 (0.933)	1.000 (0.999)
SP-K-A	0.867 (0.621)	0.985 (0.920)	0.996 (0.977)	1.000 (0.999)
SP-K-$\bar{A}$	0.829 (0.590)	0.963 (0.898)	0.991 (0.944)	1.000 (0.999)
SP-E-A	0.867 (0.622)	0.986 (0.921)	0.996 (0.979)	1.000 (0.999)
SP-E-$\bar{A}$	0.828 (0.593)	0.963 (0.897)	0.991 (0.946)	1.000 (0.999)

methods are to model misspecification.

6.7.1 Verification Model Misspecification

In practice verification probabilities are often unknown and must be estimated from the data. This is true for observational studies and even randomized studies because even though the probabilities of selection for verification may have been known by the original sampling design, the actual selection probabilities may be unknown due to drop-out, refusal, or other reasons.

In the simulations previously discussed the correct model for $P(V|T)$ is a threshold model in which the predictor is $I[T > t^{0.8}]$. Table 6.24 provides the mean estimated disease prevalence and AUC for IPW and SP when a non-threshold logit model for $P(V|T)$ is used to estimate the verification probability. Namely we consider the model where the predictor is the continuous test result T. These estimators are denoted IPW-L and SP-L-$\bar{A}$. Similar results are obtained if A is included. Clearly not robust to the model misspecification, IPW-L substantially underestimates the full data disease prevalence and overestimates AUC. Conversely, SP-L-$\bar{A}$ appears to be fairly robust to the model misspecification. However, on rare occasions when $\alpha_1 = \beta_1 = 0$, SP-L-$\bar{A}$ obtains negative estimates of disease prevalence and consequently, AUC not in the range between 0 and 1.

6.7.2 Disease Model Misspecification

In Section 6.5 verification bias was induced by selecting subjects for verification as a function of their auxiliary data. As expected, BG-$\bar{A}$ and MS-$\bar{A}$ which did not account for this dependence yielded biased results (Tables 6.15 and 6.17). This is an extreme case of model misspecification. In this case the MAR assumption is violated. That is, $P(D|T, V = 1) \neq P(D|T)$.

In the previous simulations we also investigated the performance of BG, MS, and SP when the probability of disease was fit with a logit model instead of a probit model,

Table 6.24: Mean disease prevalence (AUC) of 1000 realizations when verification depends on T. A is fixed to be $Z_1 + Z_2 + \epsilon_2$ while different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1$. Biased estimates are in bold face.

	α_1, β_1			
Method	1, 1	0.5, 0.5	1, 0	0, 0
Full Data	0.100 (0.960)	0.100 (0.876)	0.100 (0.810)	0.100 (0.499)
IPW-L	**0.090 (0.969)**	**0.079 (0.927)**	**0.075 (0.885)**	0.100 **(0.533)**
SP-L-$\bar{A}$	0.100 (0.960)	0.099 (0.879)	0.100 **(0.816)**	0.098 **(0.515)**

the true model (results not shown). Since logit and probit models are so similar, there was only negligible bias induced by using the incorrect model. Therefore, to examine the effects of misspecifying a model for the probability of disease we altered the simulation set-up slightly.

Disease status, D, was generated as a binary variable indicating whether or not both aspects of the disease process were greater than some threshold. Otherwise, the disease was not apparent and remained subclinical. In particular,

$$D = I[Z_1 > h_1, Z_2 > h_2], \quad Z_1 \sim N(0, 0.5), \; Z_2 \sim N(0, 0.5).$$

Furthermore, T, were constructed as Z_1 plus random normal error, ϵ_1 and A were generated by adding random normal error, ϵ_2 to Z_2. That is,

$$T = Z_1 + \epsilon_1, \quad \epsilon_1 \sim N(0, 0.25)$$

and

$$A = Z_2 + \epsilon_2, \quad \epsilon_2 \sim N(0, 0.25)$$

where ϵ_1 and ϵ_2 are independent.

Unlike the previous simulation set-up (Section 6.3) the true model for the probability of disease is no longer known. However, since T and A get at different components

of the disease and the disease is only apparent if both components are greater than a threshold, the correct model should contain an interaction between T and A.

Again consider an accuracy study with 1000 subjects, 10% disease prevalence, and verification as a function of T. We consider models for the probability of disease that only include T (BG-$\bar{A}$, MS-$\bar{A}$, SP-K-$\bar{A}$, and SP-E-$\bar{A}$), include both T and A (BG-A, MS-A, SP-K-A, and SP-E-A), and include T, A, and an interaction between T and A (BG-Int, MS-Int, SP-K-Int, and SP-E-Int). Even though a probit model is not necessarily the correct model, we present results when probit models are used.

Tables 6.25 and 6.26 present the mean and variance of the estimated prevalence and AUC across 1000 realizations when δ equals 0.2 (resulting in 36% of the subjects receiving disease verification) and 0.5 (resulting in 60% of the subjects receiving disease verification). Results suggest the IPW and SP estimators are unbiased. On the other hand, BG and MS estimators appear to yield biased estimates of prevalence and AUC, but the bias is substantially less than the bias in the CC estimator. As expected, there is less bias in the BG and MS estimators that include the interaction between T and A. It is also not surprising that MS estimators perform better than BG since it uses the correct disease status for those subjects in the verification group rather than estimating the probability of disease. The bias is less when $\delta = 0.5$. Similar results are obtained for $TP(c)$ and $FP(c)$ (results not provided).

6.8 Summary

Estimators that did not account for the biased sampling yielded biased estimates of disease prevalence, $TP(c)$, $FP(c)$, empirical ROC curve, and AUC. On the other hand, the methods proposed in Chapter 4 and shown in Chapter 5 to be asymptotically unbiased were also unbiased in small samples. BG and MS estimators which include auxiliary data were the most efficient. SP estimators were not as efficient, but were substantially more efficient than the IPW estimators.

Table 6.25: Mean (variance $\times 10^{-4}$) of estimated disease prevalence and AUC when verification is a function of $T = Z_1 + \epsilon_1$ and $\delta = 0.2$. A is fixed to be $Z_2 + \epsilon_2$. Biased estimates are in bold face.

	Prevalence (Full data = 0.100)	AUC (Full data = 0.802)
CC	**0.172** (4.005)	**0.723** (8.357)
BG-A	**0.106** (2.002)	**0.747** (9.535)
BG-$\bar{A}$	**0.104** (2.052)	**0.758** (10.110)
BG-Int	**0.102** (1.895)	**0.772** (9.999)
MS-A	**0.106** (2.007)	**0.757** (9.372)
MS-$\bar{A}$	**0.104** (2.065)	**0.765** (9.749)
MS-Int	**0.102** (1.918)	**0.778** (9.514)
IPW-K	0.100 (2.699)	0.805 (11.639)
IPW-E	0.100 (2.570)	0.805 (11.682)
SP-K-A	0.100 (2.328)	0.804 (13.630)
SP-K-$\bar{A}$	0.100 (2.478)	0.805 (11.320)
SP-K-Int	0.100 (2.223)	0.804 (10.917)
SP-E-A	0.100 (2.318)	0.804 (13.500)
SP-E-$\bar{A}$	0.100 (2.472)	0.805 (11.310)

Table 6.26: Mean (variance $\times 10^{-4}$) of estimated disease prevalence and AUC when verification is a function of $T = Z_1 + \epsilon_1$ and $\delta = 0.5$. A is fixed to be $Z_2 + \epsilon_2$. Biased estimates are in bold face.

	Prevalence (Full data = 0.100)	AUC (Full data = 0.802)
CC	**0.127** (1.927)	**0.780** (5.046)
BG-A	0.100 (1.240)	0.770 (5.459)
BG-$\bar{A}$	0.099 (1.266)	**0.780** (5.488)
BG-Int	0.098 (1.193)	**0.787** (5.490)
MS-A	0.100 (1.247)	**0.784** (5.323)
MS-$\bar{A}$	0.099 (1.286)	**0.788** (5.383)
MS-Int	0.100 (1.223)	0.795 (5.218)
IPW-K	0.100 (1.414)	0.803 (5.465)
IPW-E	0.100 (1.394)	0.803 (5.474)
SP-K-A	0.100 (1.297)	0.802 (5.705)
SP-K-$\bar{A}$	0.100 (1.369)	0.803 (5.359)
SP-K-Int	0.100 (1.282)	0.803 (5.291)
SP-E-A	0.100 (1.296)	0.802 (5.697)
SP-E-$\bar{A}$	0.100 (1.370)	0.803 (5.341)

ARE for disease prevalence and, to a lesser extent, TP and FP estimation appears to translate in small samples. Moreover, the mean of the variance estimator was similar to the simulation variance which can be thought of as an estimate of the true variance. And nominal 90% coverage probabilities were near 90% for the unbiased disease prevalence estimators and a majority of the $TP(c)$ and $FP(c)$ estimators suggesting the variance estimator performs well in small samples.

Clearly, in scenarios when verification did not depend on A, incorporating A improved the efficiency of the BG, MS, and SP estimators. The more informative the auxiliary data, the greater the gains in efficiency. Gains in efficiency up to 39% were observed.

An investigation of the effects of model misspecification revealed that BG and MS, which require a model for the probability of disease, yield biased results if that model is not specified correctly. MS appeared to be more robust than BG. On the other hand, IPW and SP require a model for the verification probability. If this model is incorrect, IPW results in biased results. Conversely, SP was fairly robust to this type of model misspecification in all scenarios except when there was extreme model misspecification. Therefore, in practice it is important to fit these models well.

Since BG and MS have similar efficiency and MS is more robust to model misspecification, we recommend using MS over BG. Based on efficiency results we also recommend the use of SP over IPW. Since it is easier to fit $P(V|T, A)$ than $P(D|T, A)$, we propose that future studies are designed so that verification probabilities are known or can be estimated well and then SP estimators be used to estimate accuracy. Two possible drawbacks to SP are the non-monotonicity of the estimated ROC curves and poor estimation in the upper portion of the ROC curve. Isotonic regression can be used to correct for the non-monotonicity. In studies that use pAUC corresponding to low FP rates instead of AUC the poor estimation in the upper part of the curve may not be as much of an issue. The largest drawback to SP is that the approach is not intuitive, especially to non-statisticians.

In the next chapter we analyze data from the Neonatal Hearing Screening Study. Specifically, the different estimating methods will be used to estimate prevalence of hearing impairment and accuracy of two new screening tests.

Chapter 7

NEONATAL HEARING SCREENING STUDY

7.1 Study Description

Background for the Neonatal Hearing Screening Study (NHSS) was provided in Section 1.5.1. Recall that one goal of the NHSS was to assess the accuracy of new screening tests for hearing impairment in infants. Specifically, the NHSS investigated the performance of two screening tests, distortion product otoacoustic emissions (DPOAE) and transient evoked otoacoustic emissions (TEOAE), that do not require cooperation of the infants and, therefore, can be given soon after birth.

DPOAE and TEOAE tests use electronic devices that examiners insert into the ear canal of an infant. These devices emit sounds and record continuous measures of the strength of response from the cochlea to these auditory stimuli, with lower values being more indicative of hearing loss. These tests are performed using input sounds with different frequencies and intensities.

We assessed the accuracy of DPOAE and TEOAE tests at the frequency of 2000MHz with stimulus intensity levels of 65dB and 80dB, respectively. VRA was considered the gold standard test for defining an ear to be hearing impaired. VRA is a behavioral test in which examiners observe whether each infant has a behavioral response, such as a head turn, to various sound frequencies emitted in an ear. In this analysis we define hearing impairment to be ears for which a behavioral response was only observed for auditory stimulus greater than 20dB at the 2000MHz frequency.

NHSS study protocol dictated that all infants be tested with both screening tests soon after birth and all infants were to be followed after discharge from the hospital

so that VRA could be performed at 8-12 months of age. Since tracking infants is expensive, the goal of this chapter is to analyze the performance of the methods proposed in Chapter 4 for correcting for verification bias when assessing accuracy of a test in a verification biased subset of the NHSS data. We refer to this subset of the NHSS data as the "two-phase data" because we induce verification bias through the use of a two-phase design. Since we have the full data results, this analysis is a good way to assess the ability of our methods to correct for verification bias.

7.2 Convention/Assumptions

We consider negative DPOAE and negative TEOAE test results, hereafter referred to as DPOAE and TEOAE, so that, consistent with previous chapters, larger values of the tests are more indicative of hearing loss. The bias correction methods proposed in Chapter 4 require the MAR assumption (3.1) or equivalently (3.2). With these data MAR assumes that the two-phase data can be used to obtain valid estimates of hearing impairment probabilities conditional on DPOAE and TEOAE test results for the full data. In Section 7.6.3 we assess how well this assumption appears to be met for these data.

7.3 Two-phase Data

We consider a subset of the NHSS data that resembles data that would be obtained from a two-phase design in which only a fraction of infants are tracked after discharge from the hospital so that they can receive the definitive hearing test VRA. At the first phase DPOAE and TEOAE test results are obtained on all infants. Selection for VRA testing at phase 2 depends on the phase 1 data in that all infants with a positive DPOAE test result on at least one ear and a random subset of the remaining infants are tested with VRA. We will refer to this subset of the NHSS data as the "two-phase data." Specifically, we define a positive DPOAE test to be any test result

greater than the 80th quantile of the distribution of DPOAE test results for all infants. Furthermore, VRA test results on ears corresponding to negative DPOAE tests were included in the two-phase data with probability 0.2. If one ear was selected, then both were included because it makes sense to test both ears if researchers have already used resources to track the infant.

By defining the two-phase data as we have, we can consider a scenario where verification is a function of the screening test and a scenario where verification is a function of auxiliary data. That is, we consider scenarios where (1) DPOAE is the screening test T and TEOAE test results correspond to the auxiliary data A and (2) the reverse where T=TEOAE and A=DPOAE.

7.4 Bias Correction Methods

In this dissertation we proposed several new methods for assessing the accuracy of a continuous screening test in the presence of verification bias. We apply these methods to the two-phase data where we have chosen to ignore the true hearing impairment status for a subset of infants. Specifically, we consider the following methods: an extension of Begg and Greenes' work (BG), mean score (MS), inverse probability weighting (IPW), and a semi-parametric efficient (SP). In addition, we consider a complete case (CC) approach where only data for ears with complete data are used.

7.4.1 Modelling Probability of Disease

BG, MS, and SP estimators require a model for the probability an ear is hearing impaired (i.e. diseased) conditional on the test under investigation and auxiliary data. We chose to use a logistic model.

Generalized Additive Models (GAM) were used as an exploratory tool to suggest transformations of the predictors, if necessary, to be included in the model for the probability of disease. Predictors were modelled as non-parametric smooth terms

using smoothing splines with 4 degrees of freedom. More details on fitting GAM can be found in Hastie & Tibshirani (1990). Plots of the fitted GAM models were used to suggest parametric transformations of the predictors. Each function represented in a plot, e.g. Figure 7.3 (a), is the contribution of that variable to the fitted additive predictor, the analogue of the linear predictor in generalized linear models.

7.4.2 Estimating Verification Probabilities

IPW and SP methods require estimates of verification probabilities, i.e. probability an ear in the two-phase data received VRA test, conditional on DPOAE and TEOAE test results. Since VRA results for both ears on an infant were included in the two-phase data if at least one ear was selected for verification, verification probabilities for the ears are unknown, and thus, need to be estimated. We consider empirically estimated probabilities. Specifically, empirical estimates are 1.0 for ears of infants with DPOAE test results greater than the 80th quantile of the DPOAE results, and estimates are the observed fraction of ears receiving verification below the threshold for those infants with DPOAE test results below the threshold.

7.4.3 Estimation of Accuracy and Summary Indices

The bias correction methods described above were used to estimate the prevalence of hearing impairment. Since an infant contributes data for each of two ears, we use the dependent data variance estimator (Section 5.9) to obtain variance estimates and corresponding 95% confidence intervals for disease prevalence, TP(c), and FP(c). In addition, bias-corrected estimates of TP(c) and FP(c) for each observed cutpoint c were calculated using each of the methods. Then an empirical bias-corrected ROC curve was obtained by plotting the bias-corrected TP(c) and FP(c) for all c. AUC was then estimated for each ROC curve.

Table 7.1: Number of ears and infants tested by DPOAE, TEOAE, and VRA in the full data and two-phase data.

	Ears		Infants	
	DPOAE & TEOAE	VRA	DPOAE & TEOAE	VRA
Full data	5,101	5,101	2,762	2,762
Two-phase data	5,101	2,707	2,762	1,458

7.5 Data

Table 7.1 summarizes the number of observations in the full data and two-phase data. In the full data TEOAE, DPOAE, and VRA test results were available for 5,101 ears which corresponds to 2,762 infants. The two-phase data consists of the same TEOAE and DPOAE test results, but only a subset of VRA results. In particular, the two-phase data only contains the full data VRA test results on 53.1% of the ears and 52.8% of the infants. Therefore, the two-phase data tracks and tests with VRA 48.2% fewer infants than the full data. This can result in a substantial reduction in cost.

7.6 Results

7.6.1 Full Data

Results from the NHSS have not yet been published. Of the 5,101 ears in the full data 147 (2.9%) were classified as hearing impaired by VRA. A variance estimate and 95% confidence interval for this prevalence estimate is provided in Table 7.2. TEOAE and DPOAE tests were performed on infants at an average of 38.5 weeks gestational age (range 29.6 - 53.1). That is, screening tests were on average given to infants 1.5 weeks premature.

DPOAE test results ranged from -37dB to 37.6dB while TEOAE test results

Table 7.2: Estimated prevalence, variance, and 95% confidence intervals for audiology data when verification is a function of DPOAE.

Method	Estimate	Variance	95% CI
Full Data	0.029	0.040	(0.023, 0.034)
CC	0.034	0.054	(0.027, 0.040)
BG-A	0.028	0.035	(0.023, 0.033)
BG-$\bar{A}$	0.029	0.036	(0.024, 0.034)
MS-A	0.028	0.038	(0.023, 0.034)
MS-$\bar{A}$	0.029	0.040	(0.024, 0.034)
IPW-E	0.028	0.040	(0.022, 0.032)
SP-E-A	0.027	0.040	(0.021, 0.032)
SP-E-$\bar{A}$	0.027	0.041	(0.021, 0.032)

ranged from -36.4dB to 27.4dB. Figure 7.1, a plot of empirical ROC curves for DPOAE and TEOAE, suggests that DPOAE is more accurate than TEOAE except for low false positive rates where they appear to have similar accuracy. AUC estimates corresponding to these ROC curves are 0.632 and 0.597 for DPOAE and TEOAE, respectively.

7.6.2 Two-phase Data

Of the 2,707 ears in the two-phase data for which VRA test results are available, 91 were considered hearing impaired by the VRA test. Thus, the CC estimate of the prevalence of hearing impairment in the two-phase data is 3.4% which is slightly higher than the full data estimate of 2.9%. CC estimated ROC curves are presented in Figure 7.2 along with the full data ROC curves. Clearly, CC overestimates the full data ROC curve for each screening test. Estimates of AUC corresponding to the CC ROC curves are roughly 0.04 larger than those observed with the full data (Table 7.3). This suggests that using the CC approach with the two-phase data could

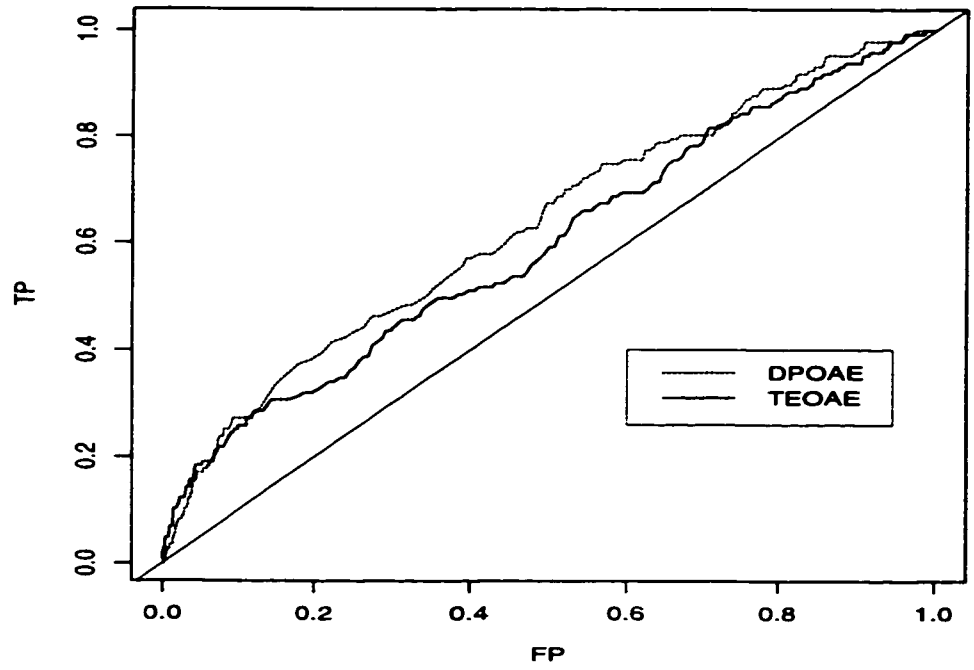

Figure 7.1: ROC curves for the TEOAE and DPOAE tests using the full data.

lead to different conclusions about the accuracy of the tests. This is not surprising since the two-phase data was obtained using a verification biased sampling procedure.

Now we investigate how the bias-correction methods proposed in this dissertation and discussed again in Section 7.4 perform on the two-phase data in which verification bias exists. The process of selecting models for disease and verification probabilities conditional on the screening test results and auxiliary data are provided in Section 7.4.

First consider the setting where T =DPOAE and we are interested in modelling

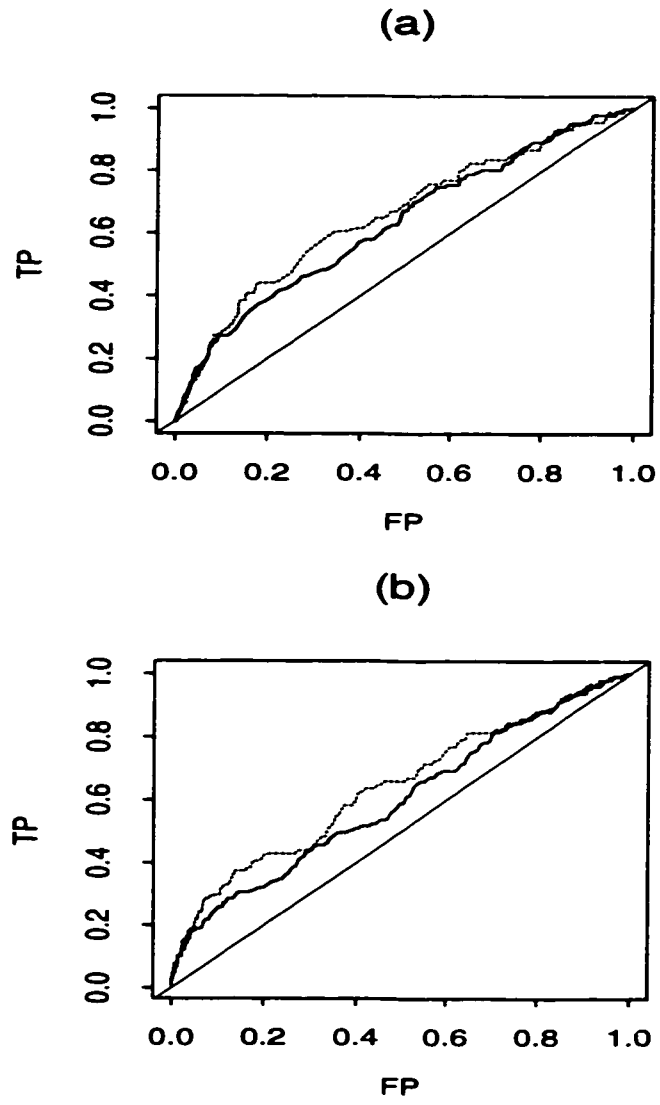

Figure 7.2: Full data (solid line) and CC (dashed lines) ROC curves for (a) DPOAE and (b) TEOAE tests.

$P(D|T)$ for the BG, MS, and SP estimators where D constitutes hearing impairment. Figure 7.3 (a) displays a plot of the GAM fit where smoothing splines are used to model T =DPOAE as a non-parametric smooth and tick marks along the x-axis represent the data. This plot suggests that a linear form of T may be adequate except for a few extreme high values of DPOAE. Therefore, we use a logistic model with the linear form of T as a predictor to estimate the probability of hearing impairment conditional on DPOAE test results. Furthermore, when we considered the model that also included auxiliary data A=TEOAE, results of GAM suggested that both $T = DPOAE$ and $A = TEOAE$ be fit as linear terms.

Different forms of the predictors were suggested by GAM results when roles of TEOAE and DPOAE were reversed, namely T=TEOAE and A=DPOAE. For example, Figure 7.3 (b), which displays a plot of the GAM fit where s(TEOAE) indicates a smoothing spline was fit to TEOAE, suggests using a piecewise linear term. Specifically, we considered a model with a linear term for I[TEOAE ≤ -5dB] and a linear term for I[TEOAE > -5dB]. In addition to the piecewise linear fit for TEOAE, we considered a linear term for $A = DPOAE$ when fitting a model for $P(D|T, A)$.

Estimates of the prevalence of neonatal hearing loss are provided in Table 7.2 for the various bias-correction estimators when applied to the two-phase data. Estimates are similar for the scenario where T =TEOAE, A =DPOAE and the scenario where T =DPOAE, A =TEOAE so we chose to present results from the latter scenario. These results suggest that the bias-correction estimators appear to do a good job of estimating the full data prevalence. Since TEOAE is only somewhat informative about hearing impairment status, it is not surprising that the variance for the estimators that include auxiliary data A =TEOAE are only slightly more efficient than those that do not include these data. Consistent with simulation results, estimates for IPW and SP are more variable than those for BG and MS. Moreover, 95% confidence intervals for all the prevalence estimators contain the full data value of 0.029.

Verification bias appears to have a large affect on AUC estimation in the two-

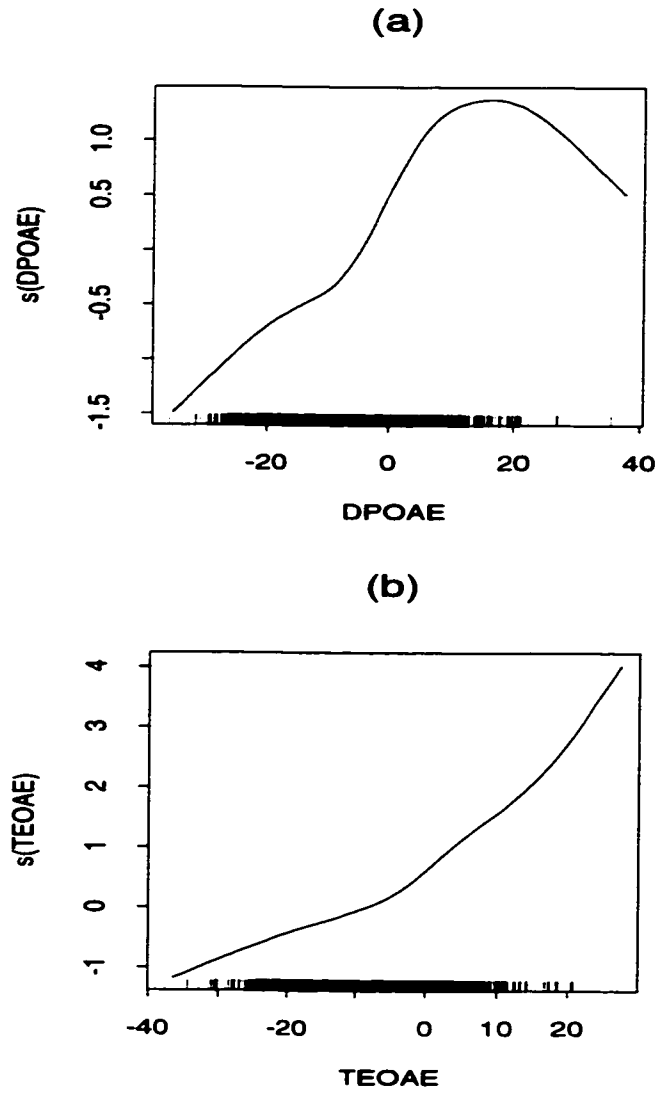

Figure 7.3: Plots of GAM fits where a smoothing spline, s(·), was fit to the predictors in a model for the probability of hearing impairment conditional on T. (a) T=DPOAE and (b) T=TEOAE.

Table 7.3: AUC estimates for DPOAE and TEOAE when verification is a function of DPOAE. Various models were used to to estimate the hearing impairment probabilities. "GAM" denotes GAM models where predictors were fit using smoothing splines. "Linear" denotes linear forms of DPOAE and TEOAE included in a logistic model; whereas, "PL" indicates a piecewise linear term was fit to TEOAE and a linear term to DPOAE.

	DPOAE		TEOAE	
Method	Linear	GAM	PL	GAM
Full Data	0.632	0.632	0.597	0.597
CC	0.659	0.659	0.641	0.641
BG-A	0.638	0.636	0.618	0.620
BG-$\bar{A}$	0.633	0.637	0.616	0.620
MS-A	0.641	0.637	0.614	0.617
MS-$\bar{A}$	0.640	0.638	0.615	0.618
IPW-E	0.624	0.624	0.608	0.608
SP-E-A	0.630	0.628	0.609	0.610
SP-E-$\bar{A}$	0.629	0.628	0.603	0.604

phase data. Estimates of the area under the ROC curves for DPOAE and TEOAE are provided in Table 7.3. As was demonstrated in Figure 7.2, CC clearly overestimates the ROC curve and the AUC. Table 7.3 also contains estimates of AUC for DPOAE and TEOAE resulting from the proposed bias-correction estimators. The columns labeled "GAM" correspond to estimators in which a GAM model with smoothing splines fit to the predictors were used to estimate conditional hearing impairment probabilities. The column labeled "Linear" denotes linear forms of DPOAE and TEOAE included in a logistic model used to estimate conditional hearing impairment probabilities; whereas, "PL" indicates the logistic model contains a piecewise linear term for TEOAE and a linear term for DPOAE.

Since GAM uses a non-parametric smooth fit of the data, by comparing results in the other columns to the results in the GAM column we are able to assess how much the parametric form of the predictors we chose affects estimates of AUC. It is reassuring that estimates from estimators that use GAM models and estimators that use the logistic models we chose yield similar results.

Clearly in Tables 7.3-7.5 AUC, TP(c), and FP(c) estimates obtained using the bias-correction estimators in the two-phase data are much closer to the full data estimates than the CC estimates. Since BG and MS estimators and IPW and SP estimators are in the same class of estimators, it is not surprising that they yield similar estimates. Estimators that use the auxiliary data are similar to those that ignore these data. This is probably due to the fact that DPOAE and TEOAE are not very accurate tests. The bias-correction methods resulted in similar AUC estimates for DPOAE. SP and IPW estimates of AUC and $TP(c)$ corresponding to TEOAE more closely resemble the full data value than the other methods. Conversely, the MS estimate of DPOAE $TP(c)$ is closest to the full data estimate.

It is apparent in Figure 7.4 that bias-corrected estimates of ROC curves for DPOAE all closely resemble the full data ROC curve. ROC curves are not presented for BG, MS, and SP that do not use auxiliary data since they are similar to

Table 7.4: TP(c) estimates (95% confidence interval) where c is such that the full data FP=0.20. Verification is a function of DPOAE.

Method	DPOAE	TEOAE
Full Data	0.381 (0.308, 0.455)	0.320 (0.247, 0.393)
CC	0.440 (0.343, 0.536)	0.429 (0.333, 0.524)
BG-A	0.362 (0.296, 0.429)	0.407 (0.343, 0.470)
BG-$\bar{A}$	0.344 (0.288, 0.401)	0.396 (0.337, 0.456)
MS-A	0.392 (0.317, 0.467)	0.356 (0.282, 0.430)
MS-$\bar{A}$	0.384 (0.310, 0.458)	0.353 (0.281, 0.426)
IPW-E	0.404 (0.306, 0.502)	0.314 (0.226, 0.402)
SP-E-A	0.405 (0.306, 0.504)	0.313 (0.225, 0.402)
SP-E-$\bar{A}$	0.405 (0.307, 0.502)	0.309 (0.221, 0.397)

Table 7.5: FP(c) estimates (95% confidence interval) where c is such that the full data FP=0.20. Verification is a function of DPOAE.

Method	DPOAE	TEOAE
Full Data	0.195 (0.186, 0.204)	0.197 (0.188, 0.206)
CC	0.217 (0.204, 0.230)	0.203 (0.190, 0.215)
BG-A	0.196 (0.185, 0.207)	0.203 (0.192, 0.214)
BG-$\bar{A}$	0.196 (0.186, 0.205)	0.203 (0.194, 0.213)
MS-A	0.195 (0.185, 0.204)	0.196 (0.187, 0.206)
MS-$\bar{A}$	0.196 (0.186, 0.205)	0.200 (0.187, 0.205)
IPW-E	0.196 (0.186, 0.206)	0.190 (0.177, 0.203)
SP-E-A	0.195 (0.185, 0.204)	0.198 (0.188, 0.207)
SP-E-$\bar{A}$	0.195 (0.185, 0.204)	0.198 (0.188, 0.207)

the curves provided. Estimated ROC curves for TEOAE (Figure 7.5) follow the full data ROC curve except for the region of FP values between 0.3 and 0.5 where they tend to overestimate the full data curve.

As mentioned in the previous chapter, although by definition ROC curves are monotone increasing, SP estimated ROC curves are not necessarily monotone. Although not an extreme example, non-monotonicity is apparent in Figure 7.6 (a). In order to yield monotone SP ROC curves, we smoothed the curves using the isotonic regression "pool adjacent violators" algorithm (Robertson et al., 1988). Details are provided in Appendix D. Figure 7.6 (b) displays the resulting smoothed SP ROC curve for DPOAE.

7.6.3 Assessing MAR Assumption

The bias-corrected estimators considered in this chapter rely on the MAR assumption (Section 7.2). In practice we recommend performing a sensitivity analysis to see how much of a change occurs in estimates by adding or removing auxiliary data in the disease and verification probability models. Although in practice we would not know the true hearing status for all subjects, with these data we do. Therefore, we can compare the estimated probability of hearing impairment obtained using the full data with those obtained using the two-phase data. Similar estimated probabilities would suggest the MAR assumption holds.

Consider the scenario when T =DPOAE. Figure 7.7 is a plot of estimated disease probabilities resulting from a linear logistic model fit to the two-phase data versus corresponding probabilities resulting from a linear logistic model fit to the full data. Estimated probabilities are close to a vertical line through the origin suggesting the MAR assumption holds. The MAR assumption appears to hold in the other scenarios as well.

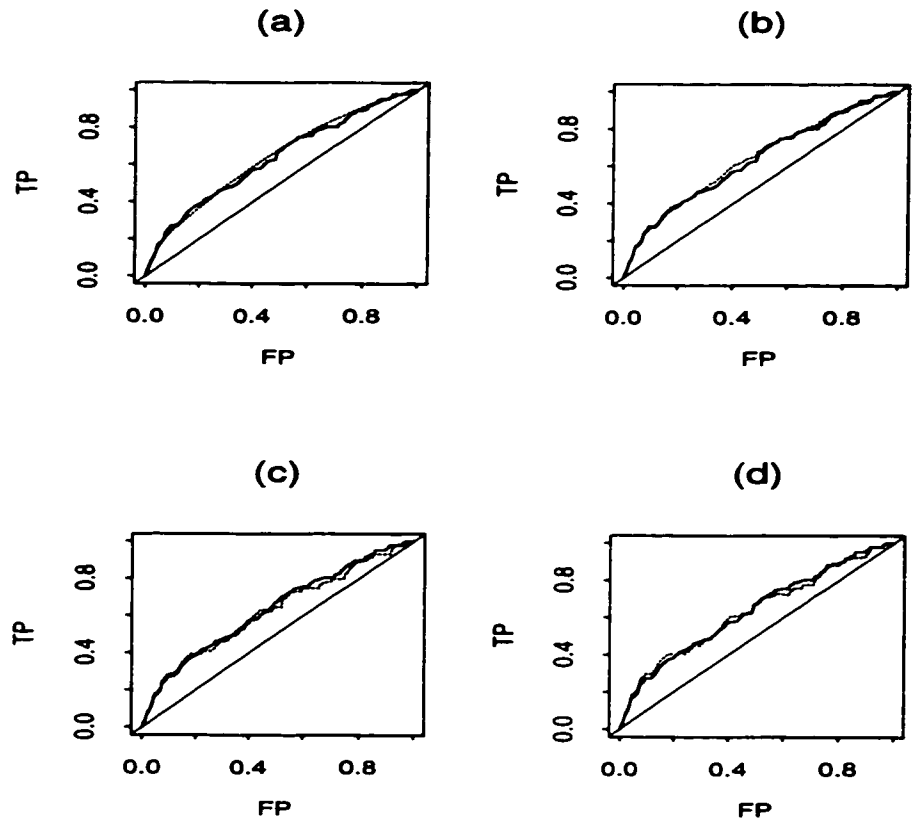

Figure 7.4: Full data ROC curves (solid lines) for the DPOAE test along with dashed line (a) BG-A, (b) MS-A, (c) IPW-E, and (d) SP-E-A ROC curves.

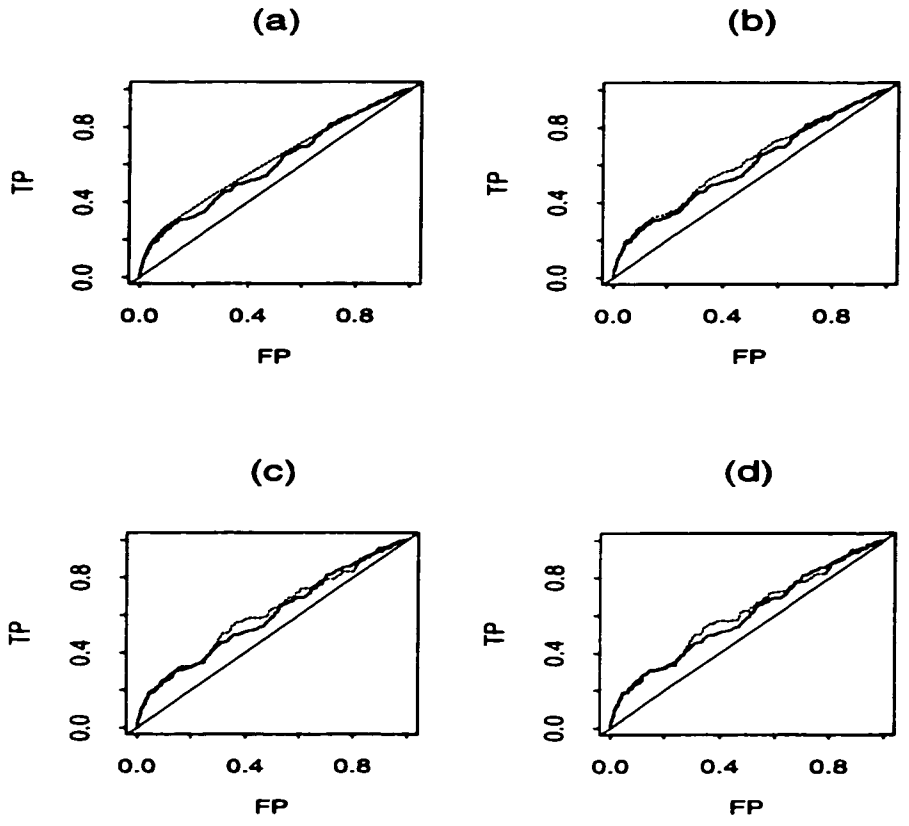

Figure 7.5: Full data ROC curves (solid lines) for the TEOAE test along with dashed line (a) BG-A, (b) MS-A, (c) IPW-E, and (d) SP-E-A ROC curves.

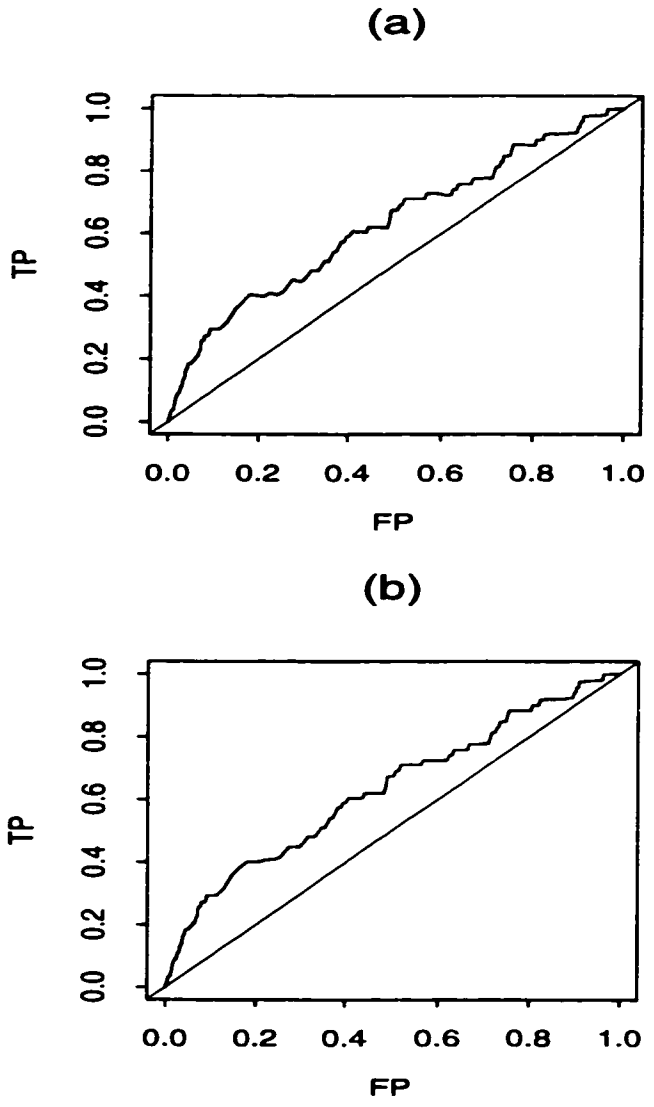

Figure 7.6: SP-E-$\bar{A}$ estimate of DPOAE ROC curve using the two-phase data. (a) Empirical ROC curve (b) Curve smoothed using isotonic regression.

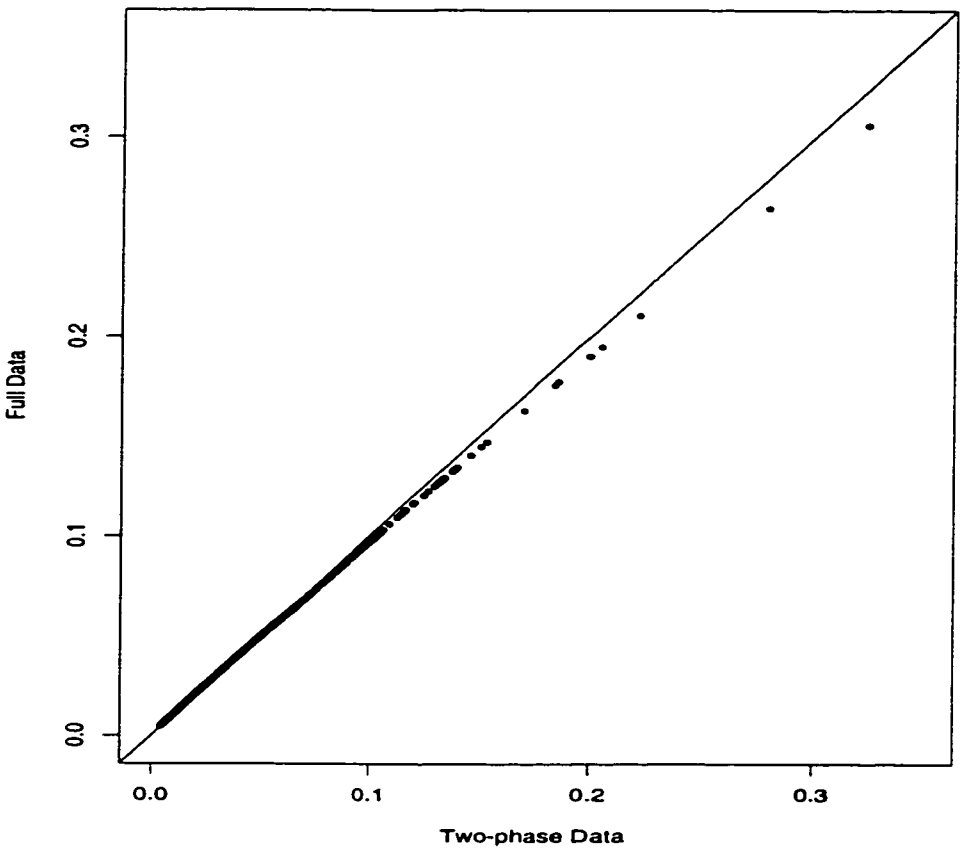

Figure 7.7: Plot of estimated disease probabilities resulting from a linear logistic model fit to the two-phase data versus corresponding probabilities resulting from a linear logistic model fit to the full data.

7.7 Discussion

When applied to the two-phase data the bias-correction methods proposed in this dissertation yielded estimates of prevalence, $TP(c)$, $FP(c)$, empirical ROC curve, and AUC that closely resemble the full data estimates. Therefore, in practice a two-phase design that does not require all infants to to receive the definitive test for hearing status could be implemented. Overall semi-parametric efficient estimators

that, unlike BG and MS estimators, do not require a correct model for the probability of disease (i.e. hearing impairment) appear to perform best with these data. The isotonic regression "pool adjacent violators" algorithm can be used in practice to yield monotone SP ROC curves.

In this chapter we chose to use GAM models fit with smoothing splines to suggest parametric logistic models. Although we attempt to find the parametric model that most resembles the GAM fit, it is likely that the parametric model will not fit the data as well as GAM. In Chapter 6 we investigated the effect of misspecifying a model for the probability of disease and found that biased results were obtained for BG and MS estimators if that model was not specified correctly. Since it is important to fit this model well and we are interested in prediction and, thus, interpretation of the coefficients is not of interest, we could use GAM directly to predict disease probabilities. The disadvantage to GAM, however, is that the asymptotic distribution theory developed in Chapter 5 cannot easily be applied because GAM cannot be written as a vector of estimating functions. Instead we could use natural splines in logistic regression which can easily be written as estimating functions. Natural splines offer a flexible approach for fitting disease and verification probability models and yield results similar to GAM with these data.

In this chapter we considered a subset of the NHSS that requires 48.2% fewer infants be tracked and tested with VRA. Moreover, the two-phase data were selected by over-sampling infants more likely to be hearing impaired as suggested by the result of the screening test, DPOAE, and, thus, induced verification bias. It was shown that the methods proposed in this dissertation can correct for this bias. The issue of determining the optimal selection criteria to be specified in the two-phase study protocol warrants more consideration. This will be discussed in more detail in the next chapter.

Chapter 8

CONCLUSIONS

8.1 Dissertation Contributions

This dissertation addresses the important and previously untackled problem of assessing the accuracy of a continuous medical diagnostic or screening test in the presence of verification bias (Zhou, 1998). Several new methods for assessing accuracy of a continuous test in the presence of verification bias were developed and compared in this dissertation. First we constructed bias-corrected estimators of TP(c) and FP(c) for each observed cutpoint c. Then an empirical bias-corrected ROC curve was obtained by plotting the bias-corrected TP(c) and FP(c) for all c. This corrected ROC curve can then be used to assess the accuracy of the test either by visual inspection or by estimating summary indices such as (TP(c), FP(c)) at a fixed cutpoint or AUC.

Verification bias has previously been shown to lead to biased operating points on an ROC curve for tests with ordinal results (Hunink et al., 1990, Hunink et al., 1993). In this dissertation we demonstrate that verification bias also leads to a shift or bias in operating points for tests with continuous results. Furthermore, we also exhibit that the proposed estimators successfully correct for this bias.

Asymptotic distribution theory for estimators of disease prevalence, TP(c), and FP(c) in the presence of verification bias was developed in Chapter 5. In practice we can use the theory developed to construct confidence intervals for $\widehat{TP}(c)$ and $\widehat{FP}(c)$ along with joint confidence regions for $(\widehat{TP}(c), \widehat{FP}(c))$. Simulation studies (Chapter 6) were employed to study small and large sample properties of the proposed estimators.

We observed that the Begg and Greenes (BG) estimator which imputes values for all subjects and the mean score (MS) estimator which conversely only imputes values for subjects missing disease status have similar properties and good efficiency. In addition, we noted that these estimators require estimates of disease probabilities conditional on the test results and auxiliary data be correct. Since mean score estimators use observed disease status for those subjects receiving verification, mean score estimators were observed to be more robust to model misspecification than the BG estimator.

The inverse probability weighting estimator (IPW) weights each observation in the verification group by the inverse of the sampling fraction (i.e. probability the subject was selected for verification). This estimator is consistent provided estimates of verification probabilities conditional on the test results and auxiliary data are correct. This estimator was observed to be less efficient than the BG and MS estimators.

We also considered a semi-parametric efficient estimator (SP) which was constructed by augmenting the inverse probability weighting estimator. Similar to the IPW estimator, the SP estimator is consistent provided verification probabilities conditional on the test results and auxiliary data are correct. By construction, the semi-parametric efficient estimator is at least as efficient as the IPW estimator. However, the semi-parametric efficient estimator was observed to be less efficient than the BG and MS estimators. One disadvantage to the semi-parametric efficient estimator is that it can result in non-monotone ROC curves. In practice, however, SP ROC curves do not appear to deviate dramatically from monotonicity and isotonic regression can be used to yield monotone SP ROC curves.

The estimators we propose are easy to implement only requiring low dimensional models to estimate disease and verification probabilities. In retrospective or observational studies where verification probabilities are unknown, we recommend the use of MS estimators since they were more robust to model misspecification than BG estimators. On the other hand, we suggest future studies are designed so that verifi-

cation probabilities are known or can be estimated reasonably well and SP estimators be used to assess accuracy.

8.2 Areas of Future Research

8.2.1 Asymptotic Distribution Theory for ROC Curves and AUC

The asymptotic distribution theory developed in Chapter 5 allows one to calculate pointwise confidence intervals for TP(c) and FP(c) and joint confidence regions for TP(c) and FP(c). We can easily extend the theory to a finite number of cutpoints by considering a vector of estimating functions corresponding to each of the cutpoints. However, this approach cannot accommodate the infinite number of cutpoints required to estimate the empirical ROC curve and AUC for continuous tests. Therefore, we must consider a different approach for the development of asymptotic distribution theory for ROC curves and AUC. The use of empirical process theory appears to be promising.

8.2.2 Covariate Effects

Often it is of interest to evaluate effects of factors that may influence test accuracy. Such factors may include characteristics of the study subjects, operating conditions for the test, or characteristics of the disease. For example, in neonatal hearing screening the gestational age of the infant may affect accuracy of the test. A regression modelling framework has been proposed for continuous tests when verification bias does not exist (Pepe, 1997, Pepe, 1998). An area of future research is to extend this framework to the setting where verification bias exists.

8.2.3 Optimal Study Design

The protocol specifying the subjects to receive disease verification is often under the control of the investigators designing the study. The dilemma is how to "opti-

mally" design the study. If the budget of the study is fixed, then an optimal design would maximize the precision of estimation. On the other hand, if the study protocol requires a fixed pre-specified precision, then the design with the minimum cost is optimal. Reilly & Pepe (1995) propose optimal sampling strategies for discrete data. Identifying the optimal two-phase study design for estimators proposed in this dissertation where we consider continuous data warrants further attention.

8.2.4 Bootstrapping Two-phase Designs

The statistical literature on bootstrapping data from two-phase designs appears to be limited. Does it make a difference if you bootstrap all subjects or if you bootstrap conditional on verification status? Clayton et al. (1998) applies both approaches to several realizations from a simulation study, but makes no comment to the validity of either approach. It would be useful to compare and determine the validity of different approaches for bootstrapping two-phase designs.

BIBLIOGRAPHY

ALONZO, T. A. & PEPE, M. S. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* **18**, 2987–3003.

BAMBER, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387–415.

BARR, J. T. & SCHUMAKER, G. E. (1984). Development of ROC curves and probability estimates for pharmacokinetic decision making: an application to theophylline toxicity. *Proceedings of the American Society of Hospital Pharmacists*, Dallas, Texas.

BATES, A. S., MARGOLIS, P. A., & EVANS, A. T. (1993). Verification bias in pediatric studies evaluating diagnostic tests. *Journal of Pediatrics* **122**, 585–590.

BEGG, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6**, 411–423.

BEGG, C. B. & GREENES, R. A. (1983). Assessment of diagnostic tests when disease is subject to selection bias. *Biometrics* **39**, 207–216.

CAMPBELL, G. (1994). General methodology I advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* **13**, 499–508.

CLAYTON, D., DUNN, G., PICKLES, A., & SPIEGELHALTER, D. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B* **60**, 71–87.

DIAMOND, G. A., ROZANSKI, A., FORRESTER, J. S., MORRIS, D., POLLOCK, B. H., STANILOFF, H. M., BERMAN, D. S., & SWAN, H. J. C. (1986). A model for assessing the sensitivity and specificity of tests subject to selection bias. *Journal of Chronic Diseases* **39**, 343–355.

DORFMAN, D. D. & ALF, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating data. *Journal of Mathematical Psychology* **6**, 487–496.

DRUM, D. E. & CHRISTACOPOULOS, J. S. (1972). Hepatic scintigraphy in clincial decision making. *Journal of Nuclear Medicine* **13**, 908–915.

FAHEY, M. T., IRWIG, L., & MACASKILL, P. (1995). Meta-analysis of Pap test accuracy. *American Journal of Epidemiology* **141**, 680–689.

FEIGL, P., BLUMENSTEIN, B., THOMPSON, I., CROWLEY, J., WOLF, M., KRAMER, B. S., CHARLES A. COLTMAN, J., BRAWLEY, O. W., & FORD, L. G. (1995). Design of the Prostate Cancer Prevention Trial (PCPT). *Controlled Clinical Trials* **16**, 150–163.

FOUTZ, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72**, 147–148.

GRAY, R., BEGG, C. B., & GREENES, R. A. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making* **4**, 151–164.

GREENES, R. A. & BEGG, C. B. (1985). Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selective verified patients. *Investigative Radiology* **20**, 751–756.

GREENHOUSE, S. W. & MANTEL, N. (1950). The evaluation of diagnostic tests. *Biometrics* .

HANLEY, J. A. (1989). Receiver operating characteristic (ROC) curve methodology. *Critical Reviews in Diagnostic Imaging* **29**, 307–335.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HORN, R. A. & JOHNSON, C. R. (1991). *Matrix Analysis.* Cambridge University Press, Cambridge.

HORVITZ, D. G. & THOMPSON, D. J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

HUI, S. L. & ZHOU, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**, 354–370.

HUNINK, M. G. M., POLAK, J. F., BARLAN, M. M., & O'LEARY, D. H. (1993). Detection and quantification of carotid artery stenosis: efficacy of various doppler velocity parameters. *American Journal of Roentgenology* **160**, 619–625.

HUNINK, M. G. M., RICHARDSON, D. K., DOUBILET, P. M., & BEGG, C. B. (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. *Medical Decision Making* **10**, 201–211.

KENNEDY, W. J. & GENTLE, J. E. (1980). *Statistical Computing.* Marcel Dekker, New York.

LITTLE, R. J. & RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley, New York.

MARSAGLIA, G. (1973). *Random Number Package: "Super-Duper".* School of Computer Science, McGill University.

MARSHALL, V., WILLIAMS, D. C., & SMITH, K. D. (1984). Diaphanography as a means of detecting breast cancer. *Radiology* **150**, 339–343.

MCCLISH, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

MCCULLAGH, P. (1980). Regression methods for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

MCNEIL, B. J., SANDERS, R., ALDERSON, P. O., HESSEL, S. J., FINBERG, H., SIEGELMAN, S. S., ADAMS, D. F., & ABRAMS, H. L. (1981). A prospective study of computed tomography, ultrasound and gallium imaging in patients with fever. *Radiology* **139**, 647–653.

NEYMAN, J. (1938). Contributions to the theory of sampling of human populations. *Journal of the American Statistical Association* **33**, 101–116.

NIH CONSENSUS STATEMENT (1993). Early identification of hearing impairment in infants and young children.

PEPE, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.

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

PEPE, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135.

PEPE, M. S., REILLY, M., & FLEMING, T. R. (1994). Auxiliary outcome data and the mean-score method. *Journal of Statistical Planning and Inference* **42**, 137–160.

PHILBRICK, J. T., HORWITZ, R. I., & FEINSTEIN, A. R. (1980). Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. *The American Journal of Cardiology* **46**, 807–812.

PICKLES, A., DUNN, G., & VÁZQUEZ-BARQUERO, J. L. (1995). Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research* **4**, 73–89.

RANSOHOFF, D. F. & FEINSTEIN, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* **299**, 926–930.

REID, M. C., LANDIS, M. S., & FEINSTEIN, A. R. (1995). Use of methodologic standards in diagnostic test research: getting better but still not good. *JAMA* **274**, 645–651.

REILLY, M. & PEPE, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

ROBERTSON, T., WRIGHT, F. T., & DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference.* John Wiley and Sons.

ROBINS, J. M., ROTNITZKY, A., & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

ROTNITZKY, A. & ROBINS, J. M. (1995). Efficient semiparametric estimation with missing outcomes and surrogate data. Technical report, Departments of Epidemiology and Biostatistics, Harvard School of Public Health.

ROTNITZKY, A., ROBINS, J. M., & SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

SCHOUW, Y. T. V. D., STRAATMAN, H., & VERBEEK, A. L. M. (1994). ROC curves and the areas under them for dichotomized tests: empirical findings for logistically and normally diagnostic test results. *Medical Decision Making* **14**, 374–381.

TENENBEIN, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* **65**, 1350–1361.

THOMPSON, M. L. & ZUCCHINI, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* **8**, 1277–1290.

WALTER, S. D. & IRWIG, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* **41**, 923–937.

WIEAND, S., GAIL, M. H., JAMES, B. R., & JAMES, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585–592.

ZHAO, L. P. & LIPSITZ, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine* **11**, 769–782.

ZHOU, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communication in Statistics: Theory and Methods* **22**, 3177–3198.

ZHOU, X. H. (1996). Nonparametric ML estimate of an ROC area corrected for verification bias. *Biometrics* **52**, 310–316.

ZHOU, X. H. (1998). Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* **7**, 337–353.

ZHOU, X.-H. & RODENBERG, C. A. (1998). Estimating an Roc curve in the presence of non-ignorable verification bias. *Communications in Statistics, Part A - Theory and Methods* **27**, 635–657.

ZWEIG, M. H. & CAMPBELL, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561–577.

Appendix A

PROOFS OF LEMMAS

A.1 Proof of Lemma 5.1

This proof aims to show that averages of the uniformly bounded $U_i(\beta)$, $\frac{\partial}{\partial\beta}U_i(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_i(\beta)$ are also uniformly bounded. Consider any function g such that $|g_i| \leq M \quad \forall\beta \in N_\delta(\beta_0)$. Then $\forall\beta \in N_\delta(\beta_0)$

$$\begin{aligned}\left|n^{-1}\sum_{i=1}^{n} g_i\right| &\leq n^{-1}\sum_{i=1}^{n}|g_i| \\ &\leq n^{-1}\sum_{i=1}^{n} M \\ &= M.\end{aligned}$$

Since we considered an arbitrary bounded function g, averages of uniformly bounded $U_i(\beta)$, $\frac{\partial}{\partial\beta}U_i(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_i(\beta)$ are also uniformly bounded.

A.2 Proof of Lemma 5.2

To prove $\frac{\partial}{\partial\beta}E(\frac{\partial}{\partial\beta}U_i(\beta))$ exists and is uniformly bounded in $N_\delta(\beta_0)$, we must show, for arbitrary $\beta \in N_\delta(\beta_0)$, $\frac{\partial}{\partial\beta}E(\frac{\partial}{\partial\beta}U_i(\beta))$ exists and $|\frac{\partial}{\partial\beta}E(\frac{\partial}{\partial\beta}U_i(\beta))| \leq M$, where M is a constant.

Fix $\beta^\star \in N_\delta(\beta_0)$. Let β_k be a sequence such that $\beta_k \to_p \beta^\star$ as $k \to \infty$. Also let $C = \frac{\partial}{\partial\beta}(\frac{\partial}{\partial\beta}U_i(\beta^\star))$ and $C_k = \frac{\frac{\partial}{\partial\beta}U_i(\beta^\star) - \frac{\partial}{\partial\beta}U_i(\beta_k)}{\beta^\star - \beta_k}$. By definition $\frac{\partial}{\partial\beta}E\left(\frac{\partial}{\partial\beta}U_i(\beta)\right)$ equals $lim_{k\to\infty}E(C_k)$ so it suffices to show that as $k \to \infty$ $E(C_k)$ exists and $|E(C_k)| \leq M$.

By the definition of the derivative, C_k converges to C as $k \to \infty$. Since C is uniformly bounded in $N_\delta(\beta_0)$, $|C| \leq M$. So for large enough k, $|C_k|$ is bounded and,

therefore, is integrable. Since C_k converges to C as $k \to \infty$ and C_k is integrable, the Dominated Convergence Theorem yields that as $k \to \infty$ $E(C_k)$ exists and $|E(C_k)| \leq M$. Hence, the proof is complete.

A.3 Proof of Lemma 5.3

In this proof we need to show that $\left|\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right|\right|$ is bounded in probability away from 0 $\forall\beta \in N_\delta(\beta_0)$ provided the following conditions are true:

i. $\frac{\partial}{\partial\beta}\left(\frac{\partial}{\partial\beta}U_i(\beta_0)\right)$ is uniformly bounded in $N_\delta(\beta_0)$

ii. $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0) \to_p E\left(\frac{\partial}{\partial\beta}U_n(\beta_0)\right)$

iii. $E\left(-\frac{\partial}{\partial\beta}U_i(\beta_0)\right)$ is positive definite.

Given $\epsilon > 0$, we can choose $N_\delta(\beta_0)$ so that $||\beta - \beta_0|| < \frac{\epsilon}{M}$ $\forall\beta \in N_\delta(\beta_0)$. By the addition and subtraction of $n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)$ and the triangle inequality,

$$\begin{aligned}\lim_{n\to\infty}\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^\star)\right|\right| &= \lim_{n\to\infty}\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^\star) - n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0) + n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right|\right| \\ &\geq \lim_{n\to\infty}\left|\left|\left(n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right)\right|\right| - \lim_{n\to\infty}\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^\star) - n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right|\right|.\end{aligned}$$

Using the mean value theorem the preceeding expression is equal to

$$\lim_{n\to\infty}\left|\left|n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta_0)\right|\right| - \lim_{n\to\infty}\left|\left|(\beta^\star - \beta_0)\frac{\partial}{\partial\beta}n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^\dagger)\right|\right| \qquad \text{(A.1)}$$

where $\beta^\dagger$ is between $\beta^\star$ and β_0. Furthermore, since the norm of the product is the product of the norms, $||(\beta^\star - \beta_0)||\left|\left|\frac{\partial}{\partial\beta}n^{-1}\frac{-\partial}{\partial\beta}U_n(\beta^\dagger)\right|\right| < \epsilon$ by (i) and the definition of $N_\delta(\beta_0)$, and $n^{-1}\frac{\partial}{\partial\beta}U_n(\beta_0) \to_p E(\frac{\partial}{\partial\beta}U_n(\beta_0))$ by (ii). This implies (A.1) is greater than or equal to $\left|\left|E\left(-\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right|\right| - \epsilon$. Finally, this expression is greater than 0 because $\left|\left|E\left(-\frac{\partial}{\partial\beta}U_i(\beta_0)\right)\right|\right|$ is positive definite by (iii). Therefore, $\left|\left|n^{-1}\frac{\partial}{\partial\beta}U_n(\beta)\right|\right|$ is bounded in probability away from 0 and the proof is complete.

Appendix B

PROPERTIES REQUIRED IN ALTERNATIVE THEORY DERIVATION FOR BG

The alternative distribution theory (Theorem 5.4) for the BG prevalence estimator requires existence and boundedness (P5) and conditional convergence (P6). Here we discuss settings when these properties are satisfied.

B.1 Existence and boundedness

Expressions for $g(\alpha, T_i, A_i)$, $\frac{\partial}{\partial \alpha} g(\alpha, T_i, A_i)$, and $\frac{\partial^2}{\partial \alpha \partial \alpha^T} g(\alpha, T_i, A_i)$ are derived in Appendix C for a logistic model. It is clear from these expressions that they are bounded if all components of $X = (1, T, A)$ and α are bounded. If we choose $N_\delta(\alpha_0)$ to be bounded, then $\forall \alpha \in N_\delta(\alpha_0)$, α will be bounded. X will be bounded provided (T, A) is bounded. Therefore, $g(\alpha, T_i, A_i)$, $\frac{\partial}{\partial \alpha} g(\alpha, T_i, A_i)$, and $\frac{\partial^2}{\partial \alpha \partial \alpha^T} g(\alpha, T_i, A_i)$ are uniformly bounded in $N_\delta(\alpha_0)$ provided $N_\delta(\alpha_0)$ is bounded and (T, A) is bounded. By assumptions A3 and A4 α, T, and A are bounded. Hence the result.

B.2 Conditional convergence

Recall that property P6 states that $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \rightarrow_d N\left(0, \not{\sum}_\alpha\right)$ where the convergence in distribution is conditional on $\{(T_i, A_i), \quad i = 1, \cdots, n\}$. We will show that P6 is satisfied when the corresponding score function and its derivatives are bounded, (T_i, A_i, D_i, V_i) is iid, and $\widehat{\alpha}$ corresponds to the MLE.

Let $S_\alpha(D_i | T_i, A_i)$ be the score contribution for the ith subject. A second-order

Taylor series expansion of $n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\widehat{\alpha}}(D_i|T_i, A_i)$ about α_0 yields:

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\widehat{\alpha}}(D_i|T_i, A_i) = n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\alpha_0}(D_i|T_i, A_i) + \frac{1}{2}(\widehat{\alpha} - \alpha_0) n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha_0} S_{\alpha_0}(D_i|T_i, A_i) + R_S \quad \text{(B.1)}$$

where

$$R_S = \frac{1}{2}(\widehat{\alpha} - \alpha_0) n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \alpha \partial \alpha^T} S_{\alpha^*}(D_i|T_i, A_i) \frac{1}{2}(\widehat{\alpha} - \alpha_0)$$

and α^* is between $\widehat{\alpha}$ and α_0. Since it can be shown that $\widehat{\alpha}$ is consistent for almost all (T, A) and the derivatives are bounded, R_S converges in probability to 0. By noting that $n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\widehat{\alpha}}(D_i|T_i, A_i) = 0$ and re-arranging (B.1) we get that

$$n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) = [-n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} S_{\alpha_0}(D_i|T_i, A_i)]^{-1} \; n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\alpha_0}(D_i|T_i, A_i) + o_p(1). \quad \text{(B.2)}$$

Finding the asymptotic distribution of the RHS of (B.2) conditional on (T_i, A_i) is, therefore, equivalent to finding the asymptotic conditional distribution of $n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0)$.

First, we will determine the asymptotic distribution of $n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\alpha_0}(D_i|T_i, A_i)$ conditional on (T, A). Given a sequence of $\{(T_i, A_i),\ i = 1, \cdots, n\}$, $S_{\alpha_0}(D_i|T_i, A_i)$ are independent with mean 0 and variance-covariance matrix equal to

$$I_i(\alpha_0) \equiv E\{S_{\alpha_0}(D_i|T_i, A_i) S_{\alpha_0}(D_i|T_i, A_i)^T | T_i, A_i\} \quad \text{(B.3)}$$

$$= -E\left\{\frac{\partial}{\partial \alpha} S_{\alpha_0}(D_i|T_i, A_i) | T_i, A_i\right\}. \quad \text{(B.4)}$$

The Lindeberg-Feller CLT for random vectors states that if

(i) $n^{-1} \sum_{i=1}^{n} I_i(\alpha_0) \to I(\alpha_0)$ and

(ii) $n^{-1} \sum_{i=1}^{n} E\{||S_{\alpha_0}(D_i|T_i, A_i)|T_i, A_i||^2 \;\; I[\; ||S_{\alpha_0}(D_i|T_i, A_i)|T_i, A_i|| > \frac{\epsilon}{2} \;]\}$

as $n \to \infty$, then $n^{-\frac{1}{2}} \sum_{i=1}^{n} S_{\alpha_0}(D_i|T_i, A_i) \to_d N(0, I(\alpha_0))$ where the convergence is conditional on the sequence $\{(T_i, A_i),\ i = 1, \cdots, n\}$. First, consider condition (i). By

the SLLN

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} I_i(\alpha_0) &\rightarrow E\{I_i(\alpha_0)\}\ \forall\ \{(T_i, A_i),\ i = 1, \cdots, n\} \text{ except a set of measure 0} \\ &= E\{E\{S_{\alpha_0}(D_i|T_i, A_i)S_{\alpha_0}(D_i|T_i, A_i)^T|T_i, A_i\}\} \\ &= E\{S_{\alpha_0}(D_i|T_i, A_i)S_{\alpha_0}(D_i|T_i, A_i)^T\} \\ &\equiv I(\alpha_0). \end{aligned} \tag{B.5}$$

Next, we will show that condition (ii) holds. Assuming a logistic model and bounded (T, A) and α, the score function is uniformly bounded so that the expectation in (ii) is 0 identically as soon as n is large enough. Since conditions (i) and (ii) are true as $n \rightarrow \infty$, the Lindeberg-Feller CLT implies that conditional on $\{(T_i, A_i),\ i = 1, \cdots, n\}$, $n^{-\frac{1}{2}}\sum_{i=1}^{n} S_{\alpha_0}(D_i|T_i, A_i) \rightarrow_d N(0, I(\alpha_0))$ for all (T, A) except for a set of measure 0.

Next, we will show that the other term on the right-hand-side of (B.2), $n^{-1}\sum_{i=1}^{n} -\frac{\partial}{\partial\alpha_0}S_{\alpha_0}(D_i|T_i, A_i)$, converges to $I(\alpha_0)$. To do this, we consider

$$\left| n^{-1}\sum_{i=1}^{n} -\frac{\partial}{\partial\alpha}S_{\alpha_0}(D_i|T_i, A_i) - n^{-1}\sum_{i=1}^{n} I_i(\alpha_0) + n^{-1}\sum_{i=1}^{n} I_i(\alpha_0) - I(\alpha_0) \right|$$

which is less than or equal to

$$\left| n^{-1}\sum_{i=1}^{n} -\frac{\partial}{\partial\alpha_0}S_{\alpha_0}(D_i|T_i, A_i) - n^{-1}\sum_{i=1}^{n} I_i(\alpha_0) \right| + \left| \sum_{i=1}^{n} I_i(\alpha_0) - I(\alpha_0) \right|$$

by the triangle-inequality. The WLLN for independent, non-identically distributed random variables implies that $|n^{-1}\sum_{i=1}^{n} -\frac{\partial}{\partial\alpha_0}S_{\alpha_0}(D_i|T_i, A_i) - n^{-1}\sum_{i=1}^{n} I_i(\alpha_0)| \rightarrow_p 0$ provided $n^{-2}\sum_{i=1}^{n} Var(\frac{\partial}{\partial\alpha_0}S_{\alpha_0}(D_i|T_i, A_i) \rightarrow 0$ as $n \rightarrow \infty$. Since the score is bounded, this condition is satisfied. $|n^{-1}\sum_{i=1}^{n} I_i(\alpha_0) - I(\alpha_0)| \rightarrow 0$ for this sequence of $\{(T_i, A_i),\ i = 1, \cdots, n\}$ because it is true for almost every sequence of (T, A) by SLLN. Hence, we have shown that $n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial\alpha_0}S_{\alpha_0}(D_i|T_i, A_i) \rightarrow_p I(\alpha_0)$ for all (T, A) except for a set of measure zero. The continuous mapping theorem implies that $\{n^{-1}\sum_{i=1}^{n} -S_{\alpha_0}(D_i|T_i, A_i)\}^{-1} \rightarrow_p I^{-1}(\alpha_0)$ for almost every sequence of (T, A).

We have shown that one term on the RHS of (B.2) converges in distribution and the other terms converge in probability so by Slutsky's theorem

$$n^{\frac{1}{2}}(\widehat{\alpha} - \alpha_0) \quad \rightarrow_d \quad N(0, I^{-1}(\alpha_0)) \tag{B.6}$$

$$\equiv \quad N(0, \not{\sum}_{\alpha}) \tag{B.7}$$

where the convergence is conditional on $\{(T_i, A_i),\ i = 1, \cdots, n\}$.

Appendix C

DERIVATION OF THE FIRST AND SECOND PARTIAL DERIVATIVES

If a logistic model is used to fit $P(D|T,A)$, then the probability that the ith subject is diseased given (T_i, A_i) is given by $g(\alpha, T_i, A_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ where $\eta_i = X_i\alpha$ and $X_i = (1, T_i, A_i)$. Using the chain rule, we can differentiate $g(\alpha, T_i, A_i)$ with respect to an arbitrary element, α_j, of α:

$$\begin{aligned}
\frac{\partial}{\partial\alpha_j}g(\alpha, T_i, A_i) &= \frac{\partial}{\partial\eta_i}g(\alpha, T_i, A_i)\frac{\partial\eta_i}{\partial\alpha_j} \\
&= \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}X_{ij} \\
&= \frac{g(\alpha, X_i)X_{ij}}{(1+\exp(\eta_i))}.
\end{aligned}$$

Similarly, the chain rule can be used to differentiate $g(\alpha, T_i, A_i)$ twice with respect to two arbitrary elements, α_j and α_k, of α:

$$\begin{aligned}
\frac{\partial^2 g(\alpha, T_i, A_i)}{\partial\alpha_j\partial\alpha_k} &= \frac{\partial}{\partial\alpha_k}\left[\frac{\partial g(\alpha, T_i, A_i)}{\partial\eta_i}\frac{\partial\eta_i}{\partial\alpha_j}\right] \\
&= \frac{\partial g(\alpha, T_i, A_i)}{\partial\eta_i}\frac{\partial^2\eta_i}{\partial\alpha_j\partial\alpha_k} + \frac{\partial\eta_i}{\partial\alpha_j}\frac{\partial}{\partial\alpha_k}\left(\frac{\partial g(\alpha, T_i, A_i)}{\partial\eta_i}\right) \\
&= 0 + \frac{\partial\eta_i}{\partial\alpha_j}\frac{\partial}{\partial\eta_i}\left(\frac{\partial g(\alpha, T_i, A_i)}{\partial\eta_i}\right)\frac{\partial\eta_i}{\partial\alpha_k} \\
&= X_{ij}X_{ik}g(\alpha, T_i, A_i)\frac{1-\exp(\eta_i)}{(1+\exp(\eta_i))^2}.
\end{aligned}$$

Appendix D

ISOTONIC REGRESSION

As noted in Section 4.9, the SP estimator of the ROC curve is not necessarily monotone. A monotone increasing estimator was derived using the "pool adjacent violators" (PAV) isotonic regression algorithm. As the name suggests, this algorithm pools data at adjacent points where the monotone increasing criterion is violated.

The first step in this algorithm is to identify the sets of adjacent points where the estimator is decreasing. Next a "current" estimator is obtained by replacing these sets of points with a weighted average of the estimator at the adjacent points. We chose to use weights equal to the number of observations contributing to each estimator. This process of identifying sets of adjacent points where the estimator is decreasing and then replacing them with a weighted average of the estimator is repeated until the resulting estimator is monotone increasing.

VITA

Todd Alonzo received a B.S. magna cum laude in Statistics from California Polytechnic State University, San Luis Obispo, California in December 1994 and an M.S. in biostatistics from the University of Washington, Seattle, Washington in August 1997.