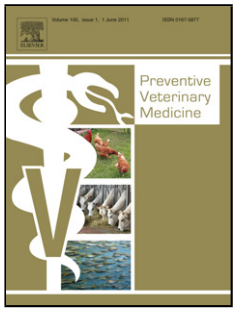

Title: Multiple imputation in veterinary epidemiological studies: a case study and simulation

Author: Ian R. Dohoo Christel R. Nielsen Ulf Emanuelson

PII: S0167-5877(16)30108-8
DOI: http://dx.doi.org/doi:10.1016/j.prevetmed.2016.04.003
Reference: PREVET 4013

To appear in: *PREVET*

Received date: 31-7-2015
Revised date: 2-3-2016
Accepted date: 4-4-2016

Please cite this article as: Dohoo, Ian R., Nielsen, Christel R., Emanuelson, Ulf, Multiple imputation in veterinary epidemiological studies: a case study and simulation.Preventive Veterinary Medicine http://dx.doi.org/10.1016/j.prevetmed.2016.04.003

Multiple imputation in veterinary epidemiological studies: a case study and simulation

Ian R. Dohooa*, Christel R. Nielsenb, Ulf Emanuelsonc

a Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI, C1A 4P3Canada
b Unit for Medical Statistics and Epidemiology, R&D Center Skåne, Skåne University Hospital, Lund, Sweden
c Department of Clinical Sciences, Swedish University of Agricultural Sciences, POB 7054, SE-75007 Uppsala, Sweden

*Corresponding author:
Tel.: +1 902 566 0640
Fax: +1 902 620 5053
E-mail address: dohoo@upei.ca

Abstract

The problem of missing data occurs frequently in veterinary epidemiological studies. Most studies use a complete case (CC) analysis which excludes all observations for which any relevant variable have missing values. Alternative approaches (most notably multiple imputation (MI)) which avoid the exclusion of observations with missing values are now widely available but have been used very little in veterinary epidemiology.

This paper uses a case study based on research into dairy producers' attitudes toward mastitis control procedures, combined with two simulation studies to evaluate the use of MI and compare results with a CC analysis. MI analysis of the original data produced results which had relatively minor differences from the CC analysis. However, most of the missing data in the original data set were in the dependent variable and a subsequent simulation study based on the observed missing data pattern and 1000 simulations showed that an MI analysis would not be expected to offer any advantages over a CC analysis in this situation. This was true regardless of the missing data mechanism (MCAR - missing completely at random, MAR - missing at random, or NMAR - not missing at random) underlying the missing values. Surprisingly, recent textbooks dealing with MI make little reference to this limitation of MI for dealing with missing values in the dependent variable.

An additional simulation study (1000 runs for each of the three missing data mechanisms) compared MI and CC analyses for data in which varying levels (n=7) of missing data were created in predictor variables. This study showed that MI analyses generally produced results that were less biased on average, were more precise (smaller SEs), were more consistent (less variability between simulation runs) and consequently were more likely to produce estimates that were close to the “truth” (results obtained from a data set with no missing values). While

the benefit of MI varied with the mechanism used to generate the missing data, MI always performed as well as, or better than, CC analysis.

Keywords: "multiple imputation", questionnaire, "dependent variable", simulation, MCAR, MAR, NMAR

1. Introduction

Despite the fact that missing data is a common problem in veterinary epidemiologic research, there is little information about the issue in the veterinary literature. For example, a search of all articles published in this journal identified only a single paper with "missing data" or "imput*" in the title (Dohoo, 2015). Of 5 other papers with one of these terms in the Abstract, only one (Hopp et al., 2001) described the use of imputation for dealing with the problem.

The standard approach to dealing with missing data has been to do a complete-case (CC) analysis (also known as listwise deletion) in which any observation with any missing values in a variable pertinent to that analysis are ignored. However, there have been substantial improvements in the ability of researchers to deal with missing data. These approaches were reviewed in one of the two previously cited papers (Dohoo, 2015). Of the options available, multiple imputation (MI) has received the most attention in the general epidemiology literature and procedures for multiple imputation are now easily accessible in multiple statistical packages. There are now several texts which deal with the use of MI to handle

missing data (Enders, 2010; Heeringa et al., 2010; Graham, 2012; van Buuren, 2012; Carpenter and Kenward, 2013; StataCorp, 2013).

1.1 Overview of multiple imputation

It is beyond the scope of this manuscript to provide a detailed description of the multiple imputation process (see texts referenced above), but a brief overview is provided for readers not familiar with the procedure. Multiple imputation consists of three general steps: creating multiple data sets with missing values filled in by the imputation process, analysing each of the generated data sets, and combining the results from the multiple analyses. The steps are as follows.

Generating imputed data sets. The process of filling in missing values with plausible replacements is done by the imputation model which uses available data to come up with good ”guesses” as to what the missing values are. The imputation model may be an explicit model with a defined statistical distribution (eg multivariate normal) or an implicit model which selects reasonable values from the existing data (eg predictive mean matching). The development of an imputation method called chained equations (also known as fully conditional specification) allows for the simultaneous imputation of missing values in multiple variables and allows for different imputation models for different variables. The process of imputation is stochastic so each generated data set will be unique. The issues of deciding which variables to use in the imputation process, which type of imputation model to use and how many imputed data sets to generate are discussed below.

Analysing imputed data sets. The analysis model is the model designed to answer the main research question of interest. The type of model and selection of variables used in this model

may differ from the imputation model(s). A complete set of coefficient estimates and their standard errors (SEs) are obtained for each generated data set.

Combining (pooling) results. The process of combining the results from the multiple data sets uses a procedure based on Rubin's rules (Rubin, 1987). This process averages the coefficient estimates but takes both the within-data set variation and between-data set variation into account in estimating the pooled SEs.

1.2 Missing data mechanisms

A key question to be addressed when dealing with missing data is "why are the data missing?", i.e. the missing data mechanism. There are three mechanisms by which data may be missing: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR - also called MNAR) and these have been explained in a recent publication - "MCAR means, as its name implies, that the probability of a value being missing in a variable is a completely random event. ... MAR means that the probability of being missing depends only on the observed data (ie the probability of being missing can be fully explained by variables recorded in the data set). ... Data are NMAR if the probability of being missing depends on unobserved data." (Dohoo, 2015).

However, it is important to keep in mind that the missing data categories are not mutually exclusive, nor are all variables in a data set likely to have values missing for the same reason(s). To extend the example from the previous publication (missing data mechanisms potentially affecting response to a question "Do you allow your dog to run free"), if people who let their dog run free are more likely to refuse to answer the question, the reason for the missingness is itself unknown so that data would be considered NMAR. However, if rural respondents were also more likely to not respond (regardless of whether or not they let their

dog run free) and the subjects rural/urban status was recorded in the data, there would be an element of MAR to the data. Finally, if some non-responses were completely random (perhaps the respondent just missed the question), then there would also be an element of MCAR to those data as well. There is no statistical way to determine if the data are MCAR, MAR or NMAR (or a combination of these), so the question bears careful reflection. Given this, it is difficult to reliably predict the expected benefits from MI for any particular variable in a study.

1.3 Background to case-study

Mastitis remains a major issue in dairy production, despite the fact that control programs have been available and implemented since the 1960s. For instance, although the lactational incidence risk of veterinary-diagnosed clinical mastitis in herds participating in the Swedish official milk-recording scheme has declined over the past 15 years it remains over 10% (Anonymous, 2014). Reasons for this are most likely manifold, but one could possibly be a poor uptake and application of suggested mastitis control measures. Research has found that it may be difficult to reach farmers with advice (Jansen et al., 2009; Jansen et al., 2010), which indicates that a lack of implementation may indeed be part of the explanation. However, very little work has been done to determine what factors influence attitudes toward and implementation of control procedures

Collection of information on implemented mastitis control measures by on-site inspection of actual procedures would provide the most accurate data, but sample size would be limited because of the cost. An alternative would be collection of information through a questionnaire, but questionnaires are often plagued by incomplete answers which leads to missing data. A questionnaire study on attitudes toward, and implementation of, mastitis control procedures was performed in Sweden in 2011 (Nielsen and Emanuelson, 2013) and,

although the study provided useful information, it was limited by a lot of missing values. This was particularly true for questions regarding the dairy producer's attitudes toward control procedures which made multivariable modelling of factors influencing their attitudes difficult.

1.4 Objectives

The objectives of this study were twofold. The first being to compare results obtained from a complete case (CC) analysis of the mastitis attitudes data (in which most missing values were in the dependent variable) to those obtained from an MI analysis, and to determine which set of results was most reliable. The second being to evaluate a variety of missing data scenarios (with an emphasis on missing predictor variables) to determine when MI was most likely to be useful, and to evaluate the impact of MI analyses as compared to CC analyses.

These objectives were to be met through the use of a case study and a set of simulation based analyses.

2. Materials and Methods

2.1 Mastitis attitudes data

Data for this study were collected through an extensive self-administered postal questionnaire sent in May 2011 to a sample of 898 Swedish dairy farmers (Nielsen and Emanuelson, 2013). The sample was drawn from the database of the Swedish official milk recording scheme (run by the Swedish Dairy Association). While the sampling frame included herds with free-stalls and an automatic milking system (AMS), herds with tie-stalls and pipeline milking, and herds with free-stalls and parlor milking, only the latter two groups were included in this study because not all attitude questions were relevant for AMS herds. The questionnaire elicited

information on the herd, and the person responsible for udder health. It also contained specific detailed questions regarding udder health and mastitis management. The questionnaire comprised a total of nine pages and 23 questions, several of which were designed with sub-questions, and completion of the questionnaire was estimated to take approximately 30 minutes. Dispatch and collection were managed by Statistics Sweden. The questionnaire was pre-tested on two researchers, one with experience of designing questionnaires and one with experience in mastitis research. The questionnaire was also reviewed for clarity by Statistics Sweden. Each questionnaire was distributed together with a lottery ticket to increase the respondents' willingness to reply. The instructions provided with the questionnaire specifically requested that the person responsible for udder health in the herd should answer the questionnaire.

2.2 Variable selection - mastitis attitudes data

Because the original data were complex and had a very large number of variables (and potentially involved 19 separate analyses – one for each of the attitude questions recorded), only a subset of variables were extracted and used in this study. The producer's attitude toward the need for keeping cows standing after milking (stand_att) was selected as the dependent variable of interest because it had the most missing values. All attitude variables were coded so that 1 meant the producer felt that the procedure (eg keeping cows standing) was not at all important and 5 meant that it was very important (Table 1).

Five potential independent variables (Table 1), which were significant in the CC analysis (described below), were retained for all analysis models (also referred to as "prediction models"). These were a continuous variable (age - the age of the producer) and 4 dichotomous variables (sex - the sex of the person responding, tie - tiestall vs freestall, vet -

use of veterinary advice for mastitis control and mcp - participation in a mastitis control program).

An additional nine management/demographic variables (designated iv1 to iv9), the producer's attitude toward wearing gloves during milking (designated gloves_att), and 19 other attitude variables (designated att1 to att19) were retained and used in the imputation model (Table 1). Variables used in the imputation model, but not considered in the analysis model are called auxiliary variables.

2.3 Complete case analysis

Both ordinal and multinomial logistic regression models were explored for determining which factors influenced a producer's attitude toward keeping cows standing after milking. Stepwise selection was used to identify a set of predictors which were statistically significant ($P<0.05$). An ordinal model was selected for all further analyses because the overall Brant test of parallel regressions was non-significant ($P>0.05$) and the test for individual predictors were non-significant ($P>0.15$) except for vet which was only borderline significant ($P=0.03$). Given the much greater simplicity of an ordinal model compared to a multinomial model (which produces a complete set of estimates for each level of the outcome), the ordinal model was preferred.

2.4 Variables used in imputation

As noted above, one of the first steps in any imputation process is to determine which variables are to be used in the imputation model. The advice provided by van Buuren (2012) is to include as many “relevant variables” as possible. Specifically, all variables for the analysis model, including the outcome variable (Moons et al., 2006; StataCorp, 2013) should

be included. A comparison of a “restrictive” (limited use of auxiliary) variables, and an “inclusive” (including many auxiliary variables) strategies clearly found the latter to be superior (Collins et al., 2001).

In this project, we also had available a number of possible predictors (iv1 - iv9) that might have been used in the analysis model but which were not included due to lack of significance. In addition, there were also a large number of variables relating to producers' attitudes toward other mastitis control procedures (gloves_att, att1-att19), which were potentially good predictors of the attitude of interest (keeping cows standing after milking) (Table 1). These were all retained as auxiliary variables.

2.5 Determining the best imputation model

The next step in the imputation procedure is to determine what imputation model is most appropriate. The variables with missing values in the analysis model included the dependent ordinal variable (producer's attitude toward keeping cows standing after milking stand_att), a continuous variable (age) and 2 dichotomous variables (sex and mcp) (Table 1). Given that the vast majority of missing values were in stand_att, we focused our attention on selecting the correct imputation model for that variable.

Three options were considered. The first two utilized ordinal or multinomial regression models to obtain predicted values for the missing observations. The third was based on predictive mean matching. In this procedure, a predicted value is computed using linear regression and then a value is randomly selected from the observed data which have values close to the predicted value, thus ensuring that only valid values are selected.

Chained equations were used for all imputations as this method maximizes the data used in the analysis (ie the imputations are not based solely on complete observations).

For each model, MI was carried out with 10 imputed datasets and the distributions of the observed values for stand_att and the imputed values were compared using a subjective visual assessment. The model which produced a distribution of imputed values which most closely matched the distribution of the observed values was identified. The whole process was repeated nine times to determine if the results of the subjective assessment were consistent.

2.6 Number of imputations

Standard advice is to impute 3-5 data sets (van Buuren, 2012; Carpenter and Kenward, 2013; StataCorp, 2013) as this will generally produce unbiased estimates with appropriate confidence intervals. However, in some situations, larger number of imputations may be required to obtain stable results. We generated 20 imputed data sets for the comparison of CC and MI analyses of the observed data, and 10 for each imputation used in the simulation studies (described below).

2.7 Comparing complete case (CC) and multiple imputation (MI) analyses

The estimates obtained from the CC and MI analyses were compared and differences noted. However, the question remained - "which set of estimates was better". To address this question, a simulation study based on the observed missing data pattern, and each of the three missing data mechanisms was carried out.

2.8 Simulation study - based on observed missing data pattern

As already described, twenty imputed data sets were generated from the original data using ordinal regression to impute stand_att and predictive mean matching for all other variables. One of these generated data sets was randomly selected to serve as a complete data set for the

simulation study. Results from an ordinal regression analysis of this complete data set were considered the ”truth”.

Starting with this complete data set, additional data sets with the same number of missing values as was observed in the original data (stand_att=74, sex=3, age=4 and mcp=4) were created. Three sets of data were created with different missing data mechanisms.

MCAR: Values were assigned to be missing in a completely random manner (using a random numbers generator)

MAR: stand_att was made missing so that the probability of being missing decreased as the value of gloves_att increased (participants who rated using gloves as important were less likely have missing data for the stand_att variable). The actual probabilities assigned to each level of stand_att were chosen so that exactly 74 observations were converted to missing. The small number of missing values in the other 3 variables were assigned MCAR (as described above).

NMAR: stand_att was made missing so that the probability of being missing decreased as the value of stand_att increased (participants who rated the procedure as important were less likely to have missing data for the stand_att variable) and the probabilities of being missing were twice as high in participants who did not seek veterinary advice (vet) on mastitis control. The actual probabilities assigned to each combination of stand_att and vet were chosen so that exactly 74 observations were converted to missing. The small number of missing values in the other 3 variables were assigned MCAR (as described above).

This process was repeated 1000 times so a total of 3000 data sets were created. Stochastic variation accounted for the variation among these data sets (within a missing data type). MI using chained equations and ordinal logistic regression for imputing stand_att and predictive

mean matching for other variables was used to analyse each dataset with the coefficient estimates and their SEs extracted and stored.

Results were summarised in three ways. First, box and whisker plots showing the distribution of the coefficient estimates from the 1000 simulations were generated. Second, the % bias for each replication within each missing data mechanism was computed as:

$$bias = \frac{observed\ value\ of\ \beta - true\ value\ of\ \beta}{true\ value\ of\ \beta} \qquad \text{Eq. 1}$$

where β is the coefficient for the predictor. Finally, the mean SE for each coefficient estimate (for each missing data mechanism) was computed.

2.9 Simulation study - missing values in predictor variables

In order to explore the potential for MI when missing values are in predictor variables, we also ran a series of simulations which started with the same complete data set (ie "the truth") and in which missing values were created for 4 of the predictors in the analysis model (sex, tie, mcp and vet). The variables age and stand_att were left complete. Within each missing data mechanism, 7 data sets were created with 27 (10%), 22, 17, 12, 7, 3 and 1 missing values for each predictor.

The missing data mechanisms were: MCAR - values were assigned to be missing in a completely random manner (using a random numbers generator); MAR - missingness was a function of gloves_att and age (these 2 variables had no missing values); NMAR - missingness was a function of stand_att and the predictor variable itself.

This process was repeated 1000 times so a total of 3000 data sets were created for each level of missing data (a total of 21,000 data sets). Stochastic variation accounted for the variation among these dat sets (within a missing data type and % missing). MI using chained equations

and predictive mean matching for all variables was used to analyse each data set with results from each analysis extracted and stored.

Results were summarised in four ways. Box and whisker plots of coefficient estimates and the mean % bias were determined as described above. In addition, the ratio of the SEs from the MI analysis to that of the CC analysis was computed and averaged. Finally, the % of the 1000 coefficient estimates that were within + 20% of "the truth" (ie the estimate from the complete data set) was determined. With the exception of the box and whisker plots, all results were presented as graphs of the parameter against the number of complete cases (which ranged from approximately 180 to 269) used in each analysis.

All simulations, imputations and analyses were carried out using Stata (Version 13) (StataCorp, 2012).

3. Results

3.1 Pattern of missing values - original data

The pattern of missing values in the variables in the original data (outcome=stand_att, predictors were: sex, age, tie, mcp and vet) is shown in Table 2. There were no missing values in tie and vet. Only 3 observations had missing values in more than one variable. There were 187 observations with no missing data.

3.2 Complete case analysis

The CC ordinal model is shown in Table 3. Females, and producer's who had a freestall barn, were on a mastitis control program and sought veterinary advice for mastitis control rated the importance of keeping cows standing higher than others. The older the producer, the lower the importance they assigned to the practise.

3.3 Determining best imputation model

For each of the three imputation models considered, the distribution of stand_att in the original observed data set and in the complete imputed data set were plotted. This process was repeated 9 times and the plots from the first iteration are shown in Figure 1. We concluded that an ordinal regression imputation model produced a distribution most similar to the observed data (the conclusion was the same when all 9 iterations were considered) and hence was selected.

3.4 Comparing CC and MI analyses

A comparsion of coefficients and their confidence intervals (CIs) from the CC and MI analyses are presented in Table 3. In general, the results between the CC and MI are quite similar and all estimates fall within the CI of the alternative estimate. The largest discrepancy is for mcp (mastitis control program) which has a coefficient of 0.62. .=0.035) in the CC analysis and 0.31 (P=0.264) in the MI analysis but this was expected as a non-significant coefficient would be expected to bounce around a lot as a result of random variation. The limited number of missing values in the predictor variables precluded a comparison of imputation methods. Consequently, these values were imputed using predictive mean matching because this has been shown to be a generally robust method for a wide variety of situations (van Buuren, 2012)

3.5 Simulation study - based on observed missing data pattern

The number of complete observations in each simulation is shown in Table 4. There were more complete observations in the MAR and NMAR data than the MCAR because with a mechanism accounting for the missingness, there were more observations with multiple

missing values and hence more left as complete cases. Table 4 also shows the number of iterations that ran in each simulation. Rarely did the imputation procedure fail to converge so the smallest number of successful iterations was 995 (out of 1000) showing that the imputation procedure appeared to be quite robust.

Box and whisker plots showing the distribution of the coefficient estimates from the 1000 simulations, with a horizontal line reflecting the true value determined from the complete data set, are shown in Figure 2. This shows that, with the exception of vet in the NMAR missing data, there was very little difference in the median estimates for any of the coefficients in any of the missing data types. In addition, overall there was only little reduction in the variability of the estimates (the 25th to 75th percentile ranges were similar, as were the length of the "whiskers") that was attributable to MI estimation.

The numerical summary of the mean % bias confirmed little beneficial effect attributable to imputation (results not shown but presented in on-line supplemental material).

The mean SE for each coefficient estimate (for each missing data mechanism) which was also computed across the 1000 replications showed there was no, or very little, reduction in the SEs of the estimates despite the fact that the sample size increased from 187 to 269 (results not shown but presented in on-line supplemental material).

Overall, there was very little, if any, benefit from imputation.

3.6 Simulation study - based on missing values in predictor variables

The number of complete observations used in these analyses is shown in Table 4. Box and whisker plots showing the distribution of the coefficient estimates from the 1000 simulations from the data sets with 27 values (10%) missing in each of 5 predictors are shown in Figure

3. With the exception of the mcp coefficient for NMAR missing data, all MI estimates were, on average, closer to the ”truth”. In addition, all MI analyses produced much less variable coefficients than the CC analyses.

The mean % bias plotted against the number of complete observations for MCAR, MAR and NMAR data are shown in Figures 4-6. They confirm that the magnitude of the bias depends on the number of missing observations (for both MI and CC analyses) and that with the exception of the MCAR results or the sex and mcp coeffiients with NMAR missing data, MI analyses always produced less biased estimates. (Numerical estimates of the % bias are not shown but are presented in on-line supplemental material)

The ratio of the SEs of the MI estimates to the CC estimates for all 3 types of missing data are shown in Figure 7. For all types, the SEs from the MI analyses were approximately 15-20% smaller than those from the CC analyses when the number of complete observations was approximately 180 (of the 269 total observations).

The SD of the coefficient estimates reflects how consistent estimates were across simulation runs. Figure 8 shows these SDs expressed as a % of the estimate (based on NMAR missing data) for MI and CC analyses. For all predictors, MI analyses produced much more consistent estimates than CC analyses. The patterns for MCAR and MAR data were very similar (results not shown but presented in on-line supplemental material)..

The proportion of coefficient estimates that were within 20% of the “truth” for each of the three missing data types is shown in Figures 9, 10 and 11 (for MCAR, MAR and NMAR, respectively). In all cases, the MI analysis outperformed the CC analysis. In many cases the difference was quite dramatic. For example, the coefficient for age was virtually always

estimated correctly (ie within 20% of the truth) by an MI analysis, but only correct about 60% of the time for a CC analysis - regardless of the missing data mechanism.

4. Discussion

The theory underpinning MI is not new. The methods were covered in a seminal book on the subject by Rubin in 1987 (Rubin, 1987) and the procedure has become much more widely used in medical epidemiology in the last 5 years. A search of PubMed found 144 articles with the term "multiple imput*" in the title in the 5-year period of 2010-2014, compared to 30 articles 10 years earlier (2000-2004). In part, this is because procedures for MI are now integrated into many major statistical programs such as Stata (-mi-), R (packages -mi- and -mice-) and SAS (Proc MI and MIANALYZE). There have also been a number of more recent texts on the subject (cited in Introduction) which have made the methods more available to new users. However, as noted in the Introduction, there has been little uptake in veterinary epidemiology.

4.1 Expected effects of MI

Based on the theory of MI, one would expect to get more precise estimates of coefficients if data are MCAR and both less biased and more precise estimates if data are MAR. It is difficult to predict what effect MI would have if data are NMAR.

There have been a number of simulation studies of the efficacy of MI reported. Lee and Carlin (2012) evaluated MI for imputing values of either a continuous or a binary predictor that had values that were MAR. Their imputation was based on multivariate normal imputation (as opposed to predictive mean matching) which assumes that the variables being imputed had a normal distribution. They found that MI worked well for the continuous variable if it was suitably transformed to have a normal distribution, but was not effective for

either the continuous variable if it had a skewed distribution, or the binary variable. This highlights the sensitivity of MI to the assumptions underlying the imputation method. Predictive mean matching (as used in this study) does not have the same distributional assumptions. Nunes et al. (2009) reported MI to be superior to CC analyses while Peyre et al. (2011) reported that MI was superior to other forms of imputation. There is considerable interest in the value of imputation in longitudinal studies (Liu et al., 2000; Graham, 2012; Donneau et al., 2015) , but this is not the focus of this manuscript.

4.2 Simulation study - based on observed missing data pattern

The results of these analyses were surprising, particularly with regard to the MAR data. Standard theory about MI suggests that we would have expected reduced bias with MAR data (and potentially some reduction with NMAR) and improved precision with all missing data mechanisms. However, no benefit from MI was observed, regardless of the missing data mechanism used to generate the missing values.

The unique feature of the original data in this study was that virtually all of the missing data were in the dependent variable. Surprisingly, the recently published texts on missing data and multiple imputation either make no distinction between imputing predictor and outcome variables (Enders, 2010; van Buuren, 2012; Carpenter and Kenward, 2013) or focus on imputing missing values of outcome values in longitudinal (repeated measures) data (Graham, 2012). The Stata reference manual (StataCorp, 2013) suggests the question of whether outcome variables should be imputed is unresolved. Crawford et al. (1995) advocated using MI when the dependent variable was missing, but more recently White and Carlin (2010) and Groenwold et al. (2012) suggest there is little advantage to MI over CC when the dependent variable is missing. von Hippel (2007) showed that imputing dependent

variable values just adds noise to the data and should not be done. A recent web posting (Allison, 2014) stated that imputation of missing values in the dependent variable (assuming MAR) only added noise to the data. Our results concur with this conclusion.

4.3 Simulation study - based on missing values in predictor variables

Predictive mean matching was chosen for imputing all the predictor variables because it is flexible (can handle both continuous and categorical variables) and is generally robust. Van Buuren (2012 - page 74) describes it as "a great all-round method with excellent properties".

We chose to present the various measures of the effect of MI ploted against the number of complete cases in the data set rather than the number of missing values assigned to each predictor (27, 22, 17, 12, 7, 3 and 1) because different missing data mechanisms produce varying numbers of observations with multiple missing values and we wanted that to be reflected in the output.

For data with MCAR missing values there was no reduction in the mean bias attributable to MI (Fig. 4) because asymptotically there is no bias generated by MCAR missing values. Observing a very small negative bias in some coefficients following MI was a surprising and unexplained finding. However, the MI estimates had smaller SEs (Fig. 7), were much less variable across simulations (Fig. 3) and were more likely to be within 20% of the "truth" (Fig. 9). Even at low levels of missing values, the CC analyses had a substantially lower levels of percentage of estimates that were close to the truth. The estimates of the coefficient for mcp did not behave in as predictable a fashion as the others because it was a non-significant predictor with a coefficient that bounced around 0 and hence was prone to large proportional changes.

For data with MAR missing values, MI analyses resulted in coefficients with lower average bias (Fig. 5), smaller SEs (Fig. 7), and which were much less variable across simulations (Fig. 3) and were more likely to be within 20% of the "truth" (Fig. 10).

For data with NMAR missing values, MI analyses resulted in coefficients with lower average bias for tie and vet, but not for the other 3 predictors (Fig. 6). However, the estimates did have, smaller SEs (Fig. 7), and were much less variable across simulations (Fig. 3 and 8) and were more likely to be within 20% of the "truth" (Fig. 11). The benefits of MI in the face of NMAR data are not as predictable. However, it has been reported that MI is robust to the NMAR mechanism if the % of missing observations is <25% and the correlation between the cause of missingness and the variable subject to missingness was <.4 (Collins et al., 2001).

5. Conclusions

MI is now a readily available tool for epidemiologists to use. The software is accessible and there are numerous learning resources to support its adoption. While the case study used in this manuscript dealt with data collected by mail survey, the methods are generally applicable to all types of observational and experimental research. Similar procedures were also used in an example based on a cross-sectional, on-farm study of infectious diseases of dairy cattle (Dohoo, 2015)

The results of the simulations showed that MI was of no use for imputation of missing values in the outcome variable. However, for missing values of predictor variables, MI analyses always performed at least as well as and, most often, substantially better than CC analyses. The proportion of observations with some missing values did not have to be very high before the benefits of MI analyses became clear. Certainly, if more than 10% of observations had

missing values, MI outperformed CC in these analyses, regardless of the missing data mechanism.

References

Allison, P., 2014. Listwise deletion its not evil. http://www.statisticalhorizons.com/listwise-deletion-its-not-evil

Anonymous, 2014. Animal health 2013/2014 [Djurhälsovård 2013/2014]. Stockholm, Sweden.

Carpenter, J.R., Kenward, M.G., 2013. Multiple Imputation and its Applications. Wiley Chichester, UK.

Collins, L.M., Schafer, J.L., Kam, C.M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 6, 330-351.

Crawford, S.L., Tennstedt, S.L., McKinlay, J.B., 1995. A comparison of anlaytic methods for non-random missingness of outcome data. J Clin Epidemiol 48, 209-219.

Dohoo, I.R., 2015. Dealing with deficient and missing data. Prev Vet Med in press.

Donneau, A.F., Mauer, M., Lambert, P., Molenberghs, G., Albert, A., 2015. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. J Biopharm Stat 25, 570-601.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Press New York, NY.

Graham, J.W., 2012. Missing Data. Springer New York, NY.

Groenwold, R.H., Donders, A.R., Roes, K.C., Harrell, F.E., Jr., Moons, K.G., 2012. Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol 175, 210-217.

Heeringa, S.G., West, B.T., Berglund, P.A., 2010. Applied Survey Data Analysis. Chapman & Hall / CRC Boca Raton, FL.

Jansen, J., Steuten, C.D., Renes, R.J., Aarts, N., Lam, T.J., 2010. Debunking the myth of the hard-to-reach farmer: effective communication on udder health. J Dairy Sci 93, 1296-1306.

Jansen, J., van den Borne, B.H., Renes, R.J., van Schaik, G., Lam, T.J., Leeuwis, C., 2009. Explaining mastitis incidence in Dutch dairy farming: the influence of farmers' attitudes and behaviour. Prev Vet Med 92, 210-223.

Lee, K.J., Carlin, J.B., 2012. Recovery of information from multiple imputation: a simulation study. Emerging Themes Epidemiol 9, 3.

Liu, M., Taylor, J.M., Belin, T.R., 2000. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. Biometrics 56, 1157-1163.

Moons, K.G., Donders, R.A., Stijnen, T., Harrell, F.E., Jr., 2006. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 59, 1092-1101.

Nielsen, C., Emanuelson, U., 2013. Mastitis control in Swedish dairy herds. J Dairy Sci 96, 6883-6893.

Nunes, L.N., Kluck, M.M., Fachel, J.M., 2009. [Multiple imputations for missing data: a simulation with epidemiological data]. Cadernos de saude publica 25, 268-278.

Peyre, H., Leplege, A., Coste, J., 2011. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 20, 287-300.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley New York, NY.

StataCorp, 2013. Stata Multiple Imputation Reference Manual. Stata Press College Stn. TX.

StataCorp, 2014. Stata Statistical Software: Release 14. Stata Corp LP College Station, Tx.

van Buuren, S., 2012. Flexible Imputation of Missing Data. Chapman & Hall/CRC Boca Raton, FL.

von Hippel, P.T., 2007. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociologicial Methodology 37, 83-117.

White, I.R., Carlin, J.B., 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 29, 2920-2931.

Figure 1

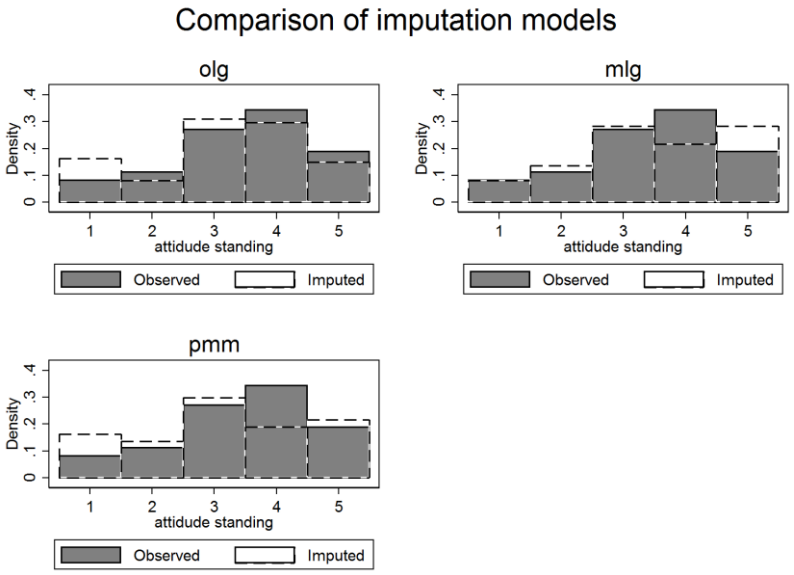

Figure 2

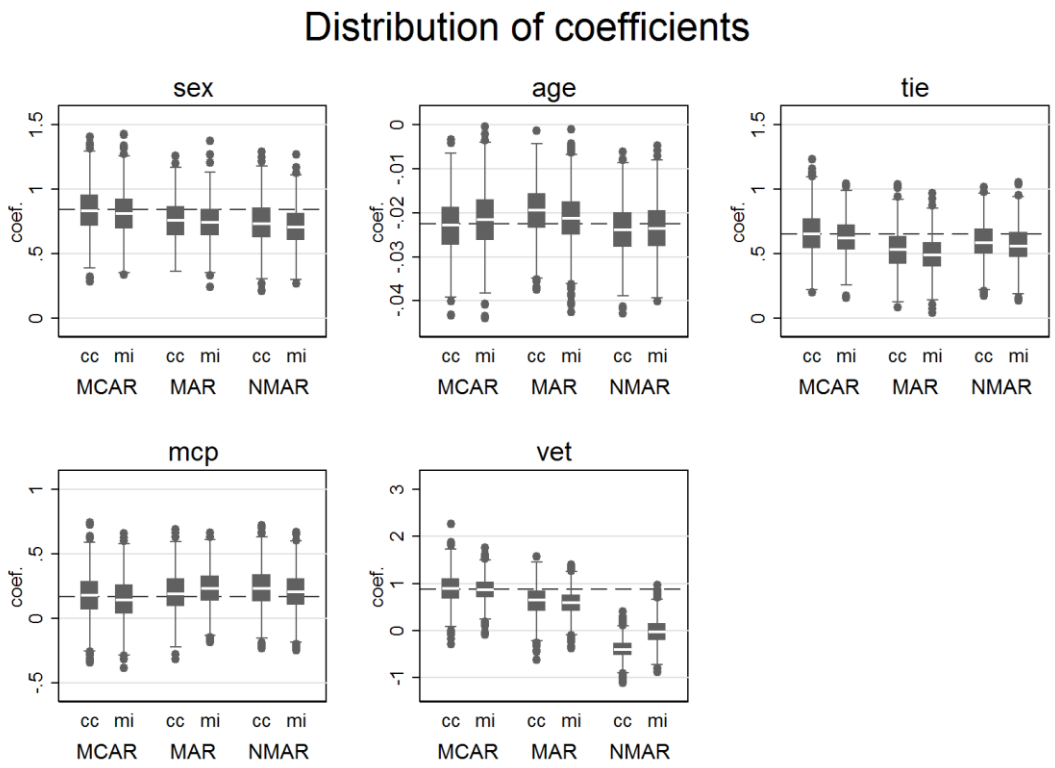

Figure 3

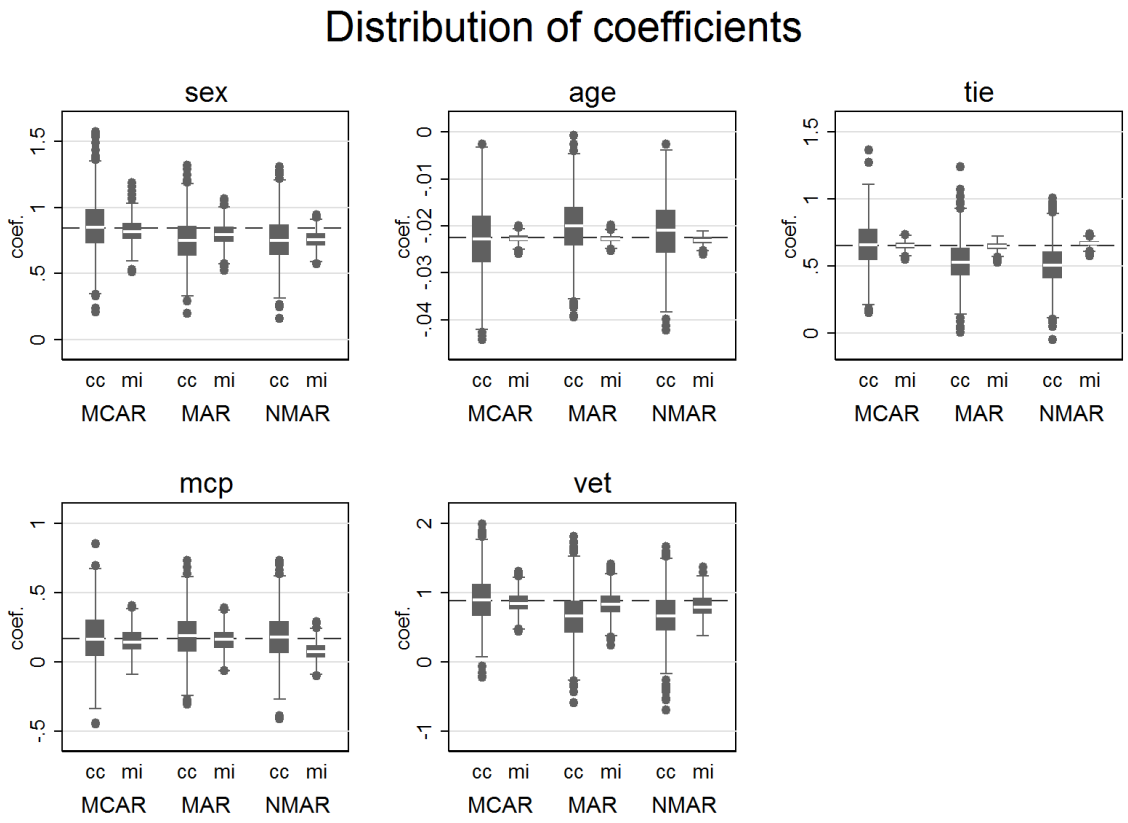

Figure 4

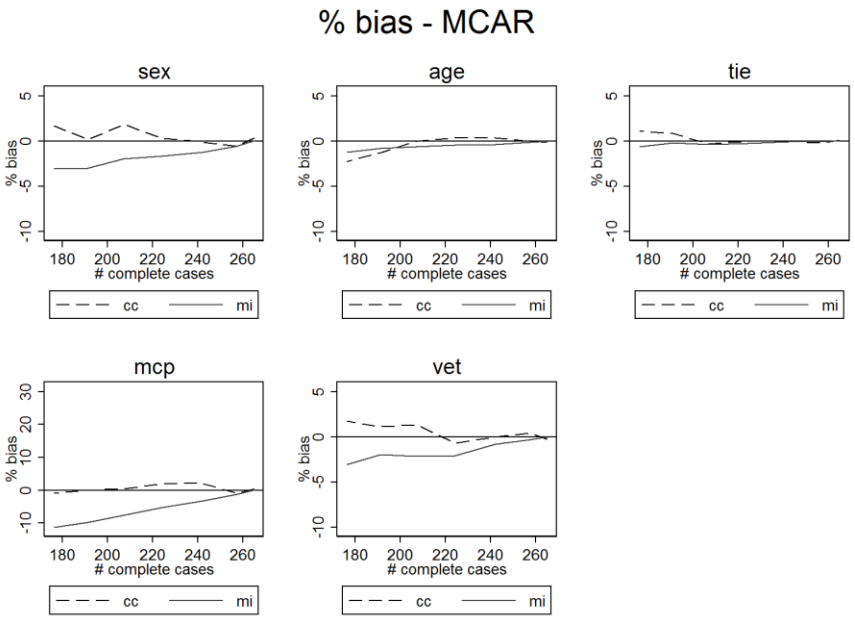

Figure 5

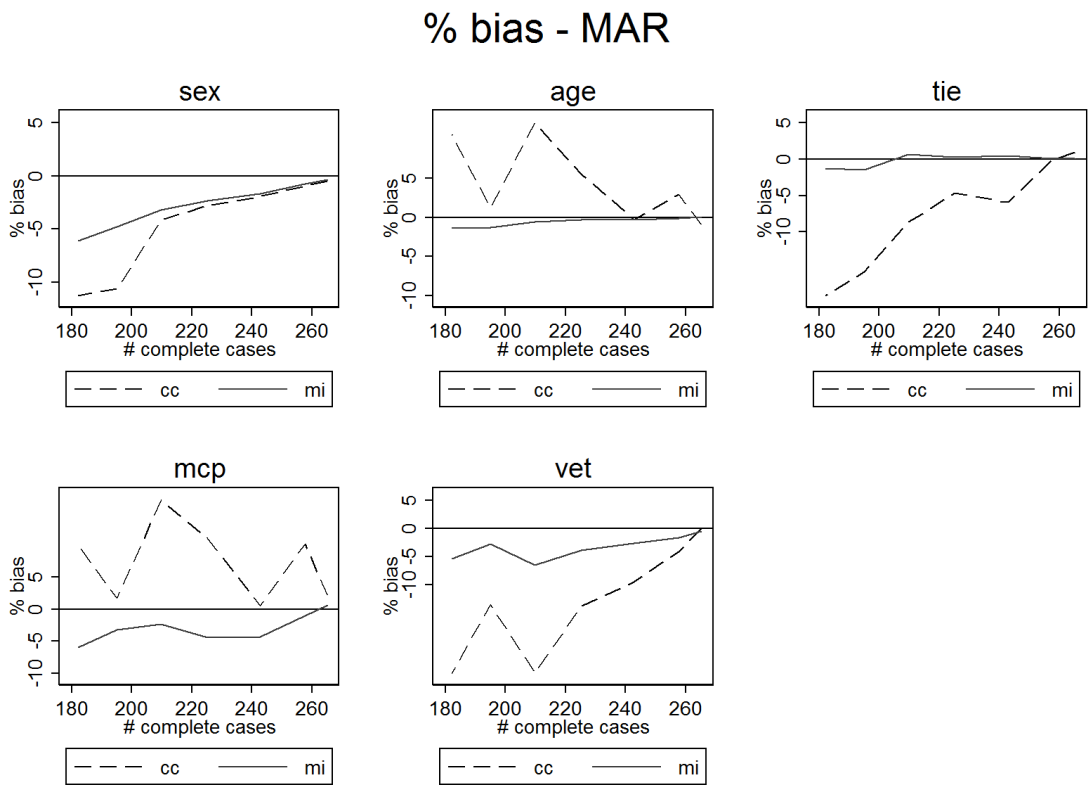

Figure 6

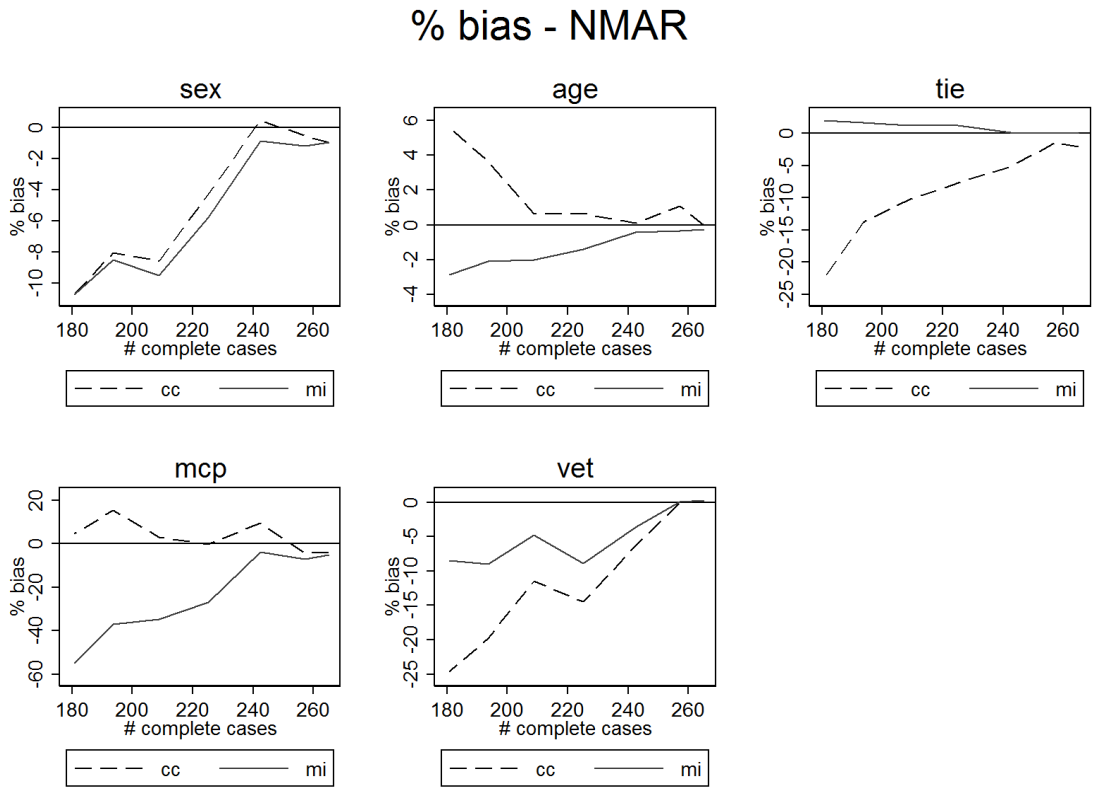

Figure 7

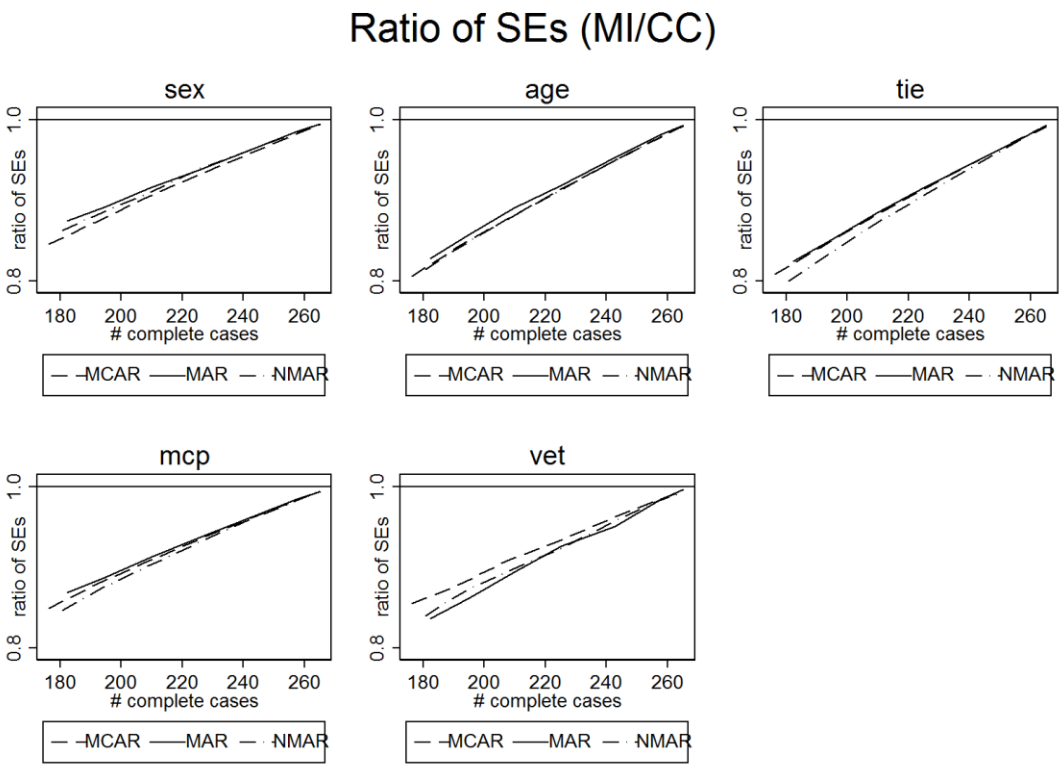

Figure 8

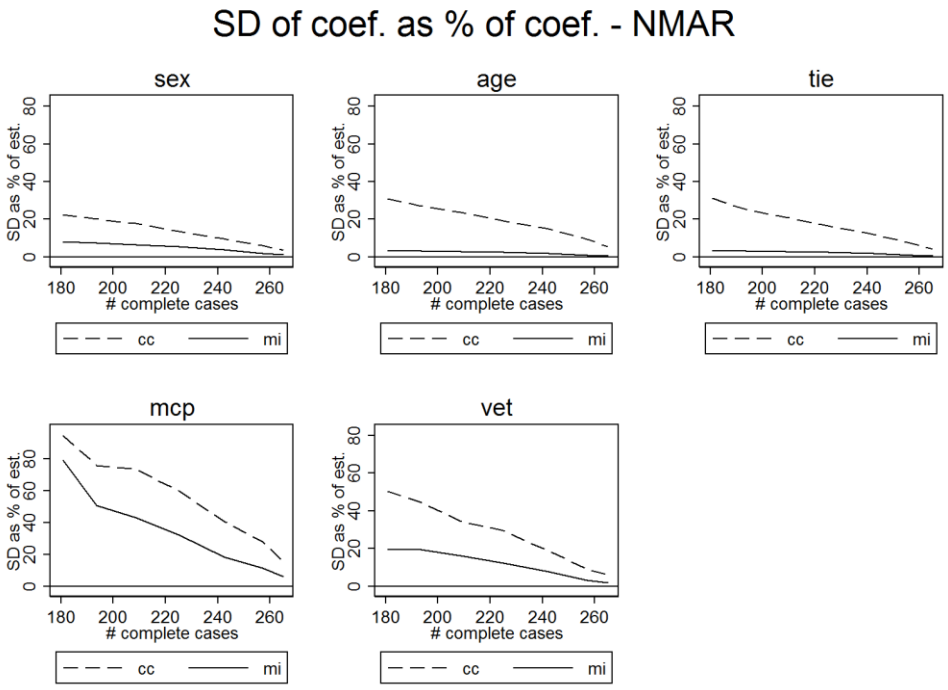

Figure 9

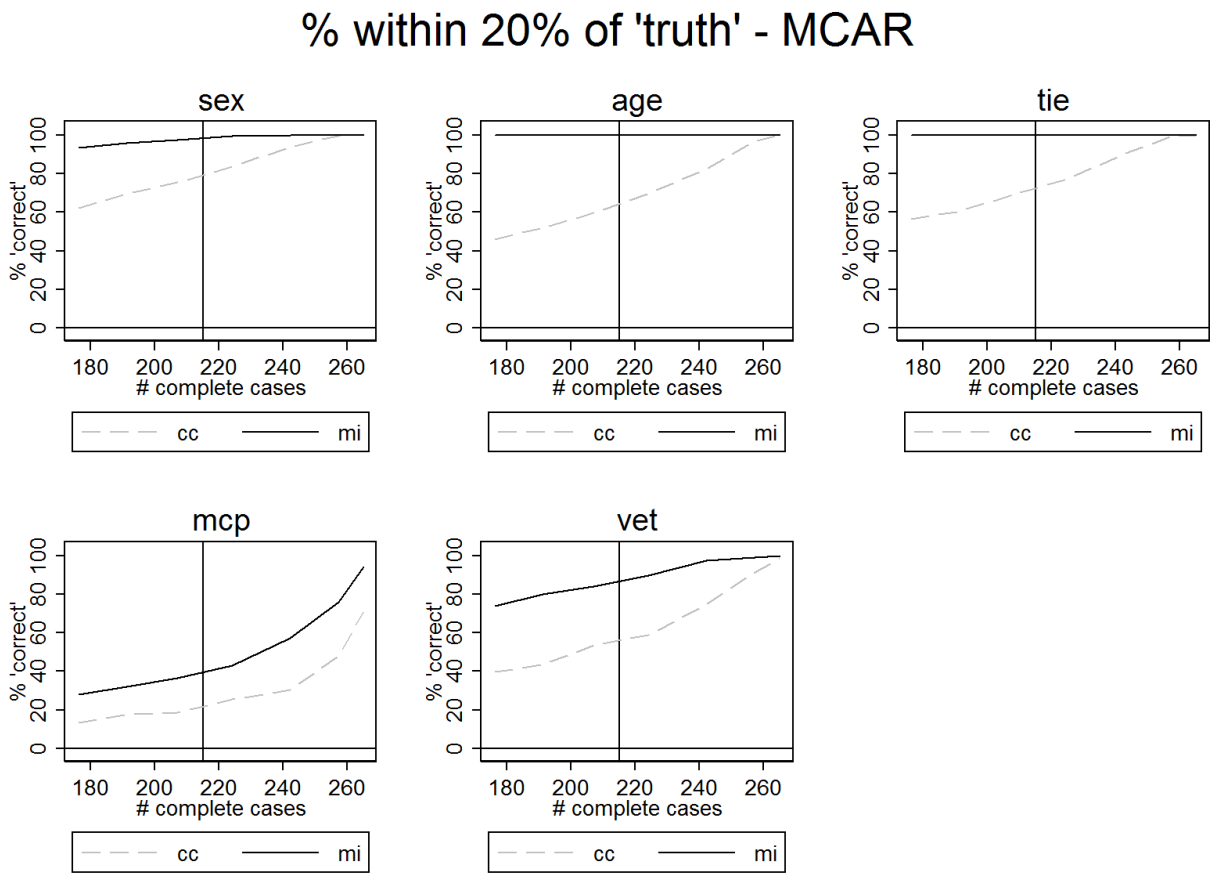

Figure 10

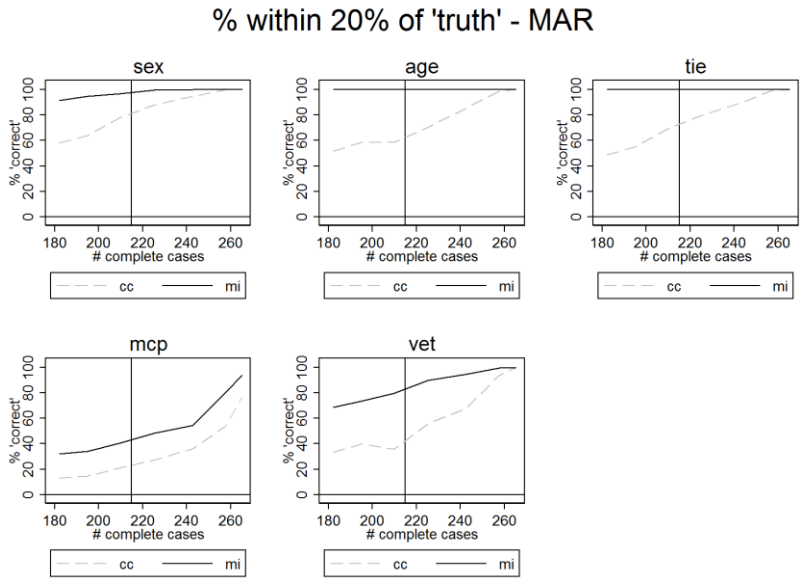

Figure 11