THE UNIVERSITY OF CALGARY

Risk Adjustment with Binary Outcomes when Covariate Information is Incomplete

by

Peter D. Faris

A DISSERTATION
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

December, 1999

Abstract

The purpose of this research was to examine the use of missing data methods applied to risk adjustment procedures for binary outcomes. Incompletely observed risk factors are problematic for researchers performing risk adjustment. The methods employed for handling missing data in these cases have an intuitive justification and have lacked underlying statistical theory. Two missing data methods were investigated. These were multiple imputation and expectation-maximization by the method of weights. Three risk adjusted estimates of relative risk were explored. Two were indirectly standardized measures, while the third was directly standardized. Estimates of variance for these measures were derived for use with missing data methods. These methods were then applied to 1995/96 data from the Alberta Provincial Program for Outcome Assessment in Coronary Heart Disease initiative, which had up to 25% of observations missing from important clinical variables. The three types of risk adjusted point estimates were similar across the different missing data methods. Directly standardized measures had the smallest variances and tended to be stable across the missing data methods.

Monte Carlo computer simulations were employed to examine the performance of the missing data methods over a variety of missing data mechanisms. For each mechanism, 500 samples of 2000 cases were generated. The variables employed included a binary outcome, and one continuous and one binary risk factor. A treatment variable with three levels was used for risk adjustment. For all but the most severe non-missing at random conditions, the bias in the risk adjusted estimates was modest. The estimates were efficient when compared to the complete data estimates. The

standard error estimates for the indirectly standardized measure performed poorly. An asymptotically unbiased variance estimate was derived using the delta method and was tested using Monte Carlo simulations.

When covariate information is missing, risk adjustment with binary outcomes can be performed using multiple imputation or expectation-maximization by the method of weights. The Monte Carlo simulations indicated that these methods work well under a variety of missing data conditions. As these methods are now becoming widely available, good diagnostic tools will need to be developed.

Acknowledgments

In this short space, it would be impossible to acknowledge all of the assistance I have received in the completion of this dissertation. I would like to thank Dr. Rollin Brant for his fine skills as a supervisor and for helping me achieve insights into some of the problems I encountered in the process of my research. Dr. Penny Brasher was instrumental in helping me start my project. The enthusiasm of Dr. Bill Ghali provided inspiration, and helped me see the relevance of my work within health care research. I would also like to thank the APPROACH initiative for the generous contribution of data and expertise.

My deepest gratitude goes to my friends, who helped me maintain some semblance of balance through the past few years. Foremost, I would like to thank Carol Scott for her support, encouragement, and wisdom. Thanks also to Stacey Page and Dave Lieske for their daily interest in my well-being. David Bright and Catherine Radimer (and Tessa too!) provided evenings of excellent conversation and entertainment. Scott and Jackie Oddie, Adele Gafka, and Monique Tenn helped me blow of steam when needed, and many others (Terryl, Jen, Roxy, Laurie-Jo, and my swimming buddies) offered companionship and maintained an active interest in my progress.

Finally, I would like to thank the Alberta Heritage Foundation for Medical Research, the William Davies scholarship fund, and Dr. Rollin Brant for financial support over the course of my studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

When researchers assess the efficacy of treatments or interventions on patient outcomes, they employ procedures to ensure that differences in outcomes across treatment groups cannot be accounted for by differences in the mix of patient characteristics among the groups. The most widely accepted manner of controlling for these differences is the randomized clinical trial, in which patients are randomly allocated to treatment arms. In many cases, however, randomization is not practical, or is not possible due to ethical considerations. In these cases, statistical methods may be applied in an attempt to control for differences in patient characteristics which may account for observed differences in outcomes. Such methods have also been applied in cases where researchers wish to compare the quality of treatment afforded to patients by different physicians or hospitals. When used in these contexts, these methods are referred to as *risk adjustment* procedures, as they are an attempt to adjust the observed outcomes for the risk of an adverse outcome presented by patients with different characteristics (Blumberg, 1986). In many cases, complete information is not available on variables which are important when adjusting for patient characteristics. The purpose of this dissertation is to examine the use of missing data methods which may be useful when conducting risk adjustment using binary outcomes.

1.1 Risk Adjustment Methods

Risk adjustment methods are based on the assumption that patient outcomes can be expressed as a function of patient attributes, random events, and the quality of treatment the patient receives (Iezzoni, 1995; Park et al., 1990). These methods have been applied to a variety of outcomes including time to death, mortality, morbidity, disease complications, quality of life, physiological functioning and costs of care (Iezzoni, 1995). Risk adjustment methods are often based on a comparison of the observed patient outcome with the outcomes that would be expected on the basis of patient risks. In the case of binary outcomes, a frequently used measure is the standardized mortality ratio (SMR) which is the ratio of observed to expected outcomes (Ash and Shwartz, 1997). Measures based on the difference between observed and expected outcomes can also be employed. These measures will be discussed in greater detail in chapter 2.

The use of risk adjustment procedures in health research has been increasing. This can be attributed to several factors. The first of these is the need to contain growing health care costs in the face of innovative and expensive treatment procedures. Further, the proliferation of new treatments has led to widespread variation in medical practices across providers, making it difficult to assess the effectiveness of these new procedures. Finally, the public is becoming more vigilant in their assessment of the health care they receive, and are demanding evaluations of the care provided by physicians and hospitals. In the United States, some states publish annual risk adjusted performance indicators for hospitals and physicians for the public (Epstein, 1995; Green and Wintfeld, 1995; Kassirer, 1994). In Canada, risk adjust-

ment has been used to compare the quality of cardiac care within provinces and across the provinces (Ghali et al., 1998).

1.2 Missing Data in Risk Adjustment

Adequate risk adjustment requires both the identification of risk factors which can affect the outcomes under investigation and specification of the correct model for the data. Further, it requires that the sources of data be of high quality. Records need to be complete and measurements need to be precise. The amount and quality of information collected can vary according by hospital type (teaching vs. non-teaching; Iezzoni et al., 1990). It can also depend on whether or not structured methods are used for data collection (Duggan et al., 1990). Beers et al. (1989) found that records tended to be more complete in inpatient facilities than in outpatient facilities.

While missing risk factors have been identified as a problem in risk adjustment, the methods which have been employed for handling missing data have generally had an intuitive justification and have lacked underlying theory or rationale. When substantial information is missing from records, researchers can be faced with the choice between removing patients with the missing data from the risk adjustment procedure, or dropping risk factors which may be important for the analysis. Another strategy employed has been to impute or infer the likely values for the missing risk factors and then conduct the analysis as if the data had been completely observed.

For example, for studies of patients in intensive care facilities, it is often assumed that in cases where the results of test procedures have not been recorded, the patient's measure falls within the normal range. The underlying logic for this procedure is that

if the patient was not tested, it was because there was no indication that the test was needed. Although this is arguably a reasonable procedure, it can be criticized on both logical and statistical grounds. There is no way of knowing that the patient would have had a normal result if the test had been conducted. Even graver errors can be committed. Blumberg (1986) reports a situation in which test results were often missing because patients died before the test could be completed. The substitution of normal results for missing data led to the anomalous result that tests in the normal range were associated with adverse outcomes.

Simple imputation procedures can also be criticized from a statistical standpoint (Little and Rubin, 1989a). Apart from biases that may occur due to errors committed in making the imputations, the use of imputed values can lead to estimates of variances that are too small. This results in smaller than expected confidence intervals and to inappropriate statistical inferences.

When risk adjustment procedures employ binary outcomes, risk adjustment can be based on statistically derived methods for handling missing data in logistic regression. The two methods examined for these purposes will be multiple imputation (MI) and the likelihood based expectation-maximization by the method of weights (EMMW). These methods will be described in chapter 3, and the performance of these methods for risk adjustment will be examined in chapters 4 and 5.

1.2.1 Missing Data in Logistic Regression

Several methods have been proposed for handling missing covariate data in logistic regression. These methods have received little use, and do not appear to have been employed for risk adjustment with binary outcomes. Although these methods are

relatively easy to implement and apply, they are not available in commonly used commercial statistical packages. Further, they are complicated by the need to specify a joint probability distribution for the covariates, and because the conditions which gave rise to the missing data often cannot be specified.

In complete data regression procedures, the probability distribution of the outcome variable is considered to be conditional on the covariates. Because of this, the probability distribution of the covariates is not considered when constructing regression models, and the observed covariate values are treated as fixed constants. For many missing data procedures, accounting for missing data requires the specification of the distribution of covariates. Rather than being treated as fixed and known, the observed values of covariates obtained in the study are treated as realizations of random variables. These variables are considered to have a joint distribution, and the parameters of this distribution are employed when accounting for the missing data. Different approaches have been taken to solve this problem. Some approaches model this distribution non-parametrically and avoid the estimation of parameters for the probability distribution (Pepe and Fleming, 1991; Brant and Tibshirani, 1991). The methods employed in this dissertation first estimate the parameters for the covariates, and then either use these parameters for the generation of multiple imputations or use them within a probability model for the outcome.

Unless the mechanism which gives rise to the missing data can be specified, valid use of missing data methods require that the data be missing at random (MAR) (Rubin, 1976). When a variable is MAR, the probability that a variable is observed can depend on the values of other observed covariates. However, the probability that a variable is missing cannot depend on the value of the missing variable or on

the value of any other missing covariate. While un-intuitive, this is the minimum condition required for valid maximum likelihood inference in the presence of missing data. Further, it is less restrictive than the condition that would seem to be required, in which the observations form a random subset of the complete data. In such a condition, the data are missing completely at random (MCAR). Missing data methods have been shown to be somewhat robust to violations of the MAR assumption (Vach and Blettner, 1991). Further, these methods are more efficient than deleting all cases with missing observations, and under appropriate conditions can be shown to be consistent and to yield asymptotically unbiased estimates of variance. However, the effectiveness of these methods within the context of risk adjustment has yet to be investigated.

1.3 The APPROACH Data

The utility of missing data methods for risk adjustment will be examined using data from the Alberta Provincial Program for Outcome Assessment in Coronary Heart Disease (APPROACH) project. The APPROACH project was initiated to assess the cost and clinical outcomes for patients undergoing angioplasty and coronary bypass surgery in the province of Alberta, Canada. Starting in 1995, physicians recorded clinical information from all patients undergoing cardiac catheterizaton. This information was recorded with the assistance of a computer program developed for the initiative. Using the program, physicians recorded information relevant to the outcomes of interest. Data collected included information on tests, symptoms, and family history. In 1995-1996, data was collected from more than 6,000 patients.

In the early stages of the study, the collection of data was often not complete, with up to 25% of the data missing on clinically important variables.

The data have been used to compare the six-month mortality rates for 4 major hospitals in Alberta. In an attempt to account for the missing clinical data, Norris et al. (1999) obtained ICD-9 discharge data for the patients. Using a computer algorithm, this discharge data was converted into diagnoses relevant to the risk adjustment. For many of the clinical variables, the diagnoses were coded as binary variables (diagnosis positive or negative). In these cases, a new variable was created which was based on both the clinical and the administrative data. If the clinical variable was missing, the administrative diagnosis was substituted for the clinical diagnosis. If the clinical variable was observed, the condition was coded as positive if either the clinical or administrative diagnosis was positive, and negative if neither was positive. For some of the variables, such as ejection fraction and coronary anatomy, there were no administrative equivalents. In these cases, a separate category was created to code for the missing observations. Following the creation of these new variables, the resulting database had complete data for 97% of the patients, and this database was successfully used for performing logistic regressions.

The strategy of creating an extra code for missing observations allowed the use of all of the variables, but did not use available information among the variables to account for the missing observations. Further, the use of such a procedure is known to introduce bias into the estimation of logistic regression coefficients, even when the observations are MCAR (Vach, 1994; Vach and Blettner, 1991). Multiple imputation and likelihood based methods would allow the use of variables with no administrative equivalents without requiring the use of an extra category to account

for missing observations. Consequently, these methods may provide a powerful tool for researchers performing risk adjustment in cases where there are no administrative equivalents for variables with missing observations. The risk adjustment procedures will be used to examine the effectiveness of the following treamtments: medical treatment, coronary artery bypass grafts (CABG), and percutaneous transluminal coronary angioplasty (PTCA). Type of treatment was chosen for the investigation of risk adjustment methods because group sizes were similar and because there appeared to be reasonably large differences in the effectiveness of the different treatments.

1.4 Purpose

The purpose of this dissertation will be to examine the use of missing data methods applied to risk adjustment procedures. This examination will be confined to the cases where the outcome is binary. The first step in examining these issues will be an investigation of risk adjustment measures and how they are estimated using logistic regression. This will be followed by a discussion of methods of handling missing data in logistic regression and risk adjustment. Missing data methods will then be used to perform risk adjustment with the APPROACH data. These methods will be used to examine the effectiveness of the treatments provided to the cases. Simulations based on these results and on the distribution of the APPROACH data will provide information regarding appropriate conditions for performing risk adjustment with missing data.

Chapter 2

Risk Adjustment Methods

Several methods of risk adjustment have been employed by researchers. These methods tend to be based on indirect standardization, where the number of deaths following treatment by a provider is compared to the number of deaths that would be expected on the basis of underlying patient risks. Risk adjustment measures which are based on indirect standardization include the standardized mortality ratio (SMR), the difference between observed and expected deaths (the $O-E$ difference), the population averaged proportion, and $Z-$scores (Shwartz et al., 1997).

In direct standardization, the observed number of deaths in a standard population is compared to the number that would have been expected if patients in the standard population had the same risk for death as patients treated by a particular provider. Methods based on direct standardization are rarely used for risk adjustment, probably because in traditional applications, variance estimates for directly standardized measures tend to be larger than those obtained using indirect standardization (Breslow and Day, 1987b). When direct and indirect standardization are used for risk adjustment, however, stratum specific estimates of risk are usually based on a regression model for the population under investigation. In this case, measures obtained using indirect and direct standardization can have the same interpretation and comparable variances.

The choice of risk adjustment methods can be guided by the underlying model for the risks associated with treatment by providers. When risks are multiplicative,

ratio measures such as the SMR are appropriate and interpretable. When underlying risks are additive, difference measures such as risk differences, are preferable. The following sections describe measures which have been employed, the interpretation that may be given to these measures, and how variance estimates may be obtained.

2.0.1 Notation and Models

Notation

The following notation and models will be useful for considering the uses and interpretation of the above measures. In our study population, we have m treatment providers h_k, with $k = 1, 2, \ldots, m$. In this population, there are N patients $i = 1, 2, \ldots, N$, with n_k patients being treated by the k^{th} provider. If we let y_{ik} be a binary random variable indicating the occurrence of death or disease, we can construct models in which an individual can be assigned a probability of death based solely on patient characteristics and which is independent of the treatment provider. This probability of death will be denoted as p_i. Further, we can denote the probability of death for this patient if he or she is treated by provider h_k as p_{ik}.

The quantity O_k, which represents the observed number of deaths for provider k will be considered to be a random variable

$$O_k = \sum_{i \in h_k} y_{ik}$$

where the y_{ik} are Bernoulli random variables

$$\begin{aligned} \Pr(y_{ik} = 1) &= p_{ik} \\ \Pr(y_{ik} = 0) &= 1 - p_{ik} \\ \mathrm{Var}(y_{ik}) &= p_{ik}(1 - p_{ik}). \end{aligned}$$

The expected value of O_k is therefore

$$\mathrm{E}(O_k) = \sum_{i \in h_k} p_{ik},$$

Assuming independence of the observations y_i for $i \in h_k$, the variance of O_k is

$$\mathrm{Var}(O_k) = \sum_{i \in h_k} p_{ik}(1 - p_{ik}).$$

The provider independent probabilities of death, p_i, can be determined on the basis of external rates or on the basis of the total observed population, and are often treated as fixed constants. The expected number of deaths will be denoted as

$$\mathrm{E}(E_k) = \sum_{i \in h_k} p_i,$$

or as

$$E_k = \sum_{i \in h_k} \hat{p}_i \tag{2.1}$$

when the p_i are estimated from the data. In general, when estimates of the p_{ik} are obtained from models which include provider effects, these estimates are subject to the constraint that

$$O_k = \sum_{i \in h_k} \hat{p}_{ik} = E_k. \tag{2.2}$$

In many situations, the characteristics which are used to determine the probability of death can be expressed as diagnostic or demographic categories. In these cases, one can construct a composite categorical variable which denotes membership in a particular joint diagnostic and demographic category. For example, one such category could denote patients who are young, male smokers with diabetes. These joint categories will be denoted as categories $l = 1, 2, \ldots, s$. Individuals i within each of these categories have probability of death p_l. Each of the n_k can be broken into

n_{kl}, where each n_{kl} represents the number of patients treated by h_k who are in risk category l. The set of the n_{kl} determines the patient mix for provider h_k. Under a binomial model, the maximum-likelihood estimate of p_{kl} is

$$\hat{p}_{kl} = \frac{1}{n_{kl}} \sum_{i \in h_{kl}} y_i$$

This leads to the following re-expressions:

$$\sum_{i \in h_{kl}} p_i = n_{kl} p_l$$

$$\sum_{i \in h_{kl}} p_{ik} = n_{kl} p_{kl}$$

$$E_k = \sum_{l=1}^{s} n_{kl} \hat{p}_l \tag{2.3}$$

$$O_k = \sum_{i \in h_k} y_i = \sum_{l=1}^{s} n_{kl} \hat{p}_{kl}, \text{ and} \tag{2.4}$$

$$\text{Var}(O_k) = \sum_{l=1}^{s} n_{kl} p_{kl} (1 - p_{kl}) \tag{2.5}$$

Definitions

The following definitions will be used for the discussion of risk adjustment measures.

Indirect Standardization. The calculation of a weighted average of the risks in a standard population. It is obtained by weighting the stratum specific risks in the standard population by the distribution of subjects treated by a given provider. Dividing the observed proportion of deaths for a provider by this weighted average yields the O/E ratio, which is also referred to as the standardized mortality ratio (SMR). The *standard population* can be an external population, the entire population of patients or the distribution obtained by combining the providers under investigation.

Direct Standardization. The calculation of a weighted average of the stratum specific risks of patients treated by a provider, where the weights are based on the distribution of patients in a standard population.

Homogeneity. Within the context of the following discussion, homogeneity will refer to the constant increase in risk across risk categories that is associated with treatment by a particular provider (Breslow and Day, 1987b; Rothman, 1986b). This increase in risk has also been referred to as proportionality of effect (Kelsey et al., 1996b), or the assumption of uniform effect (Greenland and Rothman, 1998c). For the remainder of this dissertation, the condition of constant provider effects will be referred to as *homogeneity*, as this term can be applied to both additive and multiplicative risk models. The goal of risk-adjustment is to compare the effectiveness of providers after controlling for differences in patient mix that may confound this comparison. As will be demonstrated, however, the commonly used risk-adjustment methods are sensitive to changes in patient mix unless the level of risk associated with a provider is constant across diagnostic categories. Under homogeneity, observed fluctuations in risk across diagnostic strata are attributed to random error (Kelsey et al., 1996c). The assumption of homogeneity can be tested using goodness-of-fit tests, by inspection of the risks across strata, or on the basis of *a priori* evidence. For risk-adjustment, the condition of homogeneity yields meaningful and interpretable results.

Heterogeneity. Heterogeneity refers to the condition in which the risk associated with a provider varies across diagnostic categories. Under heterogeneity, a single risk-adjusted value can no longer represent the performance of a provider. This is

because heterogeneity implies that provider performance depends on the risk categories to which patients belong. In such a case, Kelsey et al. (1996b); Breslow and Day (1987b) suggest that one should compare the performance within risk categories. Greenland and Rothman (1998c), however, take a more liberal stance towards heterogeneity. They suggest that to assume the effects are uniform, one does not need to rule out heterogeneity. Instead, they view homogeneity as a useful approximation that simplifies analysis and reporting, and which is a reasonable assumption to make provided that it is not clearly contradicted by the data or other evidence.

Multiplicative and Additive effects

Multiplicative effects. Models which assume multiplicative effects are the most frequently encountered models in epidemiology (Kelsey et al., 1996a). There are logical, mathematical and empirical grounds for justifying the use of multiplicative models. When relative risks are used as a measure of association, a confounding variable which exerts its influence through a causal variable will always have a weaker association with the outcome than will the genuine causal variable (Breslow and Day, 1980; Cornfield et al., 1959). This implies that an exposure with a strong disease association is more likely to be causal than an exposure with a weak association (Kelsey et al., 1996a). Multiplicative models have provided stable measures of association in a wide variety of populations for a wide variety of diseases, and the effects of combined exposures on an outcome are often found to be multiplicative (Breslow and Day, 1980). Multiplicative risk models are convenient to work with mathematically, as these models are linear on a log scale, and maximum likelihood estimates for the parameters in these models are more readily obtained (Bishop et al., 1975).

In the context of risk adjustment, a multiplicative model implies that treatment by provider h_k leads to a proportionate increase in risk for patients in each of the s risk categories. This proportionate increase in risk will be denoted as RR_{kl}. If effects are multiplicative, the risk for patients in category l treated by provider h_k can be modeled as

$$p_{kl} = RR_{kl}p_l \tag{2.6}$$

Under the assumption of homogeneity, the proportionate increase in risk associated with a provider is the same for all patients. This increase in risk will be denoted as RR_k. Using a multiplicative model, and assuming homogeneity, the probability of death for patients in risk category l treated by provider h_k is

$$p_{kl} = RR_k p_l \tag{2.7}$$

and the expectation for the observed number of deaths is

$$\mathrm{E}(O_k) = RR_k \sum_{l=1}^{s} n_{kl}p_l. \tag{2.8}$$

Additive effects Additive models are considered to be useful in determining the public health impact of a risk factor (Breslow and Day, 1980; Kelsey et al., 1996a), as they allow one to calculate the number of excess occurrences that are associated with a particular exposure. Additive models often lack biological plausibility (Breslow and Day, 1980), although some have argued that additive models can provide a better means of assessing causal associations (Greenland, 1998; Rothman, 1974, 1976; Rothman et al., 1980).

Under an additive model, treatment by a provider adds to the risk of death associated with patients on the basis of their diagnostic category. The additional

risk associated with treatment by provider h_k for patients in risk category l will be denoted as R_{kl}. The model for risk associated with these patients is

$$p_{kl} = R_{kl} + p_l \tag{2.9}$$

Under the assumption of homogeneity, the amount of additional risk associated with a particular provider is the same for all patients treated by the provider. This additional or excess risk will be denoted as R_k. Using a model for additive effects the probability of death for subjects in risk category l and treated by provider h_k is

$$p_{kl} = R_k + p_l \tag{2.10}$$

and the expectation for the observed number of cases is

$$\mathrm{E}(O_k) = n_k R_k + \sum_{l=1}^{s} n_{kl} p_l. \tag{2.11}$$

The choice between additive and multiplicative models can be based on empirical, logical and practical grounds (Breslow and Day, 1980, 1987a; Kelsey et al., 1996a; Rothman, 1986a). Empirical considerations include the goodness of fit of the model, and how succinctly the model can account for the data (Breslow and Day, 1980). It is desirable to capture the essential features of the data as succinctly as possible. To this end, homogeneous models are more desirable than heterogeneous models, as they do not require the calculation of parameters corresponding to interaction terms. This is an important consideration when deciding between additive and multiplicative models, since a model which is homogeneous on an additive scale will be heterogeneous on a multiplicative scale and vice versa (Greenland and Rothman, 1998b). Breslow and Day (1987a) point out, however, that it can be difficult to

discriminate between models on the basis of statistical considerations. Unless one model can be shown empirically to be much better than another, Breslow and Day suggest that 1) a-priori considerations are an important part of the decision and that 2) it is prudent to examine and present both additive and multiplicative models.

2.1 Methods Based on Indirect Standardization

As noted previously, in the face of strong evidence for heterogeneity, the use of a single summary measure to represent provider performance is questionable. Under homogeneity, risk-adjustment models can provide meaningful and interpretable results. As would be expected, difference measures, such as the $O - E$ difference are meaningful when the underlying risk model is additive. Ratio measures, such as the SMR, are more appropriate where the underlying model is assumed to be multiplicative.

Standardized risks have the general form

$$p_k = \frac{\sum_{l=1}^{s} w_l p_{kl}}{\sum_{l=1}^{s} w_l} \tag{2.12}$$

(Greenland and Rothman, 1998b). The weights w_l are determined by the distribution of patients across risk categories in the standard. For indirectly standardized risks, this distribution is based on the patient mix across risk categories for the provider under study. For provider h_k, the indirectly standardized risk of death will be denoted as

$$p_k = \frac{\sum_{l=1}^{s} n_{kl} p_{kl}}{\sum_{l=1}^{s} n_{kl}} = \frac{\sum_{l=1}^{s} n_{kl} p_{kl}}{n_k}. \tag{2.13}$$

The expected risk of death, p_{ek}, based solely on patient risk factors and ignoring the

effects of treatment by provider h_k is

$$p_{ek} = \frac{\sum_{l=1}^{s} n_{kl} p_l}{n_k}. \tag{2.14}$$

2.1.1 Ratio Measures

SMR

The SMR is a standardized risk ratio obtained by taking the ratio of indirectly standardized risks. The use of the SMR is appropriate as a risk-adjustment measure when the underlying risk model is multiplicative and when the provider effects can be assumed to be homogeneous across risk strata. When these conditions are met, SMR_k can be interpreted as an estimate of the constant proportional increase in risk associated with provider h_k. The ratio of indirectly standardized risks can be expressed as

$$\frac{p_k}{p_{ek}} = RR_k = \frac{\sum_{l=1}^{s} n_{kl} p_{kl}}{n_k} \Big/ \frac{\sum_{l=1}^{s} n_{kl} p_l}{n_k} = \frac{\sum_{l=1}^{s} n_{kl} p_{kl}}{\sum_{l=1}^{s} n_{kl} p_l}.$$

When the $\hat{p}_{kl}$ and $\hat{p}_l$ are used as estimates, by 2.3 and 2.4,

$$\widehat{RR}_k = O_k / E_k = SMR_k. \tag{2.15}$$

An estimate of the SMR can therefore be obtained by taking the ratio of the observed deaths to the number of deaths that would be expected if the stratum specific death rates associated with the provider were the same as those in the population (Greenland and Rothman, 1998a; Last, 1988).

Multiplicative Risk Model

Under a multiplicative risk model, the SMR is a weighted average of the stratum specific relative risks associated with a provider. The weights for these relative risks are determined by the patient mix and by the risks associated with the risk categories. By 2.3 and 2.6, RR_k can be expressed as

$$RR_k = \frac{\sum_{l=1}^{s} RR_{kl} n_{kl} p_l}{\sum_{l=1}^{s} n_{kl} p_l}. \tag{2.16}$$

When the RR_{kl} and p_l are estimates, the SMR_k can be expressed as a weighted average of the RR_{kl}:

$$SMR_k = \widehat{RR}_k = \frac{\sum_{l=1}^{s} \widehat{RR}_{kl} n_{kl} \hat{p}_l}{\sum_{l=1}^{s} n_{kl} \hat{p}_l}. \tag{2.17}$$

It is evident from 2.16 that for given sets of RR_{kl} and p_l, the RR_k can be sensitive to changes in patient mix. This means that two providers with identical risks associated with treatment can have different RR_k's if they differ in their mix of patients. This can be demonstrated by taking the sum of the partial derivatives of RR_k with respect the set of n_{kl}, where $l = 1, 2, 3, \ldots, s$. The relative risk for provider h_k will not be sensitive to change in patient mix for sets of RR_{kl} for which each of the partial derivatives of RR_k is equal to 0. These derivatives will equal 0 when

$$p_l(RR_{kl}/RR_k) = p_l. \tag{2.18}$$

If the RR_{kl} differ across strata, the above equality will not hold (except in the unlikely event that p_l=0). Consequently, if there is heterogeneity of the risk ratios across strata, the estimate of the SMR will be sensitive to patient mix. For an example of the effects of heterogeneity on the SMR, see section (2.4.2) below.

Homogeneity. The above equality (2.18) will hold if there is homogeneity of the relative risks across risk strata. That is, it will hold if $RR_{kl} = RR_k$ for all l. Where there is homogeneity, the RR_k will not be sensitive to patient mix, and it can be interpreted as the constant proportional increase in patient risk across risk categories that is associated with provider h_k. Under homogeneity, the $\widehat{RR}_{kl}$ are all estimates of RR_k, and fluctuations in the $\widehat{RR}_{kl}$ are attributed to random variability. Consequently, under homogeneity, $\widehat{RR}_k$ and SMR_k are both estimates of the constant proportional increase in risk associated with provider h_k.

Additive Risk Model

If the underlying risk model for treatment providers is additive, the SMR will be sensitive to patient mix and consequently is a poor choice as a risk-adjusted measure. By 2.3 and 2.11, the ratio of indirectly standardized risks would be

$$RR_k = \frac{\sum_{l=1}^{s} n_{kl}(R_{kl} + p_l)}{\sum_{l=1}^{s} n_{kl} p_l}.$$

Under homogeneity, by 2.11 this can be expressed as

$$RR_k = \frac{n_k R_k}{\sum_{l=1}^{s} n_{kl} p_l} + 1.$$

To facilitate interpretation of this measure, consider the set of weights $\mathbf{w}_k = (w_{k1}, w_{k2}, \ldots, w_{ks})$ for provider h_k, where $w_{kl} = n_{kl}/n_k$. This set of weights denotes the proportion of patients treated within each risk category. The RR_k can then be re-expressed as

$$RR_k = \frac{R_k}{\sum_{l=1}^{s} w_{kl} p_l} + 1. \tag{2.19}$$

Since SMR_k is an estimate of RR_k, it is evident that with an underlying additive risk model, an SMR will be sensitive to patient mix. With the risks associated with

patients held constant, the denominator in the first term of the above expression depends on $\mathbf{w}_k$. The term $\sum_{l=1}^{s} w_{kl}p_l$ will be large for providers which tend to treat high risk patients and it will be small for providers treating low risk patients. Consequently, if two providers that have the same additive risk associated with treatment, the provider that tends to treat higher risk patients will have a smaller RR_k than will a provider treating lower risk patients.

Variance.

The simplest means of determining the variance of the SMR is to treat the $\hat{p}_l$ as fixed and known. In this case, the variance of the SMR is

$$\text{Var}(SMR_k) = \text{Var}(O_k/E_k) = \frac{\text{Var}(O_k)}{E_k^2}$$

For the purpose of constructing confidence intervals, by 2.5 the variance is

$$\text{Var}(SMR_k) = \frac{\sum_{l=1}^{s} n_{kl}p_{kl}(1-p_{kl})}{E_k^2} \tag{2.20}$$

Under the null hypothesis, the $\hat{p}_{kl}$ are assumed to differ from the p_l due to random variation. Therefore, for hypothesis tests, the $\text{Var}(O_k)$ can be based on the p_l, and the variance of the SMR is

$$\text{Var}(SMR_k) = \frac{\sum_{l=1}^{s} n_{kl}p_l(1-p_l)}{E_k^2}. \tag{2.21}$$

The p_l are often not known and must be estimated on the basis of available data. The E_k, which are sums of the p_l, are therefore random variables. Further, if each provider is a subset of the data used to obtain the p_l, then estimates of the p_l and p_{kl} will not be statistically independent. Using a first-order Taylor series expansion,

an asymptotic expression for the variance of SMR_k is

$$\begin{aligned} \mathrm{Var}(SMR_k) &= \frac{1}{\mathrm{E}(E_k)^2}\left[\left(\frac{\mathrm{E}(O_k)}{\mathrm{E}(E_k)}\right)^2 \mathrm{Var}(O_k)\right. \\ &\quad \left. + \mathrm{Var}(E_k) - 2\left(\frac{\mathrm{E}(O_k)}{\mathrm{E}(E_k)}\right)\mathrm{Cov}(O_k, E_k)\right], \end{aligned} \tag{2.22}$$

and an approximation for the variance of SMR_k is

$$\widehat{\mathrm{Var}}(SMR_k) = \frac{1}{E_k^2}\left[\left(\frac{O_k}{E_k}\right)^2 \widehat{\mathrm{Var}}(O_k) + \widehat{\mathrm{Var}}(E_k) - 2\left(\frac{O_k}{E_k}\right)\widehat{\mathrm{Cov}}(O_k, E_k)\right].$$

Estimates of $\mathrm{Var}(O_k)$ are obtained as above. Estimates of $\mathrm{Var}(E_k)$ and $\mathrm{Cov}(O_k, E_k)$ are based on the model used for estimating the p_l. For details of how these can be derived from asymptotic likelihood theory, see appendices A and D.

The Population Averaged Proportion

The population averaged proportion is also referred to as the risk-adjusted mortality (Shwartz et al., 1997). As with the SMR, its use as a risk adjustment measure is appropriate where homogeneity and a multiplicative risk model can be assumed. Under these assumptions, the population averaged proportion can be interpreted as the proportion of persons in the treated population that would have died had all patients been treated by a given provider. The population averaged proportion will be denoted as P_k, and is calculated as

$$P_k = RR_k p \tag{2.23}$$

where $p = \sum_{k=1}^{m} \mathrm{E}(O_k)/N$. An estimate of P_k for provider h_k can be obtained by multiplying $\hat{p}$ by the SMR_k, where $\hat{p} = \sum k = 1^m O_k/N$.

As with the SMR, with an underlying multiplicative model, the $\hat{P}_k$ will be sensitive to patient mix in the presence of heterogeneity. It will also be sensitive to patient mix if the underlying model for risks is additive.

Variance Often, $\hat{p}$ and the $\hat{p}_l$ are treated as fixed constants, and the variance of $\hat{P}_k$ is estimated as

$$\widehat{\text{Var}}(P_k) = \left(\frac{\hat{p}}{E_k}\right)^2 \widehat{\text{Var}}(O_k)$$

Where p and the p_l are estimated from available data, the variance of $\hat{P}_k$ can be approximated using a first order Taylor Series expansion. In this case,

$$\begin{aligned}\widehat{\text{Var}}(\hat{P}_k) &= \left(\frac{\hat{p}}{E_k}\right)^2 \widehat{\text{Var}}(O_k) + \left(\frac{O_k}{E_k^2}\hat{p}\right)^2 \widehat{\text{Var}}(E_k) + \left(\frac{O_k}{E_k}\right)^2 \widehat{\text{Var}}(\hat{p}) \\ &\quad - 2\left(\frac{O_k}{E_k^2}\hat{p}\right)\left[\left(\frac{O_k}{E_k}\right)\widehat{\text{Cov}}(E_k,\hat{p}) + \left(\frac{\hat{p}}{E_k}\right)\widehat{\text{Cov}}(O_k,E_k) - \widehat{\text{Cov}}(O_k,\hat{p})\right].\end{aligned}$$

The variance of O_k is obtained by 2.20 or 2.21. The covariances and the variances of E_k and $\hat{p}$ depend on the model used in determining the estimates of the p_l. For further details of the asymptotic likelihood theory, see appendix A.

2.1.2 Difference Measures

Difference measures, such as the excess risk, are appropriate when the underlying model for risks is additive. When the underlying risk model is additive and when there is homogeneity of the effect for a provider, measures based on the risk difference are not sensitive to patient mix. The $O - E$ difference is a poor choice for a risk adjustment measure, as the magnitude of this difference is directly proportional to the volume of patients treated by a provider. If there is heterogeneity among the additive risks for a provider or if the underlying model for risks is multiplicative, difference measures will be sensitive to patient mix.

Risk Difference

The risk difference is also referred to as the excess risk. It is the proportion of deaths associated with treatment by a provider. If the underlying model for risks is homogeneous and additive, it can be interpreted as the constant additive increase in risk associated with treatment by a given provider. Using indirect standardization, the excess risk R_k is

$$\begin{aligned} R_k &= \frac{\sum_{l=1}^{s}(p_{kl}-p_l)n_{kl}}{\sum_{l=1}^{s} n_{kl}} \\ &= \frac{\sum_{l=1}^{s} p_{kl}n_{kl} - \sum_{l=1}^{s} p_l n_{kl}}{n_k}. \end{aligned} \tag{2.24}$$

By 2.4, 2.3 and 2.24, an estimate of the excess risk is

$$\hat{R}_k = \frac{O_k - E_k}{n_k}.$$

Additive Risk Model. If the underlying model for risks is additive, by 2.24 and 2.9 the excess risk can be expressed as

$$\begin{aligned} R_k &= \frac{\sum_{l=1}^{s}((R_{kl}+p_l)-p_l)n_{kl}}{n_k} \\ &= \frac{\sum_{l=1}^{s} R_{kl}n_{kl}}{n_k}. \end{aligned} \tag{2.25}$$

When the R_{kl} and p_l are estimates, an estimate of the excess risk is

$$\hat{R}_k = \frac{O_k - E_k}{n_k} = \frac{\sum_{l=1}^{s} \hat{R}_{kl}n_{kl}}{n_k}. \tag{2.26}$$

From 2.25, it is evident that with an additive risk model, the excess risk may be affected by patient mix. Taking the partial derivatives of R_k with respect to the n_{kl}, the R_k will not be sensitive to patient mix when

$$n_k R_{kj} = \sum_{l=1}^{s} R_{kl}n_{kl}$$

for all j where the $j = 1, 2, 3, \ldots, r$ denote membership in risk strata. The above equalities will hold when $R_{kj} = R_{kl}$ for all j, l. In other words, it will hold for each of the derivatives when there is homogeneity in the additive risks across risk strata. Under homogeneity, R_k can be interpreted as the constant additive increase in risk associated with provider h_k. The $\hat{R}_{kl}$ are estimates of R_k, and the quantity $(O_k - E_k)/n_k$ is therefore an estimate of the constant proportional increase in risk associated with provider h_k.

Multiplicative Risks If the risks associated with providers are multiplicative, by 2.24 and 2.6, the excess risk can be expressed as

$$\begin{aligned} R_k &= \frac{\sum_{l=1}^{s} RR_{kl} p_l n_{kl}}{n_k} - \frac{\sum_{l=1}^{s} p_l n_{kl}}{n_k} \\ &= \frac{\sum_{l=1}^{s} (RR_{kl} - 1) p_l n_{kl}}{n_k} \end{aligned} \tag{2.27}$$

For given RR_{kl} and p_l, and a multiplicative risk model, the excess risk will be sensitive to differences in patient mix. This will be the case even if the multiplicative risks are homogeneous. Under homogeneity, $RR_{kl} = RR_k$ and 2.27 can be expressed as

$$R_k = (RR_k - 1) \sum_{l=1}^{s} p_l w_{kl}$$

which, for given R_k and p_l will depend on the set of weights w_{kl} denoting patient mix. For example, consider two providers h_1 and h_2 with identical relative risks (i.e. $RR_1 = RR_2$). If a large proportion of patients treated by provider h_1 are in high risk categories, $\sum_{l=1}^{s} w_{kl} p_l$ will be large when compared to provider h_2, which treats low risk patients. Consequently, the expected excess risk associated with the provider h_1 will be larger than that associated with provider h_2, even though their relative risks are identical.

The difference between the observed and expected deaths is sometimes employed when performing risk adjustment. It can be interpreted as the number of excess deaths for provider h_k that can be attributed to treatment by that provider. As previously mentioned, a major weakness of this measure is that it is directly proportional to the volume of patients seen by the provider. The $O-E$ difference is closely related to the excess risk. From 2.24

$$n_k R_k = \sum_{l=1}^{s} (p_{kl} - p_l) n_{kl}.$$

With estimates of p_{kl} and p_l, by 2.4 and 2.3

$$n_k \hat{R}_k = O_k - E_k = \sum_{l=1}^{s} (\hat{p}_{kl} - \hat{p}_l) n_{kl}. \tag{2.28}$$

As with the excess risk, when risks are homogeneous and the risk model is additive, the $O-E$ is affected by patient mix if additive risks are heterogeneous or if the underlying risks are multiplicative. It is not affected by patient mix if the risks are additive and homogeneous. For providers with identical risk, the magnitude of the $O-E$ difference is directly proportional to the volume of patients treated. Consequently, providers with the same underlying risks will be ranked according to their patient volume.

Population Averaged Proportion

When the risk model for providers is additive, an alternative version of the population averaged proportion can be based on the excess risk. This alternative version is

$$P_k^+ = R_k + p = \frac{\sum_{l=1}^{s} R_{kl} n_{kl}}{n_k} + p. \tag{2.29}$$

If the risks are homogeneous, this measure will be insensitive to changes in patient mix. As with the population averaged proportion based on the relative risk, P^+ can

be interpreted as the proportion of patients in the treated population that would have died if all had been treated by a given provider.

2.2 $Z-$scores

Another measure used for risk-adjustment is the $Z-$score. $Z-$scores are not a good measure for risk-adjustment as their use confuses issues of statistical significance with the estimated magnitude of risk. Further, $Z-$score are proportional to the square root of the volume of patients treated by a given provider, and are sensitive to patient mix. The $Z-$score is the distance of the observed value from the expected value in terms of standard deviation units. For binary data, calculation of the $Z-$score is generally based on the binomial approximation to the normal distribution. Since $\text{Var}(y_i) = p_i(1-p_i)$ the variance of O_k under the null hypothesis is

$$\text{Var}(\sum_{i \in h_k} y_i) = \sum_{i \in h_k} p_i(1-p_i)$$

for continuous risk factors, or

$$\text{Var}(\sum_{i \in h_k} y_i) = \sum_{l=1}^{s} n_{kl} p_l(1-p_l)$$

for categorical data.

The $Z-$score is calculated as

$$Z_k = \frac{(O_k - E_k)}{\sqrt{\sum_{l=1}^{s} n_{kl} p_l(1-p_l)}}$$

For an underlying additive model, this can be rewritten as

$$Z_k = \frac{\sqrt{n_k}\hat{R}_k}{\sqrt{\sum_{l=1}^{s} w_{kl} p_l(1-p_l)}}. \tag{2.30}$$

If the underlying model is multiplicative, this can be expressed as

$$Z_k = \frac{\sqrt{n_k}(\widehat{RR}_k - 1)\sum_{l=1}^{s} w_{kl}p_l}{\sqrt{\sum_{l=1}^{s} w_{kl}p_l(1-p_l)}}. \tag{2.31}$$

From (2.30) and (2.31), it is evident that regardless of the underlying model, the $Z-$score will be influenced by the volume of patients treated and by patient mix. For a given patient mix, the magnitude of Z_k will be proportional to $\sqrt{n_k}$. The value of $p(1-p)$ is greatest when $p = 0.5$, and decreases as p approaches 0 or 1. Consequently, providers with a predominance of patients in low or extremely high risk categories will tend to have smaller denominators and larger values of Z.

In risk-adjustment, $Z-$scores are used to rank providers in terms of performance and to indicate providers that may be "outliers". An outlier is defined as a provider with a $Z-$score of a magnitude that is unlikely to have occurred due to chance. As such, however, it is a poor means of comparing the performance of providers, due to the sensitivity of $Z-$scores to factors which are independent of the risk posed by a given provider. Rather than using $Z-$scores, it may be advisable to separate the information regarding the risk associated with a provider from the criteria used to judge the statistical significance of the risk. A measure which is a better reflection of the risk associated with a given provider could be employed, and confidence intervals for these estimates could then be used to determine the precision of these estimates and to indicate whether chance factors could account for the obtained estimates. An "effect size" measure could be constructed by dividing the $Z-$score by $\sqrt{n_k}$. This measure would not longer be influenced by the volume of patients treated by a provider. The use of a binomial model for the variance, however, will still lead this "effect size" measure to be sensitive to patient mix.

2.3 Direct Standardization

Directly standardized risks are standardized using the patient mix from some common standard population. This standard population can be hypothetical, based on a distribution for a given year, or it may be determined by combining the patients from the providers being compared. The standardized risk associated with provider h_k is

$$p_k^* = \frac{\sum_{l=1}^{s} n_l p_{kl}}{\sum_{l=1}^{s} n_l} \tag{2.32}$$

where the n_l are the numbers within risk categories of the standard population. Ratio or difference measures can be obtained by comparing these risks to the overall standardized risk for patients in the standard population

$$p^* = \frac{\sum_{l=1}^{s} n_l p_l}{\sum_{l=1}^{s} n_l} = p. \tag{2.33}$$

Directly standardized measures have two advantages over indirectly standardized measures. First, they are not affected by differences in patient mix among providers. This is because the same standard is applied to the stratum specific risks for each provider. They are affected by the choice of standard population, however, except where the choice of measure correctly reflects the underlying model for risks, and where these risks are homogeneous across risk strata. In these cases, the directly standardized measures estimate the same quantities as those estimated by indirectly standardized measures under the condition of homogeneity; the constant additive or proportional increase in risk associated with treatment by a given provider.

Second, although directly standardized measures may be affected by the choice of standard population, providers with equivalent stratum specific risks will also

have equivalent standardized measures. This is again because the same standard population is used to weight the stratum specific risks for all providers.

One disadvantage attributed to directly standardized measures is that their variances tend to be greater those that obtained for indirectly standardized rates (Breslow and Day, 1987b). This can be especially problematic when the strata have few subjects, because the uncertainty in estimating the rates for these strata can result in high variability for the directly standardized measure. In the case of risk-adjustment, standardization is usually based on the stratum specific rates in the population receiving treatment by the providers of interest. Homogeneity is generally assumed in these models, and the stratum specific rates used for both direct and indirect standardization are based on all subjects in the population (see appendix A for details). As demonstrated in chapter 4 and chapter 5 (see section 5.2.2), the resulting variance estimates are similar for directly and indirectly standardized measures.

2.3.1 Ratio Measures

The Relative Risk.

Ratios of directly standardized risks can be used to determine relative risks associated with treatment by providers. A directly standardized ratio compares the risks which would occur if all patients in the standard population were treated by a particular provider with the actual risk in the standard population. The relative risk associated with treatment by provider h_k is

$$\frac{p_k^*}{p^*} = RR_k^* = \frac{\sum_{l=1}^{s} n_l p_{kl}}{\sum_{l=1}^{s} n_l} \Big/ \frac{\sum_{l=1}^{s} n_l p_l}{\sum_{l=1}^{s} n_l} = \frac{\sum_{l=1}^{s} n_l p_{kl}}{\sum_{l=1}^{s} n_l p_l}. \tag{2.34}$$

When the p_l and p_{kl} are estimated from the data,

$$\widehat{RR}_k^* = E_k^*/O \tag{2.35}$$

where $E_k^* = \sum_{l=1}^{s} n_l \hat{p}_{kl}$ and $O = \sum_{l=1}^{s} n_l \hat{p}_l = \sum_{i=1}^{N} y_i$, the total number of deaths in the standard population.

Multiplicative Risk Model. Under a multiplicative model, the directly standardized relative risk can be expressed as

$$RR_k^* = \frac{\sum_{l=1}^{s} RR_{kl} n_l p_l}{\sum_{l=1}^{s} n_l p_l}. \tag{2.36}$$

As noted previously, because the same standard is applied to each provider, the RR_k^* will not be sensitive to differences in patient mix among providers. They will, however, be sensitive to the choice of standard population unless there is homogeneity of the risk ratios. In this case, the choice of standard population is irrelevant, and $RR_k^* = RR_k = RR_{kl}$. When the RR_{kl} and p_l are estimated from the data,

$$\widehat{RR}_k^* = \frac{\sum_{l=1}^{s} \widehat{RR}_{kl} n_l \hat{p}_l}{\sum_{l=1}^{s} n_l \hat{p}_l}$$

Under homogeneity, the $\widehat{RR}_{kl}$ are estimates of RR_k, and $ds\widehat{RR}$ is

$$\widehat{RR}_k^* = \widehat{RR}_k \frac{\sum_{l=1}^{s} n_l \hat{p}_l}{\sum_{l=1}^{s} n_l \hat{p}_l} = \widehat{RR}_k.$$

Consequently, $\widehat{RR}_k^*$ and SMR_k are both estimates of RR_k.

Additive Risk Model. Under an additive model, RR_k^* can be expressed as

$$RR_k^* = \frac{\sum_{l=1}^{s} (R_{kl} + p_l) n_l}{\sum_{l=1}^{s} p_l n_l} = \frac{\sum_{l=1}^{s} R_{kl} n_l}{\sum_{l=1}^{s} n_l p_l} + 1. \tag{2.37}$$

While not sensitive to patient mix, this measure will be sensitive to the choice of standard population.

Variance. If the $\hat{p}_l$ are treated as fixed and known, the variance of $\widehat{RR}_k^*$ is

$$\mathrm{Var}(\widehat{RR}_k^*) = \mathrm{Var}(E_k^*/O) = \frac{\mathrm{Var}(E_k^*)}{O^2}.$$

If this variance is to be used for hypothesis testing, under the null hypothesis of no provider effects, the variance of E_k^* is

$$\mathrm{Var}(E_k^*) = \sum_{l=1}^{s} n_l p_l (1 - p_l),$$

since E_k^* is the sum of binomial random variables. For the purpose of constructing confidence intervals, the variance of E_k^* is

$$\mathrm{Var}(E_k^*) = \sum_{l=1}^{s} n_l p_{kl} (1 - p_{kl}).$$

The the p_{kl} and p_l are often estimated from available data. If each provider is a subset of the data used to estimate the p_l, the estimates of p_l and p_{kl} will not be statistically independent. The following estimate of the variance of $\widehat{RR}_k^*$ is based on a first-order Taylor series expansion:

$$\widehat{\mathrm{Var}}(\widehat{RR}_k^*) = \frac{1}{O^2}\left[\left(\frac{E_k^*}{O}\right)^2 \widehat{\mathrm{Var}}(E_k^*) + \widehat{\mathrm{Var}}(O) - 2\left(\frac{E_k^*}{O}\right)\widehat{\mathrm{Cov}}(E_k^*, O)\right]. \quad (2.38)$$

Estimates of $\mathrm{Var}(E_k^*)$ are obtained as above. Estimates of $\mathrm{Var}(O)$ and $\mathrm{Cov}(E_k^*, O)$ are based on the model used to obtain estimates of the probabilities. For details of how these estimates can be derived from asymptotic likelihood theory, see section A.2 in appendix A.

The Population Averaged Proportion.

The directly standardized risk, p_k^* can be interpreted as the proportion of patients expected to die if all patients in the standard population were treated by provider

h_k. If risks are homogeneous and multiplicative, p_k^* is equal to the population averaged proportion (P_k) obtained using indirect standardization. Under homogeneity, $RR_{kl} = RR_k$, and

$$p_k^* = \frac{\sum_{l=1}^{s} RR_k p_l n_l}{\sum_{l=1}^{s} n_l} = RR_k p = P_k.$$

An estimate of the population averaged proportion can be obtained as

$$\hat{p}_k^* = \frac{\sum_{l=1}^{s} \hat{p}_{kl} n_l}{n} = \frac{1}{n} E_k^*. \tag{2.39}$$

Variance. When the p_{kl} are estimated from the data, the variance of $\hat{p}_k^*$ can be obtained as

$$\text{Var}(\hat{p}_k^*) = \left(\frac{1}{n^2}\right) \text{Var}(\hat{E}_k^*). \tag{2.40}$$

For details on how $\widehat{\text{Var}}(\hat{p}_k^*)$ can be derived using asymptotic likelihood theory, see section A.2 in appendix A.

2.3.2 Difference Measures

Risk Difference

Directly standardized risks can also be used to measure excess risk, $p_k^* - p^*$. This is the excess risk of death that would occur if all patients in the standard population were treated by provider h_k. By 2.32 and 2.33, an estimate of the excess risk can be obtained as

$$\hat{RR}_k^* = \frac{\sum_{l=1}^{s} p_{kl} n_l}{n} - \frac{\sum_{l=1}^{s} p_l n_l}{n} = \frac{E_k^* - O}{n}$$

Multiplicative Risk Model. Under a multiplicative model, this risk difference is

$$R_k^* = p_k^* - p^* = \frac{\sum_{l=1}^{s} RR_{kl} n_l p_l}{n} - \frac{\sum_{l=1}^{s} n_l p_l}{n}$$

$$= \frac{\sum_{l=1}^{s}(RR_{kl}-1)n_l p_l}{n}. \tag{2.41}$$

If the underlying risk model is multiplicative, the excess risk will be sensitive to the choice of standard population.

Additive Risk Model. Under an additive model, the risk difference can be expressed as

$$\begin{aligned} R_k^* = \ p_k^* - p^* \ &= \frac{\sum_{l=1}^{s}(R_{kl}+p_l)n_l}{n} - \frac{\sum_{l=1}^{s} p_l}{n} \\ &= \frac{\sum_{l=1}^{s} R_{kl} n_l}{n}. \end{aligned} \tag{2.42}$$

With additive risks under homogeneity, $R_{kl} = R_k = R_k^*$, and the excess risks obtained using indirect and direct standardization are identical. When the $\hat{R}_k$ are estimated on the basis of available data,

$$\hat{R}_k^* = \frac{\sum_{l=1}^{s} \hat{R}_{kl} n_l}{n}.$$

Under homogeneity, the $\hat{R}_{kl}$ are estimates of R_k, and $\hat{R}_k^*$ is a weighted estimate of R_k.

The Population Averaged Proportion

Under a homogeneous additive model, p_k^* is equal to the population averaged proportion (P_k^*) obtained using indirect standardization, since

$$p_k^* = \frac{\sum_{l=1}^{s}(R_k+p_l)n_l}{\sum_{l=1}^{s} n_l} = R_k + p = P_k^*.$$

2.4 Logistic Models for the Probability of Death

When the proportion dying or becoming diseased is of interest, researchers typically use logistic regression to model the probability of death or desease on the basis of

patient characteristics. The $\hat{p}$'s obtained from this regression are then used to estimate the expected number of deaths associated with a given provider. Logistic regression can be used to obtain estimates which are based on both indirect and direct standardization. When indirect standardization is employed, researchers often use logistic models which do not adjust for covariate effects when estimating the risks associated with risk factors. The corresponding risk adjusted measures will be referred to as baseline model (BM) estimates. Risk adjusted measures obtained from logistic models which do account for treatment effects when estimating risks will be referred to as full model (FM) estimates. When direct standardization is employed, the corresponding logistic models also account for treatment effects. Consequently, these directly standardized measures are also full model estimates. Logistic regression can be used for categorical and/or continuous covariates. In keeping with the rest of this chapter, the following discussion will focus on categorical covariates. A description of how logistic regression can be used for risk adjustment with continuous covariates is presented in appendix A.

2.4.1 Models Adjusting for Provider Effects

For categorical covariates, logistic regression estimates are appropriate in cases where the underlying probability model is binomial. In the case where we have patients within risk categories treated at m different providers, the binomial probability for the observed number of deaths is

$$\Pr(\mathbf{O}) = \prod_{k=1}^{m}\prod_{l=1}^{s}\binom{n_{kl}}{O_{kl}}p_{kl}^{O_{kl}}(1-p_{kl})^{(n_{kl}-O_{kl})}. \tag{2.43}$$

Where $\mathbf{O}$ is a vector of the O_{kl}, which denote the observed number of deaths in each risk category for each provider. In logistic regression, the p_{kl}'s are modeled as a function of the covariates using the logistic probability distribution

$$p_{kl} = \frac{\exp(\alpha + \beta_l + \gamma_k)}{1 + \exp(\alpha + \beta_l + \gamma_k)}. \tag{2.44}$$

In the above formulation, the β_l's are estimates of the increase in risk associated with membership in a risk category relative to an arbitrary baseline risk category. Likewise, the γ_k's are the increase in risk associated with treatment by provider k relative to some arbitrary provider which serves as a baseline. Rearranging 2.44,

$$\frac{p_{kl}}{1 - p_{kl}} = \exp(\alpha + \beta_l + \gamma_k) \tag{2.45}$$

and the odds of death is therefore a multiplicative function of the effect of treatment by a provider and membership in a risk category. Where the probabilities of death are small, odds and odds ratios can be used as approximations of risks and risk ratios, and the logistic regression model approximates a multiplicative risk model.

Estimating the Coefficients. Maximum likelihood estimation is used to obtain estimates of the coefficients in the logistic regression model. The likelihood of the regression parameters is proportional to the probability of the data given these parameters. Let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be vectors of the parameters β_l and γ_k. Maximum likelihood estimates are obtained using the log of the likelihood function. For the logistic regression model described above, the log-likelihood is

$$\ell(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{O}) = \sum_{k=1}^{m} \sum_{l=1}^{s} \left[O_{kl}(\alpha + \beta_l + \gamma_k) - n_{kl} \log(1 + \exp(\alpha + \beta_l + \gamma_k)) \right]. \tag{2.46}$$

In this formulation, the $\log \binom{n_{kl}}{O_{kl}}$ are constant with respect to the log-likelihood and can be ignored when maximizing the likelihood. Maximum likelihood estimates are

obtained by taking the partial derivatives of the log of the likelihood function with respect to the regression parameters and finding the values for the parameters for which the partial derivatives are equal to zero. The partial derivatives are referred to as score equations. The score equations for α, β_l, and γ_k are

$$S_{\beta_l} = \sum_{k=1}^{p} (O_{kl} - n_{kl}p_{kl}) \tag{2.47}$$

$$S_{\gamma_k} = \sum_{l=1}^{s} (O_{kl} - n_{kl}p_{kl}) \tag{2.48}$$

$$S_{\alpha} = \sum_{k=1}^{m}\sum_{l=1}^{s} (O_{kl} - n_{kl}p_{kl}) \tag{2.49}$$

where p_{kl} is a function of the parameters as shown in 2.44.

The solutions to these equations are obtained numerically using iterative methods such as the Newton-Raphson algorithm (McCullagh and Nelder, 1989). The inverse Fisher's information matrix provides an estimate of the variance matrix of the regression parameters.

Indirect Standardization

The logistic regression described above can be used to provide estimates of death for individuals in certain risk factors treated by particular providers under study. When we obtain parameter estimates by equating equations 2.47 - 2.49 to zero, we are forcing the sum of these individual probabilities to be equal to the number of deaths associated with the provider. Consequently, these probabilities will not provide information regarding the performance of the provider. The coefficients for membership in risk categories can however be used to provide estimates of risk for patients that is independent of treatment by providers. A risk of death associated

with membership in a risk category can be estimated as

$$\hat{p}_l = \frac{\exp(\hat{\alpha} + \hat{\beta}_l)}{1 + \exp(\hat{\alpha} + \hat{\beta}_l)} \tag{2.50}$$

These risks, however, are not appropriately scaled for use with risk adjustment procedures, since the sum of these risks will not equal the number of observed deaths for the entire sample under study. Consequently, all of the risk adjusted estimates may be over or under estimated, depending on the choice of baseline provider and on the degree and nature of confounding between the risk factors and the providers. The probabilities produced by the above procedure correspond to the probabilities that are fitted for the baseline provider. For example, if the provider with the worst performance was chosen as the baseline, then for the rest of the providers, the expected death rates obtained on the basis of the risk factors would be greater than the observed number of deaths. Consequently, the baseline provider would have $O_k = E_k$, while the other providers would have $O_k < E_k$. While such a procedure would preserve the ranks of the providers, it would not provide information as to whether the providers were performing better than expected. Instead, this procedure would provide information as to whether the providers were performing better than the baseline provider.

An offset model. The data can be rescaled to allow meaningful comparisons. To re-scale, a new value for α, α_o can be obtained by performing another logistic regression where the β_l's associated with each individual are used as offsets in the model. An offset is a component in the model that is known and requires no coefficient (for details, see appendix A). The resulting fitted probabilities, $\hat{p}_{ol}$, will be scaled so that $O = E$. These probabilities are the weighted average of the $\hat{p}_{lk}$, which

are the stratum specific probabilities of death associated with the different providers. They can be interpreted as probability of death expected if all individuals in a risk stratum were treated by some super provider which treated all patients.

Risk Adjusted Estimates. The expected number of deaths for provider k are obtained as

$$E_k = \sum_{l=1}^{s} n_{kl} p_{ol},$$

and SMR_k and $\hat{P}_k$ are obtained as in 2.15 and 2.23. Estimates of variance for these measures can be obtained using the procedures detailed in appendix A (section A.1).

Direct Standardization

The full logistic regression model can be used to obtain directly standardized estimates of relative risk and the population averaged proportion. These measures are based on E_k^*, which is the number of deaths expected for all patients if they were all treated by provider k. Using the logistic regression coefficients, E_k^* can be obtained as

$$E_k^* = \sum_{l=1}^{s} n_l \frac{\exp(\alpha + \beta_l + \gamma_k)}{1 + \exp(\alpha + \beta_l + \gamma_k)}.$$

The E_k^* can then be used in the calculation of risk adjusted estimates by 2.35 and 2.39. Variance estimates for these measures are derived by asymptotic likelihood theory in appendix A (section A.2).

2.4.2 Models Ignoring Provider Effects

A common practice in risk adjustment will be referred to as baseline-model (BM) risk adjustment. In BM adjustment, the models used to estimate the p_l ignore provider

effects. A logistic regression model is fit without including providers and stratum specific estimates of risk are obtained as

$$\hat{p}_l = \frac{\exp(\alpha + \beta_l)}{1 + \exp(\alpha + \beta_l)}.$$

These $\hat{p}_l$ are then used to obtain the expected numbers of deaths for each provider, and risk adjusted estimates such as the SMR or population averaged proportion obtained using 2.15 or 2.23. Variances of these estimates are then typically obtained using 2.20 or 2.40.

A problem with this approach is that the $\hat{p}_l$, will differ from the $\hat{p}_{ol}$ obtained using the offset method described above. Further, they are obtained from a logistic regression model which does not adjust for provider effects. This has the potential to yield incorrect results, since the risk adjustment is being performed because there is a suspicion that the provider effects are confounded by patient mix on important risk factors. This practice has been defended on the grounds that it is a practical procedure and that while error prone, errors in the magnitude or variance of the risk adjusted will be of little consequence, and the correct ranking among providers will be maintained. The procedure is practical in the sense that it is computationally simple. Variables are not required to code membership in the different providers. This can be an important consideration for data sets in which many of providers may be compared. Further, the estimated probabilities may not be obtained from a single model, with sub-models being computed for different types of disorders. Finally, the BM estimates are easily obtained from the output of widely available statistical packages.

The following example demonstrates that it would be prudent to be aware that

Table 2.1: Crude and Adjusted Risks

	Provider A			Provider B			Provider C		
Risk	n	y	%	n	y	%	n	y	%
Low	500	5	1	500	10	2	5000	150	3
High	5000	250	5	2000	200	10	500	75	15
Risk Ratio			5			5			5

	Summary Risks	
	Crude Risks	Adjusted Risks
Low Risk Category	0.0275	0.0159
High Risk Category	0.07	0.0793
Risk Ratio	2.55	5

not adjusting for treatment effects can bias the risk adjusted estimates. In this example, there are three providers, A, B, and C and one risk factor with two levels (see table 2.1). The providers are strongly confounded with the distribution of patients in the risk factor; in provider A, the ratio of patients in the high vs. low risk categories is 10:1. In provider B, this ratio is 4:1 and in provider C, the ratio is 1:10. The risks associated with treatment by the providers follow a homogeneous, multiplicative model. For each provider, the relative risk associated with the risk factor is 5. Within each risk category, however, the risk associated with treatment by provider B is twice as large as that associated with treatment by provider A. The risk of treatment associated with provider C is three times as large as that associated with provider A.

The adjusted risk ratio (RR) associated with the risk factor is a weighted average of the relative risks for each provider. Using the adjusted RR and assuming homongeneity of the effects for each provider, the risk associated with the low risk

Table 2.2: Risk-Adjusted Measures

	Provider A	Provider B	Provider C
	Outcome Measures		
O	255	210	225
E_{BM}	363.75	153.75	172.5
E_{FM}	404.48	166.55	118.97
	Risk Adjusted Relative Risks		
O/E_{BM}	.701	1.366	1.304
O/E_{FM}	.630	1.261	1.891
Direct	.630	1.261	1.891

category can be obtained as

$$R_{low} = \frac{\sum_{k=1}^{3} R_{low_k}(n_{low_k} + RR \times n_{high_k})}{\sum_{k=1}^{3}(n_{low_k} + RR \times n_{high_k})}.$$

Where $k = 1, 2, 3$ refers to providers A, B, and C. The adjusted risk for the high risk category is obtained in a similar fashion.

In the low risk category, the crude risk has an upward bias due to confounding. The crude estimate of risk is 0.0275, while the adjusted estimate of risk is 0.0159. In the high risk category, the crude estimate of risk has a downward bias. The crude estimate of risk is 0.07 while the adjusted estimate is 0.0793. The bias in these estimates of risk then leads to bias in the BM risk adjusted estimates (see table 2.2). This bias is due to differences in the distribution of the risk categories across providers. For example, provider C has a large proportion of patients in the low risk category. Because there is an upward bias in the estimate of risk in the low risk category, there is an upward bias in the number of expected deaths for this provider. From table 2.2, the expected number of deaths associated with provider C based

on crude estimates is 172.5 and is much higher than the expected number of deaths based on FM adjusted estimates (118.97). When these expected deaths are then used to calculated SMRs, the direction of the bias reverses, since the expected number of deaths is placed in the denominator and because the observed number of deaths is independent of the method which is applied. Therefore, the SMR for provider C based on crude estimates has a downward bias (1.304 vs. 1.891). Also included in table 2.2 are risk adjusted estimates of risk for providers which are based on direct standardization (DS). Note that the directly standardized risks are identical to the SMRs calculated using the FM adjusted estimates of risk. An important feature of these two measures is that they maintain the relative risks among providers that were observed in table 2.1. The ratio of the risk adjusted estimates for provider B compared to provider A is 2, and the ratio comparing provider C to provider A is 3. This does not hold for the BM risk adjusted estimates calculated on the basis of risks which are not adjusted for provider effects. The ratio for provider B vs A is close to a value of 2 (1.366/.701 = 1.94), but the ratio for provider C vs A (1.304/.701 = 1.86) is not close to the correct value of 3. This is especially problematic, since the ranking among providers has not been preserved among these risks. When using baseline adjusted risks for the risk categories, the risk associated with provider C (O/E_{BM} = 1.304) is now smaller than the risk associated with provider B (O/E_{BM} = 1.366). In addition to the potential for bias in point estimates obtained using the BM of risk adjustment, results in chapter 5 also indicate that the commonly used method of obtaining variances for BM adjusted SMRs may result in biased estimates of variance.

2.4.3 Discussion

Although many risk adjustment methods are available, research in the remaining chapters will focus on measures of risk and relative risk such as the population averaged proportion and the SMR, as these methods are typically employed when using logistic regression to perform risk adjustment. When examining the use of missing data methods for risk adjustment, three risk adjustment strategies will be investigated. The first will be BM estimates, as these are often employed for risk adjustment. In keeping with the usual application of these measures, variance estimates will be obtained under the assumption that the fitted probabilities are the true probability values. These risk adjustment measures will be contrasted with two other types of full model estimates: FM estimates obtained using an offset model for the logistic regression, and directly standardized estimates. Variance calculations for the FM estimates will be based on asymptotic likelihood theory (see appendix C) and will account for variability in the fitted probabilities.

Chapter 3

Missing Data Methods for Risk Adjustment

Risk adjustment procedures often require the use of numerous variables, since many risk factors may influence patient outcomes. Although the amount of missing data in each covariate may be modest, the proportion of subjects with missing data on one or more of these covariates may be quite large. If risk adjustment is performed using only the cases with complete data, patterns of missing data in the covariates may require the exclusion of a large number of subjects. The subjects remaining in the analysis may no longer be representative of the entire data set, and the elimination of subjects also leads to a loss of efficiency, which results in inflated estimates of variability.

As noted in Chapter 2, risk adjustment estimates are often based on logistic regression models, and logistic regression can also be employed to perform risk adjustment procedures with the APPROACH data. Several methods have been employed for handling missing data in logistic regression, but for methods to be of practical use in risk adjustment, they have to meet the following five criteria.

1. **Estimates of variance.** The methods need to yield information which can be used to calculate not only risk adjusted estimates for providers, but also estimates of variability.

2. **Continuous and categorical covariates.** The methods need to be able to handle situations in which covariates are continuous or categorical, as well as

situations in which there is a mixture of continuous and categorical covariates.

3. **Large data sets.** The methods have to be capable of working with large data sets and with multiple covariates. Authors often explore the methods they propose using data sets which are of limited size and complexity. For the purposes of risk adjustment, missing data methods must be able to work with data sets containing 20 or more covariates and thousands of cases. Limitations of the methods are usually due to limitations in computer memory, storage and processing speed, although the use of many categorical covariates can lead to an inability to obtain unique parameter estimates.

4. **Rare risk factors and outcomes.** The methods need to work in situations where outcomes are rare and where adjustments are made on the basis of risk factors which have low prevalence or incidence.

5. **Availability.** The methods need to be available for use by researchers. In general, the missing data methods are not difficult to implement. However, with increasing numbers of covariates, efficient programming is not trivial. With the exception of multiple imputation (MI) methods, implementations of missing data methods are not widely available.

Before discussing missing data methods for logistic regression and risk adjustment, however, it will be useful to examine general approaches which have been used for handling missing data as well as some of the theoretical and practical issues surrounding the use of these methods.

3.1 Missing Data Methods

Many methods have been utilized for handling missing data. The most commonly used methods are ones which have been described as simple or quick procedures. Most of these methods are easily applied using standard statistical software. A major weakness of these procedures is that they are not based on statistical theory or rationale. The use of these procedures can lead to estimates that may be biased. Some of these methods are inefficient, while others provide variances estimates that are smaller than are warranted on the basis of the available data. More sophisticated methods, such as multiple imputation (MI) or the expectation-maximization (EM) algorithm can provide estimates which are unbiased and which make efficient use of the data. The following notation will be used to describe the missing data methods as well as the mechanisms under which the methods can yield valid results.

Notation

The notation of Little and Rubin (1989a); Schafer (1997a); Brand (1999), will be employed for the following discussions of missing data mechanisms and missing data methods. Let Y be an $n \times k$ matrix of complete data, where rows $i = 1, 2, \ldots, n$ denote observations for individual cases, and columns $k = 1, 2, \ldots, p$ denote observations for the variables. The observed portion of Y will be denoted as Y_{obs} and the missing portion as Y_{mis} so that $Y = (Y_{obs}, Y_{mis})$.

To specify models for the missing data, let R be an $n \times k$ matrix of indicator variables with elements corresponding to the elements of Y. The elements of R will be 0 if the corresponding elements of Y are missing and 1 of the corresponding elements of Y are observed. Probability models for the missing data mechanism

which are dependent on Y will be denoted as $\Pr(R|Y,\psi)$, where *psi* is a vector of parameters which determines the joint distribution of R conditional on Y.

3.1.1 Missing Data Mechanisms

When the mechanism which gives rise to the missing data is known or is under the control of the researcher, this mechanism can be included in the likelihood equation, and maximum likelihood estimates can be obtained for the parameters in the probability model. Although there are specialized methods for handling these cases, these methods will not be considered in this dissertation, since the missing data mechanisms for data used in risk adjustment are generally not known. Consequently, it is important to consider the conditions under which the missing data mechanism can be ignored when employing missing data methods.

Missing Completely at Random (MCAR)

If for every variable with missing data, the missing observations are a random subsample of observations for each variable, the missing data is missing completely at random (MCAR). When data are MCAR, the complete cases form a representative subset of the entire data set, and estimates based on the use of complete cases will be unbiased. Data which are MCAR may occur when observations are randomly missed due to equipment failure or transcription errors.

Another, less stringent, missing data mechanism is the stratified-MCAR condition (Greenland and Finkle, 1995). When data are stratified-MCAR the probability that observations are missing can depend on the levels of completely observed covariates. However, within the strata of the completely observed covariates, the variables with

missing data must be MCAR. For example, ejection fraction (EF) is an important clinical indicator of cardiac function which is related to the age of the patient. If age is completely observed but ejection fraction has missing observations, for missing EF observation to be stratified-MCAR, the probability that ejection fraction is missing can depend on the age of the patient. For subjects of a given age, however, the probability that ejection fraction is missing cannot depend on the value of the ejection fraction variable.

Missing at Random (MAR)

Rubin (1976) demonstrated that likelihood estimates could be consistent provided that the probability that a covariate was missing did not depend on the value of the missing covariate or on the value of any other missing covariate. The probability that an observation was missing could, however, depend on the value of other observed covariates. Rubin called this condition missing at random (MAR). The MAR condition is less stringent than either the MCAR or stratified-MCAR condition; when missing observations are MCAR or stratified-MCAR, the data are also MAR.

In many situations with missing data, the MAR assumption is clearly not met. Consider studies that ask questions of a sensitive nature, such as annual income or HIV status. The probability that a person responds to these questions is likely to depend on the true answer, and the data cannot be considered to be MAR. In the case of the APPROACH data, it is quite possible that the clinical variables are not MAR, since the probability that physicians took the time to collect information may have depended on the severity of the symptoms presented by the patient.

For the data to be MAR, all information necessary for specifying the a model for

the missing data must be contained in Y_{obs}. The inclusion or exclusion of variables in the data can affect whether or not a data set meets the MAR condition. If congestive heart failure is related to age, and congestive heart failure is stratified MCAR with respect to age, then the exclusion of age from the analysis will result in a data set which is non-MAR. This is because the probability that CHF is observed will depend on whether or not a case has CHF, regardless of whether the observation is missing. If we assume that age is completely observed, the inclusion of age in the data set restores the stratified MCAR condition, and modeling the joint distribution of age and CHF on the basis of Y_{obs} will yield valid inferences regarding CHF.

Likelihood theory. Little and Rubin (1989b) provide likelihood theory that applies to joint distributions with missing data. Let $\boldsymbol{\theta}$ be a vector of parameters that determine the joint distribution of Y. Our interest is in making inferences regarding the parameters $\boldsymbol{\theta}$ on the basis of the marginal probability density of Y_{obs}, which can be obtained by integrating the missing data out of the joint distribution of Y, or

$$f(Y_{obs}|\boldsymbol{\theta}) = \int f(Y_{obs}, Y_{mis}|\boldsymbol{\theta})dY_{mis}. \tag{3.1}$$

Following Little and Rubin (1989b), the likelihood based on Y_{obs} which ignores the missing data mechanism will be any function of $\boldsymbol{\theta}$ which is proportional to 3.1. To determine the conditions under which the missing data mechanism may be ignored, consider the probability distribution which includes the missing data mechanism. When data are incomplete, the observed data consist of Y_{obs} and R, and the likelihood of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ will be any function of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ proportional to

$$f(Y_{obs}, R|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(Y_{obs}, Y_{mis}|\boldsymbol{\theta})f(R|Y_{obs}, Y_{mis}, \boldsymbol{\psi})dY_{mis}. \tag{3.2}$$

Following Rubin (1976) and Little and Rubin (1989b), the missing data mechanism can be ignored and inference for $\boldsymbol{\theta}$ can be based on the likelihood $L(\boldsymbol{\theta}|Y_{obs})$ when

$$f(R|Y_{obs}, Y_{mis}, \boldsymbol{\psi}) = f(R|Y_{obs}, \boldsymbol{\psi}). \tag{3.3}$$

This allows the conditional distribution of the missing data to be moved outside of the integral for the joint density. It can then be factored as:

$$\begin{aligned} f(Y_{obs}, R|\boldsymbol{\theta}, \boldsymbol{\psi}) &= f(R|Y_{obs}, \boldsymbol{\psi}) \times \int f(Y_{obs}, Y_{mis}|\boldsymbol{\theta}) dY_{mis} \\ &= f(R|Y_{obs}, \boldsymbol{\psi}) f(Y_{obs}|\boldsymbol{\theta}). \end{aligned} \tag{3.4}$$

Provided that the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are distinct, 3.4 will be proportional to 3.1, and the estimates of $\boldsymbol{\theta}$ obtained by maximizing $L(Y_{obs}|\boldsymbol{\theta})$ will be the same as those obtained by maximizing $L(Y_{obs}, R|\boldsymbol{\theta}, \boldsymbol{\psi})$.

3.1.2 Quick and Simple Methods

Complete Cases

Several authors have addressed the performance of simple and commonly used methods of handling data with missing values (Little and Rubin, 1989c; Greenland and Finkle, 1995; Vach, 1994; Vach and Blettner, 1991; Brand, 1999). The most commonly used method is the complete-case method, also referred to as list-wise deletion. In this method, cases with missing values on any of the variables of interest are simply excluded from the analysis. All standard statistical packages offer this method, and most default to this method when missing values are present. Miettinen (1985) mistakenly claimed that this was the only legitimate method of handling missing data. Not only is this method inefficient, it has been demonstrated to be prone to

bias unless the data are MCAR (Little and Rubin, 1989c; Vach, 1994; Vach and Blettner, 1991). The degree of bias is generally greatest when the probability that values are missing is dependent on the outcome variable.

Additional Categories

When the covariates are categorical, researchers have attempted to add an additional category denoting whether or not observations are missing. Vach (1994) and Vach and Blettner (1991) demonstrate that this method can introduce considerable bias even when the observations are MCAR. Analogous methods using indicator variables have been used with continuous covariates (Greenland and Finkle, 1995). When covariates are continuous, each variable with missing values is replaced by two variables. One of these new variables takes on the observed values for cases without missing data, and 0 for subjects with missing data. The second variable is an indicator variable coded with a 1 if the data is missing and a 0 if the data is observed. Like the additional category method, the use of indicator variables can result in considerable bias (Greenland and Finkle, 1995). A modified indicator approach can be used to reduce this bias (Greenland and Finkle, 1995). However, the additional parameters that are estimated using this approach lead to a loss of efficiency, and simulations performed by Greenland and Finkle (1995) indicate that it is no more efficient than a complete case analysis.

Removal of Variables

When missing data is limited to a few covariates, the variables with missing data can be removed from the analysis. Such a procedure is questionable, as important information may be excluded from the analysis. Since the goal of risk-adjustment

is to examine the performance of providers after controlling for risk factors known to be related to outcomes, the removal of variables could lead to inadequate risk adjustment and invalid conclusions regarding the relative performance of providers.

Single Imputation

For single imputation methods, the values for missing data are imputed on the basis of the observed covariates, and the imputed values are used to "fill in" the data set. This "filled-in" data set is then used in subsequent analyses. There are several methods for generating the imputations. These include unconditional mean imputation, conditional mean imputation, and cold deck and hot deck methods. In general, simple imputation results in an underestimation of the standard errors associated with 1) the parameters of the joint distribution of the covariates and 2) the regression parameters associated with the covariates (Little and Rubin, 1989c). These standard estimates are not asymptotically unbiased; the degree of bias in the estimates is not affected by sample size if the proportion of cases with missing data remains constant (Little and Rubin, 1989c). Further, these methods are affected by the accuracy of the method used for filling in the data.

Mean imputation. In unconditional imputation, missing covariate values are filled in with the sample mean of the recorded values of the covariate. Since the imputed values are placed at the center of the distribution, the variance of the resulting covariate will underestimate the true variance. If the data are MCAR, the estimate of the mean for this covariate will be unbiased. Conditional imputation is a more sophisticated form of imputation. In conditional imputation, complete cases are used to form regression models, where the covariates with missing data are re-

gressed on the observed covariates. These regression models are then used to impute values for the missing data conditional on the observed covariate values. As with unconditional imputation, results based on conditional imputation underestimate the variances and covariances of the covariates (Little and Rubin, 1989c). Buck (1960) provided a method of adjusting the variances for multivariate normal data, but valid inferences based on this method require that the missing observations be MCAR.

Deck methods. Both cold and hot-deck methods involve the random selection of a value from the possible values that the missing data could assume given the subjects observed covariates. In cold deck imputation, covariate values are drawn from an external source. Selections for the missing data are drawn from the subjects who have similar patterns of response on the observed covariates. In hot deck imputation, selections are drawn from the distribution of subjects in the sample who have completely observed covariates. The term "deck" refers to the deck of computer cards which are similar to the subject with the missing data (Little and Rubin, 1989d). In general, the results of these procedures will underestimate variances, since they treat imputed values as if they are measured with certainty (Rubin, 1987a). Because the "deck" from which imputations are drawn is based on complete cases, legitimate inference requires that this deck be representative of cases with missing data. A weakness of this approach is that this deck may not represent the possible range of values from the population of interest, resulting in estimates of variance which are too small.

The problems inherent with the use of simple methods in the presence of missing data are perhaps best summarized by Little and Rubin (1989c) who state that

> . . . it is hard to recommend any of the simple methods discussed since (1) their performance is unreliable; (2) they often require ad hoc adjustments to yield satisfactory estimates, and (3) it is not easy to distinguish situations when the methods work from situations when they fail. Furthermore, the methods fail to provide simple correct answers when measures of the precision of estimates are required, as for interval estimation.

3.1.3 Multiple Imputation

Multiple imputation requires the imputation of more than one value for each missing observation. The general method for multiple imputation is to generate several "filled in" data sets, in which values are imputed to complete the data in each data set. Estimates of the parameters and their variances are then obtained by combining the estimates from each of the data sets (Little and Rubin, 1989e; Rubin, 1987a; Greenland and Finkle, 1995).

By following the methods described by Rubin (1987b) and Schafer (1997b), estimates for coefficients and their variances are easily obtained when using multiple imputation. Consider the scalar estimate Q, which can be a parameter of the imputation model or a function of the parameters in the model. In the case of risk adjustment, Q could be a logistic regression coefficient, or a risk or SMR based on the regression model. Let $\hat{Q}(Y_{obs}, Y_{mis})$ be a complete data point estimate of Q and let $U(Y_{obs}, Y_{mis})$ be a variance estimate associated with $\hat{Q}(Y_{obs}, Y_{mis})$. For brevity, these will also be denoted as $\hat{Q}$ and U. The use of multiple imputation assumes that the sample size is large enough for the normal approximation (Rubin, 1987b; Schafer, 1997b).

$$U^{-1/2}(Q - \hat{Q}) \sim N(0, 1).$$

Further, it is assumed that $\hat{Q}$ is a first order approximation to $E(Q|Y_{obs}, Y_{mis})$ and that U is a first order approximation to $V(Q|Y_{obs}, Y_{mis})$.

Obtaining MI estimates of Q and U requires the imputation of m sets of Y_{mis}. These sets of Y_{mis} are then used to obtain complete data estimates of $\hat{Q}$ and U, where these complete data estimates are

$$\begin{aligned} \hat{Q}^{(t)} &= \hat{Q}(Y_{obs}, Y_{mis}^{(t)}) \\ U^{(t)} &= U(Y_{obs}, Y_{mis}^{(t)}) \end{aligned}$$

for imputations $t = 1, 2, \ldots, m$. The MI estimate for Q, denoted as $\bar{Q}$, is obtained by taking the sample average of these complete data estimates

$$\bar{Q} = \frac{1}{m}\sum_{t}^{m} Q^{(t)}. \tag{3.5}$$

The MI estimate for the variance of $\bar{Q}$ has both a within-imputation component and a between-imputation component. The within-imputation component is obtained as the sample average of the complete data estimates:

$$\bar{U} = \frac{1}{m}\sum_{t}^{m} U^{(t)}. \tag{3.6}$$

The sample variance of the complete data point estimates is used to estimate the between imputation variance, or

$$B = \frac{1}{m-1}\sum_{t}^{m}(\hat{Q}^{(t)} - \bar{Q})^2 \tag{3.7}$$

These estimates are then combined to obtain T, an estimate of the total variance. The variance estimates are combined using

$$T = \bar{U} + (1 + m^{-1})B. \tag{3.8}$$

Inferences regarding $\bar{Q}$ are based on a t distribution with v degrees of freedom, where

$$v = (m-1)\left[1+\frac{\bar{U}}{(1+m^{-1})B}\right]^2.$$

In applications such as risk adjustment, where the fraction of missing information is moderate and the sample sizes are large, v will be large and the standard normal distribution can be used as an estimate for t_v. When Q is a vector, matrix analogues for combining the complete-data estimates can be employed. These can be useful for comparing two models for the data or when one wishes to obtain confidence regions for the vector Q.

The Fraction of Missing Information

The fraction of missing information associated with a scalar Q can be obtained using the estimates of between and within imputation variances. The relative increase in variance due to non-response is

$$r = \frac{(1+m^{-1})B}{\bar{U}}$$

and an estimate of the fraction of missing information is

$$\hat{\lambda} = \frac{r+(2/(v+3))}{r+1}.$$

Proper Imputations

Rubin (1987c) provides the conditions which must be met for valid multiple imputation inferences. Imputations which meet these criteria are proper in the sense that estimators and their variances which are based on these imputations will reflect the uncertainty in the probability model used to generate the imputations.

However, since it can be extremely difficult to determine whether imputations are proper (Schafer, 1997b), these conditions will not be discussed. Instead, the simpler Bayesianly proper approach of Schafer (1997b) will be adopted. Bayesianly proper imputations yield valid Bayesian and likelihood inferences, but may not yield proper imputations for frequentist estimates.

Bayesianly proper imputations. An important feature of proper multiple imputations is that they not only reflect uncertainty regarding Y_{mis} given the parameters in the complete-data model, but also uncertainty about the unknown model parameters. Multiple imputations are said to be Bayesianly proper if they are independent realizations of the posterior predictive distribution of the missing data under a complete data model and prior. This distribution, denoted as $\Pr(Y_{mis}|Y_{obs})$, can be obtained as the conditional predictive distribution of Y_{mis}, averaged over the observed-data posterior distribution of θ. This can be written as

$$\Pr(Y_{mis}|Y_{obs}) = \int \Pr(Y_{mis}|Y_{obs},\theta)\Pr(\theta|Y_{obs})d\theta.$$

Since this distribution does not depend on the missing data model for R, imputations which are Bayesianly proper are appropriate under the the same conditions of ignorability which are required for valid likelihood inference (Schafer, 1997b).

Generating Imputations

One method which has been employed for generating multiple imputations uses Y_{obs} to obtain a maximum likelihood estimate for the parameters of the joint distribution Y and then randomly samples $Y_{mis}|Y_{obs},\hat{\theta}$ from this distribution. This strategy was employed by Greenland and Finkle (1995). As they pointed out, however, these

imputations were not proper, because they treated $\hat{\theta}$ as fixed and did not reflect the uncertainty in the estimation of θ. Using a Markov chain Monte Carlo (MCMC) technique, Schafer (1997a) provides a means of generating imputations which are Bayesianly proper. These methods, described below, are implemented by Schafer (Schafer, 1999) and are available over the World Wide Web as stand-alone programs for the Windows 95/98/NT operating systems or as S-PLUS (MathSoft, 1999) libraries. Finally, hot deck methods are available that can provide proper imputations.

Hot Deck Imputation Hot deck methods can be used to draw multiple imputations from Y_{obs}. Generally, these imputations are not proper, and inferences based on these imputations are not valid (Rubin, 1987c). This is because the distribution of sample Y values is treated as if it is the population distribution. The variance of the imputed values will underestimate the true variance, since they have been sampled with a degree of precision not warranted on the basis of the sample data. To be proper, the imputations must reflect the uncertainty in using Y to represent the population distribution. Proper hot deck imputations can be generated using the approximate Bayesian bootstrap (ABB) described by Rubin (1987c). In the ABB, values of Y_{mis} are sampled with replacement from the distribution of Y^*_{obs}, which has itself been sampled with replacement from Y_{obs}. This approximate Bayesian bootstrap is implemented in the SOLAS (Solutions, 1999) software package. Unfortunately, the purchase price of the SOLAS implementation precluded its use in the present research.

Data augmentation Data augmentation is a form of Markov chain Monte Carlo which can be used to make imputations which are proper for likelihood inference

(Schafer, 1997c). Data augmentation can be used to make pseudo-random draws from probability distributions which are intractable or cannot be easily summarized because of missing data. In many cases, it is difficult to sample from the posterior distribution $\Pr(\theta|Y_{obs})$. Data augmentation reduces the difficulty in generating draws from the observed data posterior by augmenting Y_{obs} with Y_{mis}. Two steps are required for data augmentation. In the Imputation step or I-step, values for the missing data are sampled from the predictive distribution of Y_{mis}:

$$Y^{(t+1)} \sim \Pr(Y_{mis}|Y_{obs}, \theta^{(t)}).$$

The obtained value of $Y_{mis}^{(t+1)}$ is then used in the Posterior step or P-step to draw a new value of θ from the complete data posterior:

$$\theta^{(t+1)} \sim \Pr(\theta|Y_{obs}, Y_{mis}).$$

It is possible to use data augmentation to simulate the posterior distribution of the parameters. However, Schafer points out that this is computationally more expensive than using imputations sampled from the MCMC chain in MI procedures.

The use of data augmentation procedures for MI requires information regarding how large k must be for $\theta^{(t+k)}$ to be independent of $\theta(t)$. After a sufficient "burn in" period to allow the distribution of iterates to converge to a stationary iteration, every k^{th} iterate of θ could then be taken as an independent draw from $\Pr(\theta|Y_{obs})$.

Monitoring Convergence. While there has been theoretical work regarding the rate of convergence of Markov chains, this work does not translate into practical guidelines for knowing when convergence has occurred (Schafer, 1997c). Using simple examples, Schafer demonstrates that the rate of convergence is related to the fraction

of missing information. He also points out, however, that this relationship is difficult to formalize in a general way.

Schafer (1997b) suggests several means of monitoring the rate of convergence. These include:

1. Examination of the rate of convergence of the expectation-maximization (EM) algorithm. The EM algorithm, described in section 3.1.4, is an iterative method of obtaining maximum likelihood estimates (MLEs) in the presence of missing data.If the posterior distribution has multiple modes or is oddly shaped, the EM algorithm will converge slowly. Further, the MLEs obtained from the EM algorithm may depend on the starting point for the iterations.

2. Monitoring components or scalar functions of θ. The convergence behavior of the θ's can be monitored using time series plots and autocorrelation plots. Time series plots can be examined to determine how long it takes for the sequence to converge to an area of high density. Autocorrelation plots can be used to obtain linear trends among the parameters and can be used to detect long term trends or drifts in scalar summaries of θ. These trends or drifts indicate a slow rate of convergence to stationarity. If the time series wander, this is an indication that components of θ may be nearly or entirely inestimable from Y_{obs}. as

There are, however, problems in using plots to monitor convergence. A correlation of 0 is not the same as independence; non-linear associations may exist among the iterates. The sequence may not have converged with respect to functions or parameters which have not been examined. Finally, the posterior distribution may

be oddly shaped and the sequence of iterates may not have visited regions which might yield plausible choices for the mode. Schafer offers two possible solutions to these problems. The first is to attempt multiple runs from different starting values. The second is to examine the time-series plot of the worst linear function (WLF) of the parameters.

The WLF requires an estimate of v_1, the eigenvector associated with the largest eigenvalue of the rate matrix from the EM algorithm. Schafer asserts that of all linear functions, the asymptotic rate of missing information will be highest for $v_1^T\theta$. He shows that near the mode, $\hat{v}_1 = \theta^{(t)} - \hat{\theta}$ is approximately proportional to v_1, and suggests the use of

$$\xi(\theta) = \hat{v}_1^{(t)}(\theta - \hat{\theta})$$

as the worst linear function, where $\hat{\theta}$ is the MLE of θ obtained from the EM algorithm. In this function $\hat{\theta}$ is subtracted from θ to indicate the position of the function to the mode with respect to $\hat{v}_1$. Schafer (1997a) reports that in real-data problems this is one of the slowest functions to converge when the observed-data posterior distribution is nearly normal, but that other functions may be slower to achieve stationarity when some parameters are poorly estimated.

Multiple imputation methods show a great deal of promise. They are easily implemented; many statistical programs now contain methods of obtaining parameters for multivariate normal distributions with missing data. For the APPROACH data, the primary drawback of the use of multivariate normal distributions to generate imputations is that the APPROACH data consists primarily of categorical and dichotomous variables. Parametric methods of obtaining imputed values from discrete

distributions are available, and these will be discussed in following section 3.2.3.

3.1.4 Likelihood Based Methods

Traditional likelihood methods can sometimes be applied to $f(Y_{obs}|\theta)$. In these cases, the first and second derivatives of the log-likelihood with respect to θ

$$\ell(\theta|Y_{obs}) = \log(f(Y_{obs}|\theta))$$

can be used to obtain MLEs and estimates of variance for θ. Often, however, these derivatives are intractable. In many of these cases, the expectation-maximization (EM) algorithm provides an alternative means of obtaining MLEs.

The EM Algorithm

The EM algorithm is an iterative method of obtaining maximum likelihood estimates in the presence of missing data. For each iteration, there is an E-step and an M-step. In the E-step, the expectation of the complete data log-likelihood is obtained using the observed data and the current estimates of the parameters $\theta^{(t)}$. For the M-step, the expectation of the log-likelihood is maximized with respect to the parameters (Dempster et al., 1977). Following the notation of Schafer (1997c) and Little and Rubin (1989f), consider $Y = (Y_{mis}, Y_{obs})$ where Y_{mis} and Y_{obs} are the missing and observed components of Y.

The E-Step. The complete data log-likelihood can be factored as:

$$\begin{aligned}\ell(\theta|Y) &= \ell(\theta|Y_{obs}, Y_{mis}) \\ &= \ell(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs}, \theta)\end{aligned}$$

The expected complete data log-likelihood is:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \int \ell(\theta|Y) f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis} \\ &= \ell(\theta|Y_{obs}) + \int \ell(\theta|Y_{mis}) f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis} \\ &= \ell(\theta|Y_{obs}) + H(\theta|\theta^{(t)}) \end{aligned} \tag{3.9}$$

The M-Step. If $\theta^{(t+1)}$ is the value of θ that maximizes the expected log-likelihood, then

$$\ell(\theta^{(t+1)}|Y) \geq \ell(\theta^{(t)})$$

This result is central to the validity of the EM algorithm, and can be verified by noting that:

$$\ell(\theta|Y_{obs}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}) \tag{3.10}$$

and

$$\begin{aligned} &\ell(\theta^{(t+1)}|Y_{obs}) - \ell(\theta^{(t)}|Y_{obs}) \\ &= \left[Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})\right] - \\ &\quad \left[H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})\right] \end{aligned}$$

The M-step of the EM algorithm ensures that $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$, and Jensen's inequality can be used to show that

$$\left[H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})\right] \leq 0.$$

The observed information matrix can be obtained as the negative of the second derivative of $\ell(\theta|Y_{obs})$

$$I(\theta|Y_{obs}) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta|Y_{obs}).$$

The inverse of $I(\theta|Y_{obs})$ evaluated at the observed data MLE for θ (denoted as $\hat{\theta}$) provides a large-sample estimate of the variance-covariance matrix for $\hat{\theta}$. The primary advantage of the EM algorithm, however, is that it avoids the calculation of the derivatives of $\ell(\theta|Y_{obs})$. Applying the missing information principle (Woodbury, 1977), Louis (1982) provides a means of obtaining an estimate of $I(\theta|Y_{obs})$ which relies only on the calculation of complete-data derivatives. According to the missing information principle, by rearranging and taking the second derivative of 3.9, the observed information can be expressed as:

$$I(\theta|Y_{obs}) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta|Y_{obs}, Y_{mis}) + \frac{\partial^2}{\partial\theta^2}\log f(Y_{mis}|Y_{obs}, \theta).$$

Provided that there is sufficient regularity to pass the differentials with respect to θ through the integral signs, taking expectations with respect to $\Pr(Y_{mis}|Y_{obs}, \theta)$ yields

$$I(\theta|Y_{obs}) = -\frac{\partial^2}{\partial\theta^2}Q(\theta|\theta) + \frac{\partial^2}{\partial\theta^2}H(\theta|\theta). \tag{3.11}$$

If we refer to $-\frac{\partial^2}{\partial\theta^2}Q(\theta|\theta)$ as the complete information and $-\frac{\partial^2}{\partial\theta^2}H(\theta|\theta)$ as the missing information, then 3.11 has the useful interpretation

$$\text{observed information} = \text{complete information} - \text{missing information}.$$

Louis (1982) demonstrated that when evaluated at $\hat{\theta}$, the missing information can be expressed as

$$-\frac{\partial^2}{\partial\theta^2}H(\hat{\theta}|\hat{\theta}) = E\left[S(\theta|Y_{obs}, Y_{mis})S^{(t)}(\theta|Y_{obs}, Y_{mis})|Y_{obs}, \theta\right]\Big|_{\theta=\hat{\theta}},$$

where S denotes the score function and E denotes expectation. The observed information can then be obtained as

$$I(\hat{\theta}|Y_{obs}) = -\frac{\partial^2}{\partial\theta^2}Q(\hat{\theta}|\hat{\theta}) - E\left[S(\theta|Y_{obs}, Y_{mis})S^T(\theta|Y_{obs}, Y_{mis})|Y_{obs}, \theta\right]\Big|_{\theta=\hat{\theta}} \tag{3.12}$$

The inverse of 3.12 provides an estimate of the variance-covariance matrix of $\hat{\theta}$. Details of how this method can be employed to obtain variance estimates for logistic regression parameters are presented in appendix A.2.

Rate of Convergence

The rate of convergence of the EM algorithm is approximately linear near the mode. For scalar parameters, Dempster, Laird and Rubin (1982) demonstrated that this rate of convergence is determined by the ratio of missing information to complete information. They denote this as λ, the fraction of missing information. The rate of convergence can be approximated as

$$(\theta^{(t+1)} - \hat{\theta}) \approx \lambda(\theta^{(t)} - \hat{\theta}).$$

For vector parameters, where θ is a vector of length $k > 1$, the rate matrix D is obtained as

$$D = \left[-\frac{\partial^2}{\partial\theta^2}Q(\theta|\theta)\right]^{-1}\frac{\partial^2}{\partial\theta^2}H(\theta|\theta).$$

The rate of convergence of the EM for vector parameters is determined by λ_1, the largest eigenvalue of the rate matrix. This is the fraction of missing information in the parameter space corresponding to the direction of v_1, the eigenvector associated with λ_1.

3.1.5 Other Likelihood-Based Methods

There are other likelihood-based methods for incomplete data. In general, these methods are applicable in special cases such as two-stage research designs, or when the patterns of missing data allow the likelihood to be factored in a manner which

makes traditional maximum likelihood estimation tractable (Pepe and Fleming, 1991; Reilly and Pepe, 1995). These methods cannot be applied in typical risk-adjustment problems, where the missing data mechanisms are not known and where the missing data patterns can be complex. Consequently, they will not be considered in this dissertation.

3.1.6 Evaluation of the methods.

The suitability of the methods for performing risk adjustment can be evaluated using the four criteria outlined at the beginning of the chapter. Two methods will be considered for obtaining risk-adjusted estimates in the presence of missing data, as these methods come closest to meeting the criteria described at the beginning of this chapter. These methods are 1) multiple imputation via data augmentation, and 2) EM by the method of weights (EMMW). Both of these methods have limitations, however, and neither satisfies all of the criteria. Difficulties in meeting the criteria are generally due to problems in specifying and fitting joint probability distributions. Multiple imputation requires the specification of a joint distribution for the covariates and the outcome measure; EMMW requires the specification of a joint distribution for the covariates.

1. **Estimates of variance**. Both of the methods can be used to provide estimates of variance for parameters in logistic regression and for risk adjusted estimates. Estimates of variance are somewhat difficult to obtain when using EMMW. Details are provided later in this chapter and in appendix A.3

2. **Continuous and categorical covariates.** Schafer has implemented impu-

tation methods for continuous, and mixed categorical and continous covariates (Schafer, 1999). His software for generating imputations from log-linear categorical models is not yet complete. Provided that continuous covariates are completely observed, EMMW can handle mixed continuous and categorical covariates. It can also data sets in which all covariates are categorical. If continuous variables are split into categories, EMMW can also be used in cases where missing covariates are continuous or where both missing and observed covariates are continuous.

3. **Large data sets.** Both methods are somewhat limited in their ability to work with large data sets. In the case of multiple imputation, these limitations are due to difficulties in 1) modeling the joint distributions of covariates, and 2) to making imputations on the basis of these models. For EMMW, the limitations are due to difficulties in 1) fitting the joint distributions and 2) to the construction of an augmented data matrix which accounts for possible values the missing observations could assume. This augmented matrix can become large and unwieldy in the presence of complicated missing data problems.

4. **Rare risk factors and outcomes.** Provided that appropriate joint distributions can be modeled for the data, both methods will work with rare risk factors and outcomes. With complex joint distributions, suitable covariate models may be difficult to find. For EMMW, rare outcomes are not problematic, as these are not included when fitting a joint distribution for the covariates.

5. **Availability.** Suitable commercial implementations are not yet available for either method. IML routines for performing multiple imputations using SAS

can be obtained from SAS technical support (Sarle, 1999). As noted previously, MI methods implemented by Joe Schafer are available over the World Wide Web as stand-alone programs for the Windows 95/98/NT operating systems or as S-PLUS (MathSoft, 1999) libraries (Schafer, 1999). S-PLUS will be releasing commercial implementations of Schafer's multiple imputation methods (Schimert, 1999). Other than fitting a joint distribution for the covariates, the primary difficulty in performing EMMW lies in the construction of an augmented data matrix. Once this matrix has been obtained, the method of weights algorithm can easily be implemented using standard statistical software. Variance estimates can be obtained by implementing the calculations outlined in appendices A.2 and A.3.

The first step in describing how these methods can be applied to risk adjustment is to show how they are used for regression with missing data.

3.2 Missing Data Methods for Regression

Notation. For regression models, the distribution of a response vector Y is modeled conditionally on a covariate matrix X, with parameters β. The conditional distribution of $\Pr(Y|X,\beta)$ is assumed to follow a particular underlying probability model. It will further be assumed that Y is fully observed, since this is generally the case when employing risk-adjustment methods. The covariate matrix X will be partitioned as $X = (X_{obs}, X_{mis})$, where X_{obs} and X_{mis} denote respectively the portions of X which are observed and missing. The use of MI requires the specification of the joint distribution of Y and X_{obs}. The parameters for this distribution will be denoted

as θ, and the joint distribution will be denoted as $\Pr(Y, X_{obs}|\theta)$. EM by method of weights utilizes the joint distribution of X_{obs}. The parameters for this distribution will also be denoted as θ, and this distribution will be denoted as $\Pr(X_{obs}|\theta)$.

3.2.1 Multiple Imputation

Parameters for regression models can be obtained using multiple imputation. The first step in obtaining these estimates is to specify a distribution for $\Pr(Y, X_{obs}|\theta)$. Several data sets are then obtained by using this distribution to impute values for X_{mis}. The appropriate regression is performed on each data set, and point estimates and estimates of variance are obtained using the methods described in section 3.1.3 above.

Greenland and Finkle (1995) employed multiple imputation methods in simulations investigating methods of handling missing covariates in logistic regression. The imputations were based on a multivariate normal model in which the covariates were continuous. These simulations demonstrated the superiority of the MI methods over simple missing data methods, even though only two imputations were performed for each of the generated data sets. The authors suggest that an increase in the number of imputations would have improved the performance of the MI methods.

3.2.2 EM by the Method of Weights

Whittemore and Grosser (1986) demonstrated that the conditional distribution of Y given X_{obs} could be expressed as

$$f_{Y|X_{obs}}(Y|X_{obs}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}) = \int f_{X|X_{obs}}(X|X_{obs}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}) f_{Y|X}(Y|X, \boldsymbol{\beta}) dX. \qquad (3.13)$$

They also suggested that if $f_{X|X_{obs}}$ was known by the researcher, that 3.13 could be maximized using an EM algorithm to obtain maximum likelihood estimates for $\boldsymbol{\beta}$. This was the approach taken by Brant and Tibshirani (1991) and Ibrahim (1990) for obtaining maximum likelihood estimates in generalized linear models with missing data in the covariates. This method, EM by the method of weights (EMMW) is appropriate where variables with missing observations are discrete.

The E-Step. Denote the observed outcome and covariate values as y_i and x_i respectively for cases $i = 1, 2, \ldots, n$. The complete data log-likelihood for our logistic regression model can be specified as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \Pr(Y_i|X_i, \boldsymbol{\beta}).$$

Where there is missing data, let $x_{obs.i}$ and $x_{mis.i}$ be the rows of X_{obs} and X_{mis} corresponding to case i. Assuming the missing data mechanism is ignorable, the t^{th} expectation step of the EM algorithm is

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) &= E\{\ell(\boldsymbol{\beta}|Y, X_{obs}, \boldsymbol{\beta}^{(t)})\} \\ &= \sum_{i=1}^{n} \int \log \Pr(y_i|x_i, \boldsymbol{\beta}^{(t)}) f_{X|X_{obs}}(x_i|y_i, x_{obs.i}, \boldsymbol{\theta}) dx_i. \end{aligned}$$

In the case where all variables with missing values are discrete, the integral can be replaced by a summation sign. The probability $\Pr(x_{ij}|y_i, x_{obs.i})$ will be used in place of the density function $f_{X|X_{obs}}(x|y_i, x_{obs.i}, \boldsymbol{\theta})$. For case i, let there be $j = 1, 2, \ldots, k_i$ possible values of $x_{mis.i}$ given $x_{obs.i}$ and X_{obs}. The expectation step then takes the form

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} \log \Pr(y_i|x_{ij}, \boldsymbol{\beta}^{(t)}) \Pr(x_{ij}|y_i, x_{obs.i}, \boldsymbol{\theta})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k_i} \log \Pr(y_i|x_{ij}, \beta^{(t)}) w_{ij}.$$

The M-Step. The M-step can then be obtained as a weighted log-likelihood using standard statistical software.

Obtaining the weights. Two steps are required when obtaining the w_{ij}. A model for the joint distribution of covariates must be specified and parameters (θ) for this distribution estimated using missing data methods. Several models are possible for this joint distribution and these are described in more detail in section 3.2.3. The approach taken to model this distribution warrants concern, since the legitimacy of substituting estimates for parameters in a log-likelihood requires the estimates be consistent (Gong and Samaniego, 1981).

Secondly, an augmented data matrix must be created. This matrix will be denoted as X_{aug} and contains the covariate patterns that are possible for cases with missing observations. The following example illustrates the creation of an augmented data matrix on the basis of observed data.

Example 1.

Consider the following simple example with variables x_1, x_2 and x_3. Each of these variables has possible categories 1 *and* 2. *Let there be $j = 1, 2, \ldots, k_i$ possible covariate patterns for subject i, where $i = 1, 2, \ldots, n$. In the incomplete data matrix in table 3.1, subjects 1 through 4 have completely observed data. Subject 5 is missing an observation for variable x_3 and subject 6 is missing an observation for variables x_2 and x_3.*

If we were to augment the data by filling out the data set with all possible covariate

Table 3.1: Incomplete data matrix

	X_{obs}		
i	x_1	x_2	x_3
1	1	2	2
2	1	1	2
3	1	2	1
4	2	1	2
5	1	1	—
6	2	—	—

patterns for the missing observations, we would obtain the augmented data matrix presented in table 3.2. Note that since there are two possible covariate values for each

Table 3.2: Augmented data matrix

		X_{aug}		
i	j	x_1	x_2	x_3
1	1	1	2	2
2	1	1	1	2
3	1	1	2	1
4	1	2	1	2
5	1	1	1	1
5	2	1	1	2
6	1	2	1	1
6	2	2	1	2
6	3	2	2	1
6	4	2	2	2

variable, subject 5 has two entries in the augmented matrix. There are four possible covariate patterns for the missing observations for subject 6. Consequently, subject 6 now requires four rows in the augmented data matrix.

Denote individual rows of X_{aug} as x_{ij}. Each row is assigned a probability,

$\Pr(x_{ij}|x_{obs.i},\theta)$. The conditional probability of x_{ij} given $x_{obs.i}$ can be obtained as

$$\Pr(x_{ij}|x_{obs.i},\theta) = \frac{\Pr(x_{ij}|\theta)}{\sum_{j=1}^{k_i}\Pr(x_{ij}|\theta)}.$$

These probabilities are then used to obtain weights at each step of the EM algorithm as

$$w_{ij} = \Pr(x_{ij}|x_{obs.i},y_i,\theta) = \frac{\Pr(y_i|x_{ij},\beta^{(t)})\Pr(x_{ij}|x_{obs.i},\theta)}{\sum_{j=1}^{k_i}\Pr(y_i|x_{ij},\beta^{(t)})\Pr(x_{ij}|x_{obs.i},\theta)}.$$

3.2.3 Covariate models.

Of the many possible models for the joint distribution of covariates, three types will be considered for use with risk adjustment. These are 1) multivariate normal (MVN) models, 2) log-linear models, and 3) mixed continuous and categorical (MCC) models. Other models, such as tree-based models, will not be considered.

1. **Multivariate normal models.** Multivariate normal models have been used to model covariates when performing logistic regression using MI. The use of this distribution has been advocated even when the joint distribution is clearly not MVN (Schafer, 1997d; Greenland and Finkle, 1995). When variables are binary or ordinal, these authors recommend rounding the imputed values into the appropriate categories. Greenland and Finkle (1995) and Schafer (1997d) used simulations to demonstrate that this method yields suitable results in the presence of categorical and skewed data. A major advantage of this method for large and complex data sets is that few parameters are required for the model. Further, it is easy to implement and many implementations are already available. Schafer (1999) provides programs to fit MVN models and to generate proper imputations from these models.

2. **Log-linear models.** With categorical covariates, one can fit a multinomial model to the covariates using an EM algorithm or Newton Raphson algorithm. When there are many covariates, such a fit is problematic, as there will be a large number of parameters, many of which may be poorly estimated. The use of log-linear models provides a means of reducing the number of parameters (Bishop et al., 1975). Brant and Tibshirani (1991) used a Newton Raphson algorithm to fit log-linear models to covariates when performing EMMW. Schafer (1999) provides programs for fitting log-linear models. This program employs an EM algorithm based on Bayesian iterative proportional fitting and allows the user to specify a Dirichlet prior for the fit. A Dirichlet prior has the same functional form as the likelihood for a multinomial distribution. Let

$$L(\theta|x) \propto \theta_1^{x_1}\theta_2^{x_2}\cdots\theta_K^{x_K}$$

be a likelihood for the contingency table $x = (x_1, \ldots, x_K)$ with multinomial parameters $\theta = (\theta_1, \ldots, \theta_K)$. The Dirichlet prior for this distribution is

$$\pi(\theta) \propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_K^{\alpha_K-1},$$

where $\alpha_1, \ldots, \alpha_K$ are user specified hyperparameters. Multiplying the likelihood by the prior produces the posterior

$$P(\theta|x) = \theta_1^{x_1+\alpha_1-1}\theta_2^{x_2+\alpha_2-1}\cdots\theta^{x_K+\alpha_K-1}.$$

When finding the mode of this posterior, setting the K hyperparameters to some $c > 1$ smooths the parameter estimates toward a uniform table and adds the equivalent of $c - 1$ observations to each cell. With sparse data, parameter

solutions sometimes can not be obtained because the maximization algorithms converge to boundaries of the parameter space. With an appropriate prior, the posterior parameter distribution can be flattened, moving the solution to the interior of the parameter space (Schafer, 1997e). Unfortunately, Schafer's program for generating imputations from these models is not yet complete.

3. **Mixed continuous and categorical models.** Little and Schluchter (1985) proposed a model for mixed normal and categorical variables. This model fits a joint distribution to the categorical variables using multinomial or log-linear models. A joint MVN distribution is fit to the continuous variables conditioned on the categorical variables. Schafer (1999) provides implementations of these models as well as programs for generating imputations from these models using data augmentation. His programs allow the user to specify log-linear models for the categorical variables as well as contrasts among the means of the continuous variables within the categories. These contrasts are then used when fitting the MVN component of the model. These constraints allow the user to reduce the number of parameters required for the models (Schafer, 1997f). These programs also allow the user to specify Dirichlet priors for the categorical component of the model.

3.3 Missing Data Methods for Risk Adjustment

Logistic regressions models based on both MI and EMMW can be used to obtain risk adjusted estimates when observations are missing in the covariates. Obtaining point estimates and variances estimates is straight-forward when using MI. Point

estimates are easily obtained using EMMW, but variance estimates requires some modification of the methods outlined in appendix A.1.

3.3.1 Multiple Imputation

To obtain risk adjusted estimates using MI, one needs to impute m data sets. Logistic regression is then performed using each imputed data set and the parameters from these logistic regressions are then used to obtain risk adjusted point estimates and variance estimates by applying the procedures outlined in chapter 2 and appendix A. For each provider, the m point and variance estimates are treated as the $\hat{Q}^{(t)}$ and $U^{(t)}$ in 3.5 - 3.8 and are combined to obtain risk adjusted estimates and variances. For example, if we wish to obtain a point estimate of RR_k on the basis of m values of $\widehat{RR}_k^{(t)}$, by 3.5

$$\widehat{RR}_k = \frac{1}{m}\sum_{t=1}^{m} \widehat{RR}_k^{(t)}.$$

By combining 3.6 - 3.8, a variance estimate for $\widehat{RR}_k$ could then be obtained as

$$\widehat{\mathrm{Var}}(\widehat{RR}_k) = \frac{1}{m}\sum_{t=1}^{m} \widehat{\mathrm{Var}}(\widehat{RR}_k^{(t)}) + \left[\frac{1+m^{-1}}{m-1}\right]\sum_{t=1}^{m}(\widehat{RR}_k^{(t)} - \widehat{RR}_k)^2.$$

These procedures will work for indirectly and directly standardized measures, whether the measures are baseline-model or full-model adjusted estimates. It should be noted that as these estimates are not maximum likelihood estimates, if the imputations are based on data augmentation procedures, these estimates may not be proper.

3.3.2 EM by Method of Weights

Both baseline and full-model adjusted estimates are based on weighted averages of the risks obtained for each subject. The calculation of baseline adjusted rates does

not utilize variables to code for the treatment providers. The linear combination for the covariates of interest will be denoted as

$$\hat{\eta}_{ij} = x_{ij}^T \hat{\boldsymbol{\beta}},$$

where the x_{ij} are the rows of X_{aug}. For the purposes of calculating the point estimates, X_{aug} will be comprised of the covariates and risk factors of interest and it excludes any variables which code for the different providers. For full-model adjusted estimates, the linear combination for the covariates of interest will be denoted as

$$\hat{\eta}_{cij} = x_{ij}^T \hat{\boldsymbol{\beta}}_c,$$

where the $\hat{\boldsymbol{\beta}}_c$ are logistic regression parameters for the covariates of interest which have been obtained from a model which includes variables which code for the different providers.

Baseline-model adjusted estimates. For baseline-model adjustment, the $\hat{\eta_{ij}}$ are used to obtain $\hat{p}_{ij}$'s which correspond to rows of X_{aug}. These are

$$\hat{p}_{ij} = \frac{\exp(\hat{\alpha} + \hat{\eta}_{ij})}{1 + \exp(\hat{\alpha} + \hat{\eta}_{ij})}.$$

The $\hat{p}_{ij}$'s are then combined to get a single estimate for each subject as

$$\bar{p}_i = \sum_{j=1}^{n_j} \hat{p}_{ij} w_{ij},$$

where the w_{ij} are the weights obtained from the final iteration of the baseline-model logistic regression performed using EMMW. These weighted estimates are then used in place of the $\hat{p}_i$'s to obtain the point estimates detailed in chapter 2. For example, the expected number of deaths for provider k can be obtained as

$$\hat{E}_k = \sum_{i \in h_k} \bar{p}_i.$$

Since the outcome variable is available for all subjects, the observed number of deaths and the proportion dying in the population can be obtained in the normal manner. The E_k can then be used to calculate SMR_k and P_k using 2.15 and 2.23.

Variances for baseline-adjusted estimates are typically based on the variances of the observed number of deaths. These in turn are estimated using the estimated probability of death for each subject to estimate the variance of the outcome for that subject, or $\text{Var}(Y|X)$. When X has missing values, the variance of Y is conditional on X_{obs}, which can be obtained as

$$\text{Var}(Y|X_{obs}) = \text{Var}(E(Y|X)|X_{obs}) + E(\text{Var}(Y|X)|X_{obs}).$$

For EM by method of weights, estimates of $\text{Var}(y_i) = \text{Var}(Y|X_{obs})$ for individuals $i = 1, 2, \ldots, n$ can be obtained as

$$\begin{aligned}
\widehat{\text{Var}}(y_i) &= \left[\sum_{j=1}^{n_i} \hat{p}_{ij}^2 w_{ij} - \left(\sum_{j=1}^{n_i} \hat{p}_{ij} w_{ij}\right)^2\right] + \left[\sum_{j=1}^{n_i} \hat{p}_{ij}(1-\hat{p}_{ij}) w_{ij}\right] \\
&= \sum_{j=1}^{n_i} \hat{p}_{ij} w_{ij} - \left(\sum_{j=1}^{n_i} \hat{p}_{ij} w_{ij}\right)^2 \\
&= \bar{p}_i(1-\bar{p}_i).
\end{aligned}$$

Following the typical approach for estimating variances for the BM measures, the $\bar{p}_i$ are treated as fixed and the variance of SMR_k is obtained as

$$\begin{aligned}
\text{Var}(SMR_k) &= \text{Var}\left(\frac{O_k}{E_k}\right) \\
&= \frac{1}{E_k^2} \sum_{i \in h_k} \bar{p}_i(1-\bar{p}_i).
\end{aligned}$$

The variance of the population proportion is obtained as

$$\text{Var}(\hat{P}_k) = \frac{1}{n^2} \sum_{i \in h_k} \bar{p}_i(1-\bar{p}_i).$$

Full-model adjusted point estimates. Similar procedures are employed for indirectly standardized FM risk adjusted point estimates. Estimates of $\bar{p}_{oi}$ are obtained using the estimates of death

$$\hat{p}_{oij} = \frac{\exp(\hat{\alpha}_o + \hat{\eta}_{cij})}{1 + \exp(\hat{\alpha}_o + \hat{\eta}_{cij})}$$

from the offset model. The $\hat{\alpha}_o$ for the offset model is obtained using EMMW to ensure that $\sum_{i=1}^{n} \sum_{j=1}^{k_i} \hat{p}_{oij} = \sum_{i=1}^{n} y_i$. The estimate of death for individual i is

$$\bar{p}_{oi} = \sum_{i=1}^{n_j} \hat{p}_{oij} w_{ij}.$$

Risk adjusted point estimates are then calculated using

$$\hat{E}_k = \sum_{i \in h_k} \bar{p}_{oi}.$$

in 2.15 and 2.23.

For directly standardized measures, the calculation of risk adjusted estimates is based on the weighted probabilities of death obtained by treating all subjects as if they had been cared for by the provider in question. For individual i treated by provider k, the weighted estimate of the probability of death is

$$\bar{p}^*_{ik} = \sum_{j=1}^{n_i} \hat{p}^*_{ijk} w_{ij}.$$

The $\hat{p}^*_{ijk}$ are obtained as

$$\hat{p}^*_{ijk} = \frac{\exp(\hat{\alpha} + \beta_k + \hat{\eta}_{cij})}{1 + \exp(\hat{\alpha} + \beta_k + \hat{\eta}_{cij})},$$

where β_k is the regression parameter associated with treatment by provider k. The DS risk adjusted estimates for the RR and population averaged proportion are then

calculated using

$$E_k^* = \sum_{i=1}^{n} \bar{p}_{ik}^*$$

in 2.35 and 2.39.

Variances of the full-model estimates. For both indirectly and directly standardized full model measures, estimation of the variances requires modification of the methods outlined in chapter 2 and appendix A. In general, these methods must be adapted to account for the use of $\Pr((Y|X)|X_{obs})$ rather than $\Pr(Y|X)$ in the logistic regression. These variance estimates must account for the weights obtained in the final iteration of the EM algorithm and for the variance-covariance matrix of the the coefficients obtained using Louis's Method. The necessary changes to the variance estimates are described in detail in appendix C.

Chapter 4

Risk-adjustment using the APPROACH Data

Risk adjustment procedures are used to adjust the outcomes associated with treatment providers for the mix of patients treated by the provider. In the present investigation, risk adjustment procedures will be applied to the types of treatment provided to the cases. These treatments included medical treatment, bypass surgery (CABG), and percutaneous transluminal coronary angioplasty (PTCA). Type of treatment was chosen for risk adjustment because there was evidence with the complete cases that the risks varied across treatment groups. The clinical justification for such a choice is not strong. Although the treatments appear to vary in effectiveness, the choice of treatment involves multiple factors such as the appropriateness of the treatment given the condition of the patient, the quality of life (such as absence of chest pain) that is expected after a given form of treatment, and the willingness of the patient to allow invasive procedures. Consequently, the effectiveness of the therapies will not be discussed and attention will be focussed on the performance of the missing data methods and risk adjustment proceedures.

To account for data missing in the APPROACH project data base, Norris et al. (1999) produced a second set of diagnoses based on administrative ICD-9 discharge data. These administrative variables were then used to augment the available clinical data. Three different data sets were created from this data. In one data set, only subjects with complete information were included. In the second data set, missing variables were coded as having the reference level of risk. In the third data set,

termed the *enhanced data*, diagnoses were considered positive if they were positive in either or both of the administrative and APPROACH diagnostic variables. Some important clinical indicators did not have equivalent administrative variables. Most notable among these was ejection fraction (EF). In 1995, 27% of the APPROACH cases were missing the EF variable. To use EF in their analyses, Norris et al. (1999) used a separate category to code for missing EF observations. While allowing for the use of EF information, such a procedure is known to produce bias in regression coefficients, even when data are MCAR (Vach, 1994; Vach and Blettner, 1991). This chapter will explore the use of missing data methods as an alternative to the data enhancement technique employed by Norris et al. (1999). This examination will use APPROACH data from 1995. Only the observed data will be employed; the administrative variables will not be utilized.

4.0.3 Variables

Norris et al. (1999) compared logistic regression models based on the three data sets. They concluded that the *enhanced* data provided a better fit for one-year mortalities than the other two methods of handling the missing data. The resulting model had 30 coefficients (see table 4.1).

Covariates The number of variables used by Norris et al. (1999) in their analyses proved problematic for modelling the joint distribution of covariates for EMMW, as well as for creating an augmented data matrix. The use of categorical or mixed continuous and categorical models for the covariates required a reduction in the number of variables used in the logistic regression model. Ignoring age, the treatment

Table 4.1: Variables used by Norris et al. (1999) in their enhanced model.

Variables	Coefficient	Odds Ratio	95% Confidence Interval
INTERCEPT	-7.2		
AGE for each 10 yr.	0.32	1.4	(1.2 - 3.3)
Cerebrovascular Disease	0.75	2.1	(1.4 - 3.3)
Congestive Heart Failure	0.97	2.7	(1.9 - 3.6)
Pulmonary Disease	0.32	1.4	(0.9 - 2.0)
Renal Disease	1.72	5.6	(3.4 - 9.1)
Diabetes Mellitus	0.18	1.2	(0.8 - 1.6)
Dialysis	0.23	1.3	(0.5 - 3.3)
Hyperlipidemia	-0.27	0.8	(0.6 - 1.0)
Hypertension	0.07	1.1	(0.8 - 1.4)
Liver/GI Disease	0.00	1.0	(0.5 - 2.0)
Malignancy	-0.26	0.8	(0.4 - 1.6)
Prior CABG	0.19	1.2	(0.8 - 1.8)
Prior Myocardial Infarct	0.11	1.1	(0.8 - 1.6)
Ejection Fraction			
$< 30\% :\ > 50\%$	0.96	2.6	(1.6 - 4.4)
$30 - 50\% :\ > 50\%$	0.45	1.6	(1.0 - 2.4)
V-gram not done : $> 50\%$	1.29	3.6	(1.9 - 6.9)
missing: $> 50\%$	0.75	2.1	(1.5 - 3.1)
Coronary Anatomy			
1& 2 vessel disease: normal	0.29	1.3	(0.6 - 2.9)
2 vessel disease PLAD: normal	0.85	2.3	(0.8 - 6.5)
3 vessel disease: normal	1.17	3.2	(1.5 - 7.0)
3 vessel disease PLAD: normal	1.22	3.4	(1.5 - 7.6)
Left Main: normal	1.59	4.9	(2.2 - 11.2)
Missing: normal	0.76	2.1	(0.8 - 5.6)
Prior PTCA	-0.09	0.9	(0.6 - 1.4)
Peripheral Vascular Disease	0.14	1.5	(0.7 - 1.8)
Prior Lytic Therapy	0.35	1.4	(0.9 - 2.2)
Sex (female: male)	0.27	1.3	(1.0 - 1.8)
Clinical Indication			
Myocardial infarct: Stable Angina	- 0.03	1.0	(0.7 - 1.4)
Other: Stable angina	0.34	1.4	(0.9 - 2.2)

variable, and the outcome, the number of categories defined by the variables used by Norris et al. (1999) is $2^16 \times 4 \times 6 \times 3 = 4,718,592$. Even with restrictions on the multinomial model, acceptable fits for the categorical and mixed continuous and categorical missing data models were difficult to obtain. For the use of EMMW, the size of the required augmented data matrix greatly exceeded the capacity of the computer used for the analyses.

Although the variables included by Norris et al. (1999) were of clinical interest, the effects attributable to several of these variables were negligible. Models employing cases with complete data as well as models based on the enhanced data were examined to determine the relative importance of the variables. Decisions were based on the clinical and statistical relevance of the variables and were made in consultation with Dr. William Ghali, a physician involved with the APPROACH project. The covariates used for the missing data models are described in table 4.2.

Depending on the model used for covariates, EF and coronary anatomy (CA) were either left as categorical variables or were broken into sets of dummy variables. As there were 4 possible categories for ejection fraction (missing values were not placed in a separate category), 3 dummy variables were required to represent membership in these categories. Patients who did not receive venograms were coded as a separate category rather than as missing. This is because the cardiologist may choose not to perform a venogram when the condition of patients is poor. If no information was recorded regarding the venogram, the ejection fraction was recorded as missing. The variables were coded so that the logistic regression coefficient for each variable represented a comparison between the risk for a given EF category and the risk associated with an EF > 50. For coronary anatomy, 5 dummy variables

Table 4.2: Descriptions of the covariates used in the missing data analyses.

Variable	Description	Count	% of Total
Age	By Decade		.
	< 40	162	2.67
	40 - 50	786	12.96
	50 - 60	1362	22.46
	60 - 70	2028	33.44
	70 - 80	1463	24.12
	>80	264	4.35
CVD	Cerebrovascular Disease		.
	No	5492	90.55
	Yes	219	3.61
	Missing	354	5.84
CHF	Congestive Heart Failure		.
	No	4057	66.89
	Yes	519	8.56
	Missing	1489	24.55
PD	Pulmonary Disease		.
	No	5139	84.73
	Yes	273	4.50
	Missing	653	10.77
Creat	Creatinine		.
	No	5073	83.64
	Yes	84	1.38
	Missing	908	14.97
EF	Ejection Fraction		.
	<30%	256	4.22
	30-50%	1126	18.57
	>50%	2901	47.83
	Venogram Not Done	128	2.11
	Missing	1654	27.27
CA	Coronary Anatomy		.
	Normal	618	10.19
	1& 2 Vessel Disease	2937	48.43
	2 Vessel Disease PLAD	191	3.15
	3 Vessel Disease	1091	17.99
	3 Vessel Disease PLAD	637	10.50
	Left Main	389	6.41
	Missing	202	3.33
Sex	Male	4339	71.54
	Female	1726	28.46

were employed to represent the 6 categories. These were coded so that the logistic regression coefficients represented comparisons of the different CA risk categories with the normal CA category.

For the logistic regressions there were 14 covariates and two additional variables to code for treatment effects. The covariates in table 4.2 were included in all logistic regression analyses. For descriptive purposes, age has been broken down by decade. However, in the analyses age was treated as a continuous variable. For the logistic regressions requiring treatment effects, two binary variables were included to compare the three treatment categories. These categories were 1) Medical treatment, 2) Coronary Artery Bypass Graft (CABG), and 3) PTCA. Sex was included in all models even though there was no evidence that influenced the outcomes at the 5% level of significance. It was retained in the models because 1) it was of clinical interest, 2) it was significant in the original enhanced model and 3) the possibility existed that it might be statistically significant in one or more of the missing data models.

Outcome measure. The outcome measure for the analyses is one year mortality. Subjects dying within one year of their cardiology examination received a code of 1, while subjects surviving received a code of 0. A total of 301 of the 6065 cases (approximately 5%) died within one year of their initial angiograms.

Patterns of missing data. From table 4.2, it is evident that CHF and ejection fraction have the highest percentages of missing observations (24.6% and 27.3% respectively). This warrants concern, as these variables had large effects in the enhanced model. The APPROACH investigators were aware that the rates of missing

Table 4.3: Missing data broken down by variables of interest.

			Ejection Fraction		CHF	
Variable	Category	N in Category	N Missing	% Missing	N Missing	% Missing
Quarter	First	1516	806	53.2	503	33.2
	Second	1601	456	28.5	396	24.7
	Third	1494	212	14.2	350	23.4
	Fourth	1454	180	12.4	240	16.5
Hospital	1	1431	374	26.1	415	29.0
	2	1624	181	11.1	115	7.1
	3	1380	321	23.3	627	45.4
	4	1630	778	47.7	332	20.4
Treatment	Medical	2880	761	26.4	676	23.5
	CABG	1337	376	28.1	338	25.3
	PTCA	1848	517	28.0	475	25.7
Outcome	Alive	5764	1537	26.7	1401	24.3
	Dead	301	117	38.9	88	29.2

data varied by time and by hospital. This variation can be attributed to the time required to educate physicians about the APPROACH project, and because the project was embraced more readily at some hospitals than others. In table 4.3, the missing observations in CHF and ejection fraction are broken down by the quarter of 1995 as well as by hospital. Also included are breakdowns by type of treatment received and by outcome.

The percentages of missing observations are greatest in the first quarter and decline as the year progresses. In the first quarter, 52.2% of the cases were missing EF data and 33.2% were missing CHF data. By the end of the year, the percentages of missing data declined to 12.4% and 16.5% for EF and CHF respectively. In 1995, angiograms were performed at four Alberta hospitals. There is a large degree

of variation in the percentage of missing data by hospital. Percentages of missing ejection fraction data ranged from 11.1% to 47.7% and percentages of missing CHF data ranged from 7.1% to 45.4%.

Although hospital and quarter were related to the proportion of missing data, they were not included as variables in the analyses. In preliminary investigations with the complete cases, the inclusion of these variables did not improve the fitted models. Further, these variables did not appear to be related to 1) the outcome, 2) the observed values of the variables with missing data, or 3) to the other risk factors used in the logistic regression models. Consequently, they have been excluded when performing missing data methods and risk adjustment procedures.

4.1 Logistic Regressions with Missing Data

The application of risk adjustment methods to the data first required the development of logistic regression models. Two sets of models were employed. The first set of models did not contain treatment effects. These models were used for the calculation of baseline-model (BM) risk adjustment measures in which the effects of the covariates were not adjusted for the effects of treatment. The second set of models included treatment effects and were used to calculate full-model (FM) risk adjustment measures as well as directly standardized measures.

Adequacy of the fits

No formal methods of assessing the adequacy of logistic regression fits based on missing data methods have been addressed in the literature. For this reason, the fits will be examined using *ad hoc* methods which parallel some of the methods employed

when performing logistic regression with complete data. The methods employed will include C statistics, residual deviances (D), and by the use of Pearson residuals ($\mathbf{r_p}$). For multiple imputation models, tests of adequacy were based on the average of the fitted probabilities obtained from each of the $s = 1, 2, \ldots, m$ imputed models. For individuals $i = 1, 2, \ldots, n$, an average fitted probability was calculated as

$$\bar{p}_i = \sum_{s=1}^{m} \hat{p}_{is}. \tag{4.1}$$

For EMMW models, tests of adequacy were based on the weighted average of the fitted probabilities obtained for each subject. For each subject, the fitted probability was calculated as

$$\bar{p}_i = \sum_{j=1}^{n_i} w_{ij}\hat{p}_{ij}. \tag{4.2}$$

***C* statistics.** These fitted values were then used to obtain C statistics for the logistic regression models. For binary outcomes, the C statistic is the area under the receiver operating characteristic (ROC) curve. It is based on all possible pairs of patients, where one patient has the disease and the other does not. The C statistic is the proportion of these pairings in which the patient with the disease has a higher predicted probability of death than does the patient who does not die (Harrell and Lee, 1984)

Residual deviances. The residual deviance, or D, is twice the discrepancy between the maximum log-likelihood achievable and the log-likelihood achieved by the model under investigation (McCullagh and Nelder, 1989). For the MI and EMMW models, the achieved log-likelihood is obtained using the average fitted probabilities and weighted average fitted probabilities respectively. The residual deviances were

obtained as

$$D = -2\sum_{i=1}^{n}\left[y\log(\bar{p}_i) + (1-y)\log(1-\bar{p}_i)\right]. \tag{4.3}$$

These will be calculated for comparison purposes only and will not be used for statistical tests. Schafer (1997b) notes that for MI models, likelihood ratio test statistics must be devised that are based only on the observed data. For likelihood ratio tests, Schafer (1997c) describes a method provided by Meng and Rubin (1992).

Pearson Residuals. To further explore the adequacy of the model, the Pearson residuals from the marginal distribution of the joint EF and CHF categories were examined. These variables were chosen as they had the largest percentages of missing observations as well as the largest $\hat{\lambda}$'s in the MI model. This joint distribution also included categories for subjects with missing data on one or both of the variables.

In constructing the tables, CHF was considered to have three categories: no, yes and missing. Ejection fraction was considered to have five categories: $< 30\%$, $30-50\%$, $> 50\%$, not done, and missing. Tables were based on the cross-classification of these variables. For each category, the observed number of deaths was compared to the expected deaths from the model. For each category c_l where $l = 1, 2, \ldots, n_c$, the expected number of deaths (E_l) was obtained by summing the $\bar{p}_i$'s of the cases falling within the joint category. The observed number of deaths in the joint category c_l was denoted as O_l. Pearson residuals were obtained as the raw residual $(O_l - E_l)$ scaled by the estimated standard deviation of O_l (McCullagh and Nelder, 1989), or

$$\mathbf{r_p} = \frac{O_l - E_l}{\sqrt{\widehat{\mathrm{Var}}(O_l)}}.$$

Assuming that the observations were independent, the variance of O_l was esti-

mated as

$$\widehat{\text{Var}}(O_l) = \widehat{\text{Var}}(\sum_{i \in c_l} y_i) = \sum_{i \in c_l} \widehat{\text{Var}}(y_i).$$

When MI is used to obtain logistic regression models, the estimate (see section 3.1.3) of $\text{Var}(y_i)$ from the m models is

$$\widehat{\text{Var}}(y_i) = \frac{1}{m}\left(\sum_{s=1}^{m} \hat{p}_{is} - \sum_{s=1}^{m} \hat{p}_{is}^2\right).$$

For EMMW, the estimate of $\text{Var}(y_i)$ is

$$\begin{aligned} \widehat{\text{Var}}(y_i) &= \sum_{j=1}^{n_i} \hat{p}_{ij} w_{ij} - \left(\sum_{j=1}^{n_i} \hat{p}_{ij} w_{ij}\right)^2 \\ &= \bar{p}_i(1 - \bar{p}_i). \end{aligned}$$

For details, see C.7 in appendix C. For diagnostic tables with Pearson residuals, see tables 4.6 and 4.9.

4.1.1 Joint Distributions

As noted in chapter 3 (section 3.2.3), the application of missing data methods requires the specification of models for the joint distribution of the covariates. For analyses employing MI, multivariate normal distributions were employed. For analyses using EMMW, two types of models were employed. These were 1) mixed continuous and categorical models, and 2) multivariate normal models.

Multiple Imputation

As suggested by Schafer (1997d) and Greenland and Finkle (1995), the joint distribution which formed the basis for the imputations included not only the covariates but also the outcome measure. A multivariate normal distribution was applied to

these variables when employing MI methods. Although attempts were made to find other models to form a basis for imputations, these attempts failed due to computational difficulties posed by the number of covariates in the models. Attempted models included a log-linear model in which age was broken into categories, and a mixed continuous and categorical variables model, in which age was treated as a continuous variable. Complications generally arose because of the large number of categories defined by the variables. For example, by breaking age into 6 decades, a total of $6 \times 2^4 \times 4 \times 6 \times 2 = 4,608$ categories were defined by the covariates. With two outcome categories and three treatments included in the model, the number of categories expanded to 27,648. By placing constraints on the model, a fit could be obtained for the joint distribution of the categories. However, the programs written by Schafer (1999) were incapable of drawing imputations from a categorical model. Similar problems were encountered when attempts were made to fit mixed continuous and categorical models to the data. Treating age as a continuous variable and the remaining variables as categorical, with appropriate constraints, fits could be obtained for the joint distribution of the variables. Unfortunately the program *imp.mix* provided by Schafer was not capable of generating imputations from the mixed continuous and categorical distributions from the APPROACH data. When running the *imp.mix* program in S-PLUS (MathSoft, 1999), the program terminated abruptly with no diagnostic errors. After debugging in S-PLUS, the error appeared to occur after the S-PLUS object made a call to a FORTRAN (Free Software Foundation, 1998) routine. The problem appeared to be due to memory limitations on the computer used for the analyses: when tested with simpler models with fewer covariates and observations, the program was capable of generating imputations.

Method of Weights

Two types of covariate models were employed when fitting logistic regressions using EMMW. These were 1) mixed continuous and categorical models, and 2) MVN models. While Schafer's imputation programs could not generate imputations from the mixed continuous and categorical models, with appropriate constraints, the programs were capable of fitting mixed continuous and categorical models to the covariates. In these models, age was treated as a continuous variable, and the remaining covariates were treated as categorical. For both the mixed continuous and categorical models and the MVN models, the probabilities associated with the covariate patterns in $\tilde{\mathbf{X}}$ were then used to obtain weights at each iteration of the EMMW algorithm (see section 3.2.2).

4.1.2 Baseline-Adjusted Models

Logistic regression models used to obtain baseline-adjusted estimates did not include treatment effects. For these logistic regressions, 14 covariates were employed (see table 4.4). Age was the only continuous variable. Ages were divided by 10, and the resulting logistic regression coefficients reflect the change in outcome per 10 year increase in age. The comparisons defined by the EF and CA codes are described in section 4.0.3.

Multiple Imputation

Multiple imputations were created through the use of data augmentation procedures (see section 3.1.3). An EM algorithm was used to fit a MVN distribution to the variables. The parameter estimates for this distribution served as the starting point

for the data augmentation sequences. To ensure the legitimate use data augmentation procedures, time series plots of the sequences of parameters were examined to establish that the generated parameters were stationary about a single value (Schafer, 1997b). Correlations among parameters generated from the data augmentation sequences were examined using autocorrelation functions (ACFs) to determine 1) whether imputations selected from the series could be considered independent, and 2) the number of iterations of the DA procedure needed between draws of imputed data sets for these draws to be considered independent. Due to the large number of parameters associated with the MVN model, attention was focused on plots of the means and standard deviations. To further ensure the appropriateness of the methods, the worst linear function (WLF) of the parameters was also examined (see section 3.1.3).

Examples of the plots are provided in figures 4.1 and 4.2. Figure 4.1 contains plots of the standard deviation of the first ejection fraction variable. Of all time series plots, this series had the longest runs which deviated from a stationary point. The autocorrelations for this plot also tended to have larger values than those for other parameters. As expected, the WLF time series showed even greater deviations from a stationary point and the autocorrelations for the WLF were considerable even past 80 lags. Note that WLF was scaled so that if the time series was stationary about zero, it would be stationary about the mode of the likelihood function (Schafer, 1997b). Despite the problems apparent in the WLF, the time series plot still appeared to be stationary about zero. Similarly, although the autocorrelations appear large, after 60 lags, the majority of these correlations are less than $\rho = 0.1$. In an attempt to be conservative, imputations were obtained by sampling every 100th iterate from the

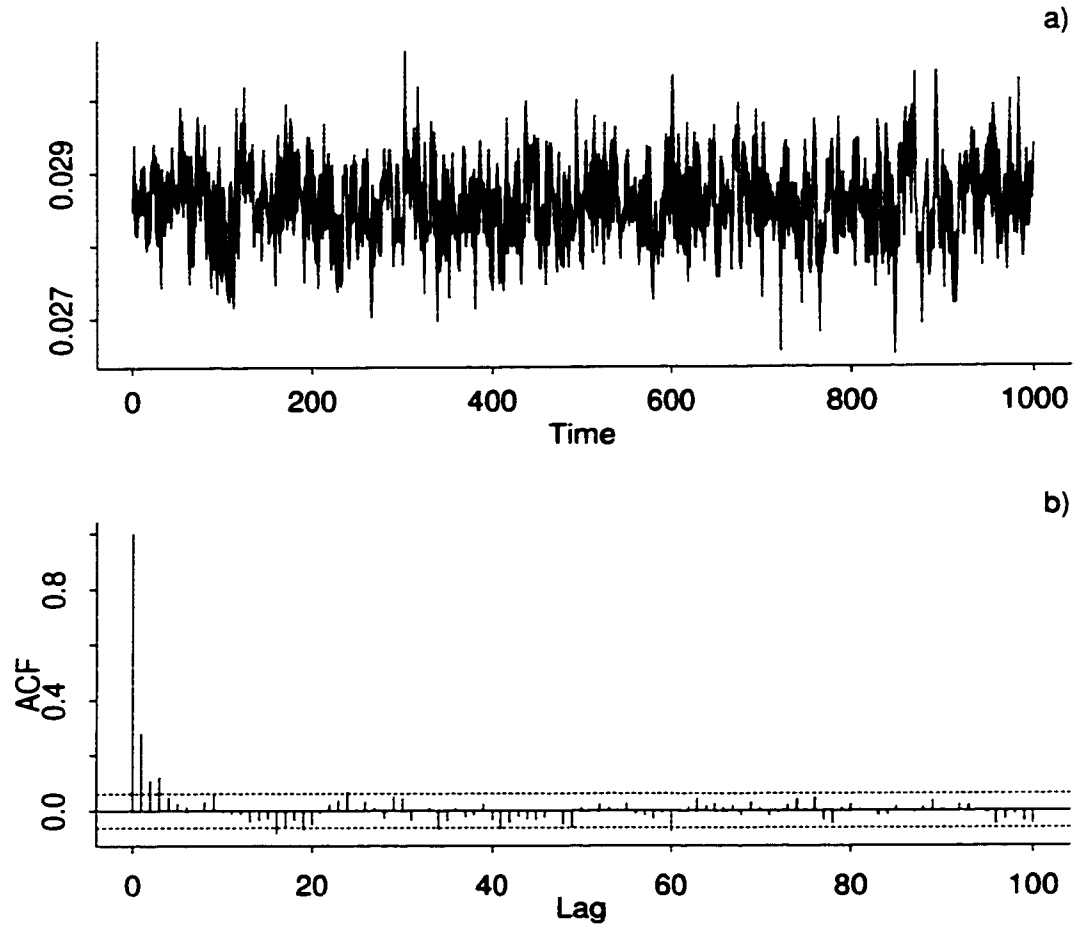

Figure 4.1: Diagnostic plots based on iterates from MVN model without treatment effects. a) Time series plot for standard deviation of the first ejection fraction variable. b) ACF for this series from iterations 100 to 1100. Dashes indicate approximate 0.05-level critical values for testing $\rho_k = \rho_{k+1} = \rho_{k+2} = \cdots = 0$.

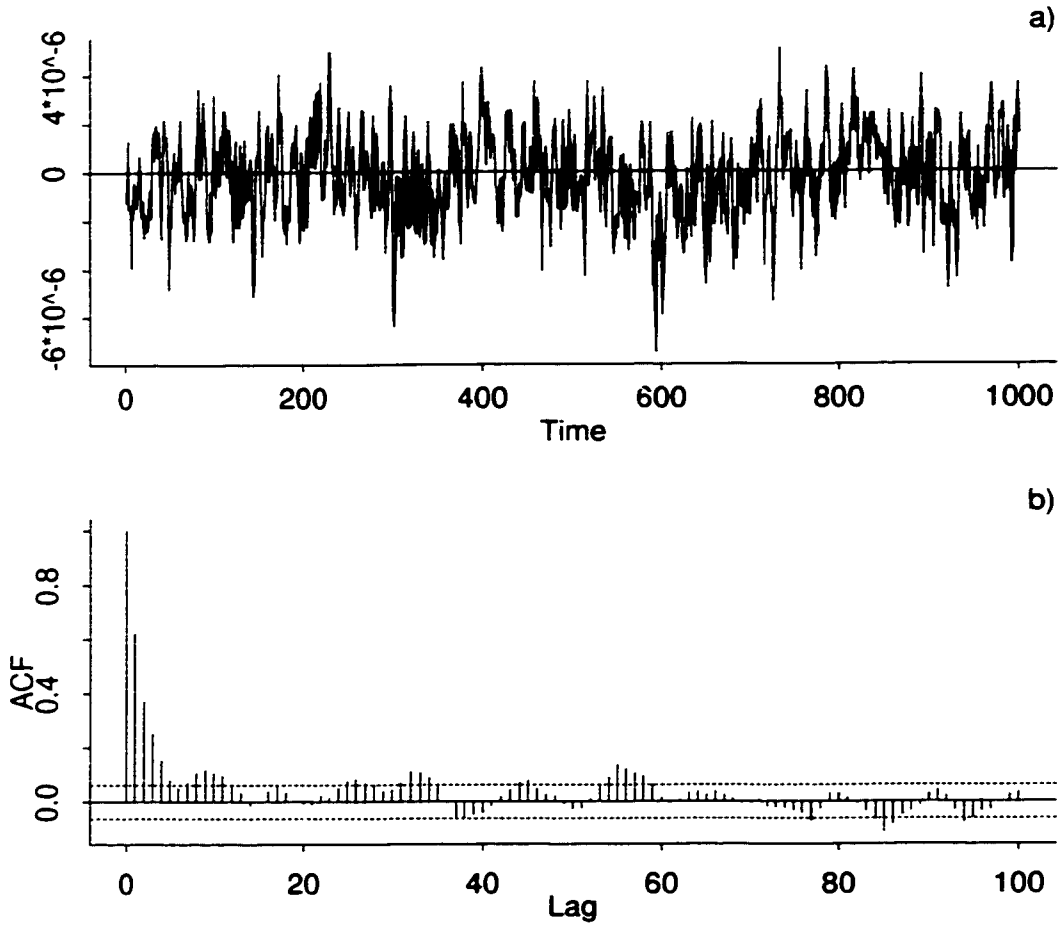

Figure 4.2: Diagnostic plots based on iterates from the MVN model without treatment effects. a) Time series plot for the worst linear function of the parameters. b) ACF for the worst linear function using iterations 100 to 1100.

series.

Ten imputations were sampled, and regression results as well as fitted values were stored for each imputation. Regression parameters and standard errors for the 10 imputations were combined using (3.5) - (3.8). The resulting regression parameters and standard errors are presented in table 4.4. Also included are estimates of the fraction of missing information for each regression parameter ($\hat{\lambda}$). These $\hat{\lambda}s$ relate the between imputation variance to the within imputation variance. They can be used to estimate the size of the obtained standard errors relative to the size of the standard errors which would have been obtained if $m = \infty$ imputations had been employed. According to Rubin (1987c), the relative efficiency of a point estimate relative obtained from m imputations relative to a point estimate based on infinite imputations is approximately $(1 + \lambda/m)^{-1}$. As CHF and ejection fraction had the largest percentages of missing observations, one would expect the $\hat{\lambda}s$ associated with these variables to be large. For CHF, the variable with the largest $\hat{\lambda}$ ($\hat{\lambda} = 0.4$), the use of $m = 10$ imputations yielded a standard error approximately $\sqrt{1 + 0.4/10} = 1.02$ times as large as the standard error which would have been obtained with $m = \infty$ imputations. For the C-statistic, residual deviance and Pearson residuals, see tables 4.5 and 4.6.

EMMW with Mixed Continuous and Categorical Model

For the first method of weights model (MW-1), the joint distribution of covariates was fit using a mixed continuous and categorical model. To reduce the number of parameters required to fit this model, a log-linear model was used for the categorical component of the model. This log-linear model fits main effects for the categorical

variables as well as one-way interactions among the variables. This component of the model required 83 parameters. Contrasts among categories were used to fit ages to the cells defined by the categories. The model for the means used a least squares approach to fit main effects for each categorical variable; no interactions terms were included. The continuous component required 14 parameters: 13 for the contrasts and 1 for the variance of age within the cells. Using Schafer's "ecm.mix" program, a maximum likelihood fit was obtained for this model. Despite the parameter restrictions, the algorithm appeared to converge to a boundary of parameter space, as many of the fitted probabilities were zero, and because many of the fitted means could not be estimated. To move the solution into the interior of the parameter space, a uniform Dirichlet prior was applied to the data (see section 3.2.3). The hyperparameters for this prior were all initially set at 1.05, but this also resulted in inestimable parameters. The magnitude of the hyperparameters was increased by increments of .05 until all parameters could be estimated. The final hyperparameters were 1.15, and the EM algorithm required 21 iterations to converge.

The use of a flat prior smooths the parameter estimates towards a table in which all probabilities are equal. The use of a hyperparameters of 1.15 is equivalent to adding 0.15 of an observation to each cell (Schafer, 1997g, p. 253), and is analogous to adding $.15 \times 768 = 115.2$ observations evenly throughout the cells. Schafer suggests that when specifying priors, the results will not be grossly distorted as long as the the number of observations added using prior information does not exceed 10-20% of the number of total observations. In the present case, the prior information amounts to $100 \times 115.2/6065 = 1.9\%$ of the total number of observations, which is well below the suggested guideline.

The joint probability of the covariates in the augmented data matrix was estimated by obtaining the probability associated with the cell membership defined by the categorical variables, and then multiplying this probability by the normal probability density associated with patient's age. These joint probabilities were then used in the EMMW algorithm as described in chapter 3 (see section 3.2.2).

Following convergence of the EM algorithm, the final weights and the augmented data matrix were used to obtain estimates of variance for the coefficients according to the method described in appendix B. The parameter estimates and their standard errors are shown in table 4.4. For the C-statistic, residual deviance and Pearson residuals, see tables 4.5 and 4.6.

EMMW with MVN Distribution

A MVN distribution for the covariates was used to obtain a second method of weights model (MW-2). Dummy variables were used to represent the variables with multiple categories. Unlike the MVN applied for the multiple imputations, the outcome measure was not included in the model. The probabilities associated with the covariates in the augmented data matrix were estimated by calculating the normal probability density defined by the fitted MVN distribution. These probabilities were then used in the EMMW algorithm (see section 3.2.2). Variances for the coefficients were obtained using Louis's method (see appendix B). The resulting logistic regression coefficients and standard errors are presented in table 4.4. For the C-statistic, residual deviance and Pearson residuals, see tables 4.5 and 4.6.

Table 4.4: Logistic regression models with no treatment effects. For the complete case (CC) analysis, n=3171.

	CC		MW-1		MW-2		MI		
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE	$100\hat{\lambda}$
Intercept	-6.93	0.773	-7.11	0.531	-6.90	0.522	-6.71	0.501	6
age	0.367	0.0982	0.380	0.0638	0.373	0.0628	0.367	0.0631	1
CVD	0.813	0.317	0.698	0.237	0.667	0.229	0.666	0.233	3
CHF	0.500	0.252	0.988	0.183	1.06	0.164	1.18	0.186	40
PD	0.414	0.356	0.213	0.239	0.323	0.225	0.285	0.229	7
creat	1.014	0.489	1.54	0.303	1.598	0.286	1.55	0.293	4
EF.1	1.38	0.294	1.28	0.246	0.895	0.232	0.992	0.233	34
EF.2	0.585	0.227	0.454	0.193	0.450	0.188	0.267	0.154	13
EF.3	1.57	0.382	1.30	0.314	1.10	0.293	1.06	0.291	9
CA.1	0.467	0.483	0.521	0.364	0.531	0.358	0.305	0.313	16
CA.2	-0.661	1.10	1.04	0.480	1.09	0.471	0.881	0.434	5
CA.3	1.10	0.496	1.30	0.369	1.33	0.364	1.12	0.314	14
CA.4	1.11	0.508	1.18	0.382	1.23	0.376	1.03	0.330	15
CA.5	1.31	0.526	1.60	0.391	1.63	0.385	1.43	0.339	12
sex	0.113	0.207	0.210	0.139	0.209	0.135	0.204	0.137	2

Table 4.5: C statistics and residual deviances for the logistic regression models.

Model	C	D
Enhanced	0.802	1982.45
MW-1	0.831	1977.94
MW-2	0.785	2063.73
MI	0.815	1982.55

Table 4.6: Standardized residuals for the marginal observed and expected values for the joint distribution of ejection fraction and CHF in models without treatment effects

Model →				MW-1		MW-2		MI	
EF	CHF	Count	O	E	z	E	z	E	z
<30%	No	112	11	13.26	-0.68	11.36	-0.12	11.02	-0.01
	Yes	99	22	23.67	-0.41	21.24	0.19	23.66	-0.41
	Missing	45	5	6.74	-0.75	5.69	-0.32	6.48	-0.66
30-50%	No	752	34	32.59	0.26	38.90	-0.82	30.42	0.67
	Yes	133	12	19.83	-2.00	23.56	-2.76	22.10	-2.47
	Missing	241	12	12.80	-0.23	12.26	-0.08	11.79	0.06
>50%	No	2264	43	53.16	-1.42	63.88	-2.68	60.76	-2.33
	Yes	86	8	7.60	0.16	9.41	-0.50	10.39	-0.81
	Missing	551	18	13.49	1.25	14.90	0.82	16.23	0.45
Not	No	64	7	7.16	-0.07	6.95	0.02	6.22	0.34
Done	Yes	28	6	8.25	-1.00	8.09	-0.93	8.34	-1.03
	Missing	36	6	6.06	-0.03	3.95	1.13	4.62	0.74
Missing	No	865	27	29.80	-0.53	29.77	-0.52	28.32	-0.26
	Yes	173	43	35.01	1.63	28.77	3.12	32.50	2.23
	Missing	616	47	31.58	2.89	22.26	5.41	28.17	3.78

4.1.3 Full-model Logistic Regressions

The calculation of risk-adjusted estimates in which the effects of covariates were adjusted for treatment effects required the use of logistic regression which included treatment effects. The procedures used in fitting these models were analogous to those used to fit the logistic regressions without treatment effects. For these models, however, the joint probability models for the covariates included the treatment categories.

Multiple Imputation

The starting point for the imputations was based on parameters fit to a MVN distribution using the "em.norm" program (Schafer, 1999). Time series plots were used to examine the rate of convergence of the Markov Monte Carlo process, and to ensure that the series for the individual parameters were stationary about a single point. Autocorrelation functions were used to determine the number of iterations required before the imputations could be considered independent. Plots were also made of the WLF of the parameters.

Examples of the plots are provided in figures 4.3 and 4.4. Figure 4.1 is a plot of the standard deviation of the first ejection fraction variable, as this parameter appeared to have the worst time series and ACF of all parameters examined. Both the ejection fraction variance and the WLF appeared to be stationary about single points. However, they both also demonstrate fairly long runs where the measures tend to deviate to one side or the other of these points. This tendency for systematic deviations is also evident in the ACF plots (figures 4.3b and 4.4b). These plots demonstrate a tendency for adjacent autocorrelations to be $>$ or $<$ 0. The auto-

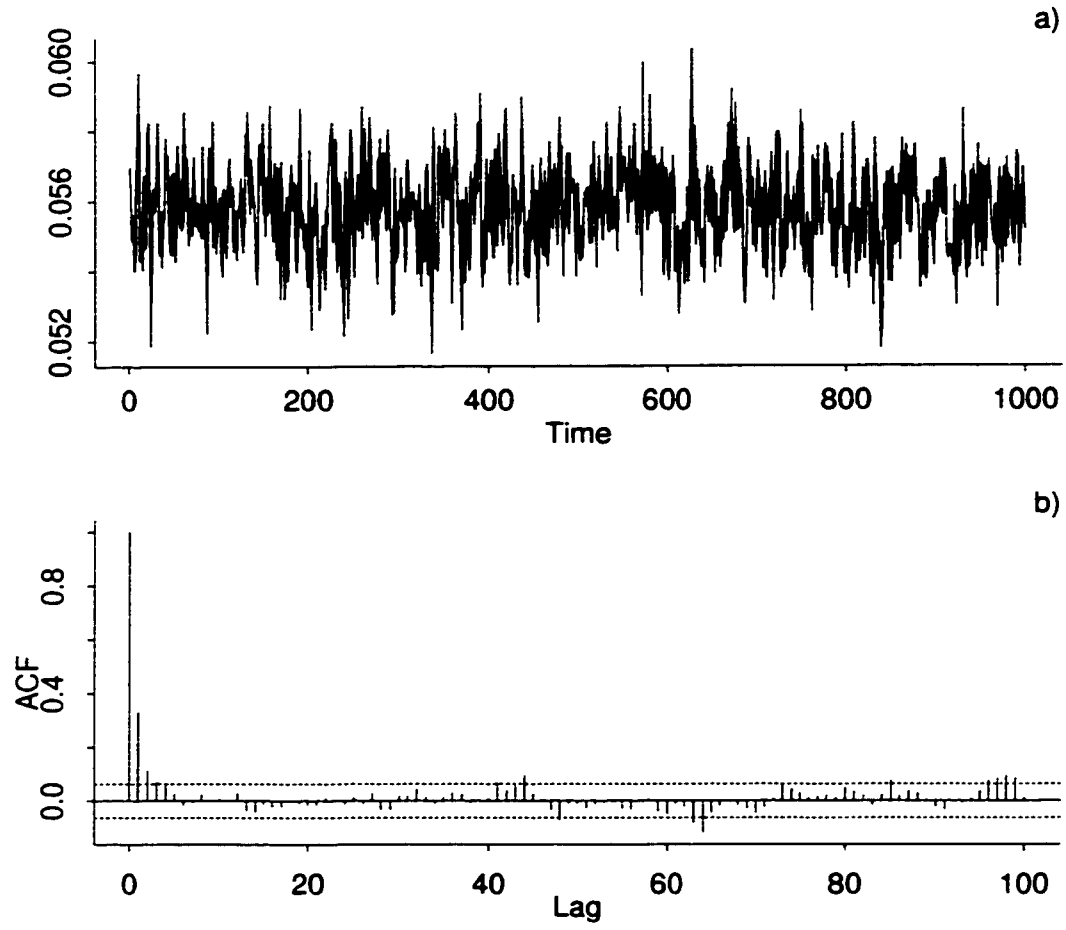

Figure 4.3: Diagnostic plots based on iterates from MVN model with treatment effects. a) Time series plot for standard deviation of the first ejection fraction variable. b) ACF for this series from iterations 100 to 1100. Dashes indicate approximate 0.05-level critical values for testing $\rho_k = \rho_{k+1} = \rho_{k+2} = \cdots = O$.

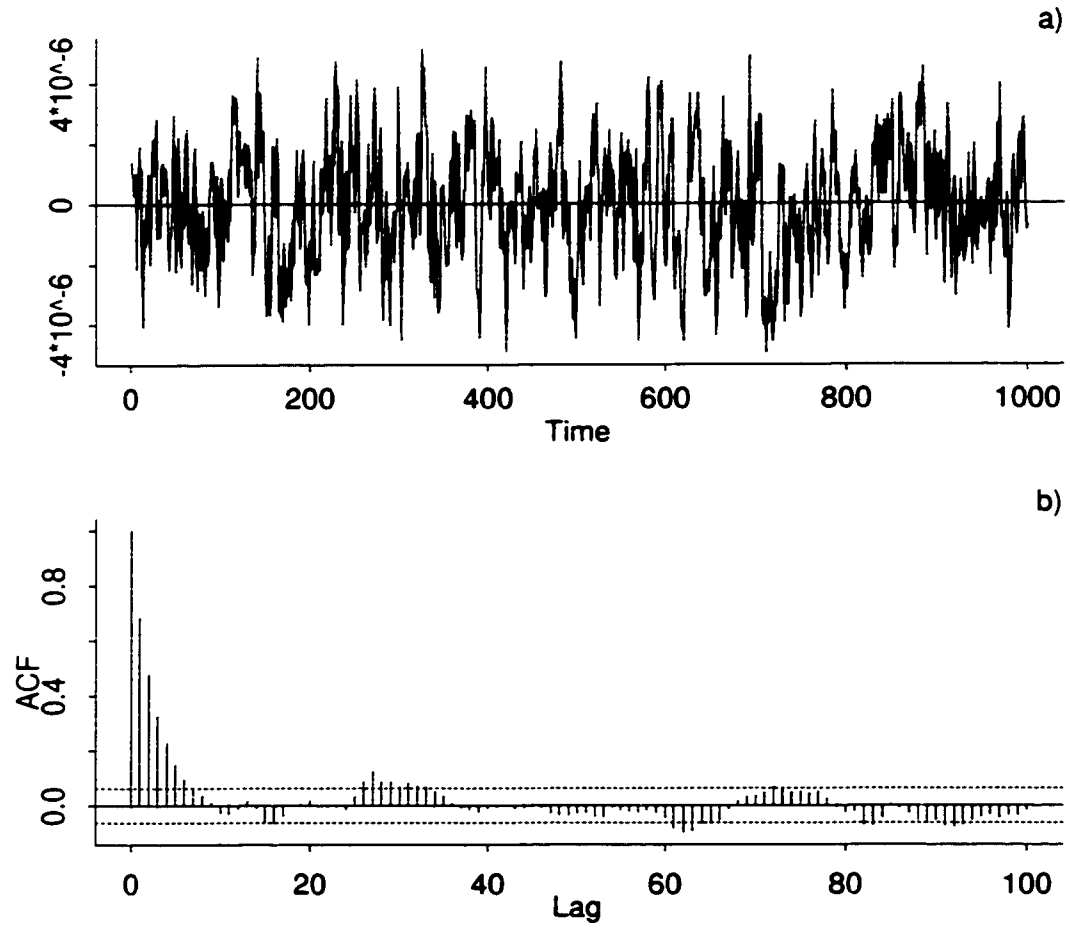

Figure 4.4: Diagnostic plots based on iterates from the MVN model with treatment effects. a) Time series plot for the worst linear function of the parameters. b) ACF for the worst linear function using iterations 100 to 1100.

correlations in these plots are also frequently greater than the 5% critical values for testing whether the correlations are equal to zero.

Despite the weaknesses evident in these plots, it should be noted that these are plots of the worst cases of the plots examined. As with the model with no treatment effects, a decision was made to sample every 100th iterate, as it was felt that this would sufficiently guard against dependence among the imputations.

Ten imputations were sampled, and regression results and fitted values were stored for each imputation. Regression parameters and variances were combined using (3.6) and (3.8). The resulting regression parameters and standard errors are presented in table 4.7. The $\hat{\lambda}$s are also included. When compared with the MI model without treatment effects (see table 4.4), the $\hat{\lambda}$s are larger for several of the variables and especially for the ejection fraction variables. The second ejection fraction variable has the largest associated $\hat{\lambda}$ ($100\hat{\lambda} = 57$). The use of $m = 10$ imputations yielded a standard error approximately $\sqrt{1 + .57/10} = 1.03$ times as large as the standard error which would have been obtained with $m = \infty$ imputations. For the C-statistic, residual deviance and Pearson residuals, see tables 4.5 and 4.6.

4.1.4 Method of Weights

Two models were fit to the data using the method of weights. In the first of these models (MW-1), a mixed continuous and categorical model was applied to the covariates. As with the method of weights without treatment effects, restrictions were placed on the model for the estimation of parameters. The categorical component of the model was fit using a log-linear model with main effects and first order interactions. The continuous portion of the model was fit using main effects for the

categorical variables. In total, 108 parameters were estimated, 91 for the categorical component of the model and 17 for the normal component. Without any prior information, estimates of means could not be obtained and many cells had probabilities of zero. With a flat prior of 1.15, all parameters could be estimated. The use of this prior is equivalent to adding .15 to each cell (Schafer, 1997g), an amount equivalent to adding $.15 \times 2304 = 345.6$ observations. This prior information amounts to $100 \times 345.6/6065 = 5.7\%$ of the total number of number of observations, well below the guideline of $10 - 20\%$ suggested by Schafer.

The joint probabilities of the covariates were used to obtain estimates of regression parameters and standard errors as described previously. These are shown in table 4.7 (MW-1). For the C-statistic, residual deviance and Pearson residuals, see tables 4.8 and 4.9.

MVN Model for Covariates

A second method of weights model, MW-2, was obtained using a MVN for the covariates. Binary variables for type of treatment were included in this model but outcome was not included. The probability densities for covariates were used to obtain estimates of regression parameters and standard errors as detailed above. The resulting coefficients and standard errors can be found in table 4.7. The C-statistic, residual deviance, and Pearson residuals are presented in tables 4.10 and 4.11.

Table 4.7: Logistic regression models with treatment effects. For complete case (CC) analyses, n=3171.

	CC		MW-1		MW-2		MI		
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE	$100\hat{\lambda}$
intercept	-6.83	0.772	-7.05	0.532	-6.80	0.522	-6.52	0.494	06
age	0.358	0.0982	0.373	0.0641	0.362	0.0629	0.355	0.0631	01
CVD	0.815	0.317	0.686	0.237	0.671	0.230	0.700	0.233	02
CHF	0.461	0.254	0.977	0.183	1.01	0.165	1.17	0.169	26
PD	0.399	0.358	0.175	0.239	0.322	0.226	0.296	0.230	08
creat	1.05	0.491	1.55	0.293	1.63	0.285	1.590	0.294	06
EF.1	1.30	0.298	1.15	0.249	0.783	0.235	0.899	0.238	37
EF.2	0.559	0.228	0.413	0.195	0.414	0.189	0.207	0.214	57
EF.3	1.51	0.388	1.24	0.321	1.04	0.298	1.01	0.322	25
CA.1	0.659	0.488	0.788	0.368	0.774	0.362	0.434	0.303	15
CA.2	-0.364	1.11	1.42	0.490	1.50	0.479	1.15	0.448	11
CA.3	1.32	0.506	1.61	0.376	1.65	0.370	1.31	0.310	16
CA.4	1.33	0.521	1.50	0.390	1.56	0.383	1.22	0.326	15
CA.5	1.52	0.544	1.92	0.402	1.98	0.395	1.64	0.340	13
sex	0.107	0.208	0.193	0.140	0.212	0.135	0.220	0.137	01
CABG	-0.317	0.248	-0.500	0.165	-0.526	0.159	-0.458	0.163	03
PTCA	-0.541	0.265	-0.640	0.175	-0.618	0.168	-0.574	0.170	02

Table 4.8: C statistics and residual deviances for the logistic regression models with treatment effects.

Model	C	D
MW-1	0.829	1954.47
MW-2	0.787	2044.36
MI	0.808	1975.84

Table 4.9: Standardized residuals for the marginal observed and expected values for the joint distribution of ejection fraction and CHF in models with treatment effects

Model →				MW-1		MW-2		MI	
EF	CHF	Count	O	E	z	E	z	E	z
<30%	No	112	11	12.55	-0.48	11.04	-0.01	11.43	-0.14
	Yes	99	22	24.01	-0.49	21.39	0.16	24.23	-0.55
	Missing	45	5	6.88	-0.82	5.91	-0.41	6.64	-0.72
30-50%	No	752	34	31.63	0.44	38.31	-0.73	32.21	0.33
	Yes	133	12	20.38	-2.12	23.78	-2.82	22.69	-2.59
	Missing	241	12	12.66	-0.19	11.97	0.01	11.94	0.02
>50%	No	2264	43	52.93	-1.39	64.18	-2.71	59.09	-2.14
	Yes	86	8	7.44	0.22	9.03	-0.37	9.24	-0.44
	Missing	551	18	13.75	1.17	14.98	0.80	15.48	0.66
Not	No	64	7	7.36	-0.15	7.29	-0.12	6.66	0.15
Done	Yes	28	6	8.17	-0.98	8.02	-0.92	8.13	-0.96
	Missing	36	6	6.23	-0.11	3.69	1.33	4.46	0.83
Missing	No	865	27	29.73	-0.52	29.84	-0.54	28.78	-0.34
	Yes	173	43	34.08	1.84	28.76	3.12	31.82	2.40
	Missing	616	47	33.23	2.53	22.80	5.24	28.20	3.80

4.1.5 Discussion

Coefficients

The estimated coefficients and standard errors for the baseline- adjusted models are presented in table 4.4. The coefficients and standard errors for the full models are presented in table 4.7. Also included in these tables are models based on the 3171 subjects with complete data. These will be referred to as complete case (CC) models.

It would be difficult to choose among the missing data models on the basis of coefficients and standard errors. Across the models, the coefficients and standard errors are reasonably similar, and without a complete data model for comparison, the accuracy and efficiency of the missing data models cannot be directly evaluated. However, the missing data models all appear to be superior to their respective

CC models. The missing data models have standard errors which are considerably smaller than those obtained for the CC models, generally being approximately 2/3 as large. Further, the coefficients for the second CA variable in the CC models are negative ($\hat{\beta} = -0.661$ for the baseline-adjusted model), a finding which would not be expected clinically as it implies that cases with 2 vessel disease and PLAD are at a lower risk than cases with normal coronary anatomy. A test of statistical significance does not provide evidence at $\alpha = .05$ this coefficient differs from zero ($\Pr(|z| > 1.96) = .548$). However, in all missing data models, the coefficients are positive and in each case, there is evidence at $\alpha = .05$ (or smaller) that the coefficients differ from zero.

4.1.6 Adequacy of the fits.

Residual deviances and *C*-statistics. The *ad hoc* methods used to examine the adequacy of the fits provide limited information regarding the relative performance for the models. For the baseline-adjusted models, table 4.5 displays the C-statistics and residual deviances for the three missing data models, as well as those obtained from the enhanced model employed by Norris et al. (1999). Note that the C-statistics and residual deviances from the MW-1 model (C=.831;D=1977.94) and the MI model (C=.815;D=1982.55) compare favorably with those obtained from the enhanced model (C=.802;D=1982.45).

Two points need to be made when comparing the measures from the missing data models with those from the enhanced model. The first is that while the enhanced model makes use of additional information from administrative data, this information may contain inaccuracies when compared to the (unobserved) clinical information.

Further, there was no administrative variable to represent ejection fraction. The use of missing data methods to account for the missing EF data may have improved the enhanced model.

The second point is that both complete data and missing data logistic regression methods are not designed to optimize the C-statistic. Consequently, while the C-statistic may provide information regarding the ability of the model to discriminate between those that do and do not die, it may not be a reasonable measure of goodness of fit. Further, the missing data fitting proceedures may yield C-statistics that are more optimistic than is justifiable on the basis of the observed data. In MI, the joint distribution of the covariates and $\mathbf{y}$ is used when generating imputed data sets. At each iteration of the EMMW algorithm, current estimates of $\Pr(\mathbf{y}|\mathbf{X}_{obs}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta})$ are incorporated into the weights to be used in the subsequent iteration. The resulting logistic regression fits may reflect random idiosyncrasies in the data to a greater degree than the fits that would be obtained using complete data. Before confidence is placed on the C-statistics obtained from missing data models, it would be prudent to use cross-validation studies to examine the performance of these measures.

Although there is no equivalent enhanced model for comparison, the pattern of C-statistics and D's among the three full-model logistic regressions is similar to the pattern among the baseline-model logistic regressions (see table 4.9). The MW-1 model appears to have the best fit (C=.829;D=1954.47) followed by the MI model (C=.808;D=1975.84) and the MW-2 model (C=.787;D=2044.36). Note that although the residual deviances for these models are smaller than their baseline model counterparts, the C-statistics are slightly smaller. This is consistent with the observation made by Harrell and Lee (1984), that increasing the number of variables

in a model can reduce the ability of the model to discriminate among outcomes.

The tables of Pearson residuals (tables 4.6 and 4.9) indicate that there are problems with the logistic regression fits obtained using missing data methods. In several of the cells of the tables, the discrepancies between observed and expected counts are large. The largest discrepancies occur in the cells where cases were missing observations on both the CHF and the EF variables. For all models, the expected counts in these cells are lower than the observed counts, indicating that the missing data procedures tended to assign (or weight) these subjects to categories for which the risks of death were too low. To a lesser degree, this also appears to have happened to cases with missing EF observations who are known to have CHF. Other notable discrepancies occurred for cases with EF $>$ 50% and a diagnosis of CHF; and for cases with and EF of 30-50% and a diagnosis of CHF. In both of these conditions, the expected counts were greater than the observed counts.

Several plausible explanations may account for these discrepancies. First, it is possible that the logistic regression model has been misspecified. Terms for a CHF $\times$ EF interaction may be needed in the model. However, a test of this possibility using cases with complete data revealed no evidence of an interaction ($\chi^2_3 = 4.61; p = 0.203$).

A second possibility is that the joint probability models have been misspecified. To some degree, this is supported by the pattern of Pearson residuals found in the table. A mixed continuous and categorical model is arguably better than a MVN model for the joint distribution of the covariates. The discrepencies for the MW-1 model, which employed a mixed continuous and categorical model, are generally smaller than those from the MW-2 and MI models, which employed MVN models

for their joint distributions. In addition to the choice of probability distribution, the relationships among the variables also have to be considered, and important interaction terms may have been excluded from the models.

Thirdly, it is possible that the data are not MAR. In the absence of information regarding the missing data mechanism, observations with missing data will tend to be placed into (or weighted to) the joint categories with the highest probabilities. Of the cases with an observed CHF variable, 89% had a negative diagnosis. Of the cases with an observed EF, 65% had an EF $> 50\%$. If the data are MAR, the information in the observed data will ensure that the cases with missing observations will tend to be placed in (or weighted toward) the categories to which they truly belong. If the data are not MAR, the inappropriate weighting or placement of cases to categories could lead to distortions in the logistic regression parameters.

It would be difficult to determine how to pursue these different possibilities. Not only are the modeling procedures unwieldy, but the problem could lie in any or all of the possibilities noted above. Even more problematic is that without complete data, it would be impossible to verify that any changes in fitting procedures would produce a more adequate fit.

Although there appear to be problems in the baseline-model fits, they will still be used in risk-adjustment procedures. This is because 1) it is difficult to know what corrective measures may be taken to improve the fits, 2) it is possible that the observed distortions will have little impact on the risk-adjusted estimates, since these distortions may be distributed fairly evenly among the providers, and 3) although flawed, the risk-adjusted estimates may be still be more efficient and less biased than those obtained using only the complete cases. The effects of different missing data

mechanisms on risk-adjusted estimates is explored using computer simulations in chapter 5.

4.2 Risk Adjustment with Missing Data

As in chapter 2, a standardized mortality ratio (SMR) will be considered to be an indirectly standardized relative risk (RR) which represents the increase in risk associated with treatment by a provider. The population averaged proportion (PAP) will be considered to be the average risk of death for cases treated by a given provider or treatment. The baseline-model (BM) measures of risk and relative risk employ indirect standardization and are based on logistic regression models which do not include the treatment variables. For brevity, only the offset model estimates will be referred to as full-model (FM) estimates. The directly standardized estimates will be referred to as DS estimates. The FM and DS measures are based on logistic regression models that include treatment effects. For details regarding the estimation of these measures, see chapters 2 and 3 as well as appendix C.

In tables 4.10 and 4.12, complete case risk estimates have been provided for purposes of comparison with the estimates obtained using missing data methods. In all cases, risk-adjusted estimates of standard errors obtained using missing data methods were smaller than those obtained using the CC logistic regressions. This is not surprising, as only 3171 subjects had complete data.

4.2.1 Risk Ratios

Standard Errors

Several observations regarding relative risks can be made on the basis of the data presented in tables 4.8 and 4.9. The standard errors for risk ratios are smaller for the full-model adjusted measures than they are for the baseline-model adjusted measures. The largest differences occur in the medical treatment condition. For the indirectly standardized measures based on MI, the standard error for the baseline-model $\widehat{RR}$ is over 50 % larger than the standard error for the full-model $\widehat{RR}$ (0.0804 *vs* 0.0523).

The standard errors for the directly standardized measures tend to be slightly smaller than those for the the full-model measures, although they are generally of comparable magnitude. For the MI model, however, the standard error for the full-model $\widehat{RR}$ is smaller than the standard error for the directly standardized $\widehat{RR}$ for those receiving medical treatment. It is also worth noting that within each of the three different treatments, the full-model $\widehat{RR}$s are more sensitive to the type of risk adjustment method than are the directly standardized and baseline-model $\widehat{RR}$s.

Point estimates

The relative risks obtained using directly standardized and full-model methods are very similar. For the PTCA treatment group, the $\widehat{RR}$s are similar regardless of the method of adjustement employed. For CABG, the full-model and directly standardized measures are very similar and are generally smaller than the baseline-model relative risks. In the medical treatment group, the $\widehat{RR}$s tend to be larger for the adjusted and direct methods of standardization than for the unadjusted measures.

In the present example, the differences in standard errors and point estimates

Table 4.10: Relative risk measures with missing data in the covariates. The full-model and baseline-model relative risks are also known as standardized mortality ratios (SMRs).

Treatment →		Medical		CABG		PTCA	
Model	Method	RR	SE	RR	SE	RR	SE
CC	BM	1.17	0.121	0.916	0.161	0.737	0.180
	FM	1.19	0.0784	0.900	0.152	0.726	0.171
	DS	1.20	0.0972	0.900	0.144	0.731	0.143
MW-1	BM	1.23	0.0784	0.858	0.0984	0.747	0.107
	FM	1.25	0.0708	0.836	0.0852	0.745	0.0877
	DS	1.26	0.0619	0.833	0.0820	0.739	0.0862
MW-2	BM	1.25	0.0797	0.844	0.0980	0.743	0.108
	FM	1.28	0.0651	0.813	0.0807	0.746	0.0875
	DS	1.28	0.0652	0.812	0.0803	0.748	0.0870
MI	BM	1.23	0.0804	0.855	0.0984	0.753	0.108
	FM	1.24	0.0523	0.831	0.0900	0.765	0.103
	DS	1.24	0.0644	0.831	0.0821	0.768	0.0890

Table 4.11: 95 % confidence intervals for relative risks associated with CABG treatment.

Treatment →		CABG		PTCA	
			95 %		99 %
Model	Method	RR	Confidence Interval	RR	Confidence Interval
MW-1	BM	0.858	0.665 - 1.05	0.747	0.471 - 1.02
	FM	0.836	0.669 - 1.00	0.745	0.519 - 0.971
	DS	0.833	0.672 - 0.994	0.739	0.517 - 0.961
MW-2	BM	0.844	0.652 - 1.04	0.743	0.464 - 1.02
	FM	0.813	0.655 - 0.971	0.746	0.520 - 0.972
	DS	0.812	0.655 - 0.969	0.748	0.524 - 0.972
MI	BM	0.855	0.662 - 1.05	0.753	0.474 - 1.03
	FM	0.831	0.655 - 1.01	0.765	0.499 - 1.03
	DS	0.831	0.670 - 0.992	0.768	0.538 - 0.998

are small enough that conclusions based on 95 or 99 % confidence intervals would not generally depend on the missing data model or on the method of estimating the relative risk. An exception to this occurs in the CABG treatment condition, where conclusions based on strict adherence to 95 % confidence intervals would differ (see table 4.9). It should be noted, however, that the magnitude of the differences in standard errors in the medical treatment condition indicates that the choice of missing data model and adjustment method has the potential to lead to very different conclusion regarding risk adjusted estimates.

4.2.2 Population Averaged Proportions

Table 4.12: Measures of population averaged proportions obtained using missing data methods.

Treatment →		Medical		CABG		PTCA	
Model	Method	PAP	SE	PAP	SE	PAP	SE
CC	BM	0.0459	0.00473	0.0358	0.00628	0.0288	0.00705
	FM	0.0467	0.00428	0.0352	0.00679	0.0284	0.00734
	DS	0.0468	0.00547	0.0352	0.00640	0.0286	0.00612
MW-1	BM	0.0610	0.00389	0.0426	0.00489	0.0371	0.00533
	FM	0.0622	0.00503	0.0415	0.00488	0.0370	0.00486
	DS	0.0627	0.00458	0.0414	0.00466	0.0367	0.00475
MW-2	BM	0.0619	0.00400	0.0419	0.00486	0.0369	0.00537
	FM	0.0634	0.00466	0.0403	0.00459	0.0370	0.00479
	DS	0.0635	0.00466	0.0403	0.00457	0.0371	0.00477
MI	BM	0.0609	0.00399	0.0424	0.00488	0.0374	0.00537
	FM	0.0615	0.00411	0.0413	0.00500	0.0380	0.00552
	DS	0.0617	0.00453	0.0413	0.00465	0.0381	0.00489

From table 4.12, it is evident that within given missing data models, the adjusted and directly standardized measures are very similar. For the medical treatment

group, the standard errors for the baseline-model risks are smaller than those for the full-model and directly standardized measures. The standard errors for the directly standardized risks are generally smaller than those for the full-model risks. An exception to this occurs in the medical treatment group for estimates obtained using the MI logistic regression models. In this case, the directly standardized estimate has the largest standard error. As with the relative risks, the standard errors obtained using the direct standardization and baseline-model adjustements are the most stable across missing data models within types of treatment.

4.2.3 Discussion

The finding that direct standardization often yielded the smallest variances may seem surprising in the light of the generally held view that directly standardized rates are less efficient than indirectly standardized rates (Breslow and Day, 1987b). It should be noted, however, that the model-based methods of obtaining directly standardized rates in the present examples differs from the method on which the assertion of inefficiency is based. In direct standardization, the typical method of calculating rates is to obtain stratum specific estimates of risk for each treatment group and then apply these estimates of risk to a common population standard. The loss of efficiency in this method arises because the entire data set is not employed to estimate common estimates of risk for the risk strata.

In the current example a model-based approach is employed, the entire data set is utilized to obtain common estimates of risk. These common estimates are then applied to a common population standard, which in this case is the distribution of patients in the observed population. The resulting rates appear to be stable and

efficient when compared to the indirectly standardized rates.

Although the standard errors estimated using missing data methods were a good deal smaller than those based on the complete cases, in the absence of the complete data, an adequate assessment of the performance of these methods is not possible. When employing missing data methods to obtain risk adjusted estimates, it would be useful to know 1) the degree of bias one might encounter in the point estimates, 2) how well the standard errors represented the variability of the point estimates, and ultimately, 3) the trustworthiness of inferences based on these point estimates and standard errors. The following chapter will use Monte Carlo simulations to address these issues.

Chapter 5

Monte Carlo Simulations

5.1 Missing Data Simulations

The purpose of this chapter is to employ Monte Carlo simulations to explore how well missing data methods for risk adjustment work under a variety of missing data conditions. The simulations were based loosely on the distribution of variables in the APPROACH data set. Due to the number of variables and the complexity of the relationships among these variables, the problem was simplified by considering a subset of the APPROACH variables. Only the performance of relative risk estimates was examined.

5.1.1 Variables

The variables chosen as a basis for the simulations were age, CHF, 6-month mortality (y), and treatment. Age was selected because it is continuous and was completely observed in the APPROACH data set. The risk factor CHF was chosen because it is binary, because a positive diagnosis was rare (11.34% of the observed diagnoses were positive), and because a large proportion of the CHF observations were missing (24.55%). These factors indicate that the use of a normal model will not be appropriate for CHF, and provide an opportunity to examine the effects of misspecifying the multivariate normal models used for the joint distributions of the variables. The simulations used the age and CHF variables to obtain risk adjusted estimates for the

medical, CABG and PTCA treatment groups. These groups will be denoted as h_1, h_2 and h_3.

5.1.2 Computations

The computer used for the simulations was a Sun^{TM} Ultra^{TM} 1 Model 140 (Sun Microsystems Computer Company) with 128 megabytes of random access memory. The computer software for the missing data methods is described in sections 4.1.1 and 4.1.1. Computer code to perform the simulations was written in S-PLUS 5 version 2 (MathSoft, 1999).

5.1.3 Generating the Random Samples

The distribution of the APPROACH variables served as a basis for the parameters used in generating data for the simulations.The degree of confounding between both age and CHF and the treatment groups was modest. An initial set of simulations was based on the observed distributions among the APPROACH variables. The results from these simulations are not reported, as they are similar to the results presented below for simulations in which a greater degree of confounding was introduced. For each case i in a given sample, observed values for covariates and treatment variables were generated randomly using a mixed continuous and categorical model.

The covariate model. The mean ages and probabilities by joint CHF$\times$treatment category are presented in table 5.1. The variance of the ages within each of the joint categories was 125, and was based on the variance observed in the APPROACH data.

Table 5.1: Distribution of covariates for the missing data simulations. The means, variance and probabilities are all based on the observed data in the APPROACH database. However, confounding was increased to provide a more rigorous test of the missing data methods. **a)** Marginal and joint distributions of age across the CHF and treatment categories. Within each joint category $c = 1, \ldots, 6$, age $\sim N(\mu_c, 125)$. **b)** Marginal and joint probability distributions for treatment and CHF.

a) Mean age by category

		CHF		
		No	Yes	
Treatment	h_1	63	68	63.72
	h_2	61	66	62.46
	h_3	65	70	65.75
		63.14	68.67	63.73

b) Joint distribution of treatment and CHF

		CHF		
		No	Yes	
Treatment	h_1	.45	.05	.5
	h_2	.19	.01	.2
	h_3	.255	.045	.3
		.895	.105	

Sampling from the covariate model. To generate the data, each of the cases were assigned to a category by randomly sampling from a multinomial distribution on the basis of the probabilities in table 5.1 b). The age of the cases were then randomly sampled from a normal distribution. The mean of this distribution corresponded with the category to which the case and assigned, the variance of this distribution was 125.

Generating the outcomes. The model for the outcomes was based on a multiplicative model with no hospital by risk factor interactions (see 2.7 in chapter 2). For each sample, two steps were employed to determine the probability of death for each subject. In the first step, a logistic model was applied to the sampled age and CHF variables. For case i, the fitted probability of death from this model was

$$p_i = \frac{\exp(-6.59 + .0488 \times age_i + 1.71 \times CHF_i)}{1 + \exp(-6.59 + .0488 \times age_i + 1.71 \times CHF_i)}. \tag{5.1}$$

The coefficients for **age** and **CHF** were based on the logistic regression models for the APPROACH data presented in chapter 4. The intercept was chosen to ensure that the expected death rate was 5%, as the six-month mortality rate in the APPROACH data was approximately 5%.

In the second step in obtaining the the probability of death, the p_i for each case was multiplied by the relative risk associated with the treatment provider h_k to which the case was assigned, or

$$p_{ik} = RR_k p_i \tag{5.2}$$

The relative risks were 1.25 for provider h_1, .85 for provider h_2, and .731 for provider h_3. The relative risks for providers h_1 and h_2 were based on the relative risks observed

Table 5.2: Regression parameters for simulations

	Coef	OR
Intercept	-6.59	
age	0.0488	1.05
CHF	1.71	5.53

for the treatments in the models in chapter 4. The relative risk for h_3 was selected so that the expected value of the p_{ik} was .05. This ensured that for each sample

$$\mathbf{E}(\sum_{i=1}^{n} y_i) = \mathbf{E}(\sum_{i=1}^{n} p_i) = \mathbf{E}(\sum_{k=1}^{3} \sum_{i \in h_k} p_{ik})$$

as this condition was necessary for obtaining risk adjusted estimates of relative risk with expectations equal to the RR_k.

For each case in a given sample, the outcome variable y_i was generated by sampling from a Bernoulli distribution with the parameter p_{ik}. For each sample, $N = 2000$ complete observations were generated according to the procedures described above. The number of observations for each provider was not fixed, but had expectations based on the proportion of subjects treated by each provider. For providers h_1, h_2 and h_3, the expected numbers of observations were 1000, 400, and 600 respectively. Observations were then deleted from the **CHF** variable according to the missing data models described below. A total of 500 samples was generated for each of these models.

5.1.4 Missing Data Models

The simulations examined missing data models in which the CHF variable was MCAR, stratified MCAR within levels of the outcome (MD_y), MCAR within levels

of age (MD_{age}), and non-MAR (NMAR) with respect to CHF. For each of the models, two levels of missingness were examined; one in which 25% of cases had missing CHF values, and one in which 40% of cases had missing values.

MCAR

For the MCAR condition, the values of CHF were randomly deleted using a Bernoulli model in which each case has a given probability of having missing data. There were two levels of missingness, with $\Pr(R = 0) = .25$ and $\Pr(R = 0) = .4$, where R is a binary (0,1) variable with 0 indicating a missing response for CHF.

Stratified MCAR

Missing dependent on y. For the models in which the missing data were dependent on the outcome, a total of 4 conditions were examined. For two of these conditions, the probability that an observation was missing CHF was 1.5 times as great for cases with $y = 1$ than for cases with $y = 0$, or

$$\frac{\Pr(R = 0|y = 1)}{\Pr(R = 0|y = 0)} = 1.5.$$

As described above, this missing data model was examined where 25% and 40% of the observations were expected to be missing from the CHF variable. The probabilities in the ratios were adjusted to reflect this requirement (see table 5.3).

For the other two conditions in which the missing data mechanism was dependent on y, the probability that observations were missing was 2 times as great for cases with $y = 1$ than for cases with $y = 0$, or

$$\frac{\Pr(R = 0|y = 1)}{\Pr(R = 0|y = 0)} = 2.$$

Table 5.3: Missing data probabilities for the MD_y conditions

% Missing	RR_{miss}	$\Pr(R = 0 \mid y = 1)$	$\Pr(R = 0 \mid y = 0)$
25	1.5	.366	.244
	2.0	.476	.238
40	1.5	.585	.390
	2.0	.762	.381

Table 5.4: Coefficients used for the MD_{age} condition.

OR_{decade}	% Missing	α	β_{age}
1.1	25	-1.69	.00953
	40	-0.997	.00953
1.2	25	-2.24	.0182
	40	-1.54	.0182

As above, this probability model was examined where there was 25% and 40% missing data. The probabilities that correspond to these models are provided in table 5.3.

Missing dependent on age. A logistic model was used for the conditions where the probability of a missing CHF observation depended on age. The coefficients for age in these models were based on the odds ratios associated with a ten unit difference in age. The odds ratio associated with a ten year increase in age was 1.1 in two of the models and 1.2 for the other two models. Intercepts for these models were obtained to ensure that the expected number of cases missing for models employing each of the coefficients was 25 and 40%. The coefficients used for these models are in table 5.4.

Table 5.5: Missing data probabilities for the NMAR conditions

% Missing	RR_{miss}	$\Pr(R = 0 \mid CHF = 1)$	$\Pr(R = 0 \mid CHF = 0)$
25	1.5	.357	.238
	2.0	.455	.227
40	1.5	.571	.381
	2.0	.727	.364

Non-missing at Random

Missing dependent on CHF. For the NMAR condition, the probability model for deleting values of CHF depended only on the value of CHF. A total of four NMAR conditions were examined. In two of these conditions, the probability that CHF was missing was 1.5 times as great for cases with $CHF = 1$ than for cases with $CHF = 0$, or

$$\frac{\Pr(R = 0 \mid CHF = 1)}{\Pr(R = 0 \mid CHF = 0)} = 1.5.$$

This missing data model was examined where there were 25% and 40% of the CHF observations expected to be missing for the total sample.

For the two other NMAR conditions, the probability that CHF was missing was 2 times as great for cases with $CHF = 1$ than for cases with $CHF = 0$, or

$$\frac{\Pr(R = 0 \mid CHF = 1)}{\Pr(R = 0 \mid CHF = 0)} = 2.$$

The probabilities of missing observations in the four different conditions are presented in table 5.5

5.1.5 Missing Data Methods

Both MI and EMMW methods were examined. For MI, multivariate normal (MVN) models were employed for the joint distributions of variables. For each sample with missing observations, three imputed data sets were generated. Estimates and standard errors were based on these three imputed data sets. For EMMW, mixed continuous and categorical models were employed for the covariates. A MVN model was not used for the EMMW analyses. This is because this method did not appear to perform well in chapter 4, and because the more appropriate mixed continuous and categorical model was available.

5.1.6 Risk-adjustment Methods

Two risk-adjustment methods were examined. These were the baseline model (BM) and direct standardization (DS). Full-model (FM) adjusted rates and standard errors were not calculated as these estimates were similar to the DS estimates in chapter 4, and because calculation of their standard errors is computationally expensive and would have greatly increased the time required to perform the simulations.

5.1.7 Parameters Examined

For each sample generated, both EMMW and MI were used to obtain BM and DS estimates of relative risk, as well as the standard errors of these relative risks. These standard errors are denoted as $s_{\widehat{RR}}$. Note that the BM adjusted relative risks are standardized mortality ratios (SMRs).

The expectations and standard deviations for the Monte Carlo relative risks were obtained under the assumption that the distribution of a relative risk is asymptoti-

cally normal. The expectation for the Monte Carlo relative risks is

$$\mu_{\widehat{RR}} = \frac{1}{n_v} \sum_{s=1}^{n_v} \widehat{RR}_{n_v}, \tag{5.3}$$

for the samples $1, \ldots, n_v$ generated for the given condition v. The standard deviations for the Monte Carlo relative risks was obtained as

$$\sigma_{\widehat{RR}} = \sqrt{\frac{1}{n_v - 1} \sum_{s=1}^{n_v} \left(\widehat{RR}_v - \mu_{\widehat{RR}}\right)^2}. \tag{5.4}$$

For each condition v, the mean of the estimates of the standard errors from each of the n_v Monte Carlo samples was obtained as

$$\bar{s}_{\widehat{RR}} = \frac{1}{n_v} \sum_{s=1}^{n_v} s_{\widehat{RR}_v}. \tag{5.5}$$

The regression coefficient for the CHF variable was also examined. The mean regression coefficient, the mean standard error of the coefficient and the 95% coverage probability for the coefficient were examined for each missing data method and missing data condition.

5.1.8 Evaluation of Methods

Relative Bias

For each condition, the means of both the DS and BM relative risks were compared with the true relative risks used in generating the data. When evaluating the bias, a measure of relative bias was obtained as the difference between the mean of the Monte Carlo relative risk and the true value divided by the true value, or

$$\mathcal{B}_{\widehat{RR}} = \frac{\mu_{\widehat{RR}} - RR}{RR}.$$

The bias was scaled in this manner to facilitate comparison of the biases across relative risks with different magnitudes.

Precision

There are two issues of concern when addressing the precision of the estimates. While smaller variance estimates reflect a more precise estimate, the variance estimates (or the corresponding standard errors) should accurately reflect the variability of the Monte Carlo estimates. If the estimates are an accurate reflection of the variability of the estimates, and if the distribution of the the point estimates is asymptotically normal, the mean of estimates of error will equal the standard deviation of the Monte Carlo point estimates.

Standard errors. For each missing data condition and method, the mean of the standard errors of the relative risks ($\bar{s}_{\widehat{RR}}$) was compared with the standard deviation of the relative risks for the given condition and method ($\sigma_{\widehat{RR}}$) to determine whether the standard errors tended to over- or under-estimate the standard deviation of the relative risks.

Coverage probabilities. Coverage probabilities were employed to ensure that inferences based on the point estimates and estimated standard errors were valid. The coverage probabilities were the proportion of the samples for which the 95% confidence interval contained the values of the relative risks used in generating the data. For each sample, the 95% confidence interval was calculated as $\widehat{RR} \pm 1.96 \times s_{\widehat{RR}}$.

The number of samples was 500 for each condition. Assuming that the distribution of the relative risks is approximately normal, and using a normal approximation to the binomial distribution, approximately 95% of the coverage probabilities would be expected to fall in the interval

$$.95 \pm 1.96\sqrt{(.95)(.05)/500)} \approx [.93 < CP < .97], \tag{5.6}$$

provided the $\widehat{RR}$ are unbiased and if the $s_{\widehat{RR}}$ do not systematically over- or under-estimate the $\sigma_{\widehat{RR}}$. This interval served as a rough guideline for assessing the coverage probabilities.

Efficiency

Mean squared errors were used to determine efficiencies of the relative risk estimates. For both the complete data and missing data estimates of relative risk, the mean squared error (MSE) was defined as

$$\text{MSE} = \sigma^2_{\widehat{RR}} + \text{bias}^2 \tag{5.7}$$

where

$$\text{bias} = \mu_{\widehat{RR}} - RR \tag{5.8}$$

and the $\sigma^2_{\widehat{RR}}$ are the Monte Carlo variances of the estimated relative risks. The relative efficiency of the missing data estimated relative risks will be defined as

$$eff(\widehat{RR}, \widehat{RR}_{complete}) = \frac{\text{MSE}_{complete}}{\text{MSE}_{missing}}. \tag{5.9}$$

If the efficiency is smaller than 1, the missing data estimate of relative risk ($\widehat{RR}$) has a larger MSE than does the complete data estimate.

A crude estimate of a critical region for the efficiencies can be obtained by treating the variances of the complete data relative risks (based on 5000 samples) as if they are the true population variances. In the absence of bias, and if the true variances for the missing data relative risks and complete data relative risks are identical, the efficiency can be expressed as

$$eff(\widehat{RR}, \widehat{RR}_{complete}) = \frac{\sigma^2_{complete}}{\hat{\sigma}^2_{\widehat{RR}}} \tag{5.10}$$

which is distributed as $1/\chi^2_{499}$. On the basis of the above assumptions, a 5% critical region for the efficiencies is $\approx$ [.887-1.137].

5.1.9 Results

Complete data estimates. The complete data Monte Carlo mean relative risks and standard errors were based on 5000 samples and can be found in table 5.6. The true relative risks used in generating the data were 1.25, .85, and .731 for h_1, h_2, and h_3 respectively. The complete data Monte Carlo mean relative risks for the DS and BM methods are comparable to the true value for h_1. The relative risks for the h_2 and h_3 conditions appear to be biased. For provider h_2, the DS estimate is biased towards the null (.857 vs .85) and the BM estimate is biased away from the null (.842 vs .85). For h_3, the mean of the DS estimates is biased away from the null (.723 vs .731).

Table 5.6: Complete data Monte Carlo mean relative risks and standard deviations, based on 5000 samples with complete data. The true values of the relative risks are 1.25, .85, and .731 for providers h_1, h_2 and h_3 respectively.

	$\mu_{\widehat{RR}}$			$\sigma_{\widehat{RR}}$		
	h_1	h_2	h_3	h_1	h_2	h_3
DS	1.25	0.857	0.723	0.0962	0.2119	0.1176
BM	1.25	0.842	0.732	0.0963	0.2121	0.1138

Missing data point estimates. For provider h_3, the means of the missing data relative risks are similar to the true values (see table 5.7) and the relative biases are small (see table 5.8). The observed differences can be attributed to random variability, since these means are based on 500 samples. For provider h_2, the means

tend to be smaller than the true values, indicating tendency to over-estimate the magnitude of the effect associated with this treatment. The degree of bias is greatest in the $MD_{CHF}; RR_{CHF} = 2$ condition with 40% of the CHF observations missing. Across all missing data conditions, the degree of bias in the indirectly standardized (BM) estimates is greater than the bias in the DS estimates. For provider h_3, the means of the relative risks tend to be larger than the true RRs, indicating a tendency to underestimate the magnitude of the true effect. The degree of bias is generally greater for the BM estimates than for the DS estimates, and the degree of bias is greatest in the $MD_{CHF}; RR_{CHF} = 2$ condition with 40% missing CHF observations. In this condition, the estimates for the MI missing data method demonstrate enough bias that the mean of the estimates for h_2 is smaller than the mean of the estimates for h_3. For all other methods and conditions, the correct ranking of the mean relative risks is preserved.

Missing data standard errors. In table 5.9 it is evident that the standard errors of the BM adjusted estimates are much larger than those obtained using the DS method of adjustment. The standard errors for provider h_2 are also larger than those obtained for providers h_1 and h_3. This is not surprising, as the expected number of observations for these providers is 400 for h_2 and 1000 and 600 for providers h_1 and h_3 respectively. For each provider, The means of the standard errors appear to be similar across the missing data conditions.

Standard deviations of the relative risks. Across all missing data conditions, missing data methods and adjustment methods, the standard deviations of the relative risks are similar (see table 5.10). While there are fluctuations in the

standard deviations, these may be due to random variation.

Bias in the standard error estimates. For the BM method of standardization, the mean standard errors all over-estimate the standard deviations of the relative risks. This bias appears to be greatest for provider h_1, where the mean standard errors are generally 40-50% larger than their respective standard deviations. The reasons for the inflation in the standard errors is discussed below (see section 5.1.11) and is explored using Monte Carlo simulations in section 5.2.

Across the missing data conditions, standard errors for the DS relative risks are similar to their respective Monte Carlo standard deviations (the $\sigma_{\widehat{RR}}$). There appears to be a tendency for the DS estimates to underestimate the standard deviations. This is most noticeable when EMMW is used as the missing data method. The greatest degree of underestimation occurs for provider h_1 in the $MD_{age}; OR_{decade} = 1.1$ condition with 40% of the CHF observations missing, where the MW $\bar{s}_{\widehat{RR}}$ is only .916 times as large as the $\sigma_{\widehat{RR}}$. This may be due to Monte Carlo error; it would be difficult to argue that the degree of bias in the $\bar{s}_{\widehat{RR}}$'s is related to the missing data mechanism, as the bias is smaller in the corresponding $OR_{decade} = 1.2$ condition.

95% coverage probabilities. The 95% coverage probabilities for the different conditions are displayed in table 5.12. The coverage probabilities for the baseline method of standardization are all $> .97$, a reflection of the tendency of the mean standard errors to be larger than the standard deviations of the relative risks when the baseline method is employed. A majority (55/84) of these coverage probabilities are $\geq .99$. The coverage probabilities for the DS method are more reasonable, with 83% (35/42) of those from the MI method falling between .93 and .97., and 76%

(32/42) from the EMMW method falling within this range. However, there appears to be a tendency for the DS coverage probabilities to be too small; 34/42 of the MI and 37/42 of the MW coverage probabilities are $< .95$.

Efficiency. The efficiencies of the relative risk estimates are close to one across most of the missing data conditions and missing data methods. There appears to be a tendency for the efficiencies to be smaller in conditions where 40% of the CHF observations are missing, although this effect is most noticeable in the $MD_{CHF}; RR_{CHF} = 2$ conditions. Efficiencies are generally highest in provider h_2 and lowest in h_3. This is not surprising since the proportions with CHF in h_1, h_2 and h_3 were .1, .05, and .15 respectively, and CHF was the only variable with missing observations.

There are anomalies in the efficiencies which warrant further investigation. In two of the conditions, the efficiencies for h_2 are larger than the upper bound for the critical region of [.887-1.137] described above. The greatest discrepancy occurs in the $MD_{CHF}; RR_{CHF} = 1.5$ condition with 40% of the CHF observations missing. In this case, 3/4 of the efficiencies for h_2 are > 1.2. It is possible that this anomaly is due to Monte Carlo error and that the 5% critical region used to evaluate the efficiencies is incorrect. Bias was ignored when constructing this region, but in tables 5.7 and 5.8 the estimated relative risks appear to be biased. To examine the possibility that the large efficiencies were due to chance, 2000 more samples were generated for this condition. The results of this simulation are presented in table 5.14. In this case, the efficiencies for h_2 are all close to 1.

Regression parameter simulations. Results for the regression parameter for CHF are presented in table 5.15. The true value of the regression parameter is 1.71. In most conditions, the mean regression parameter is close to the true value. The largest distortions occur for EMMW where the missing data mechanism depended on the outcome. In the $MD_y; RR_y = 2$ condition, the regression parameters are underestimated to a large degree (1.32 and 1.21 for the DS and BM models respectively). The only other distortions of note occur in the NMAR ($MD_{CHF}; RR_{CHF} = 2$) condition with 40% of the CHF observations missing. All of the regression parameters were underestimated in this case.

The standard errors of the coefficients are largest where the missing data mechanism depended on the outcome (RR_y=2; 40 % missing). The standard errors are also large in the NMAR (MD_{CHF}) conditions where 40% of the CHF observations were missing. The coverage probabilities are often poor. When MI was used as the missing data method, all but one of coverage probabilities for the 95% confidence intervals are $< .95$. In some cases they are much smaller than .95, most notably for the MD_y mechanism where RR_y=2 and 40% of the observations were missing. In this case, the 95% CPs from the MI method are .776 and .774 respectively for the DS and BM adjusted estimates.

5.1.10 Discussion

In all but the most extreme NMAR (MD_{CHF}) condition, the DS risk adjustment procedures appeared to work well when employing either EMMW or MI to handle missing data. The performance of the MI method was comparable to the EMMW method, even though the models for the joint distributions employed when using MI

Table 5.7: Monte Carlo mean estimates of relative risk. Each condition is based on 500 random samples. For each sample, data was deleted from the variable CHF according to the missing data mechanism.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		$MCAR$					
MI	DS	1.25	0.838	0.727	1.24	0.815	0.748
	BM	1.25	0.825	0.735	1.24	0.804	0.756
MW	DS	1.25	0.856	0.718	1.25	0.840	0.734
	BM	1.25	0.840	0.730	1.24	0.825	0.747
		$MD_y;\ RR_y = 1.5$					
MI	DS	1.25	0.845	0.732	1.25	0.819	0.741
	BM	1.25	0.832	0.741	1.25	0.807	0.749
MW	DS	1.25	0.860	0.723	1.25	0.835	0.729
	BM	1.25	0.843	0.735	1.25	0.817	0.743
		$MD_y;\ RR_y = 2$					
MI	DS	1.24	0.846	0.739	1.25	0.817	0.738
	BM	1.24	0.834	0.747	1.25	0.806	0.746
MW	DS	1.25	0.855	0.731	1.26	0.801	0.738
	BM	1.24	0.839	0.743	1.25	0.783	0.752
		$MD_{age};\ OR_{decade} = 1.1$					
MI	DS	1.25	0.835	0.734	1.24	0.805	0.749
	BM	1.25	0.822	0.742	1.24	0.794	0.756
MW	DS	1.25	0.853	0.723	1.25	0.832	0.734
	BM	1.25	0.838	0.736	1.24	0.817	0.746
		$MD_{age};\ OR_{decade} = 1.2$					
MI	DS	1.25	0.818	0.739	1.24	0.830	0.744
	BM	1.25	0.806	0.747	1.24	0.819	0.752
MW	DS	1.25	0.837	0.729	1.25	0.859	0.727
	BM	1.25	0.822	0.741	1.24	0.842	0.742
		$MD_{CHF};\ RR_{CHF} = 1.5$					
MI	DS	1.24	0.845	0.738	1.24	0.818	0.751
	BM	1.24	0.833	0.746	1.24	0.807	0.758
MW	DS	1.25	0.861	0.728	1.24	0.841	0.736
	BM	1.24	0.847	0.740	1.24	0.827	0.749
		$MD_{CHF};\ RR_{CHF} = 2$					
MI	DS	1.23	0.827	0.754	1.23	0.772	0.776
	BM	1.23	0.816	0.762	1.23	0.764	0.781
MW	DS	1.24	0.841	0.744	1.24	0.794	0.760
	BM	1.23	0.828	0.755	1.23	0.783	0.770

Table 5.8: Relative bias of the Monte Carlo mean estimates of relative risk. The relative bias was obtained as $(\mu_{\widehat{RR}} - RR)/RR$. Each condition is based on 500 random samples. For each sample, data was deleted from the variable CHF according to the missing data mechanism.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		$MCAR$					
MI	DS	0.0016	-0.0144	-0.0054	-0.0055	-0.0411	0.0241
	BM	0.0012	-0.0296	0.0064	-0.0059	-0.0543	0.0347
MW	DS	0.0039	0.0066	-0.0179	-0.0014	-0.0112	0.0042
	BM	0.0007	-0.0120	-0.0011	-0.0063	-0.0295	0.0227
		$MD_y;\ RR_y = 1.5$					
MI	DS	-0.0028	-0.0064	0.0017	-0.0017	-0.0369	0.0138
	BM	-0.0033	-0.0212	0.0135	-0.0022	-0.0507	0.0248
MW	DS	0.0003	0.0113	-0.0108	0.0033	-0.0180	-0.0024
	BM	-0.0031	-0.0077	0.0066	-0.0019	-0.0390	0.0172
		$MD_y;\ RR_y = 2$					
MI	DS	-0.0076	-0.0052	0.0108	0.0009	-0.0385	0.0098
	BM	-0.0082	-0.0190	0.0223	0.0003	-0.0522	0.0208
MW	DS	-0.0038	0.0060	-0.0002	0.0061	-0.0574	0.0097
	BM	-0.0073	-0.0129	0.0171	0.0010	-0.0783	0.0285
		$MD_{age};\ OR_{decade} = 1.1$					
MI	DS	-0.0020	-0.0180	0.0042	-0.0049	-0.0530	0.0249
	BM	-0.0024	-0.0326	0.0155	-0.0054	-0.0658	0.0344
MW	DS	0.0009	0.0036	-0.0100	-0.0010	-0.0209	0.0040
	BM	-0.0024	-0.0143	0.0066	-0.0061	-0.0383	0.0214
		$MD_{age};\ OR_{decade} = 1.2$					
MI	DS	-0.0006	-0.0372	0.0118	-0.0064	-0.0234	0.0186
	BM	-0.0010	-0.0513	0.0224	-0.0069	-0.0364	0.0290
MW	DS	0.0024	-0.0157	-0.0025	-0.0014	0.0101	-0.0044
	BM	-0.0011	-0.0328	0.0137	-0.0069	-0.0090	0.0152
		$MD_{CHF};\ RR_{CHF} = 1.5$					
MI	DS	-0.0063	-0.0058	0.0097	-0.0096	-0.0382	0.0280
	BM	-0.0067	-0.0195	0.0208	-0.0101	-0.0504	0.0379
MW	DS	-0.0033	0.0134	-0.0034	-0.0048	-0.0103	0.0073
	BM	-0.0064	-0.0036	0.0126	-0.0093	-0.0271	0.0247
		$MD_{CHF};\ RR_{CHF} = 2$					
MI	DS	-0.0138	-0.0276	0.0322	-0.0130	-0.0923	0.0623
	BM	-0.0142	-0.0399	0.0423	-0.0135	-0.1016	0.0693
MW	DS	-0.0104	-0.0105	0.0186	-0.0083	-0.0656	0.0401
	BM	-0.0133	-0.0255	0.0334	-0.0120	-0.0785	0.0539

Table 5.9: Monte Carlo means of the estimated standard errors of the relative risks ($\bar{s}_{\widehat{RR}}$). Each mean is based on 500 random samples. For each sample, data were removed from CHF according to the specified missing data mechanisms.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		$MCAR$					
MI	DS	0.0953	0.2051	0.1175	0.0969	0.2040	0.1205
	BM	0.1365	0.2478	0.1560	0.1377	0.2486	0.1575
MW	DS	0.0936	0.2114	0.1151	0.0943	0.2132	0.1159
	BM	0.1364	0.2500	0.1559	0.1374	0.2516	0.1573
		MD_y; $RR_y = 1.5$					
MI	DS	0.0965	0.2071	0.1186	0.0961	0.2035	0.1198
	BM	0.1376	0.2488	0.1568	0.1368	0.2466	0.1572
MW	DS	0.0947	0.2131	0.1161	0.0941	0.2122	0.1156
	BM	0.1377	0.2504	0.1570	0.1372	0.2481	0.1576
		MD_y; $RR_y = 2$					
MI	DS	0.0958	0.2071	0.1191	0.0967	0.2053	0.1198
	BM	0.1369	0.2486	0.1570	0.1370	0.2490	0.1568
MW	DS	0.0942	0.2125	0.1168	0.0953	0.2086	0.1174
	BM	0.1374	0.2492	0.1574	0.1380	0.2453	0.1597
		MD_{age}; $OR_{decade} = 1.1$					
MI	DS	0.0958	0.2053	0.1182	0.0963	0.2018	0.1207
	BM	0.1368	0.2479	0.1561	0.1372	0.2473	0.1580
MW	DS	0.0941	0.2118	0.1155	0.0932	0.2118	0.1160
	BM	0.1368	0.2503	0.1560	0.1369	0.2509	0.1578
		MD_{age}; $OR_{decade} = 1.2$					
MI	DS	0.0961	0.2031	0.1193	0.0972	0.2063	0.1211
	BM	0.1371	0.2468	0.1572	0.1381	0.2490	0.1586
MW	DS	0.0940	0.2095	0.1166	0.0944	0.2165	0.1160
	BM	0.1370	0.2491	0.1571	0.1380	0.2523	0.1585
		MD_{CHF}; $RR_{CHF} = 1.5$					
MI	DS	0.0964	0.2054	0.1196	0.0966	0.2012	0.1223
	BM	0.1377	0.2472	0.1579	0.1379	0.2451	0.1599
MW	DS	0.0948	0.2114	0.1173	0.0948	0.2102	0.1182
	BM	0.1377	0.2491	0.1578	0.1381	0.2480	0.1597
		MD_{CHF}; $RR_{CHF} = 2$					
MI	DS	0.0961	0.2031	0.1193	0.0959	0.1916	0.1251
	BM	0.1371	0.2468	0.1572	0.1375	0.2394	0.1623
MW	DS	0.0940	0.2095	0.1166	0.0948	0.2001	0.1217
	BM	0.1370	0.2491	0.1571	0.1376	0.2427	0.1614

Table 5.10: Monte Carlo standard deviations of the estimated relative risks. Each condition is based on 500 random samples. For each sample, data was deleted from the variable CHF according to the specified missing data mechanisms.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		MCAR					
MI	DS	0.0980	0.2198	0.1241	0.0921	0.2103	0.1167
	BM	0.0979	0.2198	0.1203	0.0920	0.2104	0.1131
MW	DS	0.0992	0.2247	0.1244	0.0940	0.2193	0.1163
	BM	0.0977	0.2218	0.1193	0.0917	0.2143	0.1106
		MD_y; $RR_y = 1.5$					
MI	DS	0.0999	0.2081	0.1241	0.0949	0.1938	0.1244
	BM	0.0997	0.2083	0.1202	0.0946	0.1939	0.1201
MW	DS	0.1009	0.2136	0.1234	0.0970	0.1998	0.1246
	BM	0.0992	0.2103	0.1182	0.0943	0.1936	0.1174
		MD_y; $RR_y = 2$					
MI	DS	0.0995	0.2049	0.1234	0.1003	0.2067	0.1211
	BM	0.0994	0.2056	0.1194	0.1001	0.2072	0.1176
MW	DS	0.1010	0.2095	0.1236	0.1019	0.2075	0.1222
	BM	0.0993	0.2058	0.1178	0.0992	0.1988	0.1157
		MD_{age}; $OR_{decade} = 1.1$					
MI	DS	0.0925	0.1995	0.1182	0.0931	0.1950	0.1207
	BM	0.0925	0.2001	0.1145	0.0930	0.1951	0.1169
MW	DS	0.0936	0.2044	0.1171	0.0946	0.2013	0.1203
	BM	0.0920	0.2026	0.1122	0.0929	0.1987	0.1151
		MD_{age}; $OR_{decade} = 1.2$					
MI	DS	0.0980	0.2119	0.1277	0.1009	0.2060	0.1215
	BM	0.0976	0.2127	0.1243	0.1008	0.2071	0.1179
MW	DS	0.0998	0.2214	0.1266	0.1028	0.2141	0.1201
	BM	0.0970	0.2174	0.1210	0.0999	0.2100	0.1142
		MD_{CHF}; $RR_{CHF} = 1.5$					
MI	DS	0.1027	0.2022	0.1241	0.0963	0.1873	0.1227
	BM	0.1025	0.2027	0.1202	0.0960	0.1878	0.1191
MW	DS	0.1038	0.2076	0.1234	0.0976	0.1957	0.1217
	BM	0.1020	0.2057	0.1183	0.0952	0.1919	0.1161
		MD_{CHF}; $RR_{CHF} = 2$					
MI	DS	0.0930	0.2055	0.1179	0.0959	0.1928	0.1258
	BM	0.0928	0.2065	0.1143	0.0958	0.1931	0.1222
MW	DS	0.0938	0.2100	0.1174	0.0980	0.1995	0.1254
	BM	0.0924	0.2082	0.1126	0.0961	0.1971	0.1194

Table 5.11: Bias in the standard error estimates. The bias was obtained by dividing the mean of the standard errors for each condition by the standard deviation of the estimated relative risks obtained for the condition, or $\bar{s}_{\widehat{RR}}/\sigma_{\widehat{RR}}$.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		MCAR					
MI	DS	0.973	0.933	0.946	1.052	0.970	1.033
	BM	1.393	1.127	1.296	1.496	1.181	1.393
MW	DS	0.943	0.941	0.926	1.003	0.972	0.997
	BM	1.396	1.127	1.307	1.498	1.174	1.422
		MD_y; $RR_y = 1.5$					
MI	DS	0.966	0.995	0.955	1.012	1.050	0.963
	BM	1.380	1.194	1.304	1.446	1.272	1.308
MW	DS	0.939	0.998	0.941	0.970	1.062	0.928
	BM	1.389	1.190	1.328	1.455	1.281	1.343
		MD_y; $RR_y = 2$					
MI	DS	0.963	1.011	0.965	0.964	0.993	0.989
	BM	1.377	1.209	1.314	1.369	1.202	1.333
MW	DS	0.933	1.014	0.945	0.935	1.006	0.961
	BM	1.384	1.211	1.336	1.391	1.234	1.380
		MD_{age}; $OR_{decade} = 1.1$					
MI	DS	1.036	1.029	1.000	0.983	0.953	0.945
	BM	1.479	1.239	1.363	1.405	1.163	1.271
MW	DS	1.006	1.036	0.987	0.934	0.956	0.916
	BM	1.488	1.235	1.390	1.412	1.154	1.304
		MD_{age}; $OR_{decade} = 1.2$					
MI	DS	1.032	1.042	0.988	0.963	1.001	0.996
	BM	1.474	1.265	1.344	1.370	1.202	1.346
MW	DS	0.994	1.041	0.969	0.918	1.011	0.966
	BM	1.474	1.254	1.364	1.381	1.201	1.388
		MD_{CHF}; $RR_{CHF} = 1.5$					
MI	DS	0.938	1.016	0.964	1.004	1.075	0.997
	BM	1.342	1.219	1.313	1.437	1.305	1.342
MW	DS	0.914	1.018	0.950	0.971	1.074	0.971
	BM	1.350	1.211	1.334	1.451	1.292	1.375
		MD_{CHF}; $RR_{CHF} = 2$					
MI	DS	1.028	0.975	1.027	1.000	0.994	0.994
	BM	1.472	1.182	1.384	1.435	1.240	1.328
MW	DS	1.005	0.981	1.011	0.967	1.003	0.971
	BM	1.482	1.180	1.403	1.431	1.231	1.352

Table 5.12: Monte Carlo coverage probabilities for relative risks. The probabilities are the proportion of the 500 95% confidence intervals $\widehat{RR} \pm 1.96 s_{\widehat{RR}}$ containing RR.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		$MCAR$					
MI	DS	0.948	0.906	0.934	0.954	0.926	0.954
	BM	0.992	0.972	0.992	0.992	0.978	0.996
MW	DS	0.940	0.918	0.930	0.944	0.936	0.944
	BM	0.992	0.972	0.990	0.994	0.976	0.996
		$MD_y;\ RR_y = 1.5$					
MI	DS	0.942	0.932	0.930	0.950	0.938	0.940
	BM	0.998	0.988	0.994	0.998	0.988	0.986
MW	DS	0.938	0.932	0.920	0.940	0.938	0.928
	BM	0.998	0.990	0.994	0.998	0.984	0.988
		$MD_y;\ RR_y = 2$					
MI	DS	0.934	0.940	0.938	0.940	0.942	0.948
	BM	0.998	0.986	0.984	0.996	0.984	0.992
MW	DS	0.936	0.942	0.936	0.938	0.938	0.942
	BM	0.998	0.988	0.986	0.998	0.984	0.996
		$MD_{age};\ OR_{decade} = 1.1$					
MI	DS	0.972	0.934	0.954	0.934	0.910	0.924
	BM	0.998	0.992	0.994	0.994	0.976	0.992
MW	DS	0.962	0.948	0.946	0.930	0.926	0.926
	BM	0.998	0.990	1.000	0.994	0.980	0.994
		$MD_{age};\ OR_{decade} = 1.2$					
MI	DS	0.958	0.942	0.952	0.926	0.934	0.944
	BM	1.000	0.994	0.996	0.988	0.982	0.994
MW	DS	0.950	0.952	0.944	0.920	0.946	0.926
	BM	0.998	0.994	0.996	0.990	0.978	0.998
		$MD_{CHF};\ RR_{CHF} = 1.5$					
MI	DS	0.930	0.938	0.938	0.944	0.940	0.948
	BM	0.988	0.988	0.990	0.994	0.992	0.992
MW	DS	0.922	0.944	0.932	0.940	0.946	0.934
	BM	0.994	0.986	0.988	0.996	0.990	0.992
		$MD_{CHF};\ RR_{CHF} = 2$					
MI	DS	0.952	0.930	0.942	0.948	0.892	0.940
	BM	0.998	0.982	0.996	1.000	0.984	0.988
MW	DS	0.954	0.926	0.938	0.940	0.910	0.950
	BM	0.998	0.980	0.998	1.000	0.984	0.992

Table 5.13: Efficiency of the relative risk estimates, obtained by dividing the MSE of the complete data relative risks by the MSE the relative risks obtained using missing data methods.

		25% Missing			40% Missing		
		h_1	h_2	h_3	h_1	h_2	h_3
		$MCAR$					
MI	DS	0.965	0.928	0.901	1.085	0.989	0.997
	BM	0.966	0.920	0.894	1.088	0.971	0.965
MW	DS	0.938	0.889	0.888	1.048	0.933	1.027
	BM	0.970	0.914	0.910	1.094	0.968	1.035
		MD_y; $RR_y = 1.5$					
MI	DS	0.927	1.037	0.902	1.028	1.167	0.892
	BM	0.930	1.031	0.890	1.034	1.143	0.878
MW	DS	0.911	0.983	0.910	0.982	1.119	0.895
	BM	0.941	1.017	0.926	1.042	1.168	0.929
		MD_y; $RR_y = 2$					
MI	DS	0.927	1.070	0.908	0.921	1.026	0.944
	BM	0.928	1.059	0.892	0.926	1.003	0.921
MW	DS	0.906	1.024	0.909	0.887	0.989	0.927
	BM	0.933	1.061	0.923	0.942	1.025	0.938
		MD_{age}; $OR_{decade} = 1.1$					
MI	DS	1.082	1.123	0.994	0.961	0.958	0.835
	BM	1.083	1.104	0.978	0.967	0.932	0.806
MW	DS	1.058	1.075	1.010	0.930	0.911	0.867
	BM	1.095	1.094	1.027	0.980	0.932	0.870
		MD_{age}; $OR_{decade} = 1.2$					
MI	DS	1.068	1.152	0.949	0.904	1.049	0.930
	BM	1.071	1.128	0.929	0.906	1.027	0.903
MW	DS	1.034	1.105	0.959	0.876	0.979	0.962
	BM	1.073	1.119	0.970	0.921	1.020	0.984
		MD_{CHF}; $RR_{CHF} = 1.5$					
MI	DS	0.873	1.099	0.899	0.984	1.245	0.898
	BM	0.875	1.089	0.882	0.988	1.215	0.866
MW	DS	0.859	1.040	0.912	0.968	1.172	0.937
	BM	0.885	1.065	0.921	1.009	1.206	0.938
		MD_{CHF}; $RR_{CHF} = 2$					
MI	DS	1.035	1.051	0.961	0.978	1.038	0.776
	BM	1.037	1.029	0.924	0.980	1.007	0.740
MW	DS	1.033	1.017	0.995	0.953	1.048	0.838
	BM	1.052	1.028	0.976	0.979	1.041	0.819

Table 5.14: Improved Monte Carlo simulation for the MD_{CHF}; $RR_{CHF} = 1.5$ condition with 40% missing observations in the CHF variable. Parameters are based on 2000 samples.

		h_1	h_2	h_3
		$\mu_{\widehat{RR}}$		
MI	DS	1.242	0.8015	0.7540
	BM	1.241	0.7912	0.7611
MW	DS	1.248	0.8237	0.7388
	BM	1.243	0.8096	0.7515
		Relative Bias		
MI	DS	-0.0065	-0.0571	0.0319
	BM	-0.0069	-0.0692	0.0417
MW	DS	-0.0013	-0.0309	0.0111
	BM	-0.0058	-0.0475	0.0285
		$\bar{s}_{\widehat{RR}}$		
MI	DS	0.0965	0.1986	0.1222
	BM	0.1376	0.2443	0.1594
MW	DS	0.0947	0.2075	0.1181
	BM	0.1378	0.2472	0.1592
		$\sigma_{\widehat{RR}}$		
MI	DS	0.0956	0.2008	0.1206
	BM	0.0955	0.2014	0.1169
MW	DS	0.0977	0.2081	0.1199
	BM	0.0953	0.2045	0.1141
		$\bar{s}_{\widehat{RR}}/\sigma_{\widehat{RR}}$		
MI	DS	1.009	0.989	1.013
	BM	1.441	1.213	1.363
MW	DS	0.969	0.997	0.985
	BM	1.446	1.209	1.396
		95% CP		
MI	DS	0.9445	0.9185	0.9500
	BM	0.9940	0.9795	0.9890
MW	DS	0.9370	0.9255	0.9445
	BM	0.9940	0.9790	0.9925
		Efficiency		
MI	DS	1.005	1.053	0.921
	BM	1.008	1.023	0.887
MW	DS	0.971	1.022	0.962
	BM	1.016	1.037	0.964

Table 5.15: Simulation results for the regression parameter associated with the binary risk factor CHF.

		25% Missing				40% Missing			
		β	SE	CP	λ	β	SE	CP	λ
		$MCAR$							
MI	DS	1.71	0.273	0.920	0.280	1.72	0.305	0.890	0.409
	BM	1.67	0.268	0.926	0.279	1.68	0.299	0.882	0.405
MW	DS	1.70	0.268	0.952	-	1.70	0.300	0.946	-
	BM	1.64	0.261	0.944	-	1.63	0.291	0.936	-
		MD_y; $RR_y = 1.5$							
MI	DS	1.70	0.290	0.898	0.351	1.66	0.345	0.882	0.513
	BM	1.66	0.284	0.896	0.346	1.62	0.339	0.868	0.514
MW	DS	1.66	0.289	0.944	-	1.56	0.350	0.946	-
	BM	1.59	0.282	0.940	-	1.47	0.338	0.924	-
		MD_y; $RR_y = 2$							
MI	DS	1.70	0.320	0.864	0.450	1.73	0.447	0.776	0.659
	BM	1.66	0.313	0.868	0.446	1.68	0.439	0.774	0.658
MW	DS	1.62	0.313	0.946	-	1.32	0.439	0.854	-
	BM	1.54	0.305	0.930	-	1.21	0.423	0.788	-
		MD_{age}; $OR_{decade} = 1.1$							
MI	DS	1.71	0.279	0.952	0.310	1.71	0.308	0.914	0.411
	BM	1.67	0.273	0.946	0.305	1.68	0.303	0.904	0.409
MW	DS	1.70	0.271	0.964	-	1.69	0.302	0.968	-
	BM	1.64	0.264	0.958	-	1.63	0.294	0.966	-
		MD_{age}; $OR_{decade} = 1.2$							
MI	DS	1.72	0.275	0.946	0.287	1.71	0.313	0.906	0.432
	BM	1.68	0.270	0.936	0.285	1.67	0.307	0.914	0.430
MW	DS	1.71	0.272	0.966	-	1.69	0.310	0.956	-
	BM	1.65	0.266	0.970	-	1.62	0.301	0.950	-
		MD_{CHF}; $RR_{CHF} = 1.5$							
MI	DS	1.69	0.288	0.906	0.289	1.69	0.334	0.886	0.378
	BM	1.65	0.283	0.894	0.288	1.66	0.328	0.886	0.375
MW	DS	1.68	0.284	0.936	-	1.68	0.337	0.956	-
	BM	1.62	0.278	0.932	-	1.61	0.328	0.940	-
		MD_{CHF}; $RR_{CHF} = 2$							
MI	DS	1.68	0.303	0.914	0.301	1.55	0.402	0.900	0.323
	BM	1.65	0.298	0.910	0.298	1.53	0.397	0.886	0.323
MW	DS	1.67	0.299	0.944	-	1.60	0.404	0.952	-
	BM	1.62	0.292	0.936	-	1.54	0.394	0.950	-

were misspecified. The standard error estimates were similar across all conditions, and the coverage probabilities were reasonable, although they had a tendency to be too narrow. The indirectly standardized (BM) estimates did not fare as well. Biases in the means of the estimated relative risks were generally larger than those observed when DS was employed. The biases in the standard errors of the coefficients were large, with the means of the standard errors consistently being larger than the standard deviations of the estimated relative risks. Further, the large coverage probabilities indicated that the the confidence intervals performed poorly. The overestimation of the standard errors and the poor performance of the coverage probabilities indicate that the method typically used for calculating the variance of the BM relative risks performs poorly. Reasons for this poor performance are discussed and explored in sections 5.1.11 and 5.2 below, and an alternative formula for calculating the variance is derived in appendix D.

The simulations also indicate, however, that caution should be used in interpreting individual regression coefficients, especially for variables with a large proportion of missing observations. The CHF risk factor demonstrated considerable bias in several of the conditions, most notably for EMMW when the missing data mechanism depended on the outcome. The coverage probabilities for the CHF coefficients deteriorated in this condition.

When using EM by method of weights, a missing data method such as the EM algorithm is used to estimate the joint distribution of the covariates. Information from y is not employed in this estimation process and if some of the covariates are MD_y, the exclusion of y from this joint distribution will make it likely that the data are NMAR (see section 3.1.1 in chapter 3). While the conditional distribution of y

given the covariates is used when obtaining the weights in the EMMW algorithm, it appears that this does not provide sufficient information to allow unbiased estimates of coefficients associated with covariates which are MD_y. Caution is therefore advised in situations where EMMW is employed to perform logistic regression when covariates are MD_y.

The coverage probabilities for the CHF variable were poor when MI was employed. The most likely explanation for this poor performance is the misspecification of the joint distribution of the variables which formed the basis for the imputations. In the present case, a MVN distribution was applied to data in which all variables but one were binary. Further, in some of these variables, positive outcomes were rare (5% in the case of y and 12% in the case of CHF). Other authors have demonstrated acceptable performance of MI in the face of misspecification of the MVN model (Schafer, 1997a; Rubin, 1987c; Greenland and Finkle, 1995). As noted by Schafer (1997a), however, this misspecification will have minimal impact if the binary variables are completely observed, as this ensures that the imputations are made conditionally on these variables. Greenland and Finkle (1995) also found a degree of bias in binary variables when the proportion of positive responses was small. In general, where more appropriate models for the joint distributions exist, it would appear to be preferable to use them, especially if the misspecified variables are not completely observed.

5.1.11 Variance of Baseline Model Estimates

The tendency of the BM variance estimates to be too large warrants further consideration, especially since the BM adjusted estimates and variances are encountered

frequently in the risk adjustment literature. When employing the BM, the variance estimate typically employed for the SMR is

$$\begin{aligned} \text{Var}(SMR) &= \text{Var}\left(\frac{O_k}{E_k}\right) && (5.11) \\ &= \frac{1}{E^2}\sum_{i\in h_k}\text{Var}(y_i). && (5.12) \end{aligned}$$

In this expression, the probabilities used in $E_k = \sum_{i\in h_k}\hat{p}_i$ are treated as fixed. When treating the $\hat{p}_i$'s as fixed, the implicit assumption is that any variance in the $\hat{p}_i$'s will be negligible when compared to the variances of the y_i's used to obtain $O_k = \sum_{i\in h_k} y_i$. As in the case of the FM adjusted estimates, however, one can treat the $\hat{p}_i$'s as random and employ the variance of a first order Taylor Series expansion of the (O_k/E_k) ratio. The resulting estimate will be referred to as the delta method estimate. As in A.11 of appendix A, the variance of the SMR can be estimated as

$$\widehat{\text{Var}}(SMR_k) = \frac{1}{E_k^2}\left[\left(\frac{O_k}{E_k}\right)^2\widehat{\text{Var}}(E_k) + \widehat{\text{Var}}(O_k) - 2\left(\frac{O_k}{E_k}\right)\widehat{\text{Cov}}(O_k, E_k)\right].$$

Details of the expressions for the variances in 5.13 can be found in appendix D. Using the delta method, the estimates for $\text{Var}(E_k)$ and $\text{Cov}(O_k, E_k)$ are the same (see D.10 and D.12 in appendix D), and 5.13 can be re-expressed as

$$\widehat{\text{Var}}(SMR_k) = \frac{1}{E_k^2}\left[(SMR_k(SMR_k - 2))\,\widehat{\text{Var}}(E_k) + \widehat{\text{Var}}(O_k)\right]. \quad (5.13)$$

A method for computing an estimate of the delta method approximation for $\text{Var}(SMR_k)$ is presented in appendix D. By subtracting 5.13 from 5.12, a first-order asymptotic estimate of the the additive bias in 5.12 is

$$\hat{B}_k = \frac{\widehat{\text{Var}}(E_k)}{E_k^2} \times (SMR_k(2 - SMR_k)). \quad (5.14)$$

From this expression the following statements can be made about the usual variance calculation employed when using the BM.

1. If SMR_k =0 or SMR_k=2, then $\hat{B}_k = 0$ and the two variance estimates 5.12 and 5.13 will be identical.

2. If 0 < SMR < 2, then $\hat{B}_k > 0$ and the usual BM variance estimate will asymptotically overestimate the true variance. For constant E_k and $\widehat{\text{Var}}(E_k)$, the degree of overestimation will be greatest when SMR_k=1.

3. If SMR > 2, then $\hat{B}_k < 0$ and the usual variance estimate will underestimate the true variance. For fixed E_k and Var(E_k), the increase in this underestimation will be quadratic with respect to increases in SMR_k beyond 2.

Note that it is unlikely that the E_k and $\widehat{\text{Var}}(E_k)$ would remain constant while SMR_k varies, and the relative sizes of these quantities will affect the degree of over or underestimation. Monte Carlo simulations were therefore performed to determine the degree of bias in the standard error estimates that might be expected when the BM method is applied to the APPROACH data. These simulations also examined the performance of the DS method as well as standard errors obtained using the delta method to estimate the BM standard errors. These standard errors will be denoted as BM_Δ.

5.2 Standard Error Simulations

5.2.1 Generating the Random Samples

The simulations were conducted to determine if the problems in estimation were related to the probability of a positive response or to the sample size. For each condition, 2000 random samples were generated. Simulations were based on complete data; no observations were deleted. For the first set of simulations, the sample size was fixed at N=2000 cases, and the expected response rates set at 5%, 10%, 25%, and 50%. The methods used to generate the samples were identical to those used in the missing data simulations (see 5.1.3). However, the parameters used to generate the the samples were based on the simulations in which there was modest confounding. The parameters for this model were obtained by using an EM algorithm to obtain a mixed continuous and categorical model for the covariates in the APPROACH data. The mean ages and probabilities by joint category are presented in table 5.16. The coefficients used for age and CHF are the same as those in the previous simulations (see 5.2) except that the intercepts in the logistic models were altered to obtain the desired expected response rates for y. The accuracy of these response rates was verified during the simulations. In the second set of simulations, the expected response rate was fixed at 5% and the sample sizes for the two conditions set at N=4000 and N=6000 cases.

5.2.2 Results

The results are presented in tables 5.2.2 through 5.2.2.

Table 5.16: Distribution of covariates for the standard error simulations. The means, variance and probabilities are all based on the observed data in the APPROACH database. **a)** Marginal and joint distributions of age across the CHF and treatment categories. Within each joint category $c = 1, \ldots, 6$, age $\sim N(\mu_c, 125)$. **b)** Marginal and joint probability distributions for treatment and CHF.

a) Mean age by category

		CHF		
		No	Yes	
Treatment	h_1	61	66	61.5
	h_2	63	68	63.5
	h_3	61	66	61.5
		61.4	66.4	61.9

b) Joint distribution of treatment and CHF

		CHF		
		No	Yes	
Treatment	h_1	.45	.05	.5
	h_2	.18	.02	.2
	h_3	.27	.03	.3
		.9	.1	

Point estimates. For completeness, the point estimates for the $\widehat{RR}_k$ are presented for the BM and DS methods of standardization (see table 5.2.2). There is little variability in these estimates across the conditions, and all are similar to the corresponding true relative risks.

Standard Errors. The means of the standard errors of the $\widehat{RR}_k$ and the standard deviations of the $\widehat{RR}_k$ are presented in table 5.18. As expected, the standard errors obtained using the typical BM method are a good deal larger than those obtained using the delta method. For all methods, the mean standard errors decrease as the expected probability of response increases. The standard errors also decrease as the sample size increases. The mean standard errors obtained using the BM_Δ and DS methods are comparable to the standard deviations of the $\hat{RR}_k$s. The BM mean estimates are always larger than their corresponding standard deviations. The relationships between the mean standard errors and the standard deviations of the $\hat{RR}_k$s was examined by taking the ratio of these quantities. The ratios are presented in table 5.19, and can be used as an indication of the biases in the standard errors.

The ratios for the BM method are generally much greater than 1, with the lone exception occurring in the condition with an expected response rate of 50%. The mean standard errors for the BM_Δ and DS methods fare much better. In general, the BM_Δ estimates appear to overestimate the standard deviation of the relative risks for provider h_3. The DS estimates tend to slightly underestimate the corresponding standard deviations.

Coverage Probabilities. The patterns observed in the mean standard errors are reflected in the coverage probabilities. The 95% CPs are presented in table 5.2.2.

Based on a normal approximation ($.95 \pm 1.96\sqrt{(.95)(.05)/2000}$), 95% of the coverage probabilities would be expected to fall in the interval [$.94 < CP < .96$]. None of the CPs from the BM method fall within this range. In the first set of simulations, 58% (7/12) of the BM_Δ CPs and 58% of the DS CPs fall within this range. In the set of simulations with larger sample sizes, 3/6 of the BM_Δ CPs fall between .94 and .96, while 4/6 of the DS CPs fall within this range. All but one of the DS coverage probabilities are $< .95$.

Table 5.17: Mean Monte Carlo relative risks. For each condition in means are based on 2000 randomly generated samples. Probabilities of death in a) are 5%, 10%, 25% and 50%, and all samples have N=2000 cases. In b) rates of death are 5%, and the samples are of size N=4000 and N=6000

a)

		$\mu_{\widehat{RR}}$			RR		
Rate	Method	h_1	h_2	h_3	h_1	h_2	h_3
5%	BM	1.25	0.849	0.699	1.25	0.85	0.692
	DS	1.25	0.847	0.700	"	"	"
10%	BM	1.25	0.849	0.695	1.25	0.85	0.692
	DS	1.25	0.847	0.696	"	"	"
25%	BM	1.25	0.851	0.691	1.25	0.85	0.690
	DS	1.25	0.847	0.692	"	"	"
50%	BM	1.24	0.858	0.693	1.25	0.85	0.688
	DS	1.24	0.852	0.696	"	"	"

b)

		$\mu_{\widehat{RR}}$			RR		
N	Method	h_1	h_2	h_3	h_1	h_2	h_3
4000	BM	1.25	0.852	0.691	1.250	0.850	0.692
	DS	1.25	0.850	0.692	"	"	"
6000	BM	1.25	0.849	0.692	1.250	0.850	0.692
	DS	1.25	0.848	0.693	"	"	"

5.2.3 Discussion

Tables 5.18 through 5.2.2 demonstrate that the standard errors obtained using the DS method are superior to those of the BM_{Δ} and BM methods. In the conditions with larger sample sizes, the DS estimates appear to work very well, although the coverage probabilities are generally smaller than .95. Consequently, the DS estimates may be a prudent choice for use in risk-adjustment, where sample sizes tend to be large. The good performance of the DS standard errors provides more evidence of the utility of direct standardization. Directly standardized estimates and standard errors are easily obtained (see A.2 in Appendix A) and also appear to yield the most stable results across missing data methods (see 4.2.1 in chapter 4). Directly standardized estimates should be seriously considered by anyone performing risk adjustment studies both in the presence of missing data or when the data are complete.

While note as good as the DS estimates in the conditions with sample sizes of N=4000 and N=6000, the BM_{Δ} standard errors are clearly superior to the BM standard errors and the performance of confidence intervals based on BM_{Δ} standard errors is comparable to those based on direct standardization. The simulation results indicate that the BM method of obtaining standard errors should be discouraged. The BM_{Δ} standard errors can be obtained without any computer programming, using the variance covariance matrix for the regression coefficients and the fitted probabilities that can be produced when performing logistic regression with most statistical packages.

Table 5.18: Means of the standard errors of the relative risks ($\bar{s}_{\widehat{RR}}$) and standard deviation of the Monte Carlo estimates of relative risk ($\sigma_{\widehat{RR}}$). In a) all relative risks and standard errors are based on random samples with N=2000 cases. Rates of death in a) are 5%, 10%, 25%, and 50%. In b) the rate of death is 5%, and relative risks and standard errors are based on random samples with N=4000 and N=6000 cases.

a)

		$\bar{s}_{\widehat{RR}}$			$\sigma_{\widehat{RR}}$		
Rate	Method	h_1	h_2	h_3	h_1	h_2	h_3
5%	BM	0.1350	0.204	0.175	0.0966	0.173	0.131
	BM_Δ	0.0996	0.182	0.150	"	"	"
	DS	0.0936	0.170	0.130	0.0969	0.174	0.131
10%	BM	0.0910	0.138	0.1180	0.0637	0.116	0.091
	BM_Δ	0.0668	0.123	0.1010	"	"	"
	DS	0.0631	0.117	0.0884	0.0637	0.117	0.091
25%	BM	0.0518	0.0787	0.0670	0.0375	0.0679	0.0527
	BM_Δ	0.0379	0.0701	0.0573	"	"	"
	DS	0.0358	0.0676	0.0511	0.0373	0.0692	0.0527
50%	BM	0.0305	0.0463	0.0394	0.0216	0.0401	0.0330
	BM_Δ	0.0222	0.0414	0.0336	"	"	"
	DS	0.0211	0.0416	0.0315	0.0214	0.0413	0.0330

b)

		$\bar{s}_{\widehat{RR}}$			$\sigma_{\widehat{RR}}$		
N	Method	h_1	h_2	h_3	h_1	h_2	h_3
4000	BM	0.0951	0.143	0.1230	0.0659	0.122	0.0924
	BM_Δ	0.0702	0.128	0.1060	"	"	"
	DS	0.0661	0.120	0.0912	0.0660	0.123	0.0924
6000	BM	0.0777	0.1170	0.1000	0.0549	0.101	0.0761
	BM_Δ	0.0573	0.1040	0.0861	"	"	"
	DS	0.0540	0.0983	0.0745	0.0549	0.102	0.0760

Table 5.19: Ratios of the mean of the estimates of the standard errors of the $\widehat{RR}$'s to the Monte Carlo standard deviation of the $\widehat{RR}$'s. Each of the $\widehat{RR}$'s and $\sigma_{\widehat{RR}}$'s is based on 2000 random samples. Probabilities of death in a) are 5%, 10%, 25% and 50%, and all samples have N=2000 cases. In b) rates of death are 5%, and the samples are of size N=4000 and N=6000.

a)

		$\bar{s}_{\widehat{RR}}/\sigma_{\widehat{RR}}$		
Rate	Method	h_1	h_2	h_3
5%	BM	1.400	1.180	1.340
	BM_Δ	1.030	1.050	1.150
	DS	0.966	0.976	0.993
10%	BM	1.43	1.190	1.290
	BM_Δ	1.05	1.060	1.110
	DS	0.99	0.995	0.972
25%	BM	1.380	1.160	1.270
	BM_Δ	1.010	1.030	1.090
	DS	0.959	0.977	0.969
50%	BM	1.410	1.16	1.190
	BM_Δ	1.030	1.03	1.020
	DS	0.987	1.01	0.956

b)

		$\bar{s}_{\widehat{RR}}/\sigma_{\widehat{RR}}$		
N	Method	h_1	h_2	h_3
4000	BM	1.44	1.180	1.330
	BM_Δ	1.07	1.050	1.140
	DS	1.00	0.976	0.986
6000	BM	1.410	1.160	1.32
	BM_Δ	1.040	1.030	1.13
	DS	0.983	0.959	0.98

Table 5.20: 95 % coverage probabilities for the estimated relative risks. These are the proportion of intervals $\widehat{RR} \pm 1.96 \times s_{\widehat{RR}}$ which contain the mean of the Monte Carlo estimates of relative risk ($\mu_{\widehat{RR}}$). Each coverage probabilities are based on 2000 random samples. Probabilities of death in a) are 5%, 10%, 25% and 50%, and all samples have N=2000 cases. In b) rates of death are 5%, and the samples are of size N=4000 and N=6000.

a)

		95% CP		
Rate	Method	h_1	h_2	h_3
5%	BM	0.996	0.977	0.992
	BM_Δ	0.957	0.961	0.979
	DS	0.939	0.937	0.945
10%	BM	0.996	0.979	0.985
	BM_Δ	0.960	0.964	0.971
	DS	0.945	0.946	0.942
25%	BM	0.993	0.976	0.990
	BM_Δ	0.947	0.960	0.970
	DS	0.939	0.940	0.946
50%	BM	0.990	0.971	0.982
	BM_Δ	0.947	0.953	0.951
	DS	0.926	0.953	0.927

b)

		95% CP		
N	Method	h_1	h_2	h_3
4000	BM	0.996	0.977	0.990
	BM_Δ	0.967	0.959	0.975
	DS	0.949	0.938	0.947
6000	BM	0.995	0.979	0.989
	BM_Δ	0.955	0.954	0.973
	DS	0.944	0.936	0.941

Chapter 6

Summary and Conclusions

Risk adjustment procedures are used in cases where researchers wish to compare the quality of treatment afforded to patients by different physicians, procedures or hospitals. This dissertation addressed methods of dealing with missing covariate information when performing risk adjustment with binary outcomes. These methods were explored using 1995/96 data from the Alberta Provincial Program for Outcome Assessment in Coronary Heart Disease (APPROACH) initiative. Norris et al. (1999) had previously accounted for missing data in this database by augmenting the data with diagnoses based on administrative discharge data. For some of the variables, such as ejection fraction (EF), there was no administrative equivalent, and extra categories were used to account for missing observations. This type of procedure does not use available information to account for possible values for the missing observations, and can result in bias in the estimation of logistic regression coefficients (Vach, 1994; Vach and Blettner, 1991). Multiple imputation (MI) and likelihood based methods can account for missing data in logistic regression, but these methods have not been applied to risk adjustment procedures. The purpose of this dissertation was to examine the use of missing data methods applied to risk adjustment procedures.

Chapter 2 described risk adjustment procedures in detail. The underlying rationale for risk adjustment methods was examined. Risk adjustment procedures typically use indirect standardization, where the number of deaths following treat-

ment by a provider is compared to the number which would be expected on the basis of underlying patient risks. An example of an indirectly standardized measure is the standardized mortality ratio, which is the ratio of observed to expected deaths for a given provider. The most commonly used methods of obtaining the expected number of deaths are baseline model (BM) estimates. Baseline models do not adjust the risks for effects associated with the providers. Methods of indirect standardization which adjust for provider effects when calculating individual patient risks will be referred to as full model (FM) estimates. A case was made for the utility of directly standardized (DS) measures. In direct standardization, the observed number of deaths in a standard population is compared with the number of deaths expected to occur in the population if all cases in the population had been treated by a given provider. Methods for obtaining variance estimates were described briefly in this chapter and are presented in greater detail in **appendix A**. An illustration was used to point out potential weaknesses in BM adjusted measures.

Chapter 3 examined missing data methods. Criteria were presented to determine whether the potential missing data methods are appropriate for use with risk adjustment. To be useful, missing data methods must 1) provide estimates of variance for the risk adjusted point estimates; 2) be capable of working with large and complex data sets; 3) be able to employ rare risk factors and outcomes; and 4) be available for use by researchers. Multiple imputation (MI) and expectation-maximization by the method of weights (EMMW) both satisfied these criteria. These methods showed some weaknesses with respect to points 2) and 4). Both methods require the specification of a joint distribution for the covariates, and this can be problematic with large and complex data sets. Suitable implementations, while

available from researchers or from the World Wide Web, do not come packaged with standard statistical software. Theoretical and practical concerns were addressed for these methods. Procedures for obtaining risk adjusted point estimates and corresponding variance estimates were presented in this chapter as well as in **appendices B and C**.

In **chapter 4**, the MI and EMMW methods described in chapter 3 were applied to a subset of the APPROACH data. For each of these methods, BM, FM, and DS estimates were obtained. The risk adjustment methods were used to examine the effectiveness of the type of treatment used for the cardiac patients. There were three possible treatments: medical treatment, coronary artery bypass graft (CABG), and percutaneous transluminal coronary angioplasty (PTCA). A multivariate normal model (MVN) was used for the joint distribution of variables to generate multiple imputations. For EMMW, MVN and mixed continuous and categorical models were used to model the joint distributions of covariates. Attempts were made to assess the goodness-of-fit of the logistic regressions, as well as to diagnose the adequacy of the missing data methods. Diagnostic tables of Pearson residuals indicated that there were some problems with the missing data methods. The most plausible explanations for these problems were 1) that models for the joint distributions of the variables were misspecified, and 2) that assumptions regarding the patterns of missing observations had been violated. However, other diagnostic measures, such as the residual deviance, indicated that the logistic fits worked well in comparison to the models developed by (Norris et al., 1999). Risk adjusted point estimates were similar across missing data methods and methods of risk adjustment. Directly standardized measures had the smallest variances, and the variance estimates of the DS and BM

measures tended to be the most stable across missing data models.

In **chapter 5**, computer simulations were employed to explore the performance of the missing data methods across a variety of conditions. The parameters of the probability model used to generate the random samples were based on the distribution of variables in the APPROACH data set. For practical reasons, simulations were based on a subset of the APPROACH variables. The variables employed were 1) congestive heart failure (CHF), a rare binary risk factor with a high proportion of missing data; 2) age, a completely observed continuous variable, 3) treatment, a categorical variable denoting the treatment providers which were to be examined using risk adjustment procedures; and 4) 6-month mortality, which was used as the outcome variable. For each simulation condition, 500 samples of 2000 cases were generated. In each sample, observations in the CHF variable were deleted on the basis of a given missing data mechanism. The missing data could be missing completely at random (MCAR), missing dependent on the outcome (MD_y), missing dependent on age (MD_{age}), or missing dependent on CHF (MD_{CHF}). The MD_{CHF} conditions violated the assumptions required for valid use of likelihood-based missing data methods. Both MI and EMMW were used to produce BM and DS risk-adjusted estimates of the effectiveness of the treatment providers. A MVN model based on the joint distribution of variables used to generate multiple imputations. A mixed continuous and categorical was used for the joint distribution of covariates when generating weights in the EMMW analyses. The risk adjusted point estimates for two of the three of the treatment groups appeared to be biased. The degree of this bias was generally greater for the BM estimates than for the DS estimates, and the degree of bias was largest in the strongest MD_{CHF} condition with the greatest proportion

of missing CHF observations. Standard errors and 95% coverage probabilities were reasonable for the DS risk adjusted estimates. The BM standard errors were all over-estimated and the 95% coverage probabilities were all $> .97$.

The logistic regression coefficient for CHF was also examined in the simulations. When using EMMW, these parameters were underestimated when the missing data mechanism was dependent on the outcome. To a lesser extent, the coefficients using both MI and EMMW were also underestimated when the missing data mechanism was dependent on CHF. Coverage probabilities were often poor, but were generally better when EMMW was used as the missing data method.

The tendency of the BM standard errors to be over-estimated was examined by comparing the formula for the typical variance estimate with the formula for a delta method estimate of variance. The degree of over-estimation was also examined using computer simulations without missing data. In these simulations, BM estimates were compared with BM estimates obtained using the delta method (BM_Δ), and with DS estimates. It was concluded that the DS standard error estimates were superior and that the typical method of obtaining BM standard errors should be abandoned in favor of DS or BM_Δ estimates.

Conclusion. In conclusion, when covariate information is missing, risk adjustment with binary outcomes can be performed using multiple imputation or EM by the method of weights. Difficulties in applying these methods generally stemmed from the need to model the joint distribution of variables in the data set. Due to the complexity of the data, the adequacy of the fit of the joint distributions can be difficult to evaluate. As it is likely that suitable multiple imputation methods will soon be included with commercially available statistical packages, the development

of good diagnostic tools will be needed to evaluate the adequacy of the missing data methods. Although cumbersome with complex data, Monte Carlo simulations can be used to examine the sensitivity of missing data methods to model misspecification and to violations of the missing at random assumption. In this dissertation, the results from Monte Carlo simulations indicated that risk adjusted estimates obtained from these methods perform well under a variety of conditions, and that with modest amounts of missing data, the efficiencies of these estimates are comparable to those obtained with complete data. Finally, the standard errors obtained using the typical method for baseline adjusted models performed poorly, and should be abandoned in favor of delta method standard error estimates. Directly standardized estimates and standard errors performed well and should be considered by researchers conducting risk adjustment with binary outcomes when covariate information is incomplete.

Appendix A

Variances of Full-Model Adjusted Estimates

A.1 Indirectly Standardized Estimates

A.1.1 Full-Model SMR

Let $\mathbf{y}$ be a binary response vector for n patients $i = 1, 2, \ldots, n$. Indicator vectors $\boldsymbol{\delta}_k$ of length n will be used to denote providers $k = 1, 2, \ldots, l$. The elements of $\boldsymbol{\delta}_k$ equal 1 if case i is a has been treated by provider h_k and 0 otherwise. Let $\mathbf{X}_o = (\mathbf{1}, \mathbf{X}_c)$, where $\mathbf{1}$ is a unit vector of length n and $\mathbf{X}_c$ is an $n \times q$ matrix containing the q covariates of interest. Let $\mathbf{X} = (\mathbf{X}_o, \mathbf{X}_h)$ be a matrix of covariates, where $\mathbf{X_h}$ is $n \times (l-1)$ matrix of covariates used to code treatment by a given provider. Rows of $\mathbf{X}$ will be denoted as x_i^T. The first step in obtain the full-model adjusted SMRs is to use logistic regression to fit $\mathbf{y}$ to $\mathbf{X}$. The resulting $(q+l)$ length vector of regression parameter estimates will be denoted as $\hat{\boldsymbol{\beta}}$. From this model, we obtain $\hat{\mathbf{p}}$, an n length vector of estimates of the probability of death. The element of $\hat{\mathbf{p}}$ for case i is

$$\hat{p}_i = \frac{\exp(x_i^T \hat{\boldsymbol{\beta}})}{1 + \exp(x_i^T \hat{\boldsymbol{\beta}})}. \tag{A.1}$$

From these estimates, we obtain the diagonal matrix

$$\mathbf{V} = \widehat{\mathrm{Var}}(\mathbf{y}), \tag{A.2}$$

with diagonal elements $v_i = \hat{p}_i(1 - \hat{p}_i)$. Let $\hat{\boldsymbol{\beta}}_c$ be a vector containing the q elements of $\hat{\boldsymbol{\beta}}$ which correspond to the covariates of interest. The vector

$$\hat{\boldsymbol{\eta}}_c = \mathbf{X}_c \hat{\boldsymbol{\beta}}_c \tag{A.3}$$

has elements $\hat{\eta}_{ci}$ which correspond to the n individual cases. The $\hat{\eta}_{ci}$ are then treated as fixed constants in a second logistic regression which does not contain $\mathbf{X}_h$. This regression will be referred to as the offset model. The offset model is used to estimate probabilities of death for each case subject to the constraint that the sum of these probabilities will be equal to the number of deaths in the sample. For case i, the estimated probability of death from the offset model is

$$\hat{p}_{o_i} = \frac{\exp(\hat{\alpha}_o + \hat{\eta}_{c_i})}{1 + \exp(\hat{\alpha}_c + \hat{\eta}_{c_i})}. \tag{A.4}$$

The intercept for the offset model, $\hat{\alpha}_o$, is obtained to satisfy the condition

$$\sum_{i=1}^{n} \hat{p}_{o_i} = \sum_{i=1}^{n} y_i.$$

Let $\hat{\mathbf{p}}_o$ be a column vector of the $\hat{p}_{o_i}$'s. The full-model adjusted estimate of the expected number of deaths for provider k is then

$$E_k = \boldsymbol{\delta}_k^T \hat{\mathbf{p}}_o \tag{A.5}$$

and the observed number of deaths for provider k is

$$O_k = \boldsymbol{\delta}_k^T \mathbf{y} \tag{A.6}$$

The full-model adjusted SMR is

$$SMR_k = \frac{O_k}{E_k}. \tag{A.7}$$

A.1.2 Delta Method Approximations

Many of the variance estimates in these appendices rely on the delta method of approximation, in which the variance is obtained from a first-order Taylor series expansion. Consider the random variable $\hat{\theta}$, which is an estimate of the true parameter

θ of the probability distribution of the random variable Z. Let $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ be a vector of n independent realizations of the random variable Z, and let $\hat{\theta}$ be a function of $\mathbf{z}$, or $\hat{\theta} = f(\mathbf{z})$. When we know the variance of $\hat{\theta}$ but we wish to obtain the variance of some function of $\hat{\theta}$, say $g(\hat{\theta})$, we can do so by taking the variance of a first order Taylor series expansion of $g(\hat{\theta})$. The Taylor series expansion of $g(\hat{\theta})$ is

$$g(\hat{\theta}) = g(\theta) + g'(\theta)(\hat{\theta} - \theta) + g''(\theta)(\hat{\theta} - \theta)^2/2 + \cdots$$

Taking the variance of the zero and first order terms of this expansion yields the approximation

$$\mathrm{Var}\left(g(\hat{\theta})\right) \approx [g'(\theta)]^2 \mathrm{Var}(\hat{\theta}). \tag{A.8}$$

In most applications, the true parameter θ is not known, and $\hat{\theta}$ is substituted for θ. To account for this substitution, the meaning of $\approx$ is broadened to indicate approximation in a probablistic sense. The resulting approximation is

$$\mathrm{Var}(g(\hat{\theta})) \approx \left[g'(\hat{\theta})\right]^2 \mathrm{Var}(\hat{\theta}). \tag{A.9}$$

Using more careful notation we have

$$n\left[\mathrm{Var}(g(\hat{\theta})) - \left[g'(\hat{\theta})\right]^2 \mathrm{Var}(\hat{\theta})\right] \to 0$$

$$\text{as } n \to \infty.$$

A.1.3 Variance of the Full-Model SMR

Let $\mathbf{V}_o$ be a diagonal matrix with elements $v_{oi} = \hat{p}_{oi}(1 - \hat{p}_{oi})$. Let

$$\hat{\boldsymbol{\beta}}_o = \begin{bmatrix} \hat{\alpha}_o \\ \hat{\boldsymbol{\beta}}_c \end{bmatrix} \tag{A.10}$$

and let

$$\hat{\boldsymbol{\eta}}_o = \mathbf{X}_o\hat{\boldsymbol{\beta}}_o$$

with elements $\hat{\eta}_{oi}$.

By first order Taylor Series expansion,

$$\begin{aligned}\text{Var}(SMR_k) &= \text{Var}(O_k/E_k) \\ &\approx \frac{1}{E_k^2}\left[\left(\frac{O_k}{E_k}\right)^2 \text{Var}(E_k) + \text{Var}(O_k) - 2\left(\frac{O_k}{E_k}\right)\text{Cov}(O_k, E_k)\right] \quad \text{(A.11)}\end{aligned}$$

The following are expressions for the variances and covariance used in obtaining the variance of the SMR:

Variance of O_k

$$\begin{aligned}\text{Var}(O_k) &= \text{Var}(\boldsymbol{\delta}_k^T\mathbf{y}) \\ &= \boldsymbol{\delta}_k^T\text{Var}(\mathbf{y})\boldsymbol{\delta}_k \quad \text{(A.12)}\end{aligned}$$

which can be estimated by substituting $\mathbf{V}$ for $\text{Var}(\mathbf{y})$.

Variance of E_k

$$\begin{aligned}\text{Var}(E_k) &= \boldsymbol{\delta}_k^T\text{Var}(\hat{\mathbf{p}}_o)\boldsymbol{\delta}_k \\ &\approx \boldsymbol{\delta}_k^T\frac{\partial\hat{\mathbf{p}}_o}{\partial\hat{\boldsymbol{\beta}}_o^T}\text{Var}(\hat{\boldsymbol{\beta}}_o)\frac{\partial\hat{\mathbf{p}}_o}{\partial\hat{\boldsymbol{\beta}}_o}\boldsymbol{\delta}_k \quad \text{(A.13)}\end{aligned}$$

The partial derivative used above can be expressed as

$$\begin{aligned}\frac{\partial\hat{\mathbf{p}}_o}{\partial\hat{\boldsymbol{\beta}}_o^T} &= \frac{\partial\hat{\mathbf{p}}_o}{\partial\hat{\boldsymbol{\eta}}_o^T}\frac{\partial\hat{\boldsymbol{\eta}}_o}{\partial\hat{\boldsymbol{\beta}}_o^T} \\ &= \mathbf{V}_o\mathbf{X}_o. \quad \text{(A.14)}\end{aligned}$$

Variance of $\hat{\boldsymbol{\beta}}_o$. The variance of $\hat{\boldsymbol{\beta}}_o$ can be expressed as the matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_o) = \begin{bmatrix} \mathrm{Var}(\hat{\alpha}_o) & \mathrm{Cov}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c)^T \\ \mathrm{Cov}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c) & \mathrm{Var}(\hat{\boldsymbol{\beta}}_c) \end{bmatrix} \tag{A.15}$$

The variance of $\hat{\boldsymbol{\beta}}_c$ is easily estimated by taking the submatrix of

$$\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}$$

which corresponds to the covariates contained $\mathbf{X}_c$. This submatrix will be denoted as

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_c) = \left[\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\right]_c \tag{A.16}$$

Variance of $\hat{\alpha}_o$. The variance of $\hat{\alpha}_o$ can be approximated by taking the variance of the first-order Taylor series expansion

$$\hat{\alpha}_o = f(\hat{\boldsymbol{\beta}}_c, \mathbf{y}) \approx f(\hat{\boldsymbol{\beta}}_c, \mathbf{p}) + (\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c)\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}} + (\mathbf{y} - \mathbf{p})\frac{\partial f}{\partial \mathbf{y}} \tag{A.17}$$

The variance of this expression can be approximated as

$$\mathrm{Var}(\hat{\alpha}_o) \approx \frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T}\mathrm{Var}(\hat{\boldsymbol{\beta}}_c)\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c} + \frac{\partial f}{\partial \mathbf{y}^T}\mathrm{Var}(\mathbf{y})\frac{\partial f}{\partial \mathbf{y}} + \frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T}\mathrm{Cov}(\hat{\boldsymbol{\beta}}_c, \mathbf{y})\frac{\partial f}{\partial \mathbf{y}} \tag{A.18}$$

Estimates of the variances in the above equation are obtained from A.2, A.16 and by the following:

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}_c, \mathbf{y}) &\approx \mathrm{Cov}\left(\frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \mathbf{y}^T}\mathbf{y}, \mathbf{y}\right) \\ &= \frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \mathbf{y}^T}\mathrm{Var}(\mathbf{y}). \end{aligned}$$

This can be estimated by

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_c, \mathbf{y}) = \left[(\mathbf{X^T V X})^{-1}\mathbf{X}^T\right]_c \mathbf{V} \qquad \text{(A.19)}$$

where $\mathbf{V}$ is used to estimate $\mathrm{Var}(\mathbf{y})$. The term for $\frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \mathbf{y}}$ contains the q rows of $\left[\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\right]$ which correspond to the covariates of interest.

The implicit function theorem can be applied to obtain the derivatives needed for the approximation of $\mathrm{Var}(\hat{\alpha}_o)$. Although $\hat{\alpha}_o = f(\hat{\boldsymbol{\beta}}_c, \mathbf{y})$, there is no closed form expression for this function. From the score equations used to obtain $\hat{\alpha}_o$, we have a function of α_o, β_c and $\mathbf{y}$ which is equal to zero at the maximum likelihood estimates of the parameters for the observed data

$$F(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c, \mathbf{y}) = \textstyle\sum_{i=1}^{n}(y_i - \hat{p}_{o_i}) = 0$$

According to the implicit function theorem

$$\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c} = -\frac{\partial F/\partial \hat{\boldsymbol{\beta}}_c}{\partial F/\partial \hat{\alpha}_o} \quad ; \quad \frac{\partial f}{\partial \mathbf{y}} = -\frac{\partial F/\partial \mathbf{y}}{\partial F/\partial \hat{\alpha}_o}$$

where

$$\begin{aligned} \partial F/\partial \hat{\boldsymbol{\beta}}_c^T &= -\sum_{i=1}^{n} \frac{\partial \hat{p}_{o_i}}{\partial \hat{\boldsymbol{\eta}}_o^T} \frac{\partial \hat{\boldsymbol{\eta}}_o}{\partial \hat{\boldsymbol{\beta}}_c^T} \\ &= -\mathbf{1}^T \mathbf{V}_o \mathbf{X}_c \end{aligned}$$

$$\begin{aligned} \partial F/\partial \mathbf{y}^T &= \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{y}} y_i \\ &= \mathbf{1} \end{aligned}$$

$$\begin{aligned} \partial F/\partial \hat{\alpha}_o &= -\sum_{i=1}^{n} \frac{\partial \hat{p}_{o_i}}{\partial \hat{\boldsymbol{\eta}}_o^T} \frac{\partial \hat{\boldsymbol{\eta}}_o}{\partial \hat{\alpha}_o} \\ &= -\sum_{i=1}^{n} v_{o_i} \end{aligned}$$

So

$$\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T} = -\left(\sum_{i=1}^n v_{o_i}\right)^{-1} \mathbf{1}^T \mathbf{V}_o \mathbf{X}_c \quad ; \quad \frac{\partial f}{\partial \mathbf{y}^T} = \left(\sum_{i=1}^n v_{o_i}\right)^{-1} \mathbf{1} \tag{A.20}$$

Substituting the above results into A.18, an estimate of $\mathrm{Var}(\hat{\alpha}_o)$ is therefore:

$$\widehat{\mathrm{Var}}(\hat{\alpha}_o) = \tag{A.21}$$
$$\left(\sum_{i=1}^n v_{o_i}\right)^{-2} \left[\mathbf{1}^T \mathbf{V}_o \mathbf{X}_c \left(\mathbf{X}^T \mathbf{V} \mathbf{X}\right)^{-1} \mathbf{X}_c^T \mathbf{V}_o \mathbf{1} + \mathbf{1}^T \mathbf{V} \mathbf{1} - 2 \cdot \mathbf{1}^T \mathbf{V}_o \mathbf{X}_c \left[\left(\mathbf{X}^T \mathbf{V} \mathbf{X}\right)^{-1} \mathbf{X}^T\right]_c \mathbf{V} \mathbf{1}\right]$$

Covariance of $\hat{\alpha}_o$ and $\hat{\boldsymbol{\beta}}_c$. The covariance of $\hat{\alpha}_o$ and $\hat{\boldsymbol{\beta}}_c$ can be approximated by taking the covariance between the right hand term of A.17 and $\hat{\boldsymbol{\beta}}_c$. This yields:

$$\mathrm{Cov}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c) \approx \frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T} \mathrm{Var}(\hat{\boldsymbol{\beta}}_c) + \frac{\partial f}{\partial \mathbf{y}^T} \mathrm{Cov}(\mathbf{y}, \hat{\boldsymbol{\beta}}_c)$$

By A.16, A.19 and A.20, this can be estimated as

$$\widehat{\mathrm{Cov}}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c) = \left(\sum_{i=1}^n v_{o_i}\right)^{-1} \mathbf{1}^T \left[\mathbf{V}\left[\mathbf{X}\left(\mathbf{X}^T \mathbf{V} \mathbf{X}\right)^{-1}\right]_c - \mathbf{V}_o \mathbf{X}_c \left[\left(\mathbf{X}^T \mathbf{V} \mathbf{X}\right)^{-1}\right]_c\right] \tag{A.22}$$

Covariance of O_k and E_k

The covariance of O_k and E_k can be approximated as

$$\begin{aligned} \mathrm{Cov}(O_k, E_k) &= \mathrm{Cov}(\boldsymbol{\delta}_k^T \mathbf{y}, \boldsymbol{\delta}_k^T \hat{\mathbf{p}}_o) \\ &\approx \boldsymbol{\delta}_k^T \mathrm{Var}(\mathbf{y}) \frac{\partial \hat{\mathbf{p}}_o}{\partial \mathbf{y}^T} \boldsymbol{\delta}_k. \end{aligned} \tag{A.23}$$

By the chain rule, the partial derivative used above can be expressed as

$$\frac{\partial \hat{\mathbf{p}}_o}{\partial \mathbf{y}^T} = \frac{\partial \hat{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T} \frac{\partial \hat{\boldsymbol{\beta}}_o}{\partial \mathbf{y}^T}$$

The expressions for the derivatives can be found in A.14 and in

$$\frac{\partial \hat{\boldsymbol{\beta}}_o}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial \alpha_o}{\partial \mathbf{y}^T} \\ \frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \mathbf{y}^T} \end{bmatrix}$$

$$= \begin{bmatrix} (\sum_i^n v_{o_i})^{-1}\mathbf{1}^T \\ \left[\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\right]_c \end{bmatrix}$$

An estimate of A.23 can be obtained by using $\mathbf{V}$ as and estimate of $\text{Var}(\mathbf{y})$.

A.1.4 Full-Model Adjusted Population Averaged Proportion

When logistic regression is used to calculate the the probability of death, the hospital adjusted population averaged proportion can be calculated as

$$\hat{P}_k = SMR_k\bar{p} = (O_k/E_k)\left[n^{-1}\mathbf{1}^T\mathbf{y}\right]. \tag{A.24}$$

A.1.5 Variance of the Population Averaged Proportion

By first order Taylor Series expansion,

$$\begin{aligned} \text{Var}(\hat{P}_k) \approx & \left(\frac{\bar{p}}{E_k}\right)^2 \text{Var}(O_k) + \left(\frac{O_k}{E_k^2}\bar{p}\right)^2 \text{Var}(E_k) + \left(\frac{O_k}{E_k}\right)^2 \text{Var}(\bar{p}) \\ & - 2\left(\frac{O_k}{E_k^2}\bar{p}\right)\left[\left(\frac{O_k}{E_k}\right)\text{Cov}(E_k,\bar{p}) + \left(\frac{\bar{p}}{E_k}\right)\text{Cov}(O_k,E_k) - \text{Cov}(O_k,\bar{p})\right]. \end{aligned} \tag{A.25}$$

Expressions for $\text{Var}(O_k)$, $\text{Var}(E_k)$ and $\text{Cov}(O_k, E_k)$ are presented in A.12, A.13 and A.23. The remaining terms can be obtained by noting that

$$\begin{aligned} \text{Cov}(\bar{p}, E_k) &= \text{Cov}(n^{-1}\mathbf{1}^T\mathbf{y}, \boldsymbol{\delta}_k^T\hat{\mathbf{p}}_o) \\ &\approx n^{-1}\mathbf{1}^T\text{Var}(\mathbf{y})\frac{\partial\hat{\mathbf{p}}_o}{\partial\mathbf{y}^T}\boldsymbol{\delta}_k \end{aligned}$$

$$\begin{aligned} \text{Cov}(O_k, \bar{p}) &= \text{Cov}(\boldsymbol{\delta}^T\mathbf{y}, n^{-1}\mathbf{1}^T\mathbf{y}) \\ &= n^{-1}\boldsymbol{\delta}_k^T\text{Var}(\mathbf{y})\boldsymbol{\delta}_k \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{p}) &= \text{Var}(n^{-1}\mathbf{1}^T\mathbf{y}) \\ &= n^{-2}\mathbf{1}^T\text{Var}(\mathbf{y})\mathbf{1} \end{aligned}$$

A.2 Directly Standardized Rates

Directly standardized rates are obtained by estimating the expected number of deaths that would have occurred if all patients in the population of interest were treated by a given provider. Let $\mathbf{C}_k$ be an $n \times (l-1)$ matrix in which each row contains the codes associated with treatment by provider k. Let $\mathbf{X}_k = (\mathbf{X}_o, \mathbf{C}_k)$ be the covariate matrix associated with provider k, where $\mathbf{X}_o$ is defined in section A.1.1. Let $\hat{\boldsymbol{\beta}}$ be the $(q+k)$ length vector of logistic regression parameters obtained by regressing $\mathbf{y}$ on $\mathbf{X}$. Define

$$\hat{\boldsymbol{\eta}}_k = \mathbf{X}_k \hat{\boldsymbol{\beta}} \tag{A.26}$$

with elements $\hat{\eta}_{ik}$.

On the basis of the covariate values associated with an individual i and the risk associated by treatment by a given provider k, an estimate of the probability of death can be obtained for each individual in the population

$$\hat{p}_{ik} = \frac{\exp(\hat{\eta}_{ik})}{1 + \exp(\hat{\eta}_{ik})}. \tag{A.27}$$

These estimates are elements of n length column vector $\hat{\mathbf{p}}_k$. The expected number of deaths in the population which would have occurred if all patients were treated by provider k is

$$E_k^* = \mathbf{1}^T \hat{\mathbf{p}}_k. \tag{A.28}$$

A.2.1 Directly Standardized Risk Ratio

The directly standardized risk ratio can be obtained as

$$\widehat{RR}_k^* = \frac{E_k^*}{O} \tag{A.29}$$

where $O = \mathbf{1}^T\mathbf{y}$.

A.2.2 Variance of the Standardized Risk Ratio

Let $\mathbf{V}_k$ be a diagonal matrix with diagonal elements $\hat{p}_{ik}(1-\hat{p}_{ik})$. An approximation for the variance of the standardized risk ratio can be obtained by a first-order Taylor Series expansion:

$$\begin{aligned}\mathrm{Var}(\widehat{RR}_k) &= \mathrm{Var}(E_k^*/O) \\ &\approx \frac{1}{O^2}\left[\left(\frac{E_k^*}{O}\right)^2 \mathrm{Var}(O) + \mathrm{Var}(E_k^*) - 2\left(\frac{E_k^*}{O}\right)\mathrm{Cov}(E_k^*, O)\right]. \quad \text{(A.30)}\end{aligned}$$

The following are expressions for the variances and covariances used in obtaining the variance of the above risk ratio.

Variance of O

The variance of O is

$$\begin{aligned}\mathrm{Var}(O) &= \mathrm{Var}(\mathbf{1}^T\mathbf{y}) \\ &= \mathbf{1}^T\mathrm{Var}(\mathbf{y})\mathbf{1}\end{aligned}$$

which can be estimated as

$$\widehat{\mathrm{Var}}(O) = \mathbf{1}^T\mathbf{V}\mathbf{1}. \quad \text{(A.31)}$$

Variance of E_k^*

The variance of E_k^* can be expressed as

$$\begin{aligned}\mathrm{Var}(E_k^*) &= \mathbf{1}^T\mathrm{Var}(\hat{\mathbf{p}}_k)\mathbf{1} \\ &\approx \mathbf{1}^T\frac{\partial\hat{\mathbf{p}}_k}{\partial\hat{\boldsymbol{\beta}}^T}\mathrm{Var}(\hat{\boldsymbol{\beta}})\frac{\partial\hat{\mathbf{p}}_k}{\partial\hat{\boldsymbol{\beta}}}\mathbf{1}. \quad \text{(A.32)}\end{aligned}$$

Applying the chain rule, the partial derivative used above can be expressed as

$$\begin{aligned}\frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T} &= \frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\eta}}_k^T}\frac{\partial \hat{\boldsymbol{\eta}}_k}{\partial \hat{\boldsymbol{\beta}}^T} \\ &= \mathbf{V}_k\mathbf{X}_k. \end{aligned} \tag{A.33}$$

By using

$$\mathrm{Var}(\hat{\beta}) = \left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1},$$

$\mathrm{Var}(E_k^*)$ can be estimated as

$$\widehat{\mathrm{Var}}(E_k^*) = \mathbf{1}^T\mathbf{V}_k\mathbf{X}_k\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}_k^T\mathbf{V}_k\mathbf{1}. \tag{A.34}$$

Covariance of O and E_k^*

A first-order approximation of the covariance of O and E_k^* is

$$\begin{aligned}\mathrm{Cov}(O, E_k^*) &= \mathrm{Cov}(\mathbf{1}^T\mathbf{y}, \mathbf{1}^T\hat{\mathbf{p}}_k) \\ &\approx \mathbf{1}^T\mathrm{Var}(\mathbf{y})\frac{\partial \hat{\mathbf{p}}_k}{\partial \mathbf{y}}\mathbf{1}.\end{aligned}$$

An expression for the partial derivative used above is

$$\begin{aligned}\frac{\partial \hat{\mathbf{p}}_k}{\partial \mathbf{y}^T} &= \frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^T} \\ &= \mathbf{V}_k\mathbf{X}_k\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\end{aligned}$$

by A.33 and by noting that $\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^T} = \left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T$. An estimate of $\mathrm{Cov}(O, E_k^*)$ is

$$\widehat{\mathrm{Cov}}(O, E_k^*) = \mathbf{1}^T\mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}_k\mathbf{V}_k\mathbf{1}. \tag{A.35}$$

A.2.3 Population Averaged Proportion

A directly standardized extimate of the population averaged proportion is

$$\hat{P}_k^* = \frac{E_k^*}{n}. \tag{A.36}$$

The variance for this estimate is

$$\text{Var}(\hat{P}_k^*) = \frac{1}{n^2}\text{Var}(E_k^*).$$

Using A.34, this can be estimated as

$$\widehat{\text{Var}}(\hat{P}_j^*) = \frac{1}{n^2}\mathbf{1}^T\mathbf{V}_k\mathbf{X}_k\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)\mathbf{X}_k^T\mathbf{V}_k\mathbf{1}. \tag{A.37}$$

Appendix B

Variance of $\hat{\beta}$ for EM by Method of Weights

The method described by Louis (1982) can be used to estimate the variance-covariance matrix of the logistic regression parameters ($\hat{\beta}$). When using the EM algorithm, the observed information matrix $I(\beta|\mathbf{X}_{obs}, \mathbf{Y})$ can be obtained by taking the second derivative of the log-likelihood $\ell(\beta|\mathbf{X}_{obs}, \mathbf{Y})$ with respect to β, or

$$\begin{aligned} \mathrm{I}(\beta|\mathbf{Y}, \mathbf{X}_{obs}) &= -\frac{\partial^2}{\partial\beta^2}\ell(\beta|\mathbf{X}_{obs}, \mathbf{Y}) \\ &= -\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta^2}\ell_i(\beta|y_i, x_{obs.i}). \end{aligned}$$

by 3.10. Louis demonstrated that the information matrix could be expressed terms of the complete data quantities:

$$\begin{aligned} \mathrm{I}(\beta|\mathbf{Y}, \mathbf{X}_{obs}) = \sum_{i=1}^{n} \big[& \mathrm{E}\left\{\mathrm{I}_i(\beta|y_i, x_i)|x_{obs.i}\right\} - \mathrm{E}\left\{\mathrm{S}_i(\beta|y_i, x_i)\right\} \\ & + \mathrm{E}\left\{\mathrm{S}_i(\beta|x_i, y_i)|x_{obs.i}\right\}\mathrm{E}\left\{\mathrm{S}_i(\beta|x_i, y_i)|x_{obs.i}\right\}^T\big] \end{aligned} \tag{B.1}$$

where

$$\begin{aligned} \mathrm{I}_i(\beta|x_i, y_i) &= -\frac{\partial^2}{\partial\beta^2}\ell_i(\beta|x_i, y_i) \\ &= v_i x_i x_i^T \\ \mathrm{S}_i(\beta|x_i, y_i) &= \frac{\partial}{\partial\beta}\ell_i(\beta|x_i, y_i) \\ &= (y_i - p_i)x_i. \end{aligned}$$

Futher $v_i = p_i(1-p_i)$, and the p_i are probabilities of death obtained from the logistic regression model. Recall that the w_{ij} from the final iteration of the EMMW logistic

regression are

$$w_{ij} = \Pr(x_{ij}|y_i, x_{obs.i}, \hat{\boldsymbol{\beta}}, \hat{\theta}). \tag{B.2}$$

Using the w_{ij}, expectations of the complete data quantities in B.1 can be obtained as

1)

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\left\{\mathrm{I}_i(\hat{\boldsymbol{\beta}}|y_i, x_i)|x_{obs.i}\right\} &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}v_{ij}x_{ij}x_{ij}^T \\ &= \tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{V}}\tilde{\mathbf{X}} \end{aligned} \tag{B.3}$$

Where $\tilde{\mathbf{X}}$ is the augmented $\mathbf{X}_{obs}$ matrix, and $\mathbf{W}$ and $\tilde{\mathbf{V}}$ are diagonal matrices containing the w_{ij} and the v_{ij} respectively.

2)

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\left\{\mathrm{S}_i(\boldsymbol{\beta}|y_i, x_i)\mathrm{S}_i(\boldsymbol{\beta}|x_i, y_i)^T|x_{obs.i}\right\} &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(y_i - p_{ij})^2 x_{ij}x_{ij}^T \\ &= \tilde{\mathbf{X}}^T\mathbf{W}\mathbf{D}^2\tilde{\mathbf{X}} \end{aligned} \tag{B.4}$$

where $\mathbf{D}$ is a diagonal matrix with elements $(y_i - p_{ij})$.

3)

$$\begin{aligned}&\sum_{i=1}^{n} \mathrm{E}\left\{\mathrm{S}_i(\boldsymbol{\beta}|x_i, y_i)|x_{obs.i}\right\} \mathrm{E}\left\{\mathrm{S}_i(\boldsymbol{\beta}|x_i, y_i)|x_{obs.i}\right\}^T = \\ &\qquad \sum_{i=1}^{n}\left[\sum_{j=1}^{n_i} w_{ij}(y_i - p_{ij})x_{ij}\right]\left[\sum_{j=1}^{n_i} w_{ij}(y_i - p_{ij})x_{ij}\right]^T \end{aligned} \tag{B.5}$$

This quantity will be denoted as $\boldsymbol{\Lambda}$.

Combining B.3, B.4 and B.5, the observed information can be expressed as

$$\mathrm{I}(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}_{obs}) = \tilde{\mathbf{X}}^T \mathbf{W}(\tilde{\mathbf{V}} - \mathbf{D}^2)\tilde{\mathbf{X}} + \boldsymbol{\Lambda}. \tag{B.6}$$

The variance-covariance matrix for the regression parameters $\hat{\boldsymbol{\beta}}$ can be obtained as $\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbf{X}_{obs})$.

Appendix C

Variances of Estimates with Missing Data in the Covariates

C.1 Indirectly Standardized Estimates

C.1.1 Full-Model SMR

The variance estimates must account for the weights obtained from the final step of the EM algorithm and for the use of Louis's method in obtaining the variance of the regression coefficients. Let $\tilde{\mathbf{X}}$ be the augmented covariate matrix. Let i denote cases $i = 1, 2, \ldots, n$ and let $\tilde{n}$ be the number of rows in $\tilde{\mathbf{X}}$. Let n_i be the number of covariate patterns for case i in the augmented covariate matrix. For cases with no missing data, $n_i = 1$. Cases with missing data have $j = 1, \ldots, n_i$ covariate patterns in $\tilde{\mathbf{X}}$, and $\tilde{n} = \sum_{i=1}^{n} n_i$. Let $\tilde{\mathbf{X}}_o = (\tilde{\mathbf{1}}, \tilde{\mathbf{X}}_c)$, where $\tilde{\mathbf{1}}$ is a unit vector or length $\tilde{n}$ and $\tilde{\mathbf{X}}_c$ is an $\tilde{n} \times q$ augmented matrix of the covariates of interest. Partition the augmented covariate matrix as $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_o, \tilde{\mathbf{X}}_h)$, where $\tilde{\mathbf{X}}_h$ is an $\tilde{n} \times (k-1)$ matrix of covariates used to code for treatment provider. Rows of $\tilde{\mathbf{X}}$ will be denoted as x_{ij}^T. Let $\tilde{\mathbf{y}}$ be an $\tilde{n}$ length outcome vector with elements corresponding to the rows of $\tilde{\mathbf{X}}$. The first step in obtaining the full-model SMRs is to use EMMW to fit $\tilde{\mathbf{y}}$ to $\tilde{\mathbf{X}}$ in a weighted logistic regression. The resulting estimated probability of death associated with covariate pattern j for subject i will be denoted as

$$\hat{p}_{ij} = \frac{\exp(x_{ij}^T \hat{\boldsymbol{\beta}})}{1 + \exp(x_{ij}^T \hat{\boldsymbol{\beta}})}.$$

The vector of the $\hat{p}_{ij}$'s will be denoted as $\tilde{\mathbf{p}}$. From these estimates, we obtain the diagonal matrix $\tilde{\mathbf{V}}$, with diagonal elements $v_{ij} = \hat{p}_{ij}(1 - \hat{p}_{ij})$. Let $\hat{\boldsymbol{\beta}}_c$ be a vector containing the q elements of $\hat{\boldsymbol{\beta}}$ which correspond to the covariates of interest. The vector

$$\tilde{\boldsymbol{\eta}}_c = \tilde{\mathbf{X}}_c \hat{\boldsymbol{\beta}}_c$$

is of length $\tilde{n}$ and has elements $\hat{\eta}_{cij}$. The $\hat{\eta}_{cij}$ are then treated as fixed constants in a second logistic regression which does not contain $\tilde{\mathbf{X}}_h$. For case i with covariate pattern j, the estimated probability of death from this model is

$$\hat{p}_{o_{ij}} = \frac{\exp(\hat{\alpha}_o + \hat{\eta}_{c_{ij}})}{1 + \exp(\hat{\alpha}_o + \hat{\eta}_{c_{ij}})}. \tag{C.1}$$

The intercept for the offset model, $\hat{\alpha}_o$, is obtained to satisfy the condition

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} \hat{p}_{o_{ij}} = 0,$$

where the w_{ij} are the weights obtained from the final iteration of the EMMW algorithm. Let $\tilde{\mathbf{p}}_o$ be a column vector of the $\hat{p}_{o_{ij}}$'s, and let W be a diagonal matrix with the w_{ij} as diagonal elements. The $n \times \tilde{n}$ indicator matrix $\mathbf{S}$ has rows $\tilde{\delta}_i^T$, where the $\tilde{\delta}_i$ are 0, 1 indicator vectors denoting the rows in $\tilde{\mathbf{X}}$ corresponding to case i. As in appendix A, θ_k is a 0, 1 indicator vector of length n denoting treatment by provider k.

The full-model adjusted estimate of the expected number of deaths for provider k is then

$$\begin{aligned} E_k &= \sum_{i \in h_k} \sum_{j=1}^{n_i} w_{ij} \tilde{p}_{oij} \\ &= \delta_k^T \mathbf{S} \mathbf{W} \tilde{\mathbf{p}}_o \qquad \text{(C.2)} \\ &= \delta_k^T \bar{\mathbf{p}}_o \qquad \text{(C.3)} \end{aligned}$$

As in appendix A, the observed number of deaths for provider k is

$$O_k = \boldsymbol{\delta}_k^T \mathbf{y} \tag{C.4}$$

The full-model adjusted SMR, obtained using EMMW is

$$SMR_k = \frac{O_k}{E_k}.$$

C.1.2 Variance of the Full-Model SMR

As in appendix A, the estimate of variance of the SMR is obtained by taking the variance of the first order Taylor Series expansion, or

$$\text{Var}(SMR_k) \approx \frac{1}{E_k^2}\left[\left(\frac{O_k}{E_k}\right)^2 \text{Var}(E_k) + \text{Var}(O_k) - 2\left(\frac{O_k}{E_k}\right)\text{Cov}(O_k, E_k)\right] \tag{C.5}$$

The following are expressions for the variances and covariances used in obtaining the variance of the SMR:

Variance of O_k

As in the complete data case, the variance of O_k can be expressed as

$$\text{Var}(O_k) = \delta_k^T \text{Var}(\mathbf{y})\delta_k$$

which can be estimated as

$$\widehat{\text{Var}}(O_k) = \boldsymbol{\delta}_k^T \widehat{\text{Var}}(\mathbf{y}|\mathbf{X}_{obs})\boldsymbol{\delta}_k. \tag{C.6}$$

Variance of y given $\mathbf{X}_{obs}$. Since the y_i are condsidered to be independent observations of the outcome, $\text{Var}(\mathbf{y}|\mathbf{X}_{obs})$ will be an $(n \times n)$ diagonal matrix with elements

$$\text{Var}(y_i|x_{obs.i}) = \text{Var}\left(\text{E}(y_i|x_i)|x_{obs.i}\right) + \text{E}\left(\text{Var}(y_i|x_i)|x_{obs.i}\right)$$

An estimate of this variance can be obtained as

$$\begin{aligned}\widehat{\mathrm{Var}}(y_i|x_{obs.i}) &= \left[\sum_{j=1}^{n_i}\hat{p}_{ij}^2 w_{ij} - \left(\sum_{j=1}^{n_i}\hat{p}_{ij}w_{ij}\right)^2\right] + \left[\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})w_{ij}\right] \\ &= \sum_{j=1}^{n_i}\hat{p}_{ij}w_{ij} - \left(\sum_{j=1}^{n_i}\hat{p}_{ij}w_{ij}\right)^2 \\ &= \bar{p}_i(1-\bar{p}_i). \end{aligned} \tag{C.7}$$

Variance of E_k

Let $\tilde{\eta}_o = \tilde{\mathbf{X}}^T\hat{\boldsymbol{\beta}}_o$.

The variance of E_k can be approximated as

$$\begin{aligned}\mathrm{Var}(E_k) &= \boldsymbol{\delta}_k^T \mathrm{Var}(\bar{\mathbf{p}}_o)\boldsymbol{\delta}_k \\ &\approx \boldsymbol{\delta}_k^T \frac{\partial \bar{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T}\mathrm{Var}(\hat{\boldsymbol{\beta}}_o)\frac{\partial \bar{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\beta}}_o}\boldsymbol{\delta}_k. \end{aligned} \tag{C.8}$$

By applying the chain rule and noting that $\bar{\mathbf{p}} = \mathbf{SW}\tilde{\mathbf{p}}_o$, the derivative used above can be expressed as

$$\begin{aligned}\frac{\partial \bar{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T} &= \frac{\partial \bar{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\eta}}_o^T}\frac{\partial \hat{\boldsymbol{\eta}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T} \\ &= \mathbf{SW}\frac{\partial \tilde{\mathbf{p}}}{\partial \tilde{\boldsymbol{\eta}}_o^T}\frac{\partial \tilde{\boldsymbol{\eta}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T} \\ &= \mathbf{SW}\tilde{\mathbf{V}}_o\tilde{\mathbf{X}}_o. \end{aligned}$$

Variance of $\hat{\boldsymbol{\beta}}_o$. As in appendix A, the variance of $\hat{\boldsymbol{\beta}}_o$ can be expressed as the partitioned matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_o) = \begin{bmatrix} \mathrm{Var}(\hat{\alpha}_o) & \mathrm{Cov}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c)^T \\ \mathrm{Cov}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c) & \mathrm{Var}(\hat{\boldsymbol{\beta}}_c) \end{bmatrix}. \tag{C.9}$$

The submatrix of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ obtained via Louis's method (see appendix B) corresponding to the covariates of interest can be used as an estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}}_c)$.

Variance of $\hat{\alpha}_o$. As in appendix A, $\text{Var}(\hat{\alpha}_o)$ can be approximated as

$$\text{Var}(\hat{\alpha}_o) \approx \frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T}\text{Var}(\hat{\boldsymbol{\beta}}_c)\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c} + \frac{\partial f}{\partial \mathbf{y}^T}\text{Var}(\mathbf{y})\frac{\partial f}{\partial \mathbf{y}} + \frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T}\text{Cov}(\hat{\boldsymbol{\beta}}_c, \mathbf{y})\frac{\partial f}{\partial \hat{\mathbf{p}}_c} \tag{C.10}$$

To obtain an estimate of C.10, $\widehat{\text{Var}}(\mathbf{y})|\mathbf{X}_{obs}$ is used to estimate $\text{Var}(\mathbf{y})$. An approximation of $\text{Cov}(\hat{\beta}_c, \mathbf{y})$ is:

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}_c, \mathbf{y}) &\approx \text{Cov}(\frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \mathbf{y}^T}\mathbf{y}, \mathbf{y}) \\ &= \frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \tilde{\mathbf{y}}^T}\frac{\partial \tilde{\mathbf{y}}}{\partial \mathbf{y}^T}\text{Var}(\mathbf{y})\end{aligned} \tag{C.11}$$

which can be estimated using

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_c, \mathbf{y}) = \left[(\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\right]_c \mathbf{S}^T(\widehat{\text{Var}}(\mathbf{y})|\mathbf{X}_{obs})$$

where $\frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \tilde{\mathbf{y}}}$ contains the q rows of $\left[(\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\right]$ which correspond to the covariates of interest and $\frac{\partial \tilde{\mathbf{y}}}{\partial \mathbf{y}^T} = \mathbf{S}^T$.

The derivatives required to approximate the variance of $\hat{\boldsymbol{\alpha}}_o$ are obtained using the implicit function theorem as outlined in appendix B. In this case, however, the solutions must account for the weights obtained from the EMMW. To account for these weights, the implicit function theorem is applied to

$$\mathbf{F}(\hat{\alpha}_o, \hat{\boldsymbol{\beta}}_c, \tilde{\mathbf{y}}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(y_i - \hat{p}_{oij}) = 0. \tag{C.12}$$

According to the implicit function theorem

$$\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c} = -\frac{\partial F/\partial \hat{\boldsymbol{\beta}}_c}{\partial F/\partial \hat{\alpha}_o} \quad ; \quad \frac{\partial f}{\partial \tilde{\mathbf{y}}} = -\frac{\partial F/\partial \tilde{\mathbf{y}}}{\partial F/\partial \hat{\alpha}_o}$$

where

$$\begin{aligned}
\partial F/\partial \hat{\boldsymbol{\beta}}_c^T &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}\frac{\partial \hat{p}_{o_i}}{\partial \hat{\boldsymbol{\eta}}_o^T}\frac{\partial \hat{\boldsymbol{\eta}}_o}{\partial \hat{\boldsymbol{\beta}}_c^T} \\
&= -\tilde{\mathbf{1}}^T\mathbf{W}\tilde{\mathbf{V}}_o\tilde{\mathbf{X}}_c
\end{aligned}$$

$$\begin{aligned}
\partial F/\partial \mathbf{y}^T &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{oij}\frac{\partial}{\partial \mathbf{y}}y_i \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial y}y_i\sum_{j=1}^{n_i} w_{ij} \\
&= \mathbf{1}
\end{aligned}$$

$$\begin{aligned}
\partial F/\partial \hat{\alpha}_o &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}\frac{\partial \hat{p}_{o_{ij}}}{\partial \tilde{\boldsymbol{\eta}}_o^T}\frac{\partial \tilde{\boldsymbol{\eta}}_o}{\partial \hat{\alpha}_o} \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}v_{oij}
\end{aligned}$$

So

$$\frac{\partial f}{\partial \hat{\boldsymbol{\beta}}_c^T} = -\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}v_{oij}\right)^{-1}\tilde{\mathbf{1}}^T\tilde{\mathbf{V}}_o\tilde{\mathbf{X}}_{oc} \quad ; \quad \frac{\partial f}{\partial \mathbf{y}^T} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}v_{oij}\right)^{-1}\mathbf{1}$$

Using the above derivatives and estimates of variance, $\mathrm{Var}(\hat{\alpha}_o)$ can be estimated as

$$\begin{aligned}
&\hat{\alpha}_o \approx \\
&\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}v_{oij}\right)^2\Big[\tilde{\mathbf{1}}^T\mathbf{W}\tilde{\mathbf{V}}_o\tilde{\mathbf{X}}_c\mathrm{Var}(\hat{\beta}_c)\tilde{\mathbf{X}}_c^T\tilde{\mathbf{V}}_o^T\mathbf{W}^T\tilde{\mathbf{1}} + \mathbf{1}^T(\mathrm{Var}(y)|x_{obs.i})\mathbf{1} - \\
&2\cdot\tilde{\mathbf{1}}^T\mathbf{W}\tilde{\mathbf{V}}_o\tilde{\mathbf{X}}_c\left[\left(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{V}}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\right]_c\mathbf{S}^T(\widehat{\mathrm{Var}}(y)|x_{obs.i})\mathbf{1}\Big]
\end{aligned}$$

Covariance of O_k and E_k

The covariance of O_k and E_k can be approximated as

$$\begin{aligned}
\mathrm{Cov}(O_k, E_k) &= \mathrm{Cov}(\boldsymbol{\delta}_k^T\mathbf{y}, \boldsymbol{\delta}_k^T\hat{\mathbf{p}}_o) \\
&\approx \boldsymbol{\delta}_k^T\mathrm{Var}(\mathbf{y})\frac{\partial \hat{\mathbf{p}}_o}{\partial \mathbf{y}^T}\boldsymbol{\delta}_k
\end{aligned} \tag{C.13}$$

Using the chain rule, the partial derivative used above can be obtained as

$$\frac{\partial \hat{\mathbf{p}}_o}{\partial \mathbf{y}^T} = \frac{\partial \hat{\mathbf{p}}_o}{\partial \hat{\boldsymbol{\beta}}_o^T} \frac{\partial \hat{\boldsymbol{\beta}}_o}{\partial \tilde{\mathbf{y}}^T} \frac{\partial \tilde{\mathbf{y}}}{\partial \mathbf{y}^T} \tag{C.14}$$

where

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{\beta}}_o}{\partial \mathbf{y}} &= \begin{bmatrix} \frac{\partial \alpha_o}{\partial \tilde{\mathbf{y}}^T} \frac{\partial \tilde{\mathbf{y}}}{\partial \mathbf{y}^T} \\ \frac{\partial \hat{\boldsymbol{\beta}}_c}{\partial \tilde{\mathbf{y}}^T} \frac{\partial \tilde{\mathbf{y}}}{\partial \mathbf{y}^T} \end{bmatrix} \\ &= \begin{bmatrix} \left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{oij} v_{oij}\right)^{-1} \mathbf{1}^T \mathbf{W} \\ \left[\left(\tilde{\mathbf{X}}^T \tilde{\mathbf{V}} \mathbf{W} \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \mathbf{W}\right]_c \end{bmatrix} \mathbf{S}^T. \end{aligned} \tag{C.15}$$

C.1.3 Full-Model Adjusted Population Averaged Proportion

When logistic regression is performed using EMMW or MI, the full-model adjusted population averaged proportion can be calculated as

$$\hat{P}_k = (O_k/E_k)\bar{p} \tag{C.16}$$

where E_k is obtained using C.3, and $\bar{p} = \mathbf{1}^T \mathbf{y}$.

C.1.4 Variance of the Population Averaged Proportion

With missing covariate data, the variance of the full-model adjusted population averaged proportion can be approximated by taking the variance of the first-order Taylor series expansion:

$$\begin{aligned} \mathrm{Var}(\hat{P}_k) \approx & \left(\frac{\bar{p}}{E_k}\right)^2 \mathrm{Var}(O_k) + \left(\frac{O_k}{E_k^2}\bar{p}\right)^2 \mathrm{Var}(E_k) + \left(\frac{O_k}{E_k}\right)^2 \mathrm{Var}(\bar{p}) \\ & - 2\left(\frac{O_k}{E_k^2}\bar{p}\right)\left[\left(\frac{O_k}{E_k}\right)\mathrm{Cov}(E_k, \bar{p}) + \left(\frac{\bar{p}}{E_k}\right)\mathrm{Cov}(O_k, E_k) - \mathrm{Cov}(O_k, \bar{p})\right]. \end{aligned} \tag{C.17}$$

Estimates for $\mathrm{Var}(O_k)$, $\mathrm{Var}(E_k)$ and $\mathrm{Cov}(O_k, E_k)$ are presented in C.6, C.8, and C.13. Estimates for the other terms can be obtained by substituting $\widehat{\mathrm{Var}}(\mathbf{y})|\mathbf{X}_{obs}$ for $\mathrm{Var}(\mathbf{y})$ in the following expressions

$$\begin{aligned}
\mathrm{Cov}(\bar{p}, E_k) &= \mathrm{Cov}(n^{-1}\mathbf{1}^T\mathbf{y}, \boldsymbol{\delta}_k^T\hat{\mathbf{p}}_o) \\
&\approx n^{-1}\mathbf{1}^T\mathrm{Var}(\mathbf{y})\frac{\partial p_o}{\partial \mathbf{y}^T}\boldsymbol{\delta}_k
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(O_k, \bar{p}) &= \mathrm{Cov}(\boldsymbol{\delta}^T\mathbf{y}, n^{-1}\mathbf{1}^T\mathbf{y}) \\
&= n^{-1}\boldsymbol{\delta}_k^T\mathrm{Var}(\mathbf{y})\boldsymbol{\delta}_k
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\bar{p}) &= \mathrm{Var}(n^{-1}\mathbf{1}^T\mathbf{y}) \\
&= n^{-2}\mathbf{1}^T\mathrm{Var}(\mathbf{y})\mathbf{1}.
\end{aligned}$$

C.2 Directly Standardized Rates

As in the case where covariates are completely observed, directly standardized rates are obtained by estimating the expected number of deaths that would have occurred if all cases in the population of interest were treated by a given provider. Let $\tilde{\mathbf{C}}_k$ be an $\tilde{n} \times (k-1)$ matrix in which each row contains the codes associated with treatment by provider k. Let $\tilde{\mathbf{X}}_k = (\tilde{\mathbf{X}}_o, \tilde{\mathbf{C}}_k)$ be the augmented covariate matrix associated with treatment by provider k. Let $\hat{\boldsymbol{\beta}}$ be the $(q+k)$ length vector of regression coefficients obtained by regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}$ using EMMW. Define

$$\hat{\boldsymbol{\eta}}_k = \tilde{\mathbf{X}}_k\hat{\boldsymbol{\beta}} \tag{C.18}$$

with individual elements $\hat{\eta}_{ijk}$. For each case $i = 1, 2, \ldots, n$, there are $j = 1, \ldots, n_i$ possible covariate combinations. The estimated probability of death for the covariate

combination j for individual i treated by provider k is

$$\hat{p}_{ijk} = \frac{\exp(\hat{\eta}_{ijk})}{1+\exp(\hat{\eta}_{ijk})}.$$

Let $\bar{\mathbf{p}}_k$ be a vector of length n with elements

$$\bar{p}_{ik} = \sum_{j=1}^{n_i} \hat{p}_{ijk} w_{ij}.$$

The expected number of deaths if the population had been treated by provider k is

$$\begin{aligned} E_k^* &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}\hat{p}_{ijk} \\ &= \sum_{i=1}^{n} \bar{p}_{ik}, \end{aligned}$$

where the w_{ij} are the weights obtained from finial iteration of the EMMW algorithm.

C.2.1 Directly Standardized Relative Risk

As in appendix A, the directly standardized relative risk is obtained as

$$\widehat{RR}_k^* = \frac{E_k^*}{O}$$

where $O = \sum_{i=1}^{n} y_i$.

C.2.2 Variance of the Standardized Risk Ratio

A first-order approximation to the variance of the standardized risk is

$$\begin{aligned} \mathrm{Var}(\widehat{RR}_k) &= \mathrm{Var}(E_k^*/O) \\ &\approx \frac{1}{O^2}\left[\left(\frac{E_k}{O}\right)^2 \mathrm{Var}(O) + \mathrm{Var}(E_k) - 2\left(\frac{E_k}{O}\right)\mathrm{Cov}(E_k, O)\right]. \quad \text{(C.19)} \end{aligned}$$

The following are expressions for the variances and covariances used above.

Variance of O

The variance of O is

$$\begin{aligned}\widehat{\mathrm{Var}}(O) &= \mathrm{Var}(\mathbf{1^T y}) \\ &= \mathbf{1}^T \mathrm{Var}(\mathbf{y})\mathbf{1}.\end{aligned}$$

An estimate of $\mathrm{Var}(O)$ can be obtained by substituting $\widehat{\mathrm{Var}}(\mathbf{y}|\mathbf{X}_{obs})$ for $\mathrm{Var}(\mathbf{y})$.

Variance of E_k.

The variance of E_k is

$$\begin{aligned}\mathrm{Var}(E_k) &= \mathrm{Var}(\tilde{\mathbf{1}}^T \mathbf{W} \hat{\mathbf{p}}_k) \\ &\approx \tilde{\mathbf{1}}^T \mathbf{W} \frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T} \mathrm{Var}(\hat{\boldsymbol{\beta}}) \frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}} \mathbf{W} \tilde{\mathbf{1}}.\end{aligned}$$

This can be estimated by using the $\widehat{\mathrm{Var}}(\hat{\beta})$ obtained from Louis's method (see appendix B). Applying the chain rule, the partial derviative used above can be expressed as

$$\begin{aligned}\frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T} &= \frac{\partial \hat{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\eta}}_k^T} \frac{\partial \hat{\boldsymbol{\eta}}_k}{\partial \hat{\boldsymbol{\beta}}^T} \\ &= \tilde{\mathbf{V}}_k \tilde{\mathbf{X}}_k\end{aligned}$$

Covariance of O and E_k^*

A first order approximation of the covariance of O and E_k is

$$\begin{aligned}\mathrm{Cov}(O, E_k) &= \mathrm{Cov}(\mathbf{1}^T \mathbf{y}, \mathbf{1}^T \bar{\mathbf{p}}_k) \\ &\approx \mathbf{1}^T \mathrm{Var}(\mathbf{y}) \frac{\partial \bar{\mathbf{p}}_k}{\partial \mathbf{y}} \mathbf{1}_k.\end{aligned}$$

An estimate of $\text{Cov}(O, E_k)$ can be obtained by substituting $\widehat{\text{Var}}(\mathbf{y})|\boldsymbol{X}_{obs}$ for $\text{Var}(\mathbf{y})$. By the chain rule, the derivative above can be expressed as

$$\begin{aligned}\frac{\partial \bar{\mathbf{p}}_k}{\partial \mathbf{y}^T} &= \frac{\partial \bar{\mathbf{p}}_k}{\partial \tilde{\mathbf{p}}_k}\frac{\partial \tilde{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^T}\frac{\partial \mathbf{y}}{\partial \tilde{\mathbf{y}}^T} \\ &= \mathbf{S}\mathbf{W}\tilde{\mathbf{V}}_k\tilde{\mathbf{X}}_k\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\mathbf{S}^T,\end{aligned}$$

Where $\tilde{\mathbf{V}}_k$ is a diagonal matrix with elements $\hat{p}_{ijk}(1-\hat{p}_{ijk})$.

C.2.3 Population Averaged Proportion

A directly standardized estimate of the population averaged proportion is

$$\hat{P}_k = \frac{E_k}{n}.$$

Variance of the Population averaged proportion

The variance of the population averaged proportion is

$$\text{Var}(\hat{P}_k) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n}\bar{p}_{ik}\right).$$

This variance can be approximated as

$$\approx \frac{1}{n^2}\mathbf{1}^T\frac{\partial \bar{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T}\text{Var}(\hat{\boldsymbol{\beta}})\frac{\partial \bar{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}}\mathbf{1} \tag{C.20}$$

where

$$\begin{aligned}\frac{\partial \bar{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T} &= \frac{\partial \bar{\mathbf{p}}_k}{\partial \tilde{\mathbf{p}}_k}\frac{\partial \tilde{\mathbf{p}}_k}{\partial \hat{\boldsymbol{\beta}}^T} \\ &= \mathbf{S}\mathbf{W}\tilde{\mathbf{V}}_k\tilde{\mathbf{X}}_k.\end{aligned}$$

The $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ from Louis's method can be used in C.20 to obtain an estimate of $\text{Cov}(O, E_k)$.

Appendix D

Taylor Series Approximation for the Baseline Model

Let y be a binary response vector for n patients $i = 1, 2, \ldots, n$. Indicator vectors $\boldsymbol{\delta}_k$ of length n will be used to denote providers $k = 1, 2, \ldots, l$. The elements of $\boldsymbol{\delta}_k$ equal 1 if case i is a has been treated by provider h_k and 0 otherwise. Let $\mathbf{X} = (\mathbf{1}, \mathbf{X}_c)$, where $\mathbf{1}$ is a unit vector of length n and $\mathbf{X}_c$ is an $n \times q$ matrix containing the q covariates of interest, but excluding the covariates used to code for treatment provider. Rows of $\mathbf{X}$ will be denoted as x_i^T. To obtain baseline-model adjusted SMRs, logistic regression is used to fit $\mathbf{y}$ to $\mathbf{X}$. The resulting $(q+1)$ length vector of regression parameter estimates will be denoted as $\hat{\boldsymbol{\beta}}$. From this model, we obtain $\hat{\mathbf{p}}$, an n length vector of estimates of the probability of death. The element of $\hat{\mathbf{p}}$ for case i is

$$\hat{p_i} = \frac{\exp(\hat{\eta})_i}{1 + \exp(\hat{\eta}_i)}, \tag{D.1}$$

where $\hat{\eta}_i = x_i^T \hat{\boldsymbol{\beta}}$. From these estimates, we obtain the diagonal matrix

$$\mathbf{V} = \widehat{\mathrm{Var}}(\mathbf{y}), \tag{D.2}$$

with diagonal elements $v_i = \hat{p}_i(1 - \hat{p}_i)$. The baseline-model adjusted estimate of the expected number of deaths for provider k is then

$$E_k = \boldsymbol{\delta}_k^T \hat{\mathbf{p}} \tag{D.3}$$

and the observed number of deaths for provider k is

$$O_k = \boldsymbol{\delta}_k^T \mathbf{y} \tag{D.4}$$

The full-model adjusted SMR is

$$SMR_k = \frac{O_k}{E_k}. \tag{D.5}$$

D.1 Variance of the Baseline-Model SMR

By first order Taylor Series expansion,

$$\begin{aligned} \mathrm{Var}(SMR_k) &= \mathrm{Var}(O_k/E_k) \\ &\approx \frac{1}{E_k^2}\left[\left(\frac{O_k}{E_k}\right)^2 \mathrm{Var}(E_k) + \mathrm{Var}(O_k) - 2\left(\frac{O_k}{E_k}\right)\mathrm{Cov}(O_k, E_k)\right] \end{aligned} \tag{D.6}$$

The following are expressions for the variances and covariance used in obtaining the variance of the SMR:

Variance of O_k

$$\begin{aligned} \mathrm{Var}(O_k) &= \mathrm{Var}(\boldsymbol{\delta}_k^T \mathbf{y}) \\ &= \boldsymbol{\delta}_k^T \mathrm{Var}(\mathbf{y}) \boldsymbol{\delta}_k \end{aligned} \tag{D.7}$$

which can be estimated by substituting $\mathbf{V}$ for $\mathrm{Var}(\mathbf{y})$.

Variance of E_k

$$\begin{aligned} \mathrm{Var}(E_k) &= \boldsymbol{\delta}_k^T \mathrm{Var}(\hat{\mathbf{p}}) \boldsymbol{\delta}_k \\ &\approx \boldsymbol{\delta}_k^T \frac{\partial \hat{\mathbf{p}}}{\partial \hat{\boldsymbol{\beta}}^T} \mathrm{Var}(\hat{\boldsymbol{\beta}}) \frac{\partial \hat{\mathbf{p}}}{\partial \hat{\boldsymbol{\beta}}} \boldsymbol{\delta}_k \end{aligned} \tag{D.8}$$

Let $\hat{\boldsymbol{\eta}}$ be an n length vector with elements $\hat{\eta}_i$. The partial derivative used above can be expressed as

$$\begin{aligned} \frac{\partial \hat{\mathbf{p}}}{\partial \hat{\boldsymbol{\beta}}^T} &= \frac{\partial \hat{\mathbf{p}}}{\partial \hat{\boldsymbol{\eta}}^T} \frac{\partial \hat{\boldsymbol{\eta}}}{\partial \hat{\boldsymbol{\beta}}^T} \\ &= \mathbf{VX}. \end{aligned} \tag{D.9}$$

By using $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}$, the variance of E_k can be estimated as

$$\widehat{\text{Var}}(E_k) = \boldsymbol{\delta}_k^T\mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}\boldsymbol{\delta}_k. \tag{D.10}$$

Covariance of O_k and E_k

The covariance of O_k and E_k can be approximated as

$$\begin{aligned}\text{Cov}(O_k, E_k) &= \text{Cov}(\boldsymbol{\delta}_k^T\mathbf{y}, \boldsymbol{\delta}_k^T\hat{\mathbf{p}}) \\ &\approx \boldsymbol{\delta}_k^T\text{Var}(\mathbf{y})\frac{\partial\hat{\mathbf{p}}}{\partial\mathbf{y}}\boldsymbol{\delta}_k.\end{aligned} \tag{D.11}$$

By the chain rule, the partial derivative used above can be expressed as

$$\frac{\partial\hat{\mathbf{p}}}{\partial\mathbf{y}^T} = \frac{\partial\hat{\mathbf{p}}}{\partial\hat{\boldsymbol{\beta}}^T}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{y}^T}$$

Noting that $\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{y}^T} = \left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T$,

$$\frac{\partial\hat{\mathbf{p}}}{\partial\mathbf{y}^T} = \mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T.$$

Using $\widehat{\text{Var}}(y) = \mathbf{V}$, the covariance of O_k and E_k can be estimated as

$$\widehat{\text{Cov}}(O_k, E_k) = \delta_k^T\mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}\delta_k. \tag{D.12}$$

Noting that this estimate is equal to the estimate of $\text{Var}(E_k)$ from D.10, an estimate of the variance of SMR_k is

$$\begin{aligned}\widehat{\text{Var}}(SMR_k) &= \frac{1}{E_k^2}\left[(SMR_k(SMR_k - 2))\,\widehat{\text{Var}}(E_k) + \widehat{\text{Var}}(O_k)\right] \\ &= \frac{1}{E_k^2}\boldsymbol{\delta}_k^T\left[(SMR_k(SMR_k - 2))\,\mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V} + \mathbf{V}\right]\boldsymbol{\delta}_k.\end{aligned} \tag{D.13}$$

Bibliography

A.S. Ash and M. Shwartz. Evaluating the performance of risk-adjustment methods: Dichotomous ourcomes. In L.I. Iezzoni, editor, *Risk Adjustment for Measuring Outcomes*, chapter 9, pages 427–69. Health Administrative Press, Chicago, 2nd. edition, 1997.

M.H. Beers, M. Munekata, and M. Storrie. The accuracy of medication histories in the hospital medical records of elderly persons. *Journal of the American Medical Association*, 38(11):1183–87, 1989.

Y.M.M Bishop, S.E. Fienberg, and P.W. Holland. *Discrete Multivariate Analysis: Theory and Practice*, chapter Maximum likelihood estimates for complete tables, page 557. The MIT Press, Cambridge, Massachusetts, 1975.

M.S. Blumberg. Risk adjusting health care outcomes. *Medical Care Review*, 43(2): 351–93, 1986.

J.P.L. Brand. *Development, Implementation and Evaluation of Multiple Imputation Stategies for the Statistical Analysis of Incomplete Data Sets.* PhD thesis, University of Rotterdam, 1999.

R. Brant and R. Tibshirani. Missing covariate values in generalized regression models. *University of Toronto Technical Reports*, 1991.

N.E. Breslow and N.E. Day. Fundamental measures of disease occurrence and association. In *Statistical Methods in Cancer Research: Volume 1 - the Analysis of Case-Control Studies*, pages 42–81. IARC, Lyon, 1980.

N.E. Breslow and N.E. Day. Fitting models to grouped data. In *Statistical Methods in Cancer Research: Volume II - the Design and Analysis of Cohort Studies*, pages 120–76. IARC, Lyon, 1987a.

N.E. Breslow and N.E. Day. Rates and rate standardization. In *Statistical Methods in Cancer Research: Volume II - the Design and Analysis of Cohort Studies*, pages 48–79. IARC, Lyon, 1987b.

S.F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. Roy. Statist. Soci*, B22:302–6, 1960.

J. Cornfield, W. Haenszel, A.M. Lilienfeld, M.B. Shimkin, and E.L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

A. P. Dempster, N. M. Laird, and D.B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, B, 39:1–38, 1977.

A.K. Duggan, B. Starfield, and C. DeAngelis. The impact on provider performance and recording of well-child care. *Pediatrics*, 85(1):104–13, 1990.

A. Epstein. Performance reports on quality-prototypes, problems and prospects. *New England Journal of Medicine*, 333:57–61, 1995.

Free Software Founation Inc. *g77 - GNU Project Fortran Compiler (V0.5.24)*, 1998.

W. Ghali, H. Quan, and R. Brant. Coronary artery bypass graft surgery in Canada: National and provincial mortality trends, 1992-1995. *Canadian Medical Association Journal*, 159(1):35–31, 1998.

G. Gong and F.J. Samaniego. Pseudo maximum likelihood estimation: Theory and application. *Annals of Statistics*, 9:861–869, 1981.

J. Green and N. Wintfeld. Report cards on cardiac surgeons-assessing New York State's approach. *New England Journal of Medicine*, 332:1229–1232, 1995.

S. Greenland. Introduction to regression models. In K.J. Rothman and S. Greenland, editors, *Modern Epidemiology*, pages 359–400. Lippincott-Raven, Philadelphia, PA, 2nd edition, 1998.

S. Greenland and W. D. Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–64, 1995.

S. Greenland and K.J. Rothman. Introduction to categorical statistics. In K.J. Rothman and S. Greenland, editors, *Modern Epidemiology*, pages 231–52. Lippincott-Raven, Philadelphia, PA, 2nd edition, 1998a.

S. Greenland and K.J. Rothman. Introduction to stratified analysis. In K.J. Rothman and S. Greenland, editors, *Modern Epidemiology*, pages 253–279. Lippincott-Raven, Philadelphia, PA, 2nd edition, 1998b.

S. Greenland and K.J. Rothman. Measures of effect and measures of association. In K.J. Rothman and S. Greenland, editors, *Modern Epidemiology*, pages 47–64. Lippincott-Raven, Philadelphia, PA, 2nd edition, 1998c.

F. E. Jr. Harrell and K. L. Lee. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3:143–52, 1984.

Ibrahim. Incomplete data in generalized linear models. *JASA*, 85(411):765–769, 1990.

L.I. Iezzoni. Risk adjustment for medical effectiveness research: An overview of conceptual and methodological considerations. *Journal of Investigative Medicine*, 43(2):136–50, April 1995.

L.I. Iezzoni, M.A. Shwartz, A.S. Moskowitz, E.S. Ash, and S. Burnside. Illness severity and costs of admissions at teaching and nonteaching hospitals. *JAMA*, 264(11):1426–31, 1990.

J.P. Kassirer. The use and abuse of practice profiles. *New England Journal of Medicine*, 330:634–635, 1994.

J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D. Thompson. Biological and statistical concepts. In *Methods in Observational Epidemiology*, pages 22–44. Oxford University Press, New York, 2nd edition, 1996a.

J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D. Thompson. Cohort studies: Statistical analysis II. In *Methods in Observational Epidemiology*, pages 167–87. Oxford University Press, New York, 2nd edition, 1996b.

J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D. Thompson. Cohort studies: Statistical analysis I. In *Methods in Observational Epidemiology*, pages 131–66. Oxford University Press, New York, 2nd edition, 1996c.

J.M. Last, editor. *A Dictionary of Epidemiology*. Oxford University Press, New York, 2nd edition, 1988.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data.* Wiley, New York, 1989a.

R.J.A. Little and D.B Rubin. *Statistical Analysis With Missing Data*, chapter 5, pages 79–96. Wiley, New York, 1989b.

R.J.A. Little and D.B Rubin. *Statistical Analysis With Missing Data*, chapter 3, pages 39–49. Wiley, 1989c.

R.J.A. Little and D.B Rubin. *Statistical Analysis With Missing Data*, chapter 4, pages 50–75. Wiley, 1989d.

R.J.A. Little and D.B Rubin. *Statistical Analysis With Missing Data*, chapter 4, pages 50–75. Wiley, New York, 1989e.

R.J.A. Little and D.B Rubin. *Statistical Analysis With Missing Data*, chapter 7, pages 127–141. Wiley, New York, 1989f.

R.J.A. Little and M.D. Schluchter. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72:497–512, 1985.

T.A. Louis. Finding the observed information when using the EM algorithm. *Journal of the Royal Statistical Society*, B44:226–33, 1982.

MathSoft. *S-PLUS 5 for UNIX.* Data Analysis Products Division, Seattle, WA, 1999.

P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall, 1989.

X.L. Meng and D.B. Rubin. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79:103–11, 1992.

O.S. Miettinen. *Theoretical Epidemiology.* Wiley, New York, 1985.

C.M. Norris, W.A. Ghali, M.L. Knudtson, C.D. Naylor, and D.L. Saunders. Dealing with missing data in observational health care outcomes analyses. *Statistics in Medicine*, 1999.

R. E. Park, R. H. Brook, J. Kosecoff, J. Keesey, L. Rubenstein, E. Keeler, K. L. Kahn, W. H. Rogers, and M. R. Chassin. Explaining variations in hospital death rates. randomness, severity of illness, quality of care [see comments]. *JAMA*, 264: 484–90, 1990.

M.S. Pepe and T.R Fleming. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113, 1991.

M. Reilly and M. S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

K.J. Rothman. Synergy and antagonism in cause-effect relationships. *American Journal of Epidemiology*, 99:385–388, 1974.

K.J. Rothman. The estimation of synergy or antagonism. *American Journal of Epidemiology*, 103:506–11, 1976.

K.J. Rothman. Measures of effect. In *Modern Epidemiology*, pages 35–40. Little, Brown, Boston, 1986a.

K.J. Rothman. Standardization of rates. In *Modern Epidemiology*, pages 41–9. Little, Brown, Boston, 1986b.

K.J. Rothman, S. Greenland, and A.M. Walker. Concepts of interaction. *American Journal of Epidemiology*, 112(4):467–70, 1980.

D. R. Rubin. Inference and missing data. *Biometrika*, 63(3):581–92, 1976.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*, chapter 1, pages 1–26. Wiley, New York, 1987a.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*, chapter 3, pages 75–112. Wiley, New York, 1987b.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*, chapter 4, pages 113–153. Wiley, New York, 1987c.

W. Sarle. Re: SAS macros for EM and data augmentation. Imputations in Data Analysis <IMPUTE@LISTSERV.NODAK.EDU>, June 1999. Communication on the IMPUTE mailing list.

J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London, first edition, 1997a.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 4, pages 89–146. Chapman and Hall, London, first edition, 1997b.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 3, pages 37–87. Chapman and Hall, London, first edition, 1997c.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 6, pages 193–238. Chapman and Hall, London, first edition, 1997d.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 8, pages 289–331. Chapman and Hall, London, first edition, 1997e.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 9, pages 333–377. Chapman and Hall, London, first edition, 1997f.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*, chapter 7, pages 239–287. Chapman and Hall, London, first edition, 1997g.

J.L. Schafer. Software for multiple imputation, July 1999. From *http://www.stat.psu.edu/~jls/misoftw.html.*

J. Schimert. Re: SAS macros for em and data augmentation. Imputatons in Data Analysis <IMPUTE@LISTSERV.NODAK.EDU>, June 1999. Communication on the IMPUTE mailing list.

M. Shwartz, A.S. Ash, and L.L. Iezzoni. Comparing outcomes across providers. In L. I. Iezzoni, editor, *Risk Adjustment for Measuring Healthcare Outcomes.* Health Administration Press, Chicago, second edition, 1997.

Statistical Solutions. *SOLAS for Missing Data Analysis 2.0.* Saugus, MA, 1999.

W. Vach. *Logistic regression with missing values in the covariates.* Springer, New York, 1994.

W. Vach and M. Blettner. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134(8):895–907, 1991.

A.S. Whittemore and S. Grosser. Modern statistical methods in disease epidemiology. chapter Regression methods for data with incomplete covariates, pages 19–34. 1986.