The Use of Multiple Imputation for Data Subject to Limits of Detection

Ofer Harel[1]*, Neil Perkins[2], and Enrique F. Schisterman[2]

[1]Department of Statistics, University of Connecticut, USA

[2]Epidemiology Branch, *Eunice Kennedy Shriver* National Institute for Child and Human Development, Rockvile, MD, USA

*Corresponding Author: ofer.harel@uconn.edu

Received: 23, July 2013 / Revised: 27, January 2014 / Accepted: 8, March 2014

ABSTRACT

Missing data due to limit of detection and limit of quantification is a common obstacle in epidemiological and biomedical research. We are interested in methodologies that provide unbiased and efficient estimates of these missing data while using popular statistical software. We describe a multiple imputation (MI) procedure for cross-sectional and longitudinal data which examines the sources of variation of hormones levels throughout the menstrual cycle conditional on specific biomarkers. We describe the rational, procedure, advantages and disadvantages of the multiple imputation procedure. We also provide a comparison to commonly used missing data procedures (complete cases analysis and single imputation). We illustrate our approach using the BioCycle data where we are interested in the effects of Vitamin E and Beta-carotene on Progesterone levels. We also evaluate the longitudinal impact of changes in Vitamin E on Progesterone levels over time. Finaly, we demonstrate the advantages of using MI over complete case analysis or naive single replacement in both cross-sectional and longitudinal analysis where measurements below the limit of quantification (LOQ) are unreported. We also illustrate that if available, inclusion of potentially demined unreliable data below the limit of detection (LOD) improves simple estimation substantially.

Keywords: Multiple imputation procedure, Complete case analysis, Longitudinal analysis, Cross sectional data

1. Introduction

As new biomarkers emerge in basic science settings, epidemiologists and statisticians are to evaluate the effectiveness and utility of these new biomarkers.

When evaluating new biomarkers, two laboratory quality control limits, limit of detection (LOD) and limit of quantification (LOQ), are commonly utilized and are an obstacle in epidemiological and other empirical studies for estimation of the health effect in humans. Although sometimes they are theoretically available, values below the LOD and/or LOQ are rarely released by the laboratories to epidemiologist causing an incomplete data situations. The LOD is defined as the lowest analyte concentration that can be distinguished with reasonable confidence from background noise, that is to say that the measurements above the LOD reflect some true level of the biomarker being present. The LOQ is defined as the value at which the coefficient of variation (standard deviation divided by the mean) of the measurement is greater than some threshold, usually 20% is commonly accepted. Measurements below the LOQ are deemed unreliable in terms of the magnitude of the biomarker measured actually being present in contrast to background noise. In general the LOD is lower than the LOQ (Currie, 1968). Observations below the LOD and LOQ are not reported and are thus missing for further data analysis. This incomplete setup might cause bias, inefficiency and in most cases will make the analyses more complex. Since these missing values are often reported without distinction of being below LOD or LOQ, henceforth we will refer solely to the LOQ unless otherwise stated.

In this paper we are interested in different methodologies to effectively analyze data subject to LOQ. One commonly used method is complete case analysis (CCA), where observations with values below the LOQ are simply eliminated. Another method is single imputation, where every value below the LOQ is replaced by a constant such as $LOQ/2$ or $LOQ/\sqrt{2}$ (Hopke *et al.*, 2001; Richardson and Ciampi, 2003; Schisterman *et al.*, 2006). More complex methodology such as maximum likelihood (ML) (Little and Rubin, 2002) and Bayesian analysis (Gelman *et al.*, 2003) have become more prominent in the last few years to account for this missingness.

While CCA and single imputation are easy to employ, subsequent analysis of CCA commonly results in biased estimation, while analysis using single imputation will result in impaired estimates of variances and covariances (Little and Rubin, 2002). Bayesian and ML methods can provide unbiased and more efficient estimation but are often dependent on strong assumptions and are more difficult to apply in practice. Here we apply multiple imputation (MI) procedures (Harel and Zhou, 2007; Little and Rubin, 2002; Rubin, 1987) to the censoring of data below the LOQ, and present it as a method that is easy to employ while maintaining unbiasedness and efficiency. Multiple imputation is a well established procedure to deal with missing data, and an extensive body of literature has established it as one of the leading methodologies to deal with incomplete data. Each missing value (or value below the LOQ) is replaced by m appropriate values resulting in m complete datasets. Then each complete or filled in dataset is analyzed independently using a CCA (standard techniques without missing data) which results in m sets of estimates and their variances. These estimates are then combined to form a final estimate reflecting the observed data and the fact that some of the values were missing. Standard MI procedures for continuous missing values are often unrestricted, meaning that imputed values are selected across the biomarker's entire distribution. Here missing biomarker levels are known to be less than the LOQ and thus will only be imputed from this lower range or restricted to this censored portion of the distribution.

Since the data are missing, we need to make some assumptions about the process in which the missing data occur in order to choose the appropriate imputation procedure. There are three assumptions which are the basis of any analysis of incomplete data (Little and Rubin, 2002; Rubin, 1976): Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR). MCAR occurs when the missing values are not related to any measure in the study. In this case, the missingness, the random variable(s) that govern the missing data process is uncorrelated with variables to be used in the analysis. In this situation, the reason values are above or below the LOD is random and has nothing to do with the outcome of interest or other measured covariates. The process that causes the missing data (i.e., missingness – the set of reasons for missing data) is independent of any variables in the data. An example for MCAR process is missing by design; in this case biomarkers are measured on a predetermined proportion (p) of samples and are not measured one the remaining samples $(1-p)$ by design. Clearly, missing values below the LOQ do not fit this scenario but rather MAR where the missingness process depends only on observed data. MNAR occurs when MAR does not hold, and the missingness process actually depends on missing values. For example, consider a response variable where values below the LOD are plausible. In addition there is a covariate in which the response will be missing with probability 80% if the covariate value is above its median and 20% if the covariate value is below its median. If the covariate is observed the missing data process can be considered as MAR because this information will be used during imputation, however, if this covariate is missing MNAR will be the right assumption because it will be dependent on information not considered during imputation.

Rubin (1976) introduced this definition in order to find the minimum condition in which one DOES NOT have to model the missingness process. In order for that to occur, two assumptions must hold. First, we have to assume the MAR assumption is valid. Second, we have to assume independence between the estimates of parameters used for imputation and those estimated in the substantive model of interest. Having these two assumptions implies ignorability, which means one can ignore the missingness model and need to deal only with the observed data. In this study, we assume that the LOQ level is known, and hence the reason for missing values is observed which leads to MAR assumption and the use of ignorable models.

In this paper, we show how MI procedures can be used to account for data below the LOQ in two different scenarios. First we consider a cross-sectional data set up, where we are interested in the effect of two biomarkers on an outcome and want to use linear regression to assess these relationships. Second, we apply similar methodology accounting for missingness to a longitudinal scenario and ask how the biomarker (over time) affects the outcome. In both scenarios we will allow measurement of biomarkers to have a positive probability for values below the LOQ, and will compare the performance of MI relative to CCA and single imputation.

Additionally, we will examine the effects of using values below the LOQ on estimation. In most cases, laboratories will have measurements below the LOQ but will prefer not to release these values to researchers due to quality control convention. Guo *et al.* (2010) showed that the measurement error below the LOQ might be actually smaller then the measurement error above the LOQ. We will argue that it will be best if one can use the values below the LOQ instead of missing values delivered by the lab albeit the assumed larger measurement error below the LOQ.

In section 2 we will describe the biomarker data, main questions and the appropriate models. In section 3 we will present the example where we apply CCA, single imputation and MI to account for missing values. We will provide some results regarding both cross-sectional and longitudinal examples with and without the availability of values below the LOQ and discuss the implications of these results. We will present a small simulation study in section 4 and conclude with some discussion in section 5.

2. Data Description and Methods

Oxygen free radicals have been associated with spontaneous abortion, infertility, birth weight, and chronic disease processes. Circulating levels of estradiol and progesterone also fluctuate during the follicular and luteal phase of the menstrual cycle in ovulatory women. However, little is known about the relation between markers of oxidative stress and antioxidant status, estrogen levels and their influence on menstrual cycle function. The BioCycle Study sought to investigate the relation between endogenous reproductive hormone levels and antioxidants including vitamin levels across the menstrual cycle in a prospective cohort of premenopausal women. A sample of 259 women, age 18 to 44 at enrollment with a self-reported cycle length from 21 to 35 days and with no known condition that might affect her cycle (e.g., being underweight, current use of exogenous hormones including hormonal contraception, current breast feeding, etc.) were recruited (Wactawski-Wende *et al.*, 2009). Blood sample collection was limited by patient burden and cost to 8 draws per cycle with two cycles per woman. In order to capture exposure at critical windows of hormonal variability, 8 visits were differentially spaced out (approximately days $2, 7, 12, 13, 14, 18, 21, 27$ of a 28 day cycle) and timed based on a home fertility monitor measuring Luteinizing hormone surge rather than simple calendar time. This monitor based timing of specimen collection allowed for a more uniform and representative hormonal profiles across the varying menstrual cycle lengths of the participants.

With multiple timed measurements across 509 total cycles from 259 women, the richness of these data allow for a wide variety of analyses. Here we are focused on answering two important questions. Does antioxidant status measured via vitamin E or Beta-carotene levels affect the maximum level of progesterone? And, do serum progesterone levels change over the cycle and what influence does vitamin E have on this change? The first question is a cross sectional analysis evaluating the relationship between maximum progesterone levels as a function of circulating vitamin E and Beta-carotene levels. The second is a longitudinal analysis as we examine change of levels and relations over time. All the measurements of progesterone, vitamin E and Beta-carotene levels were subject to LODs and LOQs (Table 2.1). Fortunately, the Biocycle study dataset includes values below the LOQ but above the LOD. Given this additional information, we are going to illustrate these analyses in two ways. First we are going to analyze these cross sectional and longitudinal relations based on conventional laboratory reporting where all values below the LOQ are missing. Second, we will replicate the analyses while incorporating measurements between the LOQ and the LOD which are usually withheld from analysis, but are available for this data.

In table 2.1 we present some summary statistics of few of the main variables. We present the mean, standard deviation and percent missing in two scenarios. First, when values below the LOQ are missing. Second, when values below the LOQ are available (values below the LOD are not available). Vitamin E has 93.7% of its values below the LOQ but only 9.8% of it values below the LOD. It is also apparent that due to this large amount of potential missing values the mean and variance of Vitamin E change dramatically if we take the values below the LOQ in consideration.

Table 2.1: Summary Statistics of Cross Sectional Data.

	Missing			With values		
Variable	Mean	Std	% missing	Mean	Std	% missing
Progesterone	11.11	5.27	0.4%	11.11	5.27	0.4%
Vitamin E	0.54	0.48	93.7%	0.18	0.17	9.8%
Beta-carotene	0.20	0.14	6.5%	0.20	0.14	6.3%
BMI	24.09	3.87	0.2%	24.09	3.87	0.2%
Age	27.38	8.23	0.2%	27.38	8.23	0.2%

2.1. Cross Sectional Analysis and MI

Conventional MI (Harel and Zhou, 2007; Little and Rubin, 2002; Rubin, 1987) is a common missing data technique which replaces every missing value with $m \geq 2$ plausible values. The procedure is comprised of three stages. First, the imputation stage, in which we create multiple, say m, ”complete” datasets by filling in missing values for each dataset independently. Second, the analysis stage, in which we analyze the m ”complete” datasets using standard CCA for datasets with no missing data. Finally, the combining results stage, in which the results from the m analyses are combined into a single, global result that takes into consideration the variability in the data and the fact that some of the values were missing.

These steps can vary widely depending on the type of missingness, imputation model and analysis model. However, most commonly used statistical software includes built in packages and functions to perform MI for many scenarios. In general, when data is continuous the imputation model will be based on a multivariate normal model and when the data is categorical the imputation will be based on multinomial model. Models for mixed, continuous and categorical, variables will be based on location models (Schafer, 1997a). Limited review of methodology and applications of MI can be found in the appendix; some other review references are also available (Harel and Zhou, 2007; Schafer and Graham, 2002). Since we are dealing with continuous variables we will conform to the norm and use the joint normal model. It has been shown that the normal assumption is quite robust to distribution misspecification and even variables which are clearly not normal result in mostly unbiased and efficient estimates (Schafer, 1997a).

In our set-up we treat values below the LOQ as missing data. Doing so leads to a conventional missing data problem, where we know that all these missing values fall in the interval $(0, LOQ)$. We assume the missingness process follows the MAR assumption and therefore we can use ignorable models. If non-ignorable models are of interest, slightly more complex modeling will be needed. However, for brevity we will not elaborate on non-ignorable models, for more details see Little and Rubin (2002). Here in the imputation stage, we use this information when filling in missing values and (under our assumptions) maximizing the amount of information used in the analysis. The assumptions required for the MI analysis are very comparable to those required by maximum likelihood analysis (normality, ignorability, etc). There is no clear guide line for the amount of missing values allowed before the procedure collapse. The advantage of MI relative to ML is that it the rate of missing information, a measure quantifying the impact of the missing values on the inference can be estimated and evaluate its effects (Harel, 2007; Harel and Zhou, 2007; Little and Rubin, 2002; Rubin, 1987).

2.2. Longitudinal MI

When variables represent the measure of specific quantity over time, we are dealing with a longitudinal structure. Again, consider our example where progesterone and vitamin E levels are measured several times along the woman's biological cycle. Suppose each column represents a variable while each line represents a subject (one can represent each subject with multiple lines, but for the purpose of MI one line per subject is required). However, there is a specific relationship between the variables.

The common imputation model (cross-sectional) is a joint normal model. This model can work to impute longitudinal data under some assumptions. In this model (under the longitudinal setup) the variables are treated as normal variables with a common joint normal distribution. With each, the measured value at time t, follows a normal distribution with mean and variance and that can be correlated with any or all other time points. This simple joint model will be appropriate for MI when the missing data are sparse cross-sectionally and the between-variable correlations dominate the within-individuals autocorrelations. More complex models have been proposed (e.g. PAN imputation or panel data (Schafer and Yucel, 2002)), but are not yet commonly used and are not available in commercial statistical software such as SAS (SAS Institute Inc., 2003). In particular PAN modeling allows missing values on the responses, though does not allow any missing values in the covariates. In our case however, the majority of missing values occur on Vitamin E which is a covariate. Therefore we chose to use the normal imputation.

After the missing values below the LOQ are imputed, the next two stages of MI follow as described in section 2.1. The m sets of ”complete” data are analyzed using the appropriate CCA for longitudinal data. Then the results from these analyses are combined to arrive at a single, final result taking into account the observed data and the missing values (values below the LOQ).

3. Progesterone and Vitamin E in Women

3.1. Cross-Sectional Data

We are interested in evaluating the effect of serum vitamin E and Beta-carotene levels on maximum progesterone levels during the menstrual cycle. For each woman, the maximum Progesterone level for the cycle is identified along with the appropriate vitamin E and Beta-carotene levels. For simplicity consider the model with fixed effects as below

$$Progesteron = \beta_0 + \beta_1(Vit\,E) + \beta_2(Beta - carotene),$$

and a random intercept for each individual as the model we are interested in evaluating in our cross sectional analysis. Recall from Table 2.1 that all of the measurements in this model are subject to LOQ and as such the laboratories commonly report these values as missing.

We will account for this missingness and compare subsequently estimated models by applying three techniques: (1) complete case analysis by naively omitting missing observations below the LOQ; (2) single imputation - every missing value is replaced by $LOQ/2$; (3) multiple imputation. After applying CCA and single imputation, the model above was estimated using proc Mixed in SAS (SAS Institute Inc., 2003). For MI we used a chain equation implemented in IVEware (Raghunathan *et al.*, 2002) an MI macro for SAS (SAS Institute Inc., 2003). Using IVEware the user can specify the interval in which the imputed values should fall. Here the LOQ of progesterone was 0.2, hence we specified that the imputed values fall between 0 and 0.2 for progesterone. After the data were imputed, the regression model specified above was estimated from all $(m = 10)$ ”complete” sets of data and the results were combined using Rubin (1987) rules through proc mianalize in SAS.

These results are presented in Table 3.1. It is clear that both single imputation and MI are preferred over CCA. CCA relies on the least amount of data resulting in estimates that appear biased (see estimates of β_2 for example, where CCA has an insignificant estimate of 3.38 with standard deviation of 6.27, compared to significant estimate of MI 6.74 with standard deviation 1.74) and have the largest standard errors. While the standard errors are similar, the magnitude of the estimated relationship is greater and thus more significant based on MI rather than a single imputed value (see estimates of β_2 for example, where single imputation has a significant estimate of 4.70 with standard deviation 1.85, compared to MI with $6.74(1.74)$). Thus antioxidant status measured via vitamin E or Beta-carotene levels appear to affect the maximum level of progesterone, with significance and magnitude which is greatest using MI.

Because of the richness of the Biocycle data set we are able to perform the same analysis while utilizing the availability of values below the LOQs (we observed progesterone 95.7%, Beta-carotene 93.6% and vitamin E 90.7%). Estimators based on CCA change markedly and standard deviation is dramatically reduced, from 6.27 with these values to 1.99 for β_2, however not enough to result in a statistically significant relationship. When comparing single imputation and MI the effects of including these data are less pronounced. However, from theoretical point of view and from the apparent results it is clear MI is superior. The Beta-carotene coefficient β_2 is 30% smaller for single imputation compared to MI. For the vitamin E coefficient there is a difference of 25% when the values below the LOQ are given and a difference of 50% when they are not.

Table 3.1: Regression Coefficients (Standard Deviation) of the Effect of Vitamin E and Beta-carotene on Progesterone.

	Missing			With values		
	CCA	Single imp	MI	CCA	Single imp	MI
β_0	10.32	10.37	9.62	10.55	10.35	9.74
	(2.12)*	(0.50)*	(0.46)*	(0.54)*	(0.50)*	(0.46)*
β_1	-1.76	-1.59	-2.12	-0.96	-1.33	-2.85
	(2.03)	(1.47)	(1.63)	(1.42)	(1.41)	(1.60)
β_2	3.38	4.70	6.74	3.47	4.65	6.71
	(6.27)	(1.85)*	(1.74)*	(1.99)	(1.85)*	(1.72)*

* Significant at 5% significance level

3.2. Longitudinal Data

Changes in vitamin E level are also of interest and its impact on Progesterone throughout the woman's biological cycle. As described in section 2, we have information about the biomarker (vitamin E) and outcome (Progestrerone) on days $2, 7, 12, 13, 14, 18, 21,$ and 27 of the cycle. Similar to the cross sectional analysis, we will illustrate the utility of MI through the relatively simple model, where the fixed effects are

$$Progesteron = \beta_0 + \beta_t (Vit\,E)_t; \quad t = 2, 7, \ldots, 27,$$

while there are subject specific (intercept) random effect.

Similar to section 3.1, we evaluate this model based on data traditionally reported above the LOQ and again when we have values below the LOQ. We are going to compare CCA, single imputation and MI. The results are presented in table 3.2. CCA and single imputation were analyzed as described above. For MI, we used IVEware (Raghunathan *et al.*, 2002) in SAS (SAS Institute Inc., 2003) as it was described in section 3.1.

Since IVEware (Raghunathan *et al.*, 2002) was not developed to deal with multilevel data, the imputation model does not take into consideration the within subject correlations. Although this is a limitation to the study, the alternative is to write a specific MCMC code for the imputation stage, which would require a different code for different models and a capable programmer. The use of the R

(R Development Core Team, 2008) package called PAN (Schafer, 1997b, 2001) which allows multilevel imputation (observations within subjects), is problematic because it allows only missing in the response variables and not in covariates. In our set-up most of the missing values were in the covariates. One may consider imputing using PAN while flipping the model (look at the responses as covariates and at the covariates as responses) however, this will cause non-conginial model (Meng, 1994) which might be problematic for the estimation. For these reasons and in order for other researchers to be able to replicate our procedures, we chose to use the currently available software (SAS) and the specific assumptions mentioned above. Using these assumptions we control for pair wise correlations between the variables but not higher level interactions.

The results in table 3.2 show that serum progesterone levels do seem to change over the cycle and are influenced by vitamin E. Again, CCA is a poor choice with coefficients clearly inconsistent with those obtained using single imputation and MI. However, if obtained, the use of data below the LOQ results in CCA estimators that is reasonable. Using the values below the LOQ, CCA results in a model similar to single imputation and MI. This might be an indicator that with the values below the LOQ present the MCAR assumption is reasonable and therefore not much difference is expected between the different estimation procedures (i.e. CCA, single imputation and MI). It is not surprising then that single imputation and MI perform similarly.

Table 3.2: Regression coefficients (Standard Deviation) of the Longitudinal Effect of Vitamin E on Progesterone.

	Missing			With values		
	CCA	Single imp	MI	CCA	Single imp	MI
β_0	2.77	2.99	3.16	3.13	3.01	3.05
	(0.42)*	(0.12)*	(0.13)*	(0.13)*	(0.12)*	(0.12)*
β_1	-1.85	-7.10	-7.05	-6.84	-6.93	-6.97
	(1.02)	(0.87)*	(0.84)*	(0.87)*	(0.85)*	(0.85)*
β_2	-2.18	-6.90	-7.00	-6.22	-6.73	-6.78
	(0.94)*	(0.81)*	(0.79)*	(0.82)*	(0.79)*	(0.79)*
β_3	-1.23	-3.89	-4.20	-3.68	-3.84	-3.89
	(0.70)	(0.65)*	(0.64)*	(0.65)*	(0.64)*	(0.64)*
β_4	-1.64	-3.49	-3.82	-3.55	-3.43	-3.55
	(0.91)	(0.79)*	(0.76)*	(0.79)*	(0.77)*	(0.78)*
β_5	-0.34	0.28	-0.43	-0.07	0.30	0.15
	(1.06)	(0.89)	(0.78)	(0.89)	(0.87)	(0.87)
β_6	4.11	15.25	12.71	13.56	14.60	14.27
	(1.16)*	(0.95)*	(1.01)*	(0.93)*	(0.90)*	(0.91)*
β_7	6.69	20.37	17.27	18.12	19.28	18.88
	(1.16)*	(0.95)*	(1.02)*	(0.94)*	(0.92)*	(0.92)*
β_8	1.67	8.53	6.54	7.31	8.11	7.87
	(1.37)	(1.10)*	(1.08)*	(1.10)*	(1.06)*	(1.07)*

* Significant at 5% significance level

4. Simulations

We conducted a brief simulation exploring the scenario in our cross-sectional example. We started by generating random data based on the log transformed distributions of Vitamin E (x) and Beta-carotene (z) and then produced dependent data with random error reflected in our progesterone data on day $7(y)$. Specifically, we generated x from a normal distribution $N(2.08, 0.252)$, z and y had the following structure,

$z = -3.187 + 0.652x + \epsilon_1$, $y = -0.484 - 0.335x - 0.11z + \epsilon_2$ where ϵ_1 $N(0, 0.652)$ and ϵ_2 $N(0, 0.592)$. Where $N(a, b)$ represent a normal distribution with mean a and variance b. Bias, standard error and mean square error (mse) for point estimates and confidence interval width and coverage probability were calculate based on 1000 iterations of samples of size $n = 500$.

Coefficients were estimated using all of the data, no censoring, and then again using the various methods to account for $10, 20$, and 50% of missing data in the x variable. The y and z variable were not censored.

Table 4.1 shows the results for the coefficient for x, the variable affected by the LOD. The bias is the least for CCA followed closely by MI. The single imputation of both 0 and $LOD/2$ showed substantial bias. With respect to standard error and mse MI performed the best, nearly as well as when all of the data was present. The small bias and standard error lead to confidence intervals with coverage far better than the unacceptable coverage due to single imputation and were markedly shorter than those generated via CCA.

Table 4.1: Simulation Results for Regression Coefficients at Risk for Missing Values.

		Point Estimate				CI	
Scenario	%miss	Mean	Std	bias	mse	Width	Cov
Complete Data	0	-0.338	0.107	-0.002	0.011	0.429	0.952
CCA	10	-0.338	0.131	-0.003	0.017	0.531	0.957
	20	-0.339	0.152	-0.004	0.023	0.621	0.956
	50	-0.328	0.252	0.007	0.063	0.994	0.959
Impute (0)	10	-0.098	0.040	0.237	0.058	0.160	0
	20	-0.076	0.031	0.259	0.068	0.121	0
	50	-0.061	0.023	0.274	0.076	0.093	0
Impute (LOD/2)	10	-0.175	0.064	0.160	0.030	0.252	0.295
	20	-0.141	0.053	0.195	0.041	0.205	0.043
	50	-0.114	0.042	0.221	0.051	0.169	0.003
MI	10	-0.344	0.111	-0.009	0.012	0.442	0.944
	20	-0.359	0.117	-0.024	0.014	0.468	0.944
	50	-0.425	0.155	-0.090	0.032	0.627	0.924

5. Conclusions

As in any missing data problem, ignoring the problem is equivalent to making a strong assumption about the data. It is well known that ignoring the missing data problem and using CCA will be valid under MCAR, but will likely be biased in other situations. In addition, the inference will be inefficient since the confidence intervals will be much larger than competing methods. Since the missingness process can not be MCAR in our situation, it is obvious that CCA performed poorly and should not be used. In both cross-sectional and longitudinal analyses we showed that the use of CCA will lead to different models (results) compared with single and multiple imputation.

Due to the fact that the missing values were missing in a specific area of the distribution (below the LOD, LOQ), it is not surprising that single and multiple imputation performed similarly with respect to the significance of the models and the estimation in general. In the cross-sectional example, while the standard errors are similar, the magnitude of the estimated relationship is greater and thus more significant based on MI rather than a single imputed value. Based on our simulations and the large volume of literature arguing that MI is superior to single imputation and the magnitude difference between the two methods we will recommend using MI when ever possible.

Guo *et al.* (2010) showed that the logic underlying not reporting values below the LOQ is flawed. They showed that the standard deviation of the predictive distribution increases with the value of the biomarker. This led them to the conclusion that ”in absolute terms, the distortion in the reported values above the LOQ is actually worse than the distortion in the values below the LOQ, which is the rationale for not reporting them.” In this manuscript we showed that the results that use the values below the LOQ are superior to those without.

It is common that laboratories will not release values below the LOQ. However, if the research community will be able demonstrate the advantage of using these values (as we have done here) the laboratories will have to conform to the demand and release these values.

Any imputation model will require several different types of assumptions. In this manuscript we assumed the data is missing at random and ignorable based on Rubin (1987) definitions. We also assumed that the cross-sectionally imputation will work for the longitudinal data as well. By doing so, we actually assumed that the variables are independent identically distributed. This will be appropriate when the between variable variation dominates the within variable autocorrelation.

We used a joint normal imputation model for the cross-sectional analysis and added the assumption that the between-variable correlations dominate the within-individuals autocorrelations in the longitudinal analysis. Although, this will be one of the limitations of this manuscript, the advantages of using this procedure are its ease, the ability to use common statistical software, and the advantages over other procedures being used for limit of detection problems.

There is not much existing work on left- or interval-censored exposures. Lynn (2001) introduced a maximum likelihood approach for left-censored HIV viral load, and compared it to ad hoc substitutions and multiple-imputation. When a left-censored exposure is subject to measurement error, Richardson and Ciampi (2003) suggested a substitution with the expectation below the LOD. They showed it will provide an unbiased estimate. Gomez *et al.* (2003) showed by simulation that midpoint substitution is flawed for a discrete-valued interval-censored baseline characteristic in a randomized trial. In linear and logistic regression, Schisterman *et al.* (2006) suggested a substitution method for handling left-censored exposures. They concluded that the values below LOD will be best if replaced with the expectation above the LOD. Lei *et al.* (2010) incorporates the approaches of Richardson and Ciampi (2003) and Schisterman *et al.* (2006) for linear regression under a likelihood-based framework.

Since the values below the LOD cannot (accurately) be measured, it is reasonable to consider that these values are a mixture of true zeros and values below the LOD. Some approaches have been developed for outcomes (Berk and Lachenbruch, 2002; Chu *et al.*, 2010), but to our knowledge not for exposures. An MI approach that allows this type of mixture was developed by Olsen and Schafer (2001).

In this manuscript we did not compared the MI results with ML and Bayesian analysis. It has been shown that under the same assumptions and vague priors the result should be quite similar. This manuscript does compare MI to the two most commonly used methods for handling data missing due to a LOD. CCA and single replacement are used not for their impressive estimation properties but rather presumably for ease of implementation. We have shown here that MI as a method to account for missing data below a LOD remains easy to employ while also achieving estimation with minimal bias and nominal coverage probability. It has been shown that procedures such as MI, ML, or Bayes will be superior in most circumstances over CCA and single imputation with MI being by far the simplest to practically employ.

Appendix

The imputation stage is the complex part of MI, since the analysis would be equivalent to a scenario of complete data and the combining rules are simple and implemented in many statistical software packages. There are many ways to implement the imputation stage. The types of variables have an important role in the imputation procedure. The most common imputation model is designed for continuous variables under multivariate normality assumption. Other models available are for categorical data, mixed model for categorical and continuous variables, and the model for multi-level data. More detailed information can be found in Schafer (1997a). It is important to specify the analysis model in advance, so that the imputation model will be as general as possible and will contain all the variables and variable relationships (interactions) one thinks are preset in the data. It was established that there is an advantage for using a comprehensive approach for the imputation model, in which one would use as many variables possible that might be related to the outcome and/or the missingness process (Collins *et al.*, 2001). If the imputation model and the analysis model are not the same we are subject to a non-congeniality case (Meng, 1994). When both models are the same (and using non-informative priors) the results of MI and ML will be very similar.

In most cases, the imputation model will not be in a closed form which will lead us to use some Markov Chain Monte Carlo (MCMC) procedures. Computations are summarized in Schafer (1997a). Lately, the use of Multiple Imputation using Chain Equations (MICE) (Van-Buuren and Oudshoorn, 1999) have become more commonly used (Horton and Lipsitz, 2001). Using MICE, the researcher represents the posterior distribution of the imputation model as a product of equations based on Bayes rule. All of which is built into standard software such as SAS (SAS Institute Inc., 2003), Splus (Schimert *et al.*, 2000), and R (R Development Core Team, 2008).

For the use of MI we must assume that with complete data, tests and intervals based on the normal approximation $(\hat{Q} - Q)/\sqrt{U} \sim N(0,1)$ would be appropriate. In the absence of Y_{mis}, we have random versions or imputations $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$ from which we calculate the imputed-data estimates $\hat{Q}^{(j)} = \hat{Q}(Y_{obs}, Y_{mis}^{(j)})$ and their estimated variances $U^{(j)} = U(Y_{obs}, Y_{mis}^{(j)}),\ j = 1, 2, \ldots, m$. The overall estimate of Q is $\bar{Q} = m^{-1}\sum \hat{Q}^{(j)}$. To obtain a standard error for $\bar{Q}$, we calculate the between-imputation variance $B = (m - 1)^{-1}\sum(\hat{Q}^{(j)} - \bar{Q})^2$ and $\bar{U} = m^{-1}\sum U^{(j)}$, the within-imputation variance. The estimated total variance is $T = (1 + m^{-1})B + \bar{U}$, and tests and confidence intervals are based on a Student's t approximation $(\bar{Q} - Q)/\sqrt{T} \sim t_\nu$, with degrees of freedom $\nu^{-1} = \frac{1}{(m-1)}\left[\frac{(1+m^{-1})B}{T}\right]^2$.

Acknowledgements

This project was partially supported by Award Number K01MH087219 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

This research was also partially supported by the Intramural program of the *Eunice Kennedy Shriver* National Institute for Child and Human Development, National Institutes of Health and the Long Range Initiatives grant, from the American Chemistry Council.

References

1. Berk, K. N. and Lachenbruch, P. A. (2002). Repeated measures with zeros. *Statistical Methods in Medical Research*, **11**(4), 303–316.

2. Chu, H., Gange, S., Li, X., Hoover, D., Liu, C., Chmiel, J., and Jacobson, L. (2010). The effect of haart on hiv rna trajectory among treatment-nave men and women: A segmental bernoulli/lognormal random effects model with left censoring. *Epidemiology*, **21**(4), S2–S34.

3. Collins, L., Schafer, J., and Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**, 330–351.

4. Currie, L. (1968). Limits for qualitative detection and quantitative determination– application to radiochemistry. *Analytical Chemistry*, **40**(3), 586–592.

5. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition.

6. Gomez, G., Espinal, A., and W. Lagakos, S. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, **22**(3), 409–425.

7. Guo, Y., Harel, O., and Little, R. (2010). How well quantified is the limit of quantification? *Epidemiology*, **21**(4), S10–S16.

8. Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, **4**, 75–89.

9. Harel, O. and Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, **26**(16), 3057–3077.

10. Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic. *Biometrics*, **57**(1), pp. 22–33.

11. Horton, N. and Lipsitz, S. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, **55**(3), 244–254.

12. Lei, N., Chu, H., Liu, C., Cole, S., Vexler, A., and Schisterman, E. (2010). Linear regression with an independent variable subject to a detection limit. *Epidemiology*, **21**(4), S17–S24.

13. Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

14. Lynn, H. S. (2001). Maximum likelihood inference for left-censored hiv rna data. *Statistics in Medicine*, **20**(1), 33–45.

15. Meng, X. (1994). Multiple imputation inference with uncongenial sources of input (with discussion). *Statistical Science*, **10**, 538–573.

16. Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, **96**(454), pp. 730–745.

17. R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

18. Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). IVEware: Imputation and Variance Estimation Software. User guide, Institute for Social Research, University of Michigan.

19. Richardson, D. B. and Ciampi, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, **157**(4), 355–363.

20. Rubin, D. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

21. Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, New York.

22. SAS Institute Inc. (2003). *SAS/STAT Software, Version 9.1*. Cary, NC.

23. Schafer, J. (1997a). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

24. Schafer, J. (1997b). Imputation of Missing Covariates Under a Multivariate Linear Mixed Model. 97-04, Dept. of Statistics, The Pennsylvania State University.

25. Schafer, J. L. (2001). *New Methods for the Analysis of Change*, chapter Multiple Imputation with PAN, pages 357–377. American Psychological Association, Washington, DC.

26. Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**(2), 147–177.

27. Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, **11**(2), pp. 437–457.

28. Schimert, J., Schafer, J., Hesterberg, T., Fraley, C., and Clarkson, D. (2000). *Analyzing Data with Missing Values in S-PLUS*. Data Analysis Products Devision, MathSoft, Seattle, WA. Software.

29. Schisterman, E. F., Vexler, A., Whitcomb, B. W., and Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology*, **163**(4), 374–383.

30. Van-Buuren, S. and Oudshoorn, C. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO Preventie en Gezondheid. TNO/VGZ/PG 99.054.

31. Wactawski-Wende, J., Schisterman, E. F., Hovey, K. M., Howards, P. P., Browne, R. W., Hediger, M., Liu, A., and Trevisan, M. (2009). Biocycle study: Design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology*, **23**(2), 171–184.