Published in final edited form as:

Stat Med. 2008 October 30; 27(24): 5039–5053. doi:10.1002/sim.3341.

Estimation of risk factor associations when the response is influenced by medication use: an imputation approach

Abstract

When the outcome of interest is a quantity whose value may be altered through the use of medications, estimation of associations with this outcome is a challenging statistical problem. For participants taking medication the treated value is observed, but the underlying “untreated” value may be the measure that is truly of interest. Problematically, those with the highest untreated values may have some of the lowest observed measurements due to the effectiveness of medications. In this paper we propose an approach in which we parametrically estimate the underlying untreated variable of interest as a function of the observed treated value, dose and type of medication. Multiple imputation is used to incorporate the variability induced by the estimation. We show that this approach yields more realistic parameter estimates than other more traditional approaches to the problem, and that study conclusions may be altered in a meaningful way by using the imputed values.

Introduction

In many epidemiological studies the outcome of interest is a quantity whose value may be altered through the use of medications. Examples include cholesterol which may be lowered by statins or other lipid lowering medications, blood pressure which is lowered by anti-hypertensive medications, or glucose which is lowered by insulin or other diabetes medications. For these participants the treated value is observed, but the underlying “untreated” value may be the measure that is truly of interest. This value may reflect lifetime exposure, and is of biologic interest when studying long term processes such as atherosclerosis. Additionally, when studying participant characteristics that were clearly present prior to medication use, such as race, gender, or genetics, the untreated value of the endpoint is clearly of interest, rather than their value under treatment. However, those with the highest untreated values may have some of the lowest observed measurements due to the effectiveness of medications.

The most common approaches to this problem are to ignore medication use, to exclude those on medication, or to adjust any regression models by including a term for medication use. These approaches produce invalid results in most practical situations. This general problem has been discussed in a handful of previous publications, particularly with regard to blood pressure and antihypertensive treatment.^{1–4}

We consider an approach in which we parametrically estimate the underlying untreated variable of interest as a function of the observed treated value. Imputation of the untreated values based on observed data is an attractive alternative in that once the data are imputed, standard complete-data techniques may be employed. Covariates are easily included in the models, and effect sizes are easily expressed. A proper analysis needs to provide a reasonable imputation model, or way of estimating the untreated values based on observed data, and a method to take into account the uncertainty in the imputed values when reporting the final parameter estimates and standard errors. In the context of a longitudinal study, we propose building an imputation model based on participants who start taking medications within the study. For these participants, we have measurements both before and after commencement of medications, and can develop a model to predict the untreated value as a function of measured covariates. Although the untreated LDL is not “missing” in the classical sense (it is more “unobservable”

at the time of measurement), we have referred to our estimation procedure as an imputation since we are using standard multiple imputation techniques to incorporate the extra variability arising from the estimation process. This also allows us to easily distinguish our model for estimating untreated LDL (the imputation model), from our model estimating the association of an exposure of interest and LDL.

To evaluate the proposed method we performed a simulation study considering several different scenarios of interest, using a range of treatment effects. We illustrate the proposed methods using data from the Multi-Ethnic Study of Atherosclerosis (MESA). We focus on LDL cholesterol for illustration as it is a well known risk factor for cardiovascular disease, and a substantial (and increasing) proportion of participants are placed on lipid lowering medications, particularly if they have other risk factors. Additionally, lipid lowering medications are very effective at lowering the levels of LDL.

Methods

(i) The Multi-Ethnic Study of Atherosclerosis (MESA) Data

MESA is a prospective cohort study designed to study the progression of subclinical cardiovascular disease. The study includes 6814 men and women aged 45–84 years who were free of clinical cardiovascular disease at entry. The participants were recruited from six U.S. communities: Baltimore, MD; Chicago, IL, Forsyth County, NC; Los Angeles County, CA, Northern Manhattan, NY; and St. Paul, MN. Each field center developed its recruitment procedures according to the characteristics of its community and available resources, including lists of residents, dwellings, and telephone exchanges. All participants gave informed consent. Details of the sampling, recruitment, and data collection have been reported elsewhere.⁵ Blood lipid measurements were obtained following an overnight fast. LDL-cholesterol is calculated in plasma specimens having a triglyceride value <400 mg/dL using the formula of Friedewald et al.⁶ Medication use was determined by questionnaire. The participant was asked to bring to the clinic containers for all medications used during the two weeks prior to the visit. The interviewer then recorded the name of each medication, the prescribed dose, and frequency of administration from the containers. The MESA study is an ongoing, multi-faceted study, with results spanning many areas of research.

(ii) Imputation Models

There were $n=487$ participants not on lipid lowering medications at baseline that commenced taking these medications between baseline and exam 2 in MESA (average time between exams=1.6 years), and who had LDL measured at baseline and exam 2. The baseline LDL measurement for these participants was assumed to be their underlying untreated LDL, and we use this “new user” subset of participants to develop a model relating treated to untreated LDL. That is, among the new users linear models of the form $LDL_{untreated} = \alpha + \beta LDL_{treated} + \lambda Z + \varepsilon$ were considered, where Z included one or more of covariates medication type, high versus low dose, age, gender and race, and ε is a standard normal error term. Stated differently, we use post-treatment LDL values (and possibly patient or dose characteristics) to estimate pre-treatment LDL. We also considered interactions between pairs of terms, but none were significant or improved the R-squared. Examination of the residuals indicated that the linear model fit the data well. The resulting model equation is then applied to the observed LDL of those on treatment at baseline to obtain an estimated untreated LDL. For participants not on treatment at baseline no imputation is performed, we simply use their observed LDL value.

To incorporate the variability in the imputed values, multiple imputation techniques are used following the algorithm of van Buuren et al.,⁷ as implemented and described by Royston.^{8–9} Essentially, rather than generating one single imputed LDL cholesterol value for each

participant on treatment, we instead generate several values. The algorithm regresses $LDL_{untreated}$ on $LDL_{treated}$ (and covariates as indicated) and takes random draws from the resultant conditional distribution of the missing values of $LDL_{untreated}$, given the observed treated values and covariates. For each realization, the corresponding set of complete data is analyzed in a standard fashion and the results are pooled using a set of rules proposed by Rubin.¹⁰ The resulting variance of the parameter estimates incorporates both a within- and between-imputation component. The estimate for a parameter of interest β based on m imputations is simply the average of the estimates from within each imputation. The estimated standard error of $\hat{\beta}$ is a function of both the within- and between- imputation variability. If $\hat{w}_k$ denotes the standard error of $\hat{\beta}_k$ then the total within imputation variance is,

$$W = \frac{1}{m} \sum_{k=1}^m \hat{w}_k^2$$

and the between-imputation variance is estimated as,

$$B = \frac{1}{(m-1)} \sum_{k=1}^m (\hat{\beta}_k - \bar{\hat{\beta}})^2.$$

The total standard error for $\hat{\beta}$ is then estimated as,

$$se(\hat{\beta}) = \sqrt{W + \frac{m+1}{m} B}.$$

For comparison with our internally developed model, we also estimated untreated LDL cholesterol values based on dose/type specific percentage reductions reported in the literature. For statins, we used the results from a meta-analysis of 164 short-term fixed-dose randomized placebo controlled trials reported by Law et al.¹¹ Effects for fibrates and resins were taken from the National Cholesterol Education Program (NCEP) Adult Treatment Panel III Report which reports a range of percent reductions for each of these three categories of therapies.¹² The midpoint of each range (13% for fibrates, and 23% for resins) was used. Effects for niacin were taken from Goldberg.¹³ Those prescribed ≥ 40 mg per day of a given statin (or ≥ 20 mg per day of Atorvastatin) were considered to be on high dose. For non-statins, high dose was considered to be >20 mg/day for resins, ≥ 1200 mg/day for fibrates, and >500 mg/day for niacin. Table 1 provides a summary of the dose/type effects used to calculate the imputed LDL cholesterols. For example, the reported effect of taking atorvastatin at 5mg per day is a 31% reduction in LDL cholesterol, hence for a participant on this dose/type combination we would multiply their observed cholesterol by $1/(1-0.31)$ to obtain their estimated untreated cholesterol.

(iii) Simulation Study

We designed our simulations to mimic a study comparing two groups of interest (which we call exposed and unexposed) with respect to LDL cholesterol. We allow a substantial proportion (roughly 15–20%) of simulated participants to be taking medication which lowers their LDL by a specified amount (which we allow to vary). For each of four scenarios considered, we simulated 1000 datasets of size $n=2000$ each. We assume the response variable

of interest is untreated LDL, generated from a normal distribution with mean $\mu|X=120+\beta X$ mg/dl, and standard deviation 30 mg/dl. The term β denotes the effect of the exposure of interest X on LDL ($X=1$ for exposed, 0 otherwise). For instance, in the applied example which follows the simulations, X is diabetes. In our simulations we have assumed that X is dichotomous for simplicity, although in principle there is no reason that X could not be continuous or categorical. An observed LDL was then generated as $LDL_{obs}=(1-\gamma)T\mu+(1-T)\mu$, where T is 1 for subjects taking lipid lowering medications, and 0 otherwise, and γ denotes the proportionate effect of treatment on average LDL. We considered treatment effects of $\gamma=5\%$, 15%, 25% and 35%. Treatment was generated from a binomial distribution with probability of success 10% for those with untreated LDL below 160 mg/dl, and 50% for those with LDL above 160mg/dl. A treatment variable (yes/no) was generated for an exam 1 and exam 2, independently. Among the new-users (not on treatment at exam 1 but on treatment at exam 2) their treated LDL at exam 2 was derived as their exam 1 LDL, plus a random draw from a $N(0,20)$ distribution, times the reduction due to treatment $(1-\gamma)$. The additional noise component addresses the issue that their untreated LDL at exam 2 is not perfectly reflected by their exam 1 LDL. The parameters were chosen to reflect the distribution of exam 1 to exam 2 change in LDL observed in MESA among those not on lipid lowering medications at either exam. This simulated treated value at exam 2 was only used in the imputation strategy.

In scenario 1, LDL was generated independently of exposure, and exposure did not influence treatment assignment. In scenario 2, exposure did not influence LDL directly ($\beta=0$) but increased the odds of treatment, by lowering the threshold for switching to the higher probability of treatment from 160 to 130 mg/dL for those exposed. In scenario 3, the exposure of interest increased LDL by an average of 5 mg/dl (i.e. $\beta=5$), but did not influence treatment assignment. Finally in scenario 4 the exposure of interest influenced LDL both directly ($\beta=5$) and indirectly via a lower threshold for treatment.

(iv) Example

We illustrate our method with an example from the MESA data, and compare our results with nine plausible alternatives. Specifically, we try the 3 common but naïve approaches to dealing with medication use: use the observed LDL without regard to medication use, adjust for medication use using an indicator variable, or exclude those on medication. Additionally, we include several alternative approaches that have been described previously, though none of these are commonly used despite clear advantages over the naïve approaches. These include fixed substitution, median regression, fixed addition, and censored normal regression. In the fixed substitution approach,¹⁴ all participants on treatment are assigned a single (usually high) value of LDL (we use 3 different choices of the fixed value: 170, 140 and 100 mg/dL). In the median regression approach, participants on treatment are assigned an arbitrary high value of LDL, and then quantile regression is used to model the association of LDL with other factors.² This approach assumes that all participants on treatment have untreated LDL above the median. In fixed addition, a single increment in LDL is added to the observed LDL for each participant on treatment. ¹⁴ We chose 43 mg/dL based on the average effect of statins reported in the literature. Finally, a censored normal model was used, which assumes that the untreated LDL is greater than or equal to the observed LDL for those on treatment, that untreated LDL may be modeled as a normal distribution, and that the distribution for those above any specific value is the same in treated and untreated individuals. ³

Results

Out of 6814 participants enrolled in the MESA study, 6701 had LDL measured at baseline, and of these 1085 (16%) were taking lipid lowering medications at baseline. We excluded from analysis 25 participants who were taking cerivastatin (Baycol) at baseline which was

subsequently withdrawn from the market, and 1 participant whose drug type was unknown, leaving 1059 treated participants at baseline. There were 38 participants that were taking more than one drug (4 participants on 3 therapies; 34 on 2 therapies; 6 out of 38 cases were different doses of the same drug). For these participants the drug with the greatest effect was assumed to be their sole therapy. This may be a conservative estimate for some as combination therapy may be more effective than monotherapy, however the number of participants is small.

Table 2 provides a comparison of those that began taking lipid lowering medications between exam 1 and exam 2 (i.e., the subset used to develop the imputation models) to those already on lipid lowering therapy at baseline (i.e., the subset to which we are applying the imputation models). These subsets were comparable in terms of age, gender, race, body mass index, smoking, diabetes status, type and dose of lipid lowering drug. The new medication users tended to have a higher rate of hypertension (67.6% versus 62.3%, $p=0.05$), and to have treated LDL cholesterol an average of 5 units lower (99.5 versus 104.5 mg/dL, $p=0.002$). Relatively few participants were taking something other than a statin (8.4% of baseline users; 10.3% of new users). Among the statins, Atorvastatin was the most common (46% of the lipid lowering use at baseline, and 46% of new medications), followed by Simvastatin (27% of baseline use, 24% of new use) and Pravastatin (12% of baseline use, 10.5% of new use). The distribution of medication type for the new medication users was very similar to the distribution for those already on lipid lowering medications at baseline.

The most basic imputation model we considered was based solely on observed treated cholesterol. The fitted equation was $LDL_{\text{untreated}} = 98.3 + 0.42LDL_{\text{treated}}$. Models which included type and dose of drug, as well as age, gender and race/ethnicity were also considered. In general we found that the various types of statins were similar in terms of their effect, and that fibrates, resins, and niacin had less of an impact on LDL than the statins but were similar to one another. Those on high dose for a given drug had more cholesterol lowering. Once these variables were controlled for adding age, gender and race to the model did not improve prediction. The R-squared for the model containing treated LDL, dose and type was 0.26. Table 3 summarizes the 3 MESA models we used as final candidates. The fitted values from these models could be used for exploratory data analysis, plotting for illustrative purposes, etc. For models and hypothesis testing, we use 5 realizations of multiply imputed values, which will be centered around these lines on average, but incorporate the variability (both in the parameter estimates and residual variation).

Table 4 and Figures 1 summarize the results of our simulation studies. For Scenario 1, exposure does not influence LDL or treatment assignment, and all the approaches are quite comparable. All provide unbiased estimates on average (point estimates at zero in the first column of Figure 1) regardless of the effectiveness of treatment. Using the imputed values was conservative, with Type I error rates ranging from 3.1% to 4.8%. The other methods had Type I errors ranging from 3.9% to 5.7%. The lower type I error rates are likely a function of the increased standard errors associated with the multiple imputation (7%–15% higher standard errors compared to the adjusted estimates; roughly equal standard errors to the approach of excluding those on medications). Scenario 2, where the exposure of interest has no direct effect on LDL but lowers thresholds for treatment, yields similarly unbiased and slightly conservative results for the imputation method, but dramatically different results for the other methods. Ignoring the medication use yields biased results, which become more biased as the impact of treatment on LDL increases. The methods of adjusting and excluding are biased in a constant way, that reflects the average amount by which those exposed but not on treatment differ from unexposed but not on treatment. They will have a lower average LDL due to the different treatment threshold. The censored normal approach produced exposure effect estimates that were biased downward when the effect of medication use on LDL levels was very large, and biased upward when the treatment effect on LDL was minimal. Type I error rates were correspondingly high.

In Scenario 3, the exposure of interest increases average LDL by 5 units, but there is no effect of exposure on treatment assignment. All methods perform reasonably well in this scenario. In terms of power to detect this effect, it was approximately 90% for the imputation approach, treatment adjustment, and censored normal, and closer to 85% for the exclude method. For models ignoring treatment, power depended on treatment effectiveness, and decreased from 92% when treatment only had a small effect on LDL, down to 79% when treatment had the largest impact. Finally, scenario 4 is a combination of scenarios 2 and 3. The method of imputing LDL performs well, and is unaffected by treatment. The methods of ignoring, adjusting, and excluding those on treatment are significantly biased, with more bias as the effectiveness of treatment increases. Power is around 82% for the imputation approach, but can be extremely low for any of the other methods even when treatment has only a small effect on LDL. For the censored normal approach, extremes of the treatment effect on LDL again cause problems. For a large effect of medication on LDL, this approach had low power as the exposure effect is underestimated on average. For a small medication effect, the approach has very high power, because the exposure effect is overestimated.

Example: LDL Cholesterol and Diabetes

A potential question of interest is whether diabetics tend to have lower or higher LDL cholesterol. In general, it has been found that diabetics do not have significantly greater levels of LDL than do non-diabetics.¹⁵ Rather, a borderline high LDL (130–160 mg/dL) in a diabetic patient is equivalent to a much higher LDL for a non-diabetic in terms of cardiovascular risk. Benefits of lipid lowering therapy in diabetics even in the absence of elevated LDL levels have been demonstrated.^{16–17} As a result, diabetics are put on lipid lowering therapy more commonly than non-diabetics, and with this therapy being initiated at lower baseline LDL levels and with lower LDL targets. The observed LDL cholesterol levels for diabetics are thus quite low as a result of these interventions, and either ignoring medication use or adjusting for medication use yields a significant negative association that is an artifact of the treatment strategy. Additionally, excluding participants on lipid lowering therapy does not overcome the problem, as we have excluded a selected subset. The remaining diabetics that were not put on medication likely have very low LDL levels (or they would have been started on medication), even relative to the non-diabetics who were not put on medication. This is analogous to Scenario 3 from our simulation studies.

In Figure 2 we illustrate the difference between using the observed and imputed data by diabetes status. Here we used simply the fitted values for the imputation, rather than a multiple imputation, to allow us to plot a single point per participant. The multiply imputed values would be centered around these on average. Using the observed values diabetics have a lower median LDL than non-diabetics (110 versus 117 mg/dl). Switching to the imputed values (for treated participants), the medians go up for both diabetics and non-diabetics (126 versus 124 mg/dl), but more so for diabetics since they were more frequently on medication. The difference between diabetes groups is no longer evident.

Figure 3 illustrates the estimated difference in mean LDL and 95% confidence intervals obtained from various models, adjusting for age, gender and race. Using any of the naïve approaches we would conclude that the LDL for diabetics is an average 6–10 mg/dL lower than for non-diabetics, with $p < 0.001$ for each comparison. Using fixed substitution, the conclusion depends entirely on the choice of fixed value. For the new users in MESA the average baseline LDL was 140 mg/dL, and we can see that using that value as the fixed substitute yields the expected null association. Substituting a higher value overestimates LDL for a large number of participants, particularly diabetics who will be on medications at lower untreated LDL levels, and hence overestimates the LDL difference between diabetics and others. Substituting a very low value (included only for illustration) doesn't change the

association much over just using the observed LDL, since 100 mg/dL is approximately the average LDL among those on medication. Using median regression yields an estimate similar to fixed substitution with a value that is too high. That is, we see a significantly higher LDL for diabetics compared to non-diabetics. This is because the assumption that untreated LDL is above the median is not met in our example of LDL. That is, participants are often placed on lipid lowering medications at low levels of LDL cholesterol, depending on their other risk factors for cardiovascular disease. Finally, using fixed addition, censored normal regression or using the imputed LDL cholesterol as proposed in this paper, we conclude that there is no association between LDL and diabetes, as expected from the historical data. For instance, using the imputation based on the observed LDL, dose and type of drug, the estimated LDL difference associated with diabetes is -1.05 (95% confidence interval (CI) -4.68 to 2.58 , $p=0.57$). The imputation based on the reported LDL effects in the literature yields a similar conclusion, with an estimated diabetes effect of -1.09 (95% CI -4.07 to 1.90 , $p=0.48$).

Discussion

It is certainly conceivable that only the current value of LDL is pertinent, rather than any untreated or underlying value, in which case the observed value may be used in analysis without regard to treatment. The situation considered here is one where the value of LDL that is of interest is that which would be observed had none of the participants taken lipid lowering medications. When relating participant characteristics present at birth such as race or gender, the untreated value is clearly of interest, rather than their value under treatment. Genetic studies are a particularly common example of this scenario. Additionally, the untreated value may reflect lifetime exposure to LDL, and is of biologic interest when studying long term processes such as atherosclerosis. In the examples presented here it is clear that in certain situations using the imputed untreated LDL results in substantially altered conclusions when compared to more traditional approaches, such as ignoring treatment, adjusting for treatment, or excluding those on treatment. Models ignoring treatment only provide an unbiased estimate of the true effect of a variable on untreated LDL in the trivial situations that either no participants are taking any lipid lowering medications, or the lipid lowering medications do not have any effect on LDL. A model which adjusts for medication use will provide unbiased estimates if lipid lowering medications lower LDL by a constant amount regardless of LDL level, but not if there is a proportionate reduction, and not if there is differential use of treatment by exposure.

For imputation of the underlying untreated LDL we found that while both dose (expressed as high versus low) and type of drug were significant in the model, and improved the R-squared, the conclusions in terms of estimating LDL *associations* were not altered by including these additional variables (over and above simply the observed treated LDL). This may be because sample sizes were limited within the data we used to construct the imputation model, and hence for certain types of medications the effects may not have been well estimated. Alternatively, this may indicate that while different imputation models may result in very different predicted untreated LDL values, the estimated associations between the imputed LDL and other variables are much more robust. Overall the R-squareds were fairly low for the imputation models (0.19–0.29), indicating that the prediction (of untreated LDL) for any particular individual may not be that accurate. Despite this, using these fitted values in subsequent models for LDL removes the bias in the coefficients of interest, by putting the subset of the population that is on treatment back in the right area of the LDL space.

In addition to the common approaches of ignoring, adjusting, or excluding considered in our study, other straightforward approaches have been proposed in the literature. Methods suggested by Hunt et al. include the addition of a fixed or random constant to the observed value, or the substitution of a fixed or random “high” value for the treated participants.¹⁴ The addition of a constant may work well in certain scenarios, but in many scenarios dose titration

to achieve target levels makes this model implausible, since absolute reduction under treatment will depend on the underlying cholesterol level.

Substitution of a fixed or random value may result in a reduction of the treated value (rather than an increase). Additionally, substitution of a fixed value reduces the variability in the data, and does not make use of the observed (treated) values. Finally, conclusions for the fixed substitution approach vary widely depending on the choice of value to substitute. The pivotal assumption for the median regression approach, that the untreated LDL for the participants on medications is above the median LDL, is not met in these examples. This approach then leads to inserting a value that is on average too high for treated participants, and overestimating LDL associations for factors related to receiving treatment.

Many of these approaches were compared and contrasted via simulation in Tobin et al.³ Similar to our study they concluded generally that the common approaches of ignoring, adjusting, or excluding should not be used, and that certain other methods (including their proposed method of censored normal regression) performed quite well depending on the scenario and assumptions made. We found the censored normal approach to work well in our simulations where treatment assignment was not differential by exposure. In this approach it is assumed that the untreated value is greater than the observed, that untreated LDL may be modeled as a normal distribution, and that the distribution for those above any specific value is the same in treated and untreated individuals. The first assumption is very reasonable, and the second assumption is common to many approaches and appears to be satisfied. The latter assumption, as pointed out by Tobin, is unlikely to be strictly true, but in many situations will be approximately true. In the situation where treatment recommendations depend on exposures of interest (such as via treatment guidelines that explicitly incorporate risk factors) this assumption is problematic. In other words, situations such as Scenario 2 and 4 in our simulations that had informative censoring by design. Under both these scenarios, the exposed group (e.g. diabetics) will have more subjects taking LDL-lowering treatment. If this treatment lowers LDL a substantial amount (as in column 4 of Table 4, or the fourth point in each subfigure of Figure 1, where the treatment lowers LDL 35%), then the exposed group will appear to have much lower LDL than they do naturally (because this group includes more subjects on treatment), and the exposure effect is underestimated (either as negative if it is truly a null effect as in scenario 2, or as something less than the true 5 mg/dL effect in scenario 4). In contrast at the opposite extreme, if treatment does very little to LDL (as in column 1 of Table 4, or the first point in each subfigure of Figure 1, where the treatment lowers LDL 5%), the censored normal approach is tending to overestimate the exposure effect. This is because the exposed group will contain a higher proportion of censored subjects, whose untreated LDL is assumed to lie above their observed value. The full distribution of untreated values above this observed value is assumed to characterize the distribution of treated participants, however since treatment is only lowering their LDL by a very small amount this overestimates their untreated LDL. As a result the exposed group (e.g., diabetics) appears to have higher untreated LDL on average than their true untreated LDL.

An advantage of the censored normal approach is that it is very easy to use, and does not require the two-stage approach of imputation and then analysis. A disadvantage of the imputation approach is that it requires the development or availability of an appropriate imputation model. An advantage of the imputation approach is that once the data are imputed, they are available for plotting and for use in future models. Additionally, the imputed values may be used as either the response or a predictor, whereas in the censored normal regression approach the treated/untreated values (e.g. LDL in our examples) are by design the response variable. We note however that using the imputed LDL as a predictor is more complex for several reasons. It is unclear whether the current or the untreated value is more relevant to risk, and this tradeoff

likely depends on how long the participant has been on treatment. Additionally, the effect of the drug on response may not be totally reflected by its effect on LDL.

The methods presented in this paper, although illustrated with LDL cholesterol, could be applied to other measurements such as blood pressure or glucose where the observed values may be strongly impacted by medication use. Any longitudinal study with sufficient sample size to observe a set of participants both before and after commencement of treatment could use these methods, and obtain imputed untreated values suited to their study population and variables of interest. Alternatively, if there are reliable effect sizes for the medications available in the literature, such as from a meta-analysis, then these may be used to impute the underlying untreated value of interest.

The effect of improperly controlling for the use of medications depends on whether there is differential use of such therapy across the groups being compared (or equivalently, across increasing values of a continuous variable). Current guidelines frequently base treatment thresholds on numerous patient characteristics. For example, cholesterol guidelines factor in smoking, hypertension, HDL cholesterol, family history of premature heart disease, age, gender and diabetes. [12] Treatment guidelines for high blood pressure consider “compelling indications”, such as heart failure, ischemic heart disease, diabetes, kidney disease, recurrent stroke, or high coronary disease risk. [18] Genetic factors, although not directly related to treatment assignment or guidelines, could be indirectly related if the gene of interest was related to any of the above risk factors or conditions. Thus, there are many scenarios for which the imputation approach we propose could provide a distinct advantage. Conclusions may be altered in a meaningful way by using the imputed values.

Acknowledgments

This research was supported by contracts N01-HC-95159 through N01-HC-95165 and N01-HC-95169 from the National Heart, Lung, and Blood Institute. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at <http://www.mesa-nhlbi.org>.

References

1. White IR, Chaturvedi N, McKeigue PM. Median analysis of blood pressure for a sample including treated hypertensives. *Statistics in Medicine* 1994;13:1635–1641. [PubMed: 7973239]
2. White IR, Koupilova I, Carpenter J. The use of regression models for medians when observed outcomes may be modified by interventions. *Statistics in Medicine* 2003;22:1083–1096.10.1002/sim.1408 [PubMed: 12652555]
3. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine* 2005;24:2911–2935.10.1002/sim.2165 [PubMed: 16152135]
4. Cook NR. An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine* 1997;16:2713–2728. [PubMed: 9421871]
5. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacobs DR, Kronmal R, Liu K, Clark Nelson J, O’Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology* 2002;156:871–881. [PubMed: 12397006]
6. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry* 1972;18:499–502. [PubMed: 4337382]
7. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999;18:681–694. [PubMed: 10204197]
8. Royston P. Multiple imputation of missing values. *The Stata Journal* 2004;4(3):227–241.

9. Royston P. Multiple imputation of missing values: Update of ice. *The Stata Journal* 2005;5(4):527–536.
10. Rubin, DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.
11. Law MR, Wald NJ, Rudnicka AR. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *British Medical Journal* 2003;326:7404.10.1136/bmj.326.7404.1423
12. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association* 2001;285:2486–2497. [PubMed: 11368702]
13. Goldberg AC. Clinical trial experience with extended release Niacin (Niaspan): dose-escalation study. *American Journal of Cardiology* 1998;82:35U–38U.
14. Hunt SC, Ellison RC, Atwood LD, Pankow JS, Province MA, Leppert MF. Genome scans for blood pressure and hypertension: the National Heart, Lung, and Blood Institute family heart study. *Hypertension* 2002;40:1–6.10.1161/01.HYP.0000022660.28915.B1 [PubMed: 12105129]
15. American Diabetes Association (Position Statement). Dyslipidemia management in adults with diabetes. *Diabetes Care* 2004;27:S68–S71. [PubMed: 14693930]
16. Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HA, Livingstone SJ, Thomason MJ, Mackness MI, Charlton-Menys V, Fuller JH. CARDS investigators. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet* 2004;364(9435):685–96. [PubMed: 15325833]
17. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002;360(9326):7–22. [PubMed: 12114036]
18. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, Jones DW, Materson BJ, Oparil S, Wright JT, Roccella EJ. the National High Blood Pressure Education Program Coordinating Committee. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* 2003;42:1206–1252. [PubMed: 14656957]

Figure 1. Simulation Study Results—Estimated Coefficients and 95% Confidence Intervals

For each scenario and method, the point estimates (indicated by x) are the average of the estimated exposure coefficients from 1000 simulated datasets, and 95% confidence intervals extend to ± 1.96 times the average of the corresponding 1000 exposure coefficient standard errors. The horizontal line indicates the true exposure effect. For each scenario estimates were obtained using one of 5 approaches: the MESA imputation method, the naïve method of ignoring medication use altogether, adjusting for medication use in the model, excluding medication users, and finally using censored normal regression. In Scenario 1 (first column) exposure was assumed to have no effect on either the response or the treatment assignment. Scenario 2 (second column), exposure was assumed to have no effect on the response, but increased the rate of treatment assignment by lowering the threshold at which treatment is recommended. In Scenario 3 (third column) exposure was assumed to increase the average response (LDL) by 5 units, but have no effect on treatment assignment. Finally, in Scenario 4 (fourth column) exposure was assumed to increase both the average response (by 5 units) and

also increase treatment assignment by lowering the threshold at which treatment is recommended.

Figure 2. Boxplot of LDL Cholesterol by Diabetes with and without imputation

Within each level of diabetes the median is shown as a white line, the shaded box extends to cover the middle 50% of the LDL range, and the whiskers extend a further 1.5 times the inter-quartile range. Outliers are indicated as individual points. For illustration, we wanted to show only a single imputed value for each participant, and the imputed values shown are the fitted values of MESA model 3. The multiple imputations would be centered around these values on average.

Figure 3. Association of LDL and Diabetes

The estimated average LDL difference between those with and without diabetes are shown as squares, with surrounding lines indicating 95% confidence intervals. Models control for age, gender and race and include all participants at the MESA baseline exam. Headings indicate which values are used for LDL cholesterol for those on treatment. The first block includes the 3 naïve approaches. The second block of estimates use previously proposed approaches. Finally, in the third block we use the estimation approaches described in this paper.

Table 1
Percent Reduction in LDL by Prescribed Dose and Type of Lipid Lowering Medication

Lipid Lowering Drug	Daily Dose (mg)				
	5	10	20	40	80
Atorvastatin	31%	37%	43%	49%	55%
Fluvastatin	10%	15%	21%	27%	33%
Lovastatin	-	21%	29%	37%	45%
Pravastatin	15%	20%	24%	29%	33%
Simvastatin	23%	27%	32%	37%	42%
Daily Dose (mg)					
Niacin	<=750	751-1250	1251-1750	1751-2250	
	3%	9%	14%	17%	

Table 2

Comparison of Baseline Users and New Users of Lipid Lowering Medications

Variable	Baseline Users (n=1059) mean $\pm$ sd or n (%)	New Users (n=487) mean $\pm$ sd or n (%)
age (years)	66.0 $\pm$ 8.9	66.1 $\pm$ 9.4
male gender	504 (47.6)	247 (50.7)
Caucasian	599 (56.6)	270 (55.4)
body mass index	28.9 $\pm$ 5.2	29.1 $\pm$ 5.4
hypertension	660 (62.3)	329 (67.6)
treated diabetes	204 (19.3)	109 (22.4)
current smoker	103 (9.7)	42 (8.6)
treated LDL (mg/dL)	104.5 $\pm$ 29.2	99.5 $\pm$ 31.2
statin use	969 (91.5)	438 (89.9)
high vs. low daily dose	356 (33.6)	167 (34.3)
Type of lipid lowering medication:		
Atorvastatin	484 (45.7)	226 (46.4)
Simvastatin	286 (27.0)	115 (23.6)
Pravastatin	126 (11.9)	51 (10.5)
Fluvastatin	46 (4.3)	16 (3.3)
Lovastatin	28 (2.6)	60 (4.6)
Fibrates	57 (5.4)	51 (4.0)
Niacin	17 (1.6)	44 (3.4)
Resins	15 (1.4)	9 (0.7)

Note: Values are taken from exam 1 for baseline users, exam 2 for new users.

Table 3

MESA Imputation Models

MESA Imputation Model	Equation to Estimate $LDL_{untreated}$	Residual Standard Error	R ²
Model 1	$98.3 + 0.42LDL_{treated}$	27.3	0.19
Model 2	$64.8 + 0.51LDL_{treated} + 27.9 \text{ statin}$	26.1	0.26
Model 3	$62.2 + 0.50LDL_{treated} + 26.8 \text{ statin} + 11.1 \text{ highdose}$	25.6	0.29

Notes: statin=1 if participant is taking a statin, 0 if participant is only taking another type of lipid lowering medication (fibrate, resin or niacin); highdose=1 if participant was prescribed a higher than average dose of the drug (see methods).

Table 4
Simulation Study Results: Proportion out of 1000 simulated datasets for which there was a significant exposure effect

Effect of Lipid Lowering Medication on LDL					
	Method	Type I Error			
		5%	15%	25%	35%
Scenario 1: exposure independent of treatment assignment and LDL	MESA imputed	0.031	0.039	0.048	0.038
	ignore	0.039	0.048	0.051	0.057
	adjust	0.043	0.047	0.054	0.047
	exclude	0.045	0.056	0.050	0.048
Scenario 2: exposure increases odds of treatment, but is independent of LDL	censored normal	0.040	0.045	0.055	0.048
	MESA imputed	0.045	0.030	0.036	0.049
	ignore	0.076	0.441	0.815	0.973
	adjust	0.322	0.409	0.388	0.436
	exclude	0.663	0.665	0.657	0.637
	censored normal	0.183	0.039	0.088	0.253
Power					
Scenario 3: exposure independent of treatment, but increases LDL	MESA imputed	0.896	0.903	0.905	0.911
	ignore	0.919	0.912	0.870	0.789
	adjust	0.904	0.919	0.907	0.914
	exclude	0.837	0.845	0.842	0.851
	censored normal	0.904	0.911	0.900	0.895
Scenario 4: exposure increases odds of treatment, and increases LDL	MESA imputed	0.819	0.842	0.826	0.819
	ignore	0.802	0.297	0.055	0.344
	adjust	0.361	0.335	0.318	0.240
	exclude	0.088	0.071	0.095	0.076
	censored normal	0.992	0.947	0.769	0.435