Multiple Imputation for Causal Inference

Irina Bondarenko* Trivellore Raghunathan†

Abstract

The potential outcome framework for causal inference is fundamentally a missing data problem with a special, the so-called file-matching, pattern of missing data. Given the large body of literature on various methods for handling missing data and associated software, it will be useful to use such methods to facilitate causal inference for routine applications. This article uses the sequential regression or chained equation imputation methodology for handling missing data to impute the potential outcomes based on the observed data. The causal inference parameters are formulated based on the models for the completed data and standard multiple imputation combining rules are applied to infer about the direct and mediated effects. Since the special pattern of missing data makes certain parameters of the joint distribution not estimable, the multiple imputation framework is modified to incorporate constraints or prior information in terms of augmented complete-data. Given the ability of the multiple imputation framework to to handle several types of variables, missing values in covariates and the availability of software for performing multiple imputations, this approach makes easier to perform causal inference from both observational and randomized studies. The methodology is illustrated through an application aimed to understand and quantify direct and mediated effect of diabetes on the cardiovascular disease using the NHANES data.

keywords: Direct Effect, Indirect Effect, Mediation, Observational Studies, Potential Outcomes, Randomized Studies

1. Introduction

Causal Inference forms a backbone of most scientific questions of interest. Though there is a considerable debate in terms of philosophical underpinnings of the cause and effect relationship, most, if not all, in statistical community (see, for example, Dawid (2000) and its discussions) have adopted the notion that causal effect of treatment or factor Z, which can take, say, two plausible values z_1 and z_2 for a subject s, measured using a variable Y is the contrast of the two possible outcomes, $Y_s(z_1)$ and $Y_s(z_2)$ for the subject s. The contrast could be measured through a difference $Y_s(z1) - Y_s(z_2)$ or the ratio $Y_s(z_1)/Y_s(z_2)$ or any other meaningful distance measure, $d(Y_s(z_1), Y_s(z_2))$ between the two potential or possible outcomes. The fundamental problem in constructing the causal effect for any particular subject is that only one of the two potential outcomes can be measured as any subject can only be measured either under z_1 or z_2. Thus, using the missing data framework, if $Y_s(z_1)$ is observed then $Y_s(z_2)$ is missing and vice versa and, hence, the result is a special pattern (the so-called"file-matching" pattern of missing data (Little and Rubin (2002))) of missing data. There is no information to estimate the correlation

*Statistician Senior,Department of Biostatistics,University of Michigan,Ann Arbor, MI 48109

†Chair and Professor of Biostatistics,Department of Biostatistics,University of Michigan,Ann Arbor, MI 48109

(partial correlation with additional pre-treatment covariates) between $Y(z_1)$ and $Y(z_2)$.

However, from the statistical perspective, the population average or expected causal effect can be measured which is defined as $T = E(Y(z_1) - Y(z_2))$ which equals $\sum_s Y_s(z_1)/N - \sum_s Y_s(z_2)/N$ for a finite population of size N (Neyman(1923), Rubin(1974) and Holland(1986)). It has been showed (Neyman (1923)) that under a completely random assignments of treatments with n subjects receiving z_1 and m subjects receiving z_2, the difference in the sample means is the unbiased estimate the population causal effect and the standard variance estimate of the difference in the sample means, $s_1^2/n + s_2^2/m$ where s_1 and s_2 are the sample standard deviations, over estimates the actual sampling variance.

The averaging of the individual causal effects can be restricted to particular subpopulation under certain conditions (Rubin(1974)). This framework has been extended to longitudinal settings (Robins(1986), Robins, Greenland and Hu (1999)) and parametric, semiparametric and nonparametric models could be used for estimating the causal effect. The conceptualization and development of these models to estimate the causal effect may also be facilitated by setting up graphical representation of regression relationships (Cox and Wermuth (2004)) among the potential outcomes.

Another important extension of this framework is to estimate the direct and indirect (or mediated) causal effects (Rubin (2004)). Suppose that the treatment can directly impact the outcome or through its effect on an intermediate variable or outcome. Suppose that the intermediate variable M is affected by the treatment, hence has two potential outcomes for the subject s, $M_s(z_1)$ and $M_s(z_2)$. A direct causal effect may be defined as $Y_s(z_1) - Y_s(z_2)$ when $M_s(z_1) = M_s(z_2)$. The population average direct causal effect may be defined as

$$D(z_1, z_2) = \frac{\sum_s [Y_s(z_1) - Y_s(z_2)]\delta_{M_s(z_1)-M_s(z_2)}}{\sum_s \delta_{M_s(z_1)-M_s(z_2)}}$$

where $\delta_A = 1$ if the event A is true and 0 otherwise and the summation is taken over all subjects in the population. Alternatively, the direct effect can be represented as the conditional expectation $E(Y(z_1) - Y(z_2)|M(z_1) = M(z_2)]$ under a joint model for $\{M(z_1), Y(z_1), M(z_2), Y(z_2)\}$. A broader definition of direct effect is through the computation of $E[Y(z_1) - Y(z_2)|M(z_1) - M(z_2)| \leq \epsilon]$ which may be useful to assess the role of the treatment directly affecting the outcome.

In the missing data analysis context it may be more convenient to express the causal parameters as regression coefficients. Defining $\Delta Y = Y(z_1) - Y(z_2)$ and $\Delta M = M(z_1) - M(z_2)$, and a regression model $\Delta Y = \beta_o + \beta_1 \Delta M + \epsilon$, the intercept term can be used to estimate the direct effect and β_1 as the summary effect on Y through the effect of Z on M.

Frankagis and Rubin (2003) developed principal stratification to isolate the total effect into direct and indirect effects where the principal strata are formed by combination of values of the potential outcomes $M_s(z_1), M_s(z_2)$. Given that the conceptualization of the direct and indirect effects involve four potential outcomes $\{M(z_1), M(z_2), Y(z_1), Y(z_2)\}$ and the observed data consists of either the pair $[M(z_1), Y(z_1)]$ or $[M(z_2), Y(z_2)]$, the estimation of the direct and indirect effect poses considerable identification problem. In the simple case with no covariates,there

is no information to estimate the covariances, $cov(M(z_1), M(z_2))$, $cov(Y(z_1), Y(z_2))$, $cov(M(z_1), Y(z_2)$, and $cov(Y(z_1), M(z_2))$, in the observed data. Some constraints have to be imposed to construct inferences for the causal parameters discussed above.

A popular constraint is the monotonicity assumption which entails, an ordering of the treatment z_1 and z_2 such that $M(z_1) \leq M(z_2)$ and $Y(z_1) \leq Y(z_2)$ for all subjects in the population. The direction of the inequalities can be changed or it can be different for M and Y depending upon the scientific context of the problem. These constraints provide information about the missing potential outcomes and may help identify the joint distribution of the four potential outcomes. See Jin and Rubin (2008).

The foregoing discussion amply illustrate that the causal inference problem can be be easily handled in practical routine application, if the missing potential outcomes could be handled as a missing data problem and are multiply imputed using an approach that identifies the joint distribution the potential outcomes conditional on the observed data. The paper discusses two potential appraaches for handling indentification problem. The monotonicity assumption allows some values to be imputed deterministically (or put some limits on the imputed values) and the rest can be imputed using some standard imputation software. The second approach is to incoporate prior information by augmenting the observed data by a small fraction of "complete-data" generated under a set of assumptions. This strategy also allows one to perform sensitivity analysis. Hence, the goal of this paper is to illustrate how the multiple imputation framework and software can be used construct causal effect estimates, total or direct.For multiple imputation, we used the sequential regression approach (Raghunathan et al (2001), Van Burren and Oudshoom (2000)) as implemented in IVEWARE (Raghunathan et al (1997)), a SAS callable routine. A similar approach is available in R and STATA environment. The over arching goal, however, is to facilitate the causal inference using easily available multiple imputation software rather than focusing on any particular software.

The motivation for this paper is to understand the role of diabetes(Z) and albuminuria(M) on cardiovascular disease(Y). Diabetes is the main risk factors for developing CVD. Albuminuria is considered to be the intermediate or mediating variable. Albuminuria is common complications of diabetes and characterized by presence of Albumin in the urine (Mogensen (1999). On the other hand Albiminuria is considered to be a significant risk factor for progression to cardiovascular disease and a predictor of cardiovascular mortality in diabetic population (Sowers at al (2001). There are a number of recent studies (Targher at al (2008), Soedamah-Muthu (2008)) that aimed to understand effects of Diabetes and Albuminuria on CVD. We used the National Health and Nutritional Examination Survey Data for this investigation. Since, this is not a randomized study, we use a large set of covariates and propensity score analysis to group subjects to achieve near randomization or balance. Analysis is performed within each propensity score class then combined across classes to yield an overall estimate.

The rest of the paper is organized into 4 sections. In Section 2, we provide the details about setting up as missing data problem, imputation of missing values and the analysis for both binary and continuous outcomes and mediating variables. In Section 3, we apply these methods to the NHANES-III data and describe the results. Section 4 conculdes with discussion of limitations and further research.

2. Methods

2.1 Imputation of potential outcomes

Here we consider a simple scenario that involves three variables Z, M, and Y, where Z is a binary treatment indicator $Z = \{0, 1\}$, M is a mediator which could be continuous or binary variable, and Y is an outcome variable which could also be a continuous or binary variable . Our goals are (1) To estimate overall causal effect of Z on Y at the population level by measuring the distance between $Y(0)$ and $Y(1)$, $d(Y(0), Y(1))$; (2) To assess if the overall effect can be decomposed into $d(Y(0), Y(1)|M(0) = M(1))$ as the "Direct effect" of Z on Y and $d(Y(0), Y(1)|M(0) \neq M(1))$ as the "indirect effect" of Z on Y and mediated through M. In any application there may be several covariates, X, to be adjusted. The key players are X, Z, $M(0)$, $M(1)$, $Y(0)$ and $Y(1)$. The data thus laid out and illustrated in Table 1 introduces identifiability problem because exactly values of all potential outcome and mediator values $M(0)$ and $Y(0)$ are missing for subjects with $Z = 1$ and the values of $M(1)$ and $Y(1)$ are missing when $Z = 0$. All four variables $M(0), M(1), Y(0)$, and $Y(1)$ are missing when Z is missing which may occur in an observational study setting. In addition some values in X may be missing for some subjects. There are several potential ways to handle the identifiability issue: (1) Impose restrictions on possible values of the potential mediator, or outcome, or both; or (2) Use informative priors to make the problem identifiable. In this article, we demonstrate both approaches. For the first scneario, we can use the monotonicity assumption, for example, $M(0) \leq M(1)$ and $Y(0) \leq Y(1)$ and is illustrated in Table 2 for binary mediator and outcome variables where the observed values are in the bold font and the values indexed with '*' are determined by the monotoncity assumption. In the case of a continuous mediator and/ or outcome the monotonicity assumptions do not allow deterministically define additional values, but imply restrictions on the plausible values of potential outcomes and mediators which is sufficient to carry out the imputations.

To proceed with the causal inference we need to make a number of assumptions. We assume the stable unit-treatment value (SUTVA) (Rubin (1974)). This assumption implies that values of the potential mediators and outcomes are not affected by the treatment assignments of other subjects in the sample. We assume that the treatment assignment are completely at random as in randomized study, and thus

$$Pr(Z = 1|M(0), M(1), Y(0), Y(1), X) = Pr(Z = 1) = p$$

If the treatment assignment is not randomized, then we assume that near randomization can be achieved by matching on the propensity score $P(Z = 1|X)$ or in other words,

$$P(Z = 1|M(0), M(1), Y(0), Y(1), X) = P(Z = 1|X)$$

In this case, following Rosenbaum and Rubin (1983) the estimated propensity score, for example, based on a logistic regression of Z on X, can be used to stratify the sample and the imputation can be carried out within each propensity score stratum. On the other hand, if the distribution of X is overlapping for the two groups $Z = 0$ and $Z = 1$, then all X could be directly used as a covariate in the imputation process.

There are many options for filling in the missing values in Table 2. We used the sequential regression approach (Raghunathan et al (2001)) as implemented in SAS callable IVEware(Raghunathan et al (1997)) to carry out the multiple imputations. The sequential regression approach involves a Gibbs sampling style iterative sampling from a sequence of conditional regression models, where the missing values in any given variable are drawn from their posterior predictive distribution corresponding to the regression model, and uses all other variables (including interaction terms) as predictors. This conditional distribution is based on a regression relating the variables being imputed and all other variables as predictors. This choice was mostly motivated by the particular application using the NHANES data set. The covariates had some missing values and were varying types such as continuous, categorical count etc, the mediator and outcome variables were also categorical or continuous and involved complex sample survey design. This general approach is also available in other software platforms MICE (R-package), and ICE (STATA).

2.2 Estimation of causal, direct and mediated effects

Once the missing values in Table 2 are multiply imputed, the estimation of causal effect is straightforward with monotonicity restrictions. For binary mediator and outcome variables, Define $I_M = 1$ if $M(0) \neq M(1)$ and 0 otherwise. Similarly, define $I_Y = 1$ if $Y(0) \neq Y(1)$ and 0 otherwise. The magnitude of the mediator M on the causal effect of Z, can be represented as

$$DE = P(Y(1) \neq Y(0)|M(0) = M(1))$$

$$ME = P(Y(1) \neq Y(0)|M(0) \neq M(1))$$

$$OR = \frac{P(Y(1) \neq Y(0)|M(0) \neq M(1))/P(Y(1) = Y(0)|M(0) \neq M(1))}{P(Y(1) \neq Y(0)|M(0) = M(1))/P(Y(1) = Y(0)|M(0) = M(1))}$$

This quantity can be estimated through a logistic regression model of I_Y on I_M, $logitPr(I_Y = 1|I_M) = \beta_o + \beta_1 I_M$ and $ME = \exp(\beta_1)$. Expressing the mediation effect in this format allows the use of the standard multiple imputation combining formula. Similarly, the probability of the change in the outcome given then there is no change in mediator due to treatment $Pr(I_Y = |I_M = 0)$ can be estimated as the direct effect $DE = (1 + exp(-\beta_o))^{-1}$.

In the case of continuous mediator M and continuous outcome Y, the regression model, $\Delta_Y = \beta_0 + \beta_1 \Delta_M + \epsilon$ discussed earlier could be used. The direct effect is then the intercept β_o and the magnitude of mediation (M) in a pathway between Z and Y can be expressed as β_1 per unit increase in Δ_M. Note, that this simple model assumes that the effect of mediation is additive and does not depend on $M(0)$ or $Y(0)$. Both variables can be added to the list of predictors and potential interactions can also be considered.

When the monotonicity restrictions are not imposed, there are multiple ways to quantify direct and mediated effects. For example,

$$DE_+ = \frac{P(Y = (0,1)|M(0) = M(1))}{P(Y = (1,0)|M(0) = M(1))},$$

measures the direct effect of the treatment on a positive relative to a negative outcome (or vice versa). Alternatively, the strength of the mediated relative to

direct effects can be expressed in terms of the following two relative odds,

$$OR_{+} = \frac{P(Y=(0,1)|M=(0,1))/P(Y=(1,0)|M=(0,1))}{P(Y=(0,1)|M(0)=M(1))/P(Y=(1,0)|M(0)=M(1))},$$

and

$$OR_{-} = \frac{P(Y=(0,1)|M=(1,0))/P(Y=(1,0)|M=(1,0))}{P(Y=(0,1)|M(0)=M(1))/P(Y=(1,0)|M(0)=M(1))}$$

3. Application to NHANES Data

As briefly discussed in the introduction, we applied the proposed methodology to the estimation of the causal effect of diabetes on Cardiovascular disease (CVD), based on NHANES III data. The Third National Health and Nutrition Examination Survey (NHANES III), 1988-94, contains data for 33,994 persons ages 2 months and older who participated in the survey. For our investigation we selected subjects 25 years and older. Our goal is to estimate a direct effect of diabetes, assess if causal effect of diabetes is mediated by albuminuria, and quantify this mediation. Diabetes was defined based on self-reporting, medication use, and an elevated 8-hours fasting glucose $\leq$ 126. A subject was classified to having CVD if the respondent had previously experienced one of the following: heart attack, stroke, or congestive heart failure. Albuminuria (ALB) is characterized by a presence of albumin in the urine and is measured as a ratio of Albumin-to creatine. It's a common in clinical practice to classify individuals into three categories based on this ratio. These three categories are: normal ($\leq 3mg/g$), micro albuminuria($> 3mg/g$,$\leq 30mg/g$), and macro-albuminuria($> 30mg/g$).

Given that NHANES III is an observational study and exposure to diabetes is known to be associated with a number of risk factors, we matched like subjects based on the probability of having diabetes, conditional on the covariates. From the NHANES III we selected a number of demographic variables and are listed in Table 3. For the purpose of the illustratione we limited the sample to the complete cases (N=11505) though we also performed the analysis that simultaneously imputed the missing covariates, an advantage of the multiple imputation framework. The results were quite similar.

We utilized logistic models to estimate propensity scores. Based on the estimated propensity of having diabetes, the subjects were stratified into quintiles. Due to a very small number of observed CVD cases in the first two low propensity quintiles, these two were combined resulting in four strata. We estimated the mediation effect of albuminuria expressed as (1) a ratio (continuous), and (2) dichotomized at 3mg/g. We log-transformed albumin-creatinine ratio to make normality assumptions more plausible.

For both continuous and binary definitions of Albuminuria, we estimated direct and mediated effects by applying (a) monotonicity restrictions $Alb(0) \leq Alb(1)$, and $CVD(0) \leq CVD(1)$ or (b) using augmented data. To augment the data, we sample 1% from each stratum and filled out missing values for the mediator and outcome based on hypothesized pattern. We explored a number of patterns and chose two to present in this article. After being added to the original data, these 'completed' data serve as priors and labeled as Prior1 and Prior2.

Prior1 was defined based on the hypothesis that missing values of mediator, and outcome are equal to the observed values ($ALB(0) = ALB(1), CVD(0) = CVD(1)$). For the continuous mediator we used a less restricting form of this prior assuming monotonicity for the mediator, and inequality for outcomes ($ALB(0) \leq ALB(1)$, $CVD(0) = CVD(1)$).

Prior 2 was based on the hypothesis that potential values of Albumin and CVD under diabetes are independent of the corresponding values Albumin and CVD with no diabetes exposure, given observed covariates $\{(ALB(0), CVD(0)\} \perp \{(ALB(1), CVD(1))|X)\}$

Next, the missing values of potential outcome and mediator were multiply imputed (N=50) conditional on the observed data. The estimates and standard errors of the causal parameters were pooled across the four propensity score strata for each imputed data set and then combined using the multiple imputation combining rules.

The first two estimates characterize causal effects of diabetes on any change in the CVD status relative to no change. DE is the probability of change in CVD status when no change in Albuminuria occurs; OR is the odds ratio of change in CVD associated with 1-unit change in Albuminuria. These two measurements allow us to compare estimates under restricted and unrestricted scenarios. However, they do not reflect a nature of causal effect for unrestricted scenario. To describe the direction of causal effect that is the change for better, we defined two more parameters: $DE+$ is odds of having CVD concordant with diabetes exposure vs being discordant, given no change in mediator, and $OR+$ odds ratio of concordant to discordant change CVD status due to concordant change in mediator. Table 4 shows summary of results for the four sets of estimates. All scenarios suggest that probability of change in CVD status caused by exposure to diabetes for subjects with no change in Albuminuria is close to 0.3. For example, for dichotomized Albuminuria and under monotonicity restrictions for 27% of population (95% CI 14-46%) change in diabetes status alone, may lead to change in the CVD status. For the unrestricted scenario and Prior1 33% of the population (95% CI 18-55%) are estimated to experience change in CVD status, when exposed to diabetes, but no change in Albumin values.

Estimated OR shows that subjects with a change in Albuminuria status are more likely to experience change in CVD status when exposed to diabetes, but the effect of change in Albumin is not significant. Assuming monotonicity doesn't change the results meaningfully. Odds of concordant CVD and Albuminuria pair vs the discordant one, caused by exposure to diabetes ($DE+$) is estimated between 2.21 to 2.63, depending on the prior assumptions and statistically significant (at 5% level of significance) for all scenarios. Estimates for $OR+$, though being positive for all scenarios, did not suggest a significant increase in odds of accordant CVD due to positive change in Albuminuria status. In sum, our analyses suggest that effect of diabetes on CVD is largely direct. We found a slight increase in risk of CVD due change in the Albuminuria levels caused by exposure to diabetes. However, this increase in risk is not significant.

4. Discussion

The advancements in methodolgy for analyzing data with some missing values and its implementation through user-friendly software provides a useful framweork for causal inference. The causal inference problem formulated through potential out-

come framework is essentially a missing data problem. However, some parameters of the joint distribution of the potential outcomes are not estimable neccesitating either restrictions or prior information. The most popular restriction, monotonicity restriction, lead to some deterministic imputation and the rest of the missing data can be imputed using any appropriate multiple imputation software. If the monotonicity restriction is deemed to be inappropriate, then some prior distribution may be needed to create imputations. The prior distribution can be introduced through augmented complete data under differing assumptions. This strategy also allows exploring sensitivity of inferences to prior distribution assumptions.

Many causal parameters of interest can be expressed as regression coefficients or ratios of simple proportions and can be easily estimated from the completed-data analysis. The standard multiple imputation combining rules can be applied. Once the data sets are completed, many different strategies can be used to infer about the direct and indirect causal effects of exposure to risk factors.

In this paper, we used the chained equations or sequential regression approach for imputing the missing values. However, almost any other approach can be used to impute the missing values. For example, semiparametric or nonparametric approaches (hotdeck, Bayesian Bootstrap etc) approaches can be used. The missing values in covariates can also be simutaneously handled in the analysis.

We have conducted a limited simulation study to evaluate the repeated sampling properties of the estimates. Though not shown, the estimates of causal parameters are unbiased and the interval estimates have desirable coverage properties. Further research is needed on this methodology and evaluation of different approaches for performing multiple imputations.

Aknowlegments

The research was partially supported by the NIH grants, 5-RO1-CA-129101-02, 5-RC1-RR-028366-02.

5. References

1. Cox, D.R., Wermuth, N., (2004), "Causality: a Statistical View ", *Internat. Statist. Rev.*, 72(3), 285–305.

2. Dawid, A. P. (2000), "Causal Inference without Counterfactuals", *J. Amer. Statist. Ass.* 95, 407-448.

3. Frangakis, C., Rubin D.B., (2002), " Principal stratification in causal inference," *Biometrics*, 58, 21–29.

4. Holland, P.W, (1986)," Statistics and Causal Inference (with discussion)", *Journal of the American Statistical Association*, 81, 945–970.

5. Jin, H., Rubin, D.B., (2008), "Principal stratification for causal inference with extended partial compliance," *Journal of the American Statistical Association*, 103, 101–111.

6. Little, R.J.A. and Rubin, D.B.,(2002), " Statistical Analysis with Missing Data", 2nd edition. New York: Wiley.

7. Mogensen, G.E., (2003), " Microalbuminuria and hypertension with focus on type 1 and type 2 diabetes", *Journal of Internal Medicine*, 254, 45–66.

8. Neyman, J. (1923), "On the Application of Probability Theory to AgriculturalExperiments. Essay on Principles. Section 9," *Statistical Science* 5(4), 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.

9. Raghunathan,T.E., Van Hoewyk,J., Solenberger,P.,(1997),"IVEWARE: Imputation and Variance Estimation Software," *http://www.isr.umich.edu/src/smp/ive*

10. Raghunathan, T. E., Lepkowski, J. E., Van Hoewyk, J. and Solenberger, (2001)"A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, 27, 85–95.

11. Robins, J.M., (1986). "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect,"*Mathematical Modelling*, 7, 1393–1512.

12. Robins, J. Greenland,S., F-C Hu , (1999)" Estimation of the Causal Effect of a Time-varying Exposure on the marginal Mean of a repeated Binary outcome", *JASA*, 94, 702–704 (with Discussion)

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

14. Rubin, D.B. , (1974)."Estimating causal effects of treatments in randomized and nonrandomized studies", *Journal of Educational Psychology*, 66(5), 688–701

15. Rubin, D.B., (2004),"Direct and Indirect causal Effects via Potential Outcomes (with discussion)", *Scandinavian Journal of Statistics*, 31, 161–170, 195–198

16. Soedamah-Muthu, SS., Visseren, FL., Algra, A., van der Graaf Y, SMART Study Group. (2008), "The impact of Type 2 diabetes and microalbuminuria on future cardiovascular events in patients with clinically manifest vascular disease from the Second Manifestations of ARTerial disease (SMART) study", *Diab Med*, 25(1), 51–57.

17. Sowers, J.R., Epstain, M.,Frohlich, E.D.,(2001), "Diabetes, Hypertension, and Cardiovascular Disease. An Update", *Hypertension*, 37, 1053–1059

18. Targher, G., Bertolini, L., Zenari, L., Lippi, G., Pichiri. I., Zoppini,. G, Muggeo, M., Arcaro, G., (2008), " Diabetic retinopathy is associated with an increased incidence of cardiovascular events in Type 2 diabetic patients", *Diabet Med*, 25(7),882–883.

19. Van Buuren, S., Oudshoom, C. G. M., (2000), "MICE: Multivariate imputation by chained equations", *web.inter.nl.net/users/S.van.Buuren/mi/*

Table 1: Observed Data Structure

Z	$M(0)$	$M(1)$	$Y(0)$	$Y(1)$
0	Observed	Missing	Observed	Missing
1	Missing	Observed	Missing	Observed

Table 2: Observed Data Structure without and with Montonicity Restriction. The items indicated with * are dterministic imputations

Z	$M(0)$	$M(1)$	$Y(0)$	$Y(1)$
0	0	Missing	0	Missing
0	1	1*	0	Missing
0	0	Missing	1	1*
0	1	1*	1	1*
1	0*	0	Missing	1
1	Missing	1	Missing	1
1	0*	0	0*	0
1	Missing	1	0*	0

Table 3: Albuminuria and CVD by Diabetes status based on observed data

		No CVD	CVD
No Diabetes (N=10280)	No Albuminuria	85.4%	5.5%
	Albuminuria	7.7%	1.5%
Diabetes (N=1225)	No Albuminuria	51.4%	11.4%
	Albuminuria	26.8%	10.4%

Table 4: Distribution of covariates

		Diabetes $N = 10280$	No Diabetes $N = 1225$
Race	Caucasians	46.4	39.4
	African-Americans	27.5	28.7
	Hispanic-Latino	22.9	29.4
	Others	3.2	2.4
Education	Less then HS	36.1	56.4
	HS	31.9	25.6
	More then HS	32.0	18.0
Gender	Male	47.0	46.5
	Female	53.0	53.5
Medical coverage	Yes	85.2	89.6
Current smoker	Yes	26.2	17.1
Blood cholesterol ever taken	Yes	57.1	71.8
Age	Mean (Std)	49.7 (17.5)	62.3 (14.1)
BMI	Mean (std)	27.1 (5.6)	30.1 (6.2)
Income-to-poverty ratio	Mean (std)	2.6 (1.8)	2.2 (1.7)

Table 5: Direct and Mediated effects of Diabetes on CVD.

Mediator	*Method*	DE (95% CI)	OR (95% CI)	DE+ (95% CI)	OR+ (95% CI)
Binary mediatior	Restricted	0.27(0.14,0.46)	1.26(0.50,3.18)	N/A	N/A
	Prior 1	0.29(0.14,0.51)	1.49 (0.57,3.93)	2.21(1.09,4.50)	1.43(0.12,17.70)
	Prior 2	0.29(0.15,0.48)	1.56 (0.55,4.42)	2.30(1.36,3.90)	1.36(0.24,7.71)
Continuous mediator	Restricted	0.28(0.11,0.55)	1.05(0.77,1.43)	N/A	N/A
	Prior 1	0.34(0.18,0.55)	1.04(1.82,1.31)	2.55(1.40,4.63)	1.09(0.76,1.55)
	Prior 2	0.33(0.17,0.53)	1.05(0.85,1.30)	2.63(1.50,4.62)	1.03(0.74,1.43)