
Health Services & Outcomes Research Methodology 2:291–315, 2001

Examining the Impact of Missing Data on Propensity Score Estimation in Determining the Effectiveness of Self-Monitoring of Blood Glucose (SMBG)

RALPH D'AGOSTINO JR., WEI LANG, MICHAEL WALKUP, AND TIMOTHY MORGAN
Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Winston-Salem, NC

ANDREW KARTER
Kaiser Permanente, Oakland, CA

Received December 12, 2000; revised December 20, 2001; accepted December 21, 2001

Abstract. In many health services applications, research to determine the effectiveness of a particular treatment cannot be carried out using a controlled clinical trial. In settings such as these, observational studies must be used. Propensity score methods are useful tools to employ in order to balance the distribution of covariates between treatment groups and hence reduce the potential bias in treatment effect estimates in observational studies. A challenge in many health services research studies is the presence of missing data among the covariates that need to be balanced. In this paper, we compare three simple propensity models using data that examine the effectiveness of self-monitoring of blood glucose (SMBG) in reducing hemoglobin A1c in a cohort of 10,566 type 2 diabetics. The first propensity score model uses only subjects with complete case data ($n = 6{,}687$), the second incorporates missing value indicators into the model, and the third fits separate propensity scores for each pattern of missing data. We compare the results of these methods and find that incorporating missing data into the propensity score model reduces the estimated effect of SMBG on hemoglobin A1c by more than 10%, although this reduction was not clinically significant. In addition, beginning with the complete data, we artificially introduce missing data using a nonignorable missing data mechanism and compare treatment effect estimates using the three propensity score methods and a simple analysis of covariance (ANCOVA) method. In these analyses, we find that the complete case analysis and the ANCOVA method both perform poorly, the missing value indicator model performs moderately well, and the pattern mixture model performs even better in estimating the original treatment effect observed in the complete data prior to the introduction of artificial missing data. We conclude that in observational studies one must not only adjust for potentially confounding variables using methods such as propensity scores, but one should also account for missing data in these models in order to allow for causal inference more appropriately to be applied.

Keywords: missing data, pattern mixture models, propensity scores, self-monitoring of blood glucose, diabetes

1. Introduction

In many health services applications, research to determine the effectiveness of a particular treatment cannot be carried out using a controlled clinical trial. In settings such as these, observational studies must be used to make inference concerning the effectiveness of a particular non-randomized treatment. The treated and non-treated (i.e., control) groups in these observational studies may have substantial differences in observed covariates, and

these differences can lead to biased estimates of treatment effects unless properly handled. Propensity score methods are popular tools used for balancing the distribution of the covariates in the two groups to reduce this bias. This method has been shown to confer a greater reduction in bias than standard adjustment methods, such as ANCOVA, in many circumstances. (D'Agostino Jr, 1998; Rubin and Thomas, 2000) In order to estimate the propensity score, defined as the conditional probability of being treated given the observed covariates, we must model the distribution of the treatment indicator given the observed covariates. An additional complication that often occurs, particularly when using health services administrative data, is that missing data may be present among the covariates and in some cases the pattern of the missing covariates may be prognostically important. When this occurs, the propensity score needs to be modeled conditional on both the observed values of the covariates and the patterns of missing data.

The goal of this paper is to compare several propensity score models that were fit using real data to examine the efficacy of self-monitoring of blood glucose (SMBG) for a diabetic population in which the covariate information is missing for many participants. We compare three propensity score models; the first is based on using only participants that had complete covariate data and the other two consider straightforward approaches for handling missing data that require very little statistical programming and can be implemented using standard statistical packages such as SAS. More sophisticated and computer intensive approaches for handling missing data are not included in this paper although we briefly describe two of these. The first is based on a method we recently described in the literature (D'Agostino Jr, and Rubin, 2000) that uses a model-based approach for handling missing data and the second uses imputation techniques to handle missing data. (Robins and Wang, 2000; Allison, 2000; Rubin, 1987)

In addition to comparing the methods using real data, we compare the three propensity score methods and a standard ANCOVA adjustment approach for estimating treatment effects using the complete case data after we artificially introduce missing data. Since we know the "true" treatment effect in this example, we are able to determine more accurately which method performs best.

The next section of this paper presents details of the real data example that examines the effectiveness of self-monitoring of blood glucose. This is followed by some additional background concerning propensity score methods including notation. We then present the methods used for analyzing this data followed by the results from our three propensity score models and the comparison of methods using the data with artificially created missing values. A brief discussion follows, and the paper concludes with a description of current and future methodologic research topics in this area.

2. Background of Self-Monitoring of Blood Glucose (SMBG)

Self-monitoring of blood glucose (SMBG) is considered one of the cornerstones of diabetes care and is widely recommended, (American Diabetes Association, 1999) despite the lack of evidence regarding the effectiveness of SMBG in improving glycemic control. Proponents consider SMBG useful for achieving and maintaining near-normal blood

glucose levels, providing feedback to the health care provider and patient regarding therapeutic effectiveness, helping patients adjust insulin dosaging, diet and exercise regimens, and aiding in detection and prevention of asymptomatic hypoglycemia and extreme hyperglycemia. (American Diabetes Association, 1994, 1998; Faas, Schellevis and Van Eijk, 1997) Given the growing clinical consensus that SMBG is an important component of care, the American Diabetes Association (ADA) Provider Recognition Program (co-sponsored by the National Committee for Quality Assurance) now recognizes the proportion of diabetic patients that perform SMBG as a performance measure when assessing the quality of managed care plans.

Currently, there is little evidence regarding the effectiveness of this very costly practice, especially for those persons with type 2 diabetes treated with oral agents only. (Faas, Schellevis and Van Eijk, 1997; Evans, Newton and Ruta et al., 1999) Given that SMBG is universally recommended for pharmacologically-treated diabetic patients, randomized assignment to SMBG is no longer ethical, leaving observational assessment as the only option. In order to study the effectiveness of SMBG, we examine data from the Northern California Kaiser Permanente Diabetes Registry (Martin, Selby and Zhang, 1995; Selby, Ray and Zhang et al., 1997; Selby, Ettinger and Swain et al., 1999; Karter, Ferrara and Darbinian et al., 2000) that was collected on 10,566 persons with diabetes between 1994 and 1997 who treated their diabetes using oral agents. The goal of this research is to compare methodologic approaches that will allow investigators to study the effectiveness of the use of SMBG as recommended by American Diabetes Association Clinical Practice Recommendations (American Diabetes Association, 1999).

The data used in these analyses are derived from health surveys (83% response rate) and Kaiser Permanente administrative records. Covariate data are missing for 3,879 (37%) participants. The exposure of interest is the average number of glucometer strips redeemed at the Kaiser Permanente pharmacies during a 12 month baseline period (1/1/96–12/31/96). Using this measure of strip utilization (U), patients are dichotomized into those who perform SMBG within ADA guidelines (at least once daily) and those who do not. To accommodate for occasional missed monitoring days, patients are categorized as monitoring daily if $U \geq 0.75$ per day versus less than daily if $0.75 > \mathrm{U}$ per day. This cutpoint for SMBG use is based on previous research in the area of utilization of daily monitoring techniques. (Karter, Ferrara and Darbinian et al., 2000; Karter, Ackerson and Darbinian et al., 2001) The outcome of interest, glycemic control, is based on the latest laboratory measure of hemoglobin A1c (HbA1c) ascertained during the year subsequent to the measurement of strip utilization (1997). Normal values of HbA1c range between 4–6.8% with higher values of HbA1c indicating that a person is not in glycemic control.

In addition to the measure of SMBG and HbA1c, there are nine other patient characteristics that we consider in our analyses. The first seven include: gender, smoking status (current, former, or never), drinking status (abstain, < 3 drinks per day, 3+ drinks per day), education (high school or less, some college, college graduate or more), race (Asian Pacific, Black, Caucasian, Hispanic, Native American, Multi-ethnic (2 or more of the previous) and other), participant's age, and duration of diabetes. The last two are measures of neighborhood level socioeconomic status (SES) that were obtained by geocoding each

member's address and mapping it to its census (1990) block group and determining the associated average annual per capita income and proportion in a working class profession.

In this study since participants are not randomly assigned to the treatment (using SMBG within guidelines—SMBG adherent group) or control groups (SMBG non-adherent group), large differences may exist between treated and control groups on observed covariates, which can lead to biased estimates of treatment effects.

3. Propensity Score Background

We propose to use propensity score methods to balance the distribution of the covariates in the two groups to reduce this imbalance using a weighted analysis based on subclassification defined by propensity scores. In order to estimate propensity scores, which are the conditional probabilities of being treated (being in the SMBG adherent group) given a vector of observed covariates, we will model the distribution of the treatment indicator given these observed covariates. We also specifically address the problem of calculating propensity scores when covariates have missing values.

Since their introduction, propensity scores have been used in observational studies in many fields to adjust for imbalances on pretreatment covariates, X, between a treated group, indicated by $Z = 1$, and a control group indicated by $Z = 0$. (D'Agostino Jr., 1998; Rosenbaum and Rubin, 1983; Rubin, 1997) Propensity scores are a one-dimensional summary of multidimensional covariates, X, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates, X, are balanced in expectation across the two groups. Typically, matched sampling (Heckman, Ichimura and Smith et al., 1996; Lytle, Blackstone and Loop et al., 1999; Rosenbaum and Rubin, 1985) or subclassification (Barker 2nd, Chang and Gutin et al., 1998; Conners Jr., Speroff and Dawson et al., 1996; Rosenbaum and Rubin, 1984) on estimated propensity scores is used, often in combination with model-based adjustments (Lieberman, Cohen and Lang et al., 1996; Rich, 1998; Gu and Rosenbaum, 1993).

The propensity score for an individual is the probability of being treated conditional on the individual's covariate values. To estimate propensity scores for individuals, one must model the distribution of Z given the observed covariates, X. There is a large technical literature on propensity score methods with complete data. (Rubin and Thomas, 1992a, 1992b, 1996; Rubin, 1978) In practice, however, typically some covariate values will be missing, and so it is not clear how the propensity score should be estimated. Often, the missingness itself may be predictive about which treatment is received in the sense that the treatment assignment mechanism is ignorable (Rubin, 1986) given the observed values of X and the observed pattern of missing covariates but not ignorable given only the former.

In this paper we consider three approaches to handling the missing data. The first approach simply ignores the missing data and employs traditional complete data techniques using only participants that contain fully observed covariates. The second approach we consider is to include indicator variables for the covariates that contain missing values and then fit a model that includes both the observed data and the missing value indicators. For categorical variables this second approach is equivalent to adding an extra category for

"missing" and for continuous variables this approach implies creating an additional indicator variable for the missing values and then setting the covariate values to zero where missing data had existed.[1] This approach has an advantage of using the entire sample of participants in the analyses, including those with missing covariate data. The third approach we consider uses a "pattern mixture" model (Rosenbaum and Rubin, 1984; Little, 1993) for propensity score estimation with missing covariate data. For this approach (Rosenbaum and Rubin, 1984) a "generalized" propensity score is defined as the probability of treatment assignment given both the observed covariates and the pattern of missing data. We estimate this generalized propensity score by estimating a separate propensity score model using the subset of covariates fully observed for each pattern of missing data. In ideal situations, we would estimate a separate propensity score model for each specific pattern of missing data which could be interpreted as fitting a stratified model that stratifies the overall propensity score model by the missing data pattern. However, in many applied examples where there are not enough observations available to estimate separate propensity scores for each pattern of missing data, one may need to use methods that smooth across patterns of missing data and therefore estimate propensity scores using fewer parameters (Olkin and Tate, 1961). These smoothing methods would employ the general location model (D'Agostino Jr. and Rubin, 2000) to model the relationship between the covariates and the treatment indicator, and then use the EM (Dempster, Laird and Rubin, 1977) and/or ECM (Meng and Rubin, 1993) algorithms to obtain maximum likelihood estimates for the parameters of the model. Alternative smoothing approaches utilize data across several missing data patterns simultaneously to estimate propensity scores for patterns with few observations. In the applied example we present, we use this simple smoothing method to allow estimation of propensity scores for each pattern of missing data. As with the missing indicator approach, the pattern mixture approach allows all participants to be included in the analyses, including those with missing covariate information.

4. Notation

4.1. Propensity Scores with Complete Data

With complete data, the propensity score for subject i ($i = 1, \ldots, N$) is the conditional probability of receiving a particular treatment ($Z_i = 1$) versus control ($Z_i = 0$) given a vector of observed covariates, x_i:

$$e(x_i) = Pr(Z_i = 1 | X_i = x_i), \tag{1}$$

where it is assumed that, given the X's, the Z_i are independent:

$$Pr(Z_1 = z_1, ..., Z_N = z_N | X_1 = x_1, ..., X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i}. \tag{2}$$

For a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, if the treatment assignment is strongly ignorable given the covariates. (Rosenbaum and Rubin, 1983) Thus, matching, subclassification, or covariance adjustment on the propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable. This occurs when the treatment assignment, Z, and the potential outcomes, Y, are conditionally independent given the covariates $X : Pr(Z|X, Y) = Pr(Z|X)$. In the applied example presented in this manuscript, this assumption would suggest that the probability of being adherent to recommended SMBG guidelines (the treatment assignment) is independent of a patient's hemoglobin A1c value (the potential outcome) once we adjust for background characteristics measured on that patient. Whether this assumption is met for this example is a topic that the investigators are still pursuing, as more potential covariates are being added to X in order to strengthen the plausibility for the strong ignorability of the treatment assignment given the expanded set of covariates. For further discussions concerning the importance of the treatment assignment for causal inference, one can refer to recent publications by Rubin (1991, 1997).

In this paper, we use subclassification based on the estimated propensity score to estimate the treatment effect. We now briefly describe this method. Once propensity scores are estimated, we order participants based on their propensity scores from smallest to largest and create deciles of participants. Deciles are chosen since we have such a large data set ($n = 10{,}566$) and we wish to create subclasses of participants with very similar propensity scores, however in many other applications using quintiles may be sufficient for creating similar subclasses. Within each decile, we separate the participants by SMBG status. Next, we assign weights to the participants based on the distribution of SMBG status within each decile. For instance, if there were 100 persons in a decile and 10 were SMBG adherent while 90 were SMBG non-adherent, we would assign a weight of 5 to each of the 10 SMBG adherent participants and a weight of $5/9$ to each of the 90 SMBG non-adherent participants. The formula to determine the weight for each participant is: weight = (total number of participants in decile) $\times$ (0.5)/(number in SMBG group). Using this weighting scheme, the total weight of each decile would be equal and in this example equal to $100((5 \times 10) + (5/9 \times 90) = 100)$.

These weights can be considered as sampling weights rather than precision weights and are used to adjust the distribution of propensity scores between the SMBG adherent and the SMBG non-adherent groups, so that the weighted distribution for the propensity score deciles is the same for the two groups. Thus, in the above illustration, with no weights, the 10 persons within SMBG guidelines would provide 10% ($10/100$) of the weight for the decile and the 90 persons out of SMBG guidelines would provide 90% ($90/100$) of the weight for the decile. In contrast, after using the sampling weights, the 10 persons within guidelines now have 50% of the weight for the decile ($10 \times 5 = 50$—each person now is weighted as 5 people), and the 90 persons out of guidelines would also have 50% of the weight for the decile ($90 \times 5/9 = 50$—each person now is weighted as $5/9$ of a person). In this manner, the participants who are SMBG adherent or non-adherent provide equal weight in estimating treatment effects within each decile based on the estimated propensity

scores. Using these weights in the 10×2 table created by the propensity score deciles is analagous to producing adjusted mean estimates from a general linear model using a technique such as LSMEANS (least-squares means) in SAS. Thus, each participant's HgA1c value is given appropriate weight conditional on their observed covariates, with the result being that certain individuals receive weights larger than or smaller than one depending on their propensity score and SMBG status.

Using these weights from each propensity score model, we then estimate the residual imbalance (differences in covariates) and the treatment effect (difference in the outcome of interest, HbA1c) between treatment groups (SMBG adherent vs. non-adherent).

4.2. Propensity Score Estimation with Incomplete Data

Let the response indicator be $R_{ij}(j = 1, ..., T)$, which is one when the value of the j-th covariate for the i-th subject is observed and zero when it is missing; R_{ij} is fully observed by definition. Also, let $X = (X_{obs}, X_{mis})$, where $X_{obs} = \{X_{ij}|R_{ij} = 1\}$ denotes the observed parts and $X_{mis} = \{X_{ij}|R_{ij} = 0\}$ denotes the missing components of X.

The generalized propensity score for subject i, which conditions on all of the observed covariate information, is

$$e_i^* = e_i^*(X_{obs,i}, R_i) = Pr(Z_i = 1|X_{obs,i}, R_i), \tag{3}$$

With missing covariate data and strongly ignorable treatment assignment given X_{obs} and R, the generalized propensity score, e_i^* in (3), plays the same role as the usual propensity score, e_i in (1) with no missing covariate data (Conners, Speroff and Dawson et al., 1996). Treatment assignment is strongly ignorable given (X_{obs}, R) if $Pr(Z|X, Y, R) = Pr(Z|X_{obs}, R)$. If in addition, the missing data mechanism is such that $Pr(R|X, Z) = Pr(R|X_{obs})$, then $Pr(Z|X, Y, R) = Pr(Z|X_{obs})$, and R itself can be ignored in the modeling. It is important to emphasize that, just as with propensity score modeling with no missing data, the success of a propensity score estimation method is to be assessed by the quality of the balance in the (X_{obs}, R) distributions between the treated and control groups (SMBG adherent/non-adherent) that has been achieved by adjustment using it. Consequently, the usual concerns with the fit of a particular model are not relevant if such balance is achieved.

5. Methods

We fit three propensity score models to the data. The first propensity score model included only participants that had complete data ($n = 6{,}687$). For the categorical covariates with more than two levels, we created indicator variables for the additional levels (i.e., for race we created six indicator variables and used Caucasians as the reference group). The second propensity score model, using all 10,566 participants, added seven missing value indicator variables to the model for the seven variables that included missing values. We imputed a

value of zero for the missing values for each of these covariates. By including the indicator variables with the variables themselves we are able to estimate parameters for the variables based on the observed data and then additional parameters for the missing value indicators. The third propensity score method estimated a separate propensity score model for each of the 36 patterns of missing data. All participants with data observed for a specific pattern were included in the specific model, however only the propensity scores corresponding with the persons with that specific pattern were retained for further analyses. Thus, individuals with complete data were used in all propensity score models, however their propensity scores were based only on the one propensity score model that was fit using those with complete data ("completer"). This approach for estimating the pattern mixture model is one simple form of smoothing across patterns of missing data.

In addition to the above models and in order to understand further the impact missing data has on treatment effect estimates, we introduced missing data into the subset of participants with complete data ($n = 6{,}687$) and then re-estimated the treatment effects using the same three models described above. In addition to the three propensity score models, we fit a simple ANCOVA model using the complete data to estimate the treatment effect. In the ANCOVA model we estimated the SMBG effect adjusting for all nine covariates using the PROC GLM procedure in SAS.

For these models, we created missing data on four of the nine covariates: smoking status, race, age and duration of diabetes. Of the 6,687 participants with complete data, we introduced missing data for 1,523 participants on some combination of these four covariates. In order to create a mixture of ignorable and non-ignorable missing data, we allowed roughly 20% of patients within SMBG guidelines to have some combination of race, smoking status and/or duration of diabetes to be missing. Participants who were older (> 65 years), within SMBG guidelines and healthier (HgA1c $< 6.3\%$) had a 56% chance of having their age and race be missing, whereas participants who were younger (< 59 years), out of SMBG guidelines and less healthy (HgA1c $> 8.3\%$) had a 56% chance of having their age, smoking status, and duration of diabetes be missing. Using this approach the missing data could be considered to be both type I nonignorable (i.e., age which is missing dependent on the unobserved value of age) and type II nonignorable (i.e., duration of diabetes which is missing dependent on other partially observed missing data) as defined by Baker (2000).

After we estimated propensity scores for each of the models above, we calculated deciles based on the estimated scores. Within each decile, we separated the participants by SMBG status. Next, we assigned weights to the participants based on the distribution within each decile using the methods described above.

Using these weights from each model, we then estimated the differences (i.e., residual imbalances) in the nine covariates after propensity score adjustment for the first set of models. We then compared the different models' performance in reducing the bias (imbalance in covariate values) between the SMBG adherent participants and the SMBG non-adherent participants. We finally examined the impact these adjustments had when comparing the outcome of HbA1c between SMBG users and non-users for all propensity score models and the ANCOVA model.

Table 1. Table of observed proportions for categorical covariates

		Participants with complete data ($N = 6687$)			Participants with some missing covariates			
		SMBG in Guidelines			SMBG in Guidelines			
Covariate		No	Yes	p-value	No	Yes	p-value	N*
Gender	Female	.46	.54	< 0.001	.46	.54	< 0.001	3879
Smoking status	Current	.11	.08		.11	.06		
	Former	.41	.48		.38	.43		
	Never	.48	.44	< 0.001	.51	.51	0.002	1919
Drinking	Abstain	.51	.56		.57	.60		
	< 3 per day	.47	.43		.41	.39		
	3+ per day	.02	.01	0.005	.02	.01	0.307	1452
Education	High school or less	.44	.43		.53	.47		
	Some college	.31	.31		.27	.28		
	College graduate+	.25	.26	0.844	.20	.25	0.016	1919
Race	Caucasian	.58	.69		.54	.69		
	Black	.12	.08		.15	.10		
	Hispanic	.08	.07		.11	.07		
	Asian Pacific	.13	.09		.12	.08		
	Native American	.01	.01		.00	.00		
	Other	.00	.00		.01	.01		
	Multi-ethnic	.08	.06	< 0.001	.07	.05	< 0.001	1965

*N denotes the sample size used in the analysis, which includes participants with missing values in other covariates.

6. Results

6.1. Descriptive Statistics

Table 1 displays descriptive statistics for the five categorical variables broken down by SMBG status (adherent/non-adherent). The first 3 columns of this table show statistics for the 6,687 participants with complete data, and the last 4 columns show statistics for the 3,879 participants with missing data. For instance, gender was observed on all 10,566 participants, thus the value for N* (last column) equals 3,879—the number of people with some missing data, whereas race was missing on 1,914 participants thus the value of N* is 1,965 ($3{,}879 - 1{,}914 = 1{,}965$). As can be seen by this table there are modest imbalances on the categorical variables by SMBG status for participants with complete data. Significantly more women than men (54% vs. 46%), higher numbers of Caucasians (69% vs. 58%), more abstainers (from drinking) (56% vs. 51%) and more former smokers (48% vs. 41%) are adherent. Similar patterns exist for those with incomplete data, with a few notable differences. There does not appear to be a significant difference among drinking characteristics for those with incomplete data. However, significantly fewer

SMBG adherent people have a high school education or less (47% vs. 53%). The former smoker difference between groups remained (43% for those within guidelines vs. 38% for those out of guidelines) and the race difference became slightly more pronounced with 69% of Caucasians adherent vs. 54% non-adherent.

Table 2 displays descriptive statistics for the continuous variables broken down by SMBG status (adherent/non-adherent). The first four columns of this table show statistics for the 6,687 participants with complete data, and the last five columns show statistics for the 3,879 participants with missing data. From this table, we see that among those with complete data, the proportion of individuals in working class professions were significantly lower (62% vs. 64%) for those who were adherent. In addition, the adherent individuals were significantly older (62 vs. 61 years). Per capita income was not significantly different between the two groups, however the adherent participants had marginally higher income ($17,727 vs. $17,378). There was no significant difference in the duration of diabetes between the two groups. For the participants with missing data, the three significant trends all increased. For the proportion of individuals in working class professions the difference increased to 3% (62% vs. 65%); for age the difference increased to 2 years (62 vs. 60 years); and for income the difference increased to $936 ($17,423 vs. $16,487, $p = 0.047$).

We next compared the missing value indicators for the seven variables that contained missing data between the adherent and non-adherent participants (Table 3). For all seven we found highly significant differences, where significantly more people had observed data if they were within guidelines than those out of guidelines. These differences ranged from five to seven percent.

When we compared the distribution of the covariates between those with complete data and those with missing data we found several significant differences (Table 4). There were significant differences for smoking status (more never smokers had missing data), drinking status (more abstainers had missing data), education (more participants with high school or less education had missing data), race (fewer Caucasians had missing data), % working class (higher percent of working class participants had more missing data), duration of diabetes (those with longer durations had more missing data), and per capita income (those with missing data had lower incomes).

In summary, these descriptive statistics indicate that the SMBG adherent and non-adherent groups differed on both covariate distributions and missing data patterns and therefore in order to estimate the true treatment effect we need to adjust for both of these differences.

6.2. Propensity Score Model results

6.2.1. Model 1 – Completers Only

The results from the logistic regression model used for estimating the propensity scores can be found in Table 5 From this model we see that among the nine covariates included, which resulted in 17 slope parameters, only duration of diabetes was not significant. For some of the categorical covariates, one or more levels of the variable were not significant, although as a whole the variable was significant (i.e., smoking status where the indicator for former smokers was highly significant while the indicator for current smoker was not).

Table 2. Means (Standard Deviations), standardized differences in percent for continuous covariates in both groups

	SMBG in Guidelines								
	Participants with complete data ($N = 6687$)				Participants with some missing covariates				
Covariate	Yes Mean (SD)	No Mean (SD)	Standardized Difference (%)	p-value	Yes Mean (SD)	No Mean (SD)	Standardized Difference (%)	p-value	N*
% Working Class	.64 (.13)	.62 (.13)	9	.0008	.65 (.13)	.62 (.13)	18	.0056	1201
Duration of Diabetes	8.1 (8.4)	8.2 (8.6)	−2	.5494	8.8 (9.5)	8.5 (8.8)	4	.5411	1517
Age	61 (11)	62 (11)	−15	.0001	60 (12)	62 (11)	−17	.0001	3879
Per Capita Income	17378 (7013)	17727 (6987)	−5	.0712	16487 (7114)	17423 (6922)	−13	.0474	1209

*N denotes the sample size used in the analysis, which includes participants with missing values in other covariates.

Table 3. Table of missing-value indicators (proportion missing)

	SMBG within Guidelines		
Covariate	No	Yes	*p*-value
Smoking Status	.20	.14	< .0001
Drinking Status	.25	.17	< .0001
Education	.20	.14	< .0001
Race	.20	.13	< .0001
% Working Class	.27	.22	< .0001
Duration of Diabetes	.24	.17	< .0001
Per Capita Income	.27	.22	< .0001

This table indicates that these covariates remained statistically significant in a multivariable model, which suggests that in order to control their potential bias a multivariable technique such as propensity score modeling must be used rather than using bivariate techniques.

When we compared propensity scores between SMBG adherent and non-adherent participants, we found that the standardized difference in percent was 36% between the two groups. Here the standardized difference in percent is defined as the difference in the

Table 4. Table for difference in the distribution of covariates between individuals with complete data and those with missing data

Covariate		Complete	Missing	*p*-value
Gender	Female	.48	.48	.75
Smoking	Current	.10	.09	
Status	Former	.43	.39	
	Never	.47	.51	< .0046
Drinking	Abstain	.52	.58	
	< 3 per day	.46	.41	
	3+ per day	.02	.02	.0015
Education	High school or less	.44	.51	
	Some college	.31	.28	
	College graduate +	.25	.21	< .0001
Race	Caucasian	.61	.58	
	Black	.11	.13	
	Hispanic	.08	.10	
	Asian Pacific	.12	.11	
	Native American	.00	.00	
	Other	.00	.01	
	Multi-ethnic	.07	.07	< .0001
		Mean (SD)	Mean (SD)	
Age		61.0 (11.2)	60.8 (11.6)	.5155
% Working Class		.63 (.13)	.64 (.13)	.0335
Duration of Diabetes		8.12 (8.50)	8.76 (9.31)	.0147
Per Capita Income		17472 (7007)	16718 (7076)	.0006

Table 5. Parameter estimates from propensity score model using complete data only

		Parameter	Std. Error	Wald Chi Square	*p*-value
Intercpt		0.58	0.344	2.87	0.0902
Gender	Female	0.40	0.060	44.90	0.0001
Smoking	Current	0.064	0.105	0.369	0.5433
Status	Former	−0.297	0.062	23.25	0.0001
Drinking	0 < drinks < 3 per day	0.164	0.060	7.50	0.0062
	3+ per day	0.472	0.240	3.85	0.0498
Education	Some college	−0.104	0.068	2.32	0.1281
	College graduate +	−0.173	0.076	5.24	0.0221
Race	Black	0.577	0.103	31.44	0.0001
	Hispanic	0.296	0.112	6.952	0.0084
	Asian Pacific	0.467	0.097	23.25	0.0001
	Native American	0.303	0.350	0.750	0.3865
	Other	−0.154	0.606	0.065	0.7991
	Multi-ethnic	0.3121	0.115	7.37	0.0066
Age		−0.010	0.00275	13.43	0.0002
% Working Class		0.9369	0.325	8.30	0.0040
Duration of Diabetes		0.00059	0.00339	0.030	0.8616
Per Capita Income		0.000012	.0000061	3.605	0.0576

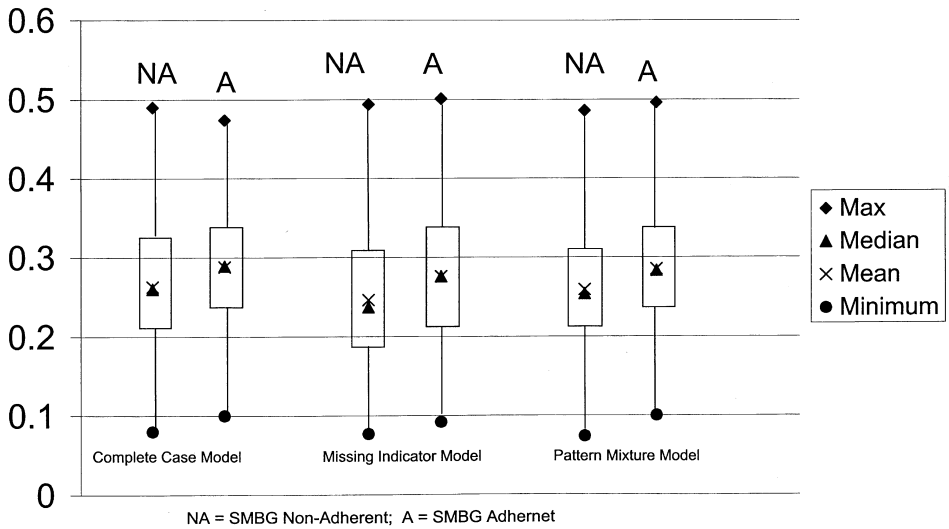

Figure 1. Box-plots comparing propensity score distributions between the SMBG adherent and non-adherent participants for each of the three models fit.

mean value of the propensity score between the non-adherent and adherent groups divided by the average standard deviation of the two groups $100 \times [\bar{x}_{NA} - \bar{x}_A / \sqrt{(s^2_{NA} + s^2_A)/2}]$. Figure 1 contains box-plots that describe the propensity score distributions in the SMBG adherent and non-adherent groups for each of the three propensity score models. For the complete data model, the distribution of propensity scores for the 4,883 participants who were SMBG non-adherent participants ranged from 0.08 to 0.49 with a median of 0.26 (mean 0.263, standard deviation 0.071) whereas the 1,804 participants who were SMBG adherent had propensity scores between 0.10 and 0.47 with a median of 0.29 (mean 0.288, standard deviation 0.069).

6.2.2. Missing Value Indicator Model

The second propensity score model we fit included missing-value indicator variables for each covariate that contained missing data, thus increasing the number of parameters estimated from 17 to 24. This model exhibited several differences from the complete data model above. First, an obvious difference is that when all 10,566 participants were included in this model the overall SMBG rate dropped from 27.0% to 25.4%. This is due to the fact that among the 3,879 additional participants included in this model, only 22.6% were within SMBG guidelines, whereas 27.0% of the completers were within guidelines. The estimates for the parameters for the individual covariates were relatively the same in both magnitude and significance when compared to the completer model, however in this model several missing value indicators were significant predictors in the propensity score model. These include the missing value indicators for smoking, drinking and race.

We found that the standardized difference in percent for the estimated propensity score was 40% between the two groups. The distribution of propensity scores for the 7,886 participants who were SMBG non-adherent ranged from 0.08 to 0.49 with a median of 0.24 (mean 0.246, standard deviation 0.074) whereas the 2,680 participants who were SMBG adherent had propensity scores between 0.09 and 0.50 with a median of 0.28 (mean 0.276, standard deviation 0.076) (See Fig. 1). When we compare these values to the model from the completers, we see that the overall distributions were similar, however there was slightly larger variability around the mean for the incomplete data model than for the completer model.

6.2.3. Pattern Mixture Model Approach for Propensity Score Estimation

This approach to propensity score estimation fitted a separate propensity score model for each of the 36 distinct missing data patterns. By fitting these 36 models, we estimated a total of 440 parameters. One benefit of this approach, as compared to the previous two, is that the covariates were allowed to have different estimated parameters depending on the pattern of missing data. Thus, age, which was fully observed, had 36 different estimates for its relationship with SMBG use, each conditional on only the set of observed covariates within a particular missing data pattern.

When we compared the distributions of the propensity scores themselves from this approach (Fig. 1), we found that for the 7,886 participants who were SMBG non-adherent the scores ranged from .07 to .49 with a median of .25 (mean .259, standard deviation .067) whereas the 2,680 participants who were SMBG adherent had propensity scores

between .10 and .50 with a median of .28 (mean .285, standard deviation .068). When we compare these values to the previous models, we again see that the overall distributions were similar, however there was slightly less variability around the mean for this pattern mixture model than for the other models.

6.3. Comparison of Models in Achieving Balance

All three models performed well in reducing the imbalance between the SMBG adherent and non-adherent participants for the nine covariates of interest. In fact, for all variables in all models, there were no significant differences on any covariates after adjustments were made for propensity score decile. Table 6. displays the reduction in bias for each of the nine variables by comparing the original differences in proportions (for categorical variables) or means (for continuous variables) and the corresponding χ^2 or t statistics followed by the same statistics calculated after adjusting for each of the three propensity score models. The largest reduction in bias occurred for the Race variable. Here we notice that before propensity score adjustment, the χ^2 statistic that described the relationship between SMBG adherers and non-adherers was 66 (for complete case participants) and 101 (for all participants with Race observed). In the unadjusted analyses, 69% of the SMBG adherent participants were Caucasian whereas only 58% of the SMBG non-adherent participants were Caucasian (11% difference in proportions). This difference was reduced to less than 1% after propensity score adjustment for all three models (0.4% for Completer Model, 0.09% for Missing Indicator Model, and 0.13% for the Pattern Mixture Model).

6.4. Comparison of Models for Hemoglobin A1c Between SMBG Adherent and Non-Adherent Groups

The most appropriate outcome measurement to determine the effectiveness of SMBG is HbA1c, a measure of glycemic control that is regularly monitored among patients with diabetes. We first compared the unadjusted levels of HbA1c between SMBG adherent and non-adherent participants for all participants ($n = 10{,}566$) using a simple t-test. We found a highly significant difference (9.09 for SMBG non-adherent versus 8.25 for SMBG adherent, $p < 0.001$). Next, we compared the unadjusted levels of HbA1c between SMBG adherent and non-adherent participants with complete data ($n = 6{,}687$) again using a t-test and found a highly significant difference (9.05 for SMBG non-adherent versus 8.22 for SMBG adherent, $p < 0.001$). We then estimated the difference for the participants with complete data using a weighted analysis where the weights were estimated from the propensity score model using complete data. This analysis also found a highly significant difference between guidelines adherence levels (9.01 for SMBG non-adherent versus 8.26 for SMBG adherent, $p < 0.001$). Although the adjustment for propensity scores did not change the ultimate conclusion, that SMBG use is related to better glycemic control, we did find that the estimated difference between adherent and non-adherent was reduced from 0.83 to 0.75 which is equivalent to an 10%

Table 6. Comparison of differences in proportions* or means** and corresponding χ^2 and *t*-statistics for covariates before and after propensity score adjustment

		Completer Model				Missing Data Models					
		Initial Bias		Bias After Adjustment		Initial Bias		Bias After Adjustment			
								Indicator Model		Pattern Mixture Model	
Covariate		% Diff.*	χ^2	% Diff.	χ^2	% Diff.	χ^2	% Diff.	χ^2	% Diff.	χ^2
Gender	Female	8	33	0.2	0.04	8	48	0.54	0.31	0.90	0.85
Smoking	Current	−3	25	0.2	0.05	−3	34	−0.29	0.24	−0.7	1.7
Status	Former	7		0.1		6		0.31		−0.4	
	Never	−4		0.1		−3		−0.02		1.1	
	Abstain	4	11	0.6	0.22	4	13	0.37	0.13	−0.29	0.11
Drinking	< 3 per day	−4		−0.6		−3		−0.33		0.33	
	3+ per day	−1		0		−1		−0.05		−0.05	
	High school or less	−1	.34	−0.1	0.12	−2	3.4	0.20	0.11	−0.69	1.2
Education	Some college	0		−0.2		0		−0.32		−0.3	
	College graduate +	1		0.4		1		0.13		1.0	
Race	Caucasian	11	66	−0.4	0.32	12	101	0.09	0.24	0.13	0.98
	Black	−4		0.1		−4		−0.17		0.16	
	Hispanic	−1		−0.1		−2		−0.05		−0.34	
	Asian Pacific	−4		0.3		−4		0.17		0.20	
	Native American	0		0		0		−0.05		−0.06	
	Other	0		0		0		−0.02		−0.08	
	Multi-ethnic	−2		0		−1		0.03		−0.02	

Table 6. continued

	Mean** Diff.	T-stat	Mean Diff.	T-stat	Mean Diff.	T-stat	Mean Diff.	T-stat	Mean Diff.	T-stat
Age (years)	1.7	5.4	0.12	0.47	1.7	6.9	0.1	0.46	0.02	0.07
% Working Class	−1.2	3.4	−0.001	0.25	2.1	3.2	0.01	0.22	0.2	0.47
Duration of Diabetes (years)	0.14	0.6	0.1	0.47	0.6	3.2	0.02	0.09	−0.04	0.24
Per Capita Income ($)	348	1.8	62	0.36	1205	5.6	61	0.32	95	.60

*% Difference is the difference in the % observed in the SMBG adherent group minus the % observed in the SMBG non-adherent group. For Gender 54% of women were in SMBG adherent and 46% of women were SMBG non-adherent thus the difference is 8%.

**Mean Difference is the difference in the mean values observed in the SMBG adherent group minus the percent observed in the SMBG non-adherent group. For example, for Age the average age in the SMBG adherent group was 62.2 years and the average age in the SMBG non-adherent group was 60.5 years thus the difference is 1.7 years (62.2 – 60.5 = 1.7).

reduction ((0.83 − 0.75)/0.83) in the treatment effect. This result is consistent with the propensity score model that found that SMBG non-adherent participants had a significantly different profile based on their background covariates, characteristics which, through their association with poorer glycemic control, partially explained (biased) the unadjusted contrast.

When we compared the HbA1c results using the models that incorporated missing data we found similar results. When all 10,566 participants were considered in the analyses, the t-test comparing SMBG adherent and non-adherent again showed a significant difference (9.09 for SMBG non-adherent participants versus 8.25 for SMBG adherent participants, $p < 0.001$). The indicator model predicted a difference of 9.05 versus 8.30 and the pattern mixture model predicted a difference of 9.05 versus 8.29. These two models showed an 11% and 10% reduction in the estimated treatment effects, respectively. Figure 2 contains 95% confidence intervals for the estimated treatment effect for the four models and for the unadjusted comparison for the entire sample. As can be seen, all four models have similar estimates for the difference between the SMBG adherent group and non-adherent group and comparable confidence intervals. The unadjusted difference (.84% in HgA1c) is slightly larger than the adjusted differences, indicating that without proper adjustment we might overestimate the beneficial effects of SMBG use on HgA1c values.

Although the reductions in the treatment effects after propensity score adjustment do not alter the clinical relevance of findings gleaned from unadjusted comparisons, they do support the notion that SMBG effectiveness may be overestimated due to confounding bias. In addition, as research continues, additional covariates may become available that would be useful to include in the propensity score models such as general markers for the severity of diabetes. As stated earlier, the assumption that the treatment assignment is

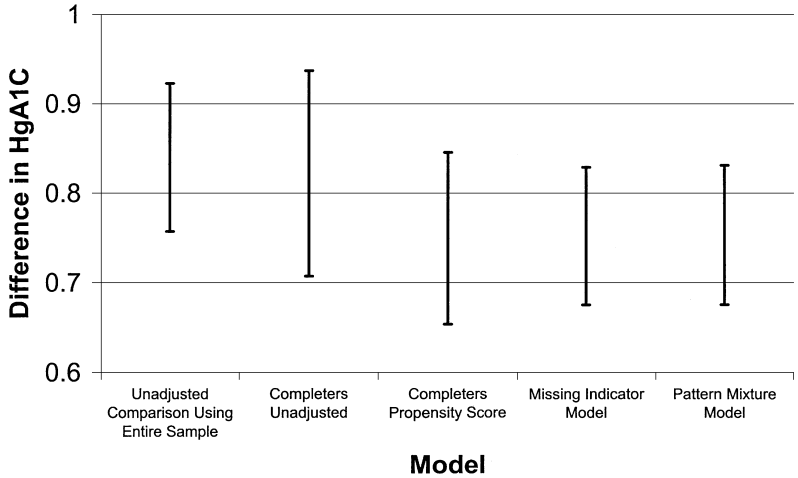

Figure 2. 95% confidence intervals for estimated treatment effect (difference in Hemoglobin A1c) between SMBG non-adherent and adherent groups unadjusted and after adjustment for each of the 3 propensity score models and for unadjusted comparison using the entire sample ($N = 10{,}566$).

strongly ignorable in this example is still being debated. We are optimistic that as additional prognostically important covariates are identified and added to the propensity score models this assumption will be more clearly satisfied.

6.5. Comparison of Models for Hemoglobin A1c Between SMBG Adherent and Non-Adherent Groups Using Artificially Created Missing Data

Next, we compared the estimated treatment effects for the three models and the ANCOVA model using the participants with complete data after we introduced missing data. The goal of these analyses was to compare treatment effect estimates for the three propensity score models and the ANCOVA model where we knew the "true treatment effect." Here the true treatment effect estimate was 0.75% in HgA1c based on the complete data ($n = 6{,}687$) propensity score model above (HgA1c of 9.01% for the SMBG non-adherent group and 8.26% for the SMBG adherent group).

Using the complete data propensity score model ($n = 5{,}164$), we estimated the treatment effect to be 0.40% (8.69% for the SMBG non-adherent group and 8.29% for the SMBG adherent group). The ANCOVA model was also fit using only those participants with complete data since PROC GLM in SAS automatically excludes all records that contain any missing values in its calculations. When we estimated the difference in HgA1c adjusting for all the covariates included in the propensity score model, we found the treatment effect estimate to be 0.38% (8.69% for the SMBG non-adherent group and 8.31% for the SMBG adherent group). Using the missing indicator propensity score model, we estimated the treatment effect to be 0.94% (9.19% for the SMBG non-adherent group and 8.25% for the SMBG adherent group). Finally, using the pattern mixture approach, we estimated the treatment effect to be 0.80% (9.04% for the SMBG non-adherent group and 8.24% for the SMBG adherent group).

When we compare the four models we notice that the complete data model and the ANCOVA model significantly underestimate the true treatment effect, reducing the effect size by 47% (0.75 to 0.40) and 49% (0.75 to 0.38), respectively. On the other hand, the missing indicator model overestimates the treatment effect by 25% (0.75 to 0.94). The pattern mixture model, however, provides an estimate of the treatment effect that was only 7% above the true treatment effect (0.75 to 0.80).

Figure 3 contains 95% confidence intervals for the mean difference in HgA1c for each of the four models, as well as the true treatment effect estimate based on the complete data prior to the addition of missing values. As can be seen in this plot, the confidence interval based on the pattern mixture model covers the true treatment effect, while the confidence intervals using the other three methods are either above or below the true value.

7. Discussion

We have described briefly several simple propensity score models that can be used to help estimate the effectiveness of SMBG. In a nonrandomized study of the effect of an

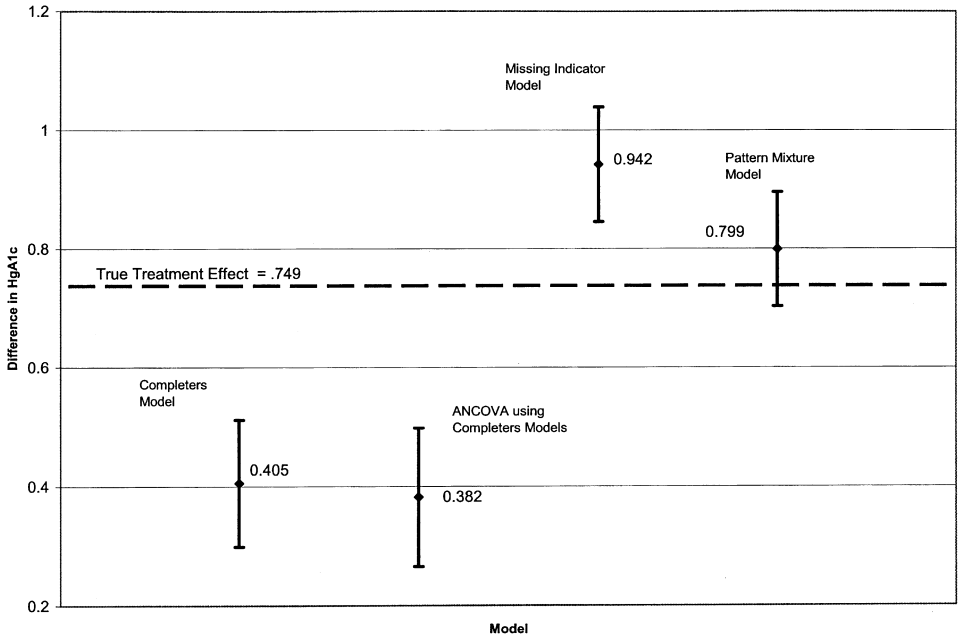

Figure 3. 95% confidence intervals for the estimated treatment effect (difference in Hemoglobin A1c) compared with the true treatment effect estimate using data with artificially created missing values.

intervention, there should always be concern that any observed effect on the primary outcome could have been explained by differences in the characteristics of the subjects who received the intervention versus those who did not rather than by the intervention itself. In our example, there were significant differences in the characteristics of the two groups, and by using propensity score methodology, these differences were appropriately adjusted for in making group comparisons on HbA1c. However, the unbiasedness and generality of this estimate assumes either that there are no missing data or makes strong assumptions concerning the mechanism of how missing data occurred. Based on the descriptive data, it is clear that in this applied example there are several characteristics concerning the participants that are different depending upon whether certain covariates are observed or missing. This suggests that ignoring the missing data could lead to incorrect estimates of treatment effects depending upon which outcome is compared. In addition, the population to which one can generalize the inference is more restricted when only participants with complete data are used.

When we compare the two missing data methods, we see only small differences in their performance for reducing the bias (imbalance of covariate values) between the SMBG adherent and non-adherent groups. Both models performed very well in this regard. In addition, both models produced nearly identical results for estimating the SMBG effect on HbA1c. Thus, for this particular applied example either method could be used, and so the simpler method of including indicator variables may be preferred. However, in general, the

indicator method assumes that the slope for all covariates is the same for every pattern of missing data. For instance, there is only one parameter estimate for age when determining its relationship to SMBG use. The pattern mixture approach allows for this restriction about parameters' estimation to be relaxed such that a different parameter estimate is allowed for every pattern of missing data. Thus, for this example, there would be 36 different estimates for the parameter corresponding to age, one for each pattern of missing data. There may be situations where a very limited number of covariates contain missing data, while the rest are complete. In such cases, the pattern mixture approach would not be computationally cumbersome and may lead to more precise estimates relative to the missing indicator approach. While the example included in this paper illustrates qualitatively the importance of different methodologic approaches, the quantitative differences between model outcomes was modest. However there clearly exist situations where quantitative differences will be marked, and thus methodologic choices become more critical.

In order to demonstrate this we introduced missing data into the subset of participants who had complete data at the beginning of the analyses. Using these participants we could determine what the "true" treatment effect estimate should be and then compare this to estimates using the methods proposed as well as using a simple ANCOVA approach. In these analyses, it is clear that only the pattern mixture approach resulted in treatment effect estimates that resembled the truth. Interestingly, the treatment effect estimates were either underestimated (using complete cases or ANCOVA) or over-estimated (using missing-value indicators), thus suggesting that the type of bias introduced into the estimation of treatment effects may not be easily predictable. This artificial example clearly illustrates a situation where ignoring the missing data or using a naïve approach to handle missing data would lead to incorrect inference. It should be noted that if missing data were introduced in a different manner there is the possibility that the results may have been different, however the goal of this illustration was to demonstrate how ignoring missing data can lead to biased treatment effect estimates. In addition, there do exist other alternatives for handling missing data using ANCOVA that we have not described in this manuscript (Vach and Blettner, 1991) and using these alternative methods may have led to more accurate treatment effects estimates for the ANCOVA model than were presented in this manuscript. Nevertheless, the goal of this manuscript was to clearly show the importance of incorporating the information from missing data into propensity score estimation in order to appropriately estimate treatment effects in observational studies and demonstrate that ignoring missing data can lead to biased results. Future manuscripts may focus on determining whether there are specific situations where using more advanced missing data techniques for ANCOVA models will lead to treatment effect estimates that are comparable to those provided using the pattern mixture approach outlined in this manuscript for propensity score estimation.

As mentioned in the introduction, the methods proposed in this paper are simple and easy to implement using standard statistical software. Thus, although the missing value indicator model is likely to create biased estimates, as shown in the example where we created missing data, it is a simple method that many investigators consider using to handle missing data. Although the real data example did not demonstrate clear differences between that method and the pattern mixture method, we strongly discourage investigators

from using the missing indicator approach unless there is no other option available. In fact, there is a growing literature that indicates that using the missing indicator approach typically leads to biased results. (Ibrahim, Lipitz and Chen, 1999)

Clearly other more sophisticated approaches may also provide accurate treatment effect estimates. One such approach would be to use some form of imputation to handle missing data. The simplest form of this approach would be to use a mean imputation algorithm, where the mean value from the observed data would be imputed into the missing data. The disadvantage of such a method is that it would create estimates that may underestimate the true variability of models. More sophisticated imputation strategies such as multiple imputation would likely create better propensity score models and be useful in estimating treatment effects. We currently are exploring the advantages and disadvantages of such an approach when compared to the pattern mixture approach.

Another area of future research is to be able to compare the different parameter estimates generated by the pattern mixture approach to determine whether there is significant variability among estimates. If there is little variability in the parameter estimates across patterns of missing data, then a reduction in the number of parameters needed to be estimated can be considered.

D'Agostino Jr. and Rubin (2000) considered an approach that allowed fewer parameters to be fit than the full pattern mixture approach uses. This approach involves constructing a log-linear model to describe the relationship between the missing value indicators and other parameters in the model. In general, this may be considered the preferable approach, however it involves more extensive programming to implement than was used in this current example. At this time, we are developing software so that these models can be used more readily with existing software packages such as SAS.

8. Conclusions

Much work has been done that suggests using methods such as propensity score modeling is valuable when trying to make causal inference in observational research. This paper points out that when using these methods one must account for missing data as well in order to make valid causal inference. In the same way that ignoring a potentially confounding variable in the analyses may lead to biased treatment effect estimates, ignoring information contained in missing data patterns may also lead to biased treatment effect estimates. We presented two simple approaches to handling missing data in propensity score modeling and refer the reader to more sophisticated approaches as well. When we introduced missing data into a subset of the participants that contained complete data, we showed that the treatment effect estimates can be poorly estimated using both complete data techniques (either propensity score models or ANCOVA models) and naïve missing data methods such as the missing data indicator model. However, the pattern mixture approach was able to accurately estimate the true treatment effect in this situation. Current simulations are underway to better understand under what other circumstances the patterns of missing data have more or less impact on estimation of treatment effects.

Notes

1. Mathematically it can be shown that as long as one includes both the missing value indicator and the covariate in the model, the parameter estimate for the covariate stays the same regardless of what value is imputed for the missing data. In fact, the parameter estimates for the covariates are the same as that obtained from the complete-case model although the propensity scores are different due to the inclusion of missing indicator. The impact of choosing the number zero to be imputed only changes the estimate of the parameters for the missing value indicators not the parameters for the observed data.

Acknowledgments

This work was supported in part by National Cancer Institute grant 1 RO1 CA 79934 and NHLBI grant RO1 HL 40619. The authors wish to thank the reviewers and editors for their helpful comments. They also wish to acknowledge and thank their families, in particular, Carey for her assistance in preparing this manuscript.

References

American Diabetes Association: Clinical Practice Recommendations 1999, *Diabetes Care*, 22(Supplement 1), pp. S77–S78, 1999.

Paul Allison, D, "Multiple imputation for missing data: a cautionary tale," *Sociological Methods & Research*, 28, pp. 301–309, 2000.

American Diabetes Association, Self-monitoring of blood glucose. *Diabetes Care*, 17, pp. 81–86, 1994.

American Diabetes Association: Clinical Practice Recommendations, *Diabetes Care*, 21 (Supplement 1), pp. S1–S98, 1998.

American Diabetes Association: Clinical Practice Recommendations, *Diabetes Care*, 22, 1999.

S. G. Baker, "Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable and all or none compliance," *JASA*, 95, pp. 43–50, 2000.

F. G. Barker 2nd, S. M. Chang, P. H. Gutin, M. K. Malec, M. W. McDemott, M. D. Prados and C. B. Wilson, "Survival and functional status after resection of recurrent glioblastoma multiforme," *Neurosurgery*, 42, pp. 709–720, 1998.

A. F. Conners Jr., S. A. Speroff and N. V. Dawson et al., "The effectiveness of right heart catheterization in the initial care of critically ill patients. suppor*t* investigators," *Journal of the American Medical Association*, 276, pp. 889–997, 1996.

R. B. D'Agostino Jr., "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group," *Statistics in Medicine*, 17, pp. 2265–2281, 1998.

R. B. D'Agostino Jr. and D. B. Rubin, "Estimating and using propensity scores with partially missing data," *JASA*, 95, pp. 749–759, 2000.

A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 39, pp. 1–38, 1977.

J. M. Evans, R.W. Newton, D. A. Ruta, T. M. MacDonald, R. J. Stevenson and A. D. Morris, "Frequency of blood glucose monitoring in relation to glycemic control: observational study with diabetes database," *BMJ*, 319, pp. 83–86, 1999.

A. Faas, F. G. Schellevis and J. T. Van Eijk, "The efficacy of self-monitoring of blood glucose in NIDDM subjects. A criteria-based literature review," *Diabetes Care*, 20, pp. 1482–1486, 1997.

X. S. Gu and P. R. Rosenbaum, "Comparison of multivariable matching methods; Structures, distances, and algorithms," *Journal of Computational and Graphical Statistics*, 2, pp. 405–420, 1993.

J. J. Heckman, H. Ichimura, J. Smith and P. Todd, "Sources of selection bias in evaluating social programs: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method," *Proceedings of the National Academy of Sciences of the United States of America*, 93, pp. 13416–13420, 1996.

J. G. Ibrahim, S. R. Lipitz and M. Chen, "Missing covariates in generalized linear models when the missing data mechanism is nonignorable," *Journal of the Royal Statistical Society*, Series B 61, Part1 1, pp. 173–190, 1999.31 D. B. Rubin, *Multiple imputation for nonresponse in surveys*, John Wiley & Sons, 1987.

A. J. Karter, A. Ferrara, J. Darbinian, L. M. Ackerson and J. V. Selby, "Self-monitoring of blood glucose: language and financial barriers in a managed care population with diabetes," *Diabetes Care*, 23(4), pp. 477–483, 2000.

A. J. Karter, L. M. Ackerson, J. Darbinian, R. B. D'Agostino Jr., J. Liu and J. V. Selby, "Self-monitoring of blood glucose and glycemic control in a large managed care population: The Northern California Kaiser Permanente Diabetes Registry," *American Journal of Medicine*, 111, pp. 1–9, 2001.

E. Lieberman, A. Cohen, J. Lang, R. B. D'Agostino Jr., S. Datta and F. D. Frigoletto Jr., "The association of epidural anesthesia with cesarean delivery in nulliparas," *Obstetrics and Gynecology*, 88, pp. 993–1000, 1996.

R. J. A. Little, "Pattern-mixture models for multivariate incomplete data," *JASA*, 88, pp. 125–134, 1993.

B. W. Lytle, E. H. Blackstone, F. D. Loop, P. L. Houghtaling, J. H. Arnold, P. M. Mccarthy and D. M. Cosgrove, "Two internal thoracic artery grafts are better than one," *Journal of Thoracic and Cardiovascular Surgery*, 117, pp. 855–872, 1999.

T. L. Martin, J. V. Selby and D. Zhang, "Physician and patient prevention practices in NIDDM in a large urban managed-care organization," *Diabetes Care*, 18, pp. 1124–1132, 1995.

X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: a general framework," *Biometrika*, 80, pp. 267–278, 1993.

I. Olkin and R. F. Tate, "Multivariate correlation models with mixed discrete and continuous variables," *Annals of Mathematical Statistics*, 32, pp. 448–465, 1961.

S. S. Rich. "Analytic options for asthma genetics," *Clinical and Experimental Allergy*, 28, pp. 108–110, 1998.

James M. Robins and Wang, "Naisyin inference for imputation estimators," *Biometrika*, 87, pp. 113–124, 2000.

P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, pp. 41–55, 1983.

P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *JASA*, 79, pp. 516–524, 1984.

P. R. Rosenbaum and D. B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *American Statistician*, 39, pp. 33–38, 1985.

D. B. Rubin, "Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse," *ASA Proceedings of Survey Research Methods Section*, pp. 20–28, 1978.

D. B. Rubin, "Basic ideas of multiple imputation for nonresponse," *Survey Methodology* 12, pp. 37–47, 1986.

D. B. Rubin, "Practical implications of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 47, pp. 1213–1234, 1991.

D. B. Rubin, "Estimating causal effects from large data sets using propensity scores," *Annals of Internal Medicine*, 127, pp. 757–763, 1997.

D. B. Rubin and N. Thomas, "Affinely invariant matching methods with ellipsoidal distributions," *The Annals of Statistics*, 20, pp. 1079–1093, 1992a.

D. B. Rubin and N. Thomas, "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika*, 79, pp. 797–809, 1992b.

D. B. Rubin and N. Thomas, "Matching using estimated propensity scores; Relating theory to practice," *Biometrics*, 52, pp. 249–264, 1996.

D. B. Rubin and N. Thomas, "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association (JASA)*, 95, pp. 573–586, 2000.

J. V. Selby, B. Ettinger, B. Swain and J. B. Brown, "First 20 months' experience with use of Metformin for type 2 diabetes in a large health maintenance organization," *Diabetes Care*, 22, pp. 38–44, 1999.

J. V. Selby, G. T. Ray, D. Zhang and C. J. Colby, "Excess costs of medical care for patients with diabetes in a managed care population," *Diabetes Care*, 20, pp. 1396–1402, 1997.

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, pp. 895–907, 1991.