Methodology and Statistics for the Social and Behavioural Sciences
Utrecht University, the Netherlands

MSc Thesis Emmeke Aarts
TITLE: A novel method to obtain the treatment effect assessed for a completely randomized design: Multiple imputation of unobserved potential outcomes.
May 2010

Supervisors:
Prof. Dr. S. van Buuren
Dr. L.E. Frank

Preferred journal of publication: Statistics in Medicine
Word count: 8,844

A novel method to obtain the treatment effect assessed for a completely randomized design: Multiple imputation of unobserved potential outcomes

Emmeke Aarts[a*†], Stef van Buuren[b,a] and Laurence E. Frank[a]

Abstract

In this paper a novel method to obtain the treatment effect called multiple imputation of unobserved potential outcomes is evaluated by comparing its performance to the Student's t-test and ANCOVA. The novel method originates from Rubins potential outcomes framework which explicitly defines that each unit has a possible active treatment and control treatment outcome. Since every unit can only be assigned to one treatment, one of the potential outcomes is unobserved for each unit. To approximate calculating the treatment effect, the unobserved potential outcomes are multiple imputed. To start at the basics of group comparisons, the paper is restricted to completely random groups. The method is evaluated by a series of simulations using a realistic, empirical synthetic population. Overall, the results show that the novel method performs up to standards: bias is negligible compared to the standard error and 95 per cent confidence interval coverage is above 90 per cent. Also, the novel method is more efficient and powerful than the frequently used Student's t-test when the relation between the covariates and potential outcomes is linear. Multiple imputation of potential outcomes performs approximatly equally well as classical ANCOVA. At a small sample size, the novel method is somewhat more powerful when the assumption of parallel slopes is violated, but it is slightly less efficient than ANCOVA for all used properties of covariates, sample sizes and effect sizes.

Keywords: potential outcomes framework, counterfactuals, ANCOVA, missing data, causal effect

[a]Department of Methodology and Statistics, University of Utrecht, the Netherlands
[b]TNO Quality of Life, Leiden, the Netherlands
*Correspondence to: Emmeke Aarts, Department of Methodology and Statistics, University of Utrecht, Heidelberglaan 1, 3584 CS, the Netherlands
†E-mail: E.Aarts@students.uu.nl

1 Introduction

One of the most common purposes of psychological and medical research is to examine the effect of a treatment. Methods commonly used to obtain the treatment effect in a completely randomized design are testing the simple mean difference with the Student's t-test or testing adjusted means with analysis of covariance (ANCOVA). A comparative review of the performance of various methods to obtain the treatment effect is provided by Schafer and Kang (2008). In their paper, one novel method to obtain the treatment effect is not assessed: multiple imputation (MI) of unobserved potential outcomes. The objective of this paper is to fill this void by evaluating this novel method.

The approach to multiply impute potential outcomes originates from Rubin's potential outcomes framework (Little & Rubin, 2000; Rubin, 2005) and has been put forward by Rubin on several occasions (e.g. Rubin, 2004, 2006) but has not been assessed yet. The potential outcomes framework proposes a structure for causal inference where every unit has two possible outcomes: an active and control treatment group outcome. Because every unit can only be assigned to one treatment, one of the potential outcomes is unobserved for each unit. Rubin suggests an extensive missing data perspective in applied and theoretical statistical problems. According to this perspective, a scientific problem should be viewed as one where the scientific answer could be calculated if some missing data were available, instead of statistically inferred. To approximate calculating the treatment effect, the unobserved potential outcomes are multiple imputed. MI is the preferred method to obtain the unobserved potential outcomes because repeating imputations not only produces estimates that are approximately unbiased and efficient, but reflects the uncertainty of these estimates as well (Rubin, 1987).

The novel method proposed in this paper is almost identical to Bayesian causal inference, where the inference of the treatment effect follows from the predictive distribution of the unobserved potential outcomes (Rubin, 1978). This Bayesian approach to causal inference has recently been applied by Dominici, Zeger, Parmigiani, Katz, and Christian (2006) and by Jin and Rubin (2008). Bayesian causal inference however requires specialist programming every time an analysis is executed. Also, a good understanding of Bayesian statistics in general is required. Because MI of unobserved potential outcomes is more user friendly and easier to understand for researchers not familiar with Bayesian statistics, the focus of this paper will be on the former.

MI of unobserved potential outcomes uses the information of the covariates in order to impute the unobserved potential outcome and thus be able to utilize both the observed outcomes and unobserved potential outcomes. Thereby, it increases power and efficiency compared to methods that only make use of the observed outcomes, simply because it uses more information. One prerequisite to decrease the standard error (SE) of and variability between the estimates is however that the imputations of the unobserved potential outcomes must be precise enough not to induce much uncertainty. It is expected that multiply imputing unobserved potential outcomes is more efficient and powerful than the Student's t-test, since this frequently used method in a completely randomized design only makes use of the observed outcomes. Yet using covariates to increase efficiency and power is not new. A well-known method that utilizes this principle as well with completely randomized groups is Analysis of covariance (ANCOVA). If the used data meets the assumptions of ANCOVA, there will most likely not be a large difference between the performance of MI of potential outcomes and ANCOVA.

The assumptions of classical ANCOVA to ensure the validity of the estimated treat-

ment effect are however quite rigid: the outcome needs to vary linearly with the covariates with identical slopes for both treatment groups. Because MI of unobserved potential outcomes is a more flexible model, the two previously mentioned assumptions do not have to be met. As a result, MI of unobserved potential outcomes is expected to provide a much closer approximation to the true model than ANCOVA in cases where the data deviates from these ANCOVA requirements. Therefore, it is hypothesized that MI of unobserved potential outcomes outperforms classical ANCOVA in terms of power and efficiency when the data deviates from the ANCOVA requirements.

The objective of this paper is to investigate whether and under which circumstances MI of unobserved potential outcomes improves the quality of obtaining the treatment effect compared to current commonly used methods. In this paper, quality of causal inference is operationalized as power and efficiency. The used comparison methods are the Student's t-test and the classical ANCOVA. Also, it is investigated under which circumstances in terms of number of covariates used and strength of relationship between the variables the novel method works best. In addition, the method is illustrated through an application of MI of unobserved potential outcomes. To start at the basics of group comparisons, the paper is restricted to completely random groups.

The paper is organized as follows. In Section 2, the potential outcomes framework is further described to clarify the building blocks for the novel method. Section 3 outlines the method and application of MI of potential outcomes. In Section 4, MI of potential outcomes is evaluated by comparing the novel method to commonly used methods through simulation studies for various conditions. The application of the novel method is provided in Section 5 and all results are discussed in Section 6.

2 The potential outcomes framework

The potential outcomes framework explicitly defines that each unit has a possible active treatment and control treatment outcome (Neyman, 1923; Rubin, 1974b, 1978). This framework provides the building blocks for MI of unobserved potential outcomes, which defines the treatment effect as the individual difference between both potential outcomes. This definition of the treatment effect according to the potential outcomes framework is often buried under notation of commonly used methods (Little & Rubin, 2000). Also, by making the definition of the treatment effect explicit, basic assumptions that need to be met for causal inference are made more explicit than they usually are (Holland, 1986). Therefore, a description of the potential outcomes framework and its assumptions are provided below.

The average treatment effect

The point of view that each unit has multiple outcomes was introduced by Neyman (1923), who used it to define the treatment effect in the context of a completely randomized experiment in a hypothetical agricultural example. Rubin elaborated on this by proposing a similar viewpoint within the context of nonrandomized, observational studies (Rubin, 1974b, 1978), producing a framework extending the idea beyond randomized experiments and randomization-based inference. This framework is commonly referred to as Rubin's causal model (RCM) (Holland, 1986) and is frequently used in statistics and epidemiology (Höfler, 2005; Rubin, 2005).

Let T_i denote the treatment assigned to the i^{th} unit for $i = 1, ..., N$ where N denotes the number of units. The assigned treatment can either be $T_i = 1$ when the unit is actively treated or $T_i = 0$ when the unit is not actively treated. Furthermore, let Y_{i1} denote the response when the i^{th} unit is assigned to the active treatment, Y_{i0} denote the response when the i^{th} unit is assigned to the control treatment. Furthermore, X_{ik} for $k = 1, ..., K$ denotes the matrix of covariates which are assumed to be independent of treatment. In Table 1, a hypothetical example to clarify the potential outcomes framework is given for a sample of 6 units. The potential outcomes for a set of units can be depicted as an N $\times$ 2 matrix, as presented in Table 1. Besides the potential outcomes, several covariates, the treatment indicator and the unit level treatment effect are given for the sample as well.

The golden standard as stated by the RCM to obtain the estimate of interest (Rubin, 2005), the average treatment effect (ATE), is the averaged difference between the potential outcomes for every unit:

$$ATE = E(D_i) = E(Y_{i1} - Y_{i0}), \tag{1}$$

where D_i denotes the difference between the potential outcomes. When hypothetically all potential outcomes are observed, the ATE is estimated by

$$\widehat{ATE}_1 = \frac{1}{N}\sum_{i=1}^{N} D_i = \frac{1}{N}\sum_{i=1}^{N} Y_{i1} - \frac{1}{N}\sum_{i=1}^{N} Y_{i0}. \tag{2}$$

Applying Equation 2 to the potential outcomes of the hypothetical example shown in Table 1 yields an estimated $\widehat{ATE}_1$ of - 0.50. Equation 2 is considered ideal because the estimate is unbiased and reasonably efficient regardless of the distribution of the data.

The golden standard of $\widehat{ATE}_1$ can however not be obtained. Only one of the potential outcomes can be observed for a unit, as units can be assigned to one treatment condition only. This dilemma is referred to as the fundamental problem of causal inference (Holland, 1986). Therefore, techniques for causal inference are in essence missing-data methods. The conventional way to deal with this problem is to aggregate the data of the actively treated units and the control treated units to a group mean. The treatment effect is then inferred instead of calculated by taking the difference between the observed group means. Hence,

Table 1: Illustration of the potential outcomes framework, where mean individual treatment effect denotes the ATE_1

	Covariates			Treatment	Potential outcomes		Individual treatment effect
Units	X_{i1}	...	X_{ik}	T_i	Y_{i1}	Y_{i0}	D_i
1	1	...	20	0	2*	2	0*
2	2	...	31	0	2*	4	- 2*
3	2	...	18	1	6	5*	1*
4	1	...	39	0	4*	5	- 1*
5	1	...	40	1	2	3*	- 1*
6	2	...	26	1	5	5*	0*
Mean	1.5	...	29	0.5	3.5*	4*	- 0.5*

*Note:** Values not actually observed.

the ATE is estimated by

$$\widehat{ATE}_2 = \frac{1}{n_1}\sum_{i=1}^{N} T_i Y_{i1} - \frac{1}{n_0}\sum_{i=1}^{N}(1-T_i)Y_{i0}, \tag{3}$$

where n_1 denotes the sample size of the treatment group and n_0 denotes the sample size of the control group. When Equation 3 is applied to the example, an $\widehat{ATE}_2$ of 0.67 is now obtained. Note that only the observed values are now used, values complemented with an asterisk are not brought into the computation.

The precision and statistical significance of the $\widehat{ATE}_2$ can be assessed using the t-distributed 95 per cent confidence interval (CI),

$$\widehat{ATE}_2 \pm t_{.05/2,N-1}\ s_{\bar{Y}_1-\bar{Y}_0}, \tag{4}$$

where

$$s_{\bar{Y}_1-\bar{Y}_0} = \sqrt{\frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1+n_0-2}} \tag{5}$$

is the pooled standard deviation of the two treatment groups, where s_1^2 is the variance of the active treatment and s_0^2 is the variance of the control treatment (Miller, 2004). The t-distributed CI is used because the mean and standard deviation are assumed unknown. In the following sections, a method based on MI of the unobserved potential outcomes is presented. Using this method, the hypothesis is that the obtained 95 per cent CI shows better properties compared to only using the observed outcomes.

Basic assumptions of the ATE

In order for the $\widehat{ATE}$ to be an unbiased estimate of ATE, a few basic assumptions need to be met. First of all, the stable unit treatment value assumption (SUTVA) (Rubin, 1980, 1990). SUTVA even applies if hypothetically all potential outcomes are observed for each unit. SUTVA assumes that the treatment effect for any unit does not depend on the treatment assignment of other units and there are no other (hidden) versions of the active and control treatment. An example of an (hidden) extra treatment would be when the dose of the active treatment is mistakenly not the same for every unit, for example a high and a low dose. In this case there would actually be three treatments instead of two: low dose active treatment, high dose active treatment and control treatment. When the SUTVA assumption is met, it truly can be stated that every unit has only two potential outcomes because there are only two distinct possible treatments for every unit.

When only the observed potential outcomes are available, extra assumptions regarding the relationship between treatment assignment T_i and the potential outcomes Y_{i1} and Y_{i0} are required (Rubin 1978, 2005). For $\widehat{ATE}_2$ to be an unbiased estimate of ATE, the two treatment groups must consist of a common set of units from the same population. Since the treatment assignment determines which units make up the active and control treatment group, the mechanism behind the treatment assignment determines if this assumption is met. This is known as the treatment assignment mechanism (Rubin, 1978), which can be seen as a mechanism that determines which of the potential outcomes are unobserved taking into account the covariates and the potential outcomes itself. The treatment assignment mechanism can be stated as the probability of the treatment as-

signment T_i being active treatment or control treatment given the covariates X_i and the potential outcomes Y_{i1} and Y_{i0}, which can be written as

$$Pr(T|X, Y_1, Y_0). \tag{6}$$

Equation 6 can be further refined, since some of the potential outcomes for both the active and control treatment are observed (Y_1^{obs} and Y_0^{obs}), and some are not (Y_1^{mis} and Y_0^{mis}):

$$Pr(T|X, Y_1^{obs}, Y_0^{obs}, Y_1^{mis}, Y_0^{mis}). \tag{7}$$

In Equation 7 the actual value of $Pr(T)$ is unknown, because a part of the equation is unobserved. In other words, the unobserved potential outcomes are missing not at random (Rubin, 1976). Since we do not have all information of the full mechanism that causes units to be assigned to either treatment group, we cannot compare the two groups unless other unverifiable assumptions are made about the missingness. For further information on variables that are missing not at random, see Schafer and Graham (2002).

In the special case of a complete randomized experiment, the treatment assignment T_i is independent of the potential outcomes Y_{i1} and Y_{i0}, observed or unobserved, and the covariates X_i (Rubin, 1978), that is

$$Pr(T|X, Y_1, Y_0) = Pr(T). \tag{8}$$

The two groups can be seen as a completely random common set of units, and thus making the difference between the observed group means an unbiased estimate of $\widehat{ATE}_1$ without needing any extra requirements (Rubin, 1974b). This also implies that the unobserved potential outcomes are missing completely at random (Rubin, 1976). It is theoretically important that $Pr(T)$ is not 0 or 1 (Rubin, 1978), so each unit has a chance to be selected in either treatment. Otherwise, the comparison between the potential outcomes would be of no meaning. The treatment effect itself is not assumed to be a constant value for all units, the effect of treatment can vary over units. This implies that the correlation between potential outcomes can deviate from 1.

3 Multiple Imputation of unobserved potential outcomes

The above described $\widehat{ATE}_2$ does not make use of the measured covariates. This is inefficient, because in case of a complete random experiment covariates can be used to increase efficiency of the estimate (Little & Rubin, 2000). MI of unobserved potential outcomes does utilize the information provided by the covariates, which will be described next.

Novel method: making use of covariates by multiple imputation

A straightforward way to think about missing data is to consider how to multiply impute them (Rubin, 2004, 2006). The novel method multiply imputes the unobserved potential outcomes.

Suppose that for the active treated units only the observed potential outcomes Y_{i1}^{obs} are linearly regressed on the covariate measures X_i. By doing this, a model for Y_{i1} is obtained. Assuming the treatment assignment mechanism is ignorable, draws can be generated from

the posterior distribution of $Y_{i1}^{obs}|X_i$ by methods developed in Rubin (1987, p. 167). The draws can be obtained regardless of treatment assignment. The same can be done for the control treatment outcome. Using the observed values for the control group, a model for Y_{i0} is obtained by linear regressing Y_{i0}^{obs} on the covariates X_i. Next, the model is used to obtain a draw from the posterior distribution of $Y_{i0}^{obs}|X_i$ for all units. Now, for every unit a draw for both active and control treatment is available, besides the observed value for one of the two treatments.

Whether a separate model for each variable with missingness must be specified, or if just one joint model for the missingness is specified, depends on the method used for the MI. What the draw consists of and how it is used to obtain the imputed value, depends on the model used for the MI. Although fully conditional specification (FCS) introduced by van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006) in combination with predictive mean matching (PMM), proposed by Little (1986) is used exclusively in this paper, MI of potential outcomes could be applied using another method or model to perform MI as well.

With FCS, a separate conditional model for each variable with missingness based on the remaining observed variables is specified. Next, each variable is completed by imputing a value based on the variable-specific model for each missing value. The imputations are carried out by iterating over all conditional models, where each iteration consists of one cycle through all variables with missingness. A considerable advantage of FCS is its flexibility. The imputation model for the missingness in the data consists of multiple univariate densities, making it for example uncomplicated to combine different measurement levels (van Buuren et al., 2006; van Buuren, 2007).

With PMM, the imputed value for the unobserved potential outcome is obtained as follows. For both potential outcomes separately, coefficients for the slopes and variances and covariances are drawn from the posterior distribution $Y_{i1}^{obs}|X_i$ and $Y_{i0}^{obs}|X_i$. Using these drawn estimates, the predicted values for both the observed outcomes and unobserved potential outcomes are computed. Next, every predicted score of an unobserved potential outcome has a few matches with the closest predicted values of observed outcomes. From the observed outcomes that match the unobserved potential outcome, one observed value is randomly selected and imputed as observed value for the unobserved potential outcome. Since only values that are actually observed in the data are imputed, PMM has potential to preserve non-linear relations even if these are not included in the structural part of the imputation model (Little, 1988).

With the imputed values, a new complete dataset is created in which the observed outcomes remain and the gaps of the unobserved potential outcomes are filled with imputed potential outcomes. This process is repeated m times, creating m complete datasets where the observed outcomes are equivalent and the imputed unobserved potential outcomes slightly differ in each dataset. It should be noted that the number of units used to estimate the model needs to be sufficient in order to prevent the scenario that the same value is imputed each time when using PMM (Little, 1988).

When MI of the unobserved potential outcomes is completed, $\widehat{ATE}_1$ can be computed for every dataset using Equation 2, since the unit level treatment effect is available for each unit. Next, the m versions of $\widehat{ATE}_1$ are pooled to obtain one single outcome measure by

$$\overline{\widehat{ATE}_1} = \frac{1}{m}\sum_{j=1}^{m}\widehat{ATE}_{1j}, \tag{9}$$

where m is the number of replications and $\widehat{ATE}_{1j}$ is the estimated $\widehat{ATE}_1$ in replication j. The pooled SE of $\overline{ATE}_1$ is obtained by

$$S.E._{(\overline{ATE}_1)} = \sqrt{\frac{1}{m}\sum_{j=1}^{m} s_j^2 + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{j=1}^{m}\left(\widehat{ATE}_{1j} - \overline{ATE}_1\right)^2}, \qquad (10)$$

where s_j is the estimated SE in replication j. Both equations are derived from Rubin's (1987) classical equations to pool an estimate and its SE from multiple datasets. Now, using the pooled $\overline{ATE}_1$ and its pooled SE, the precision and significance of $\overline{ATE}_1$ can again be assessed using the t-distributed 95 per cent CI,

$$\overline{ATE}_1 \pm t_{.05/2,N-1} S.E._{\left(\overline{ATE}_1\right)}. \qquad (11)$$

Assumptions of MI of potential outcomes

Besides the assumptions specified for the estimation of the ATE, various other assumptions need to be made to obtain $\overline{ATE}_1$ from the observed data. First of all, a separate conditional model is specified for the unobserved active and control treatment outcomes. Therefore the relation between the potential outcomes is not taken into account. This implies that the partial correlation between the potential outcomes Y_{i1} and Y_{i0} given the covariates X_i, $\rho_{Y_1 Y_0 \cdot X}$, is assumed to be zero. This assumption cannot be checked, for the reason that $\rho_{Y_1 Y_0 \cdot X}$ cannot be estimated from the data (Rubin, 1974a). This might seem a problem, but the estimation of the ATE actually is not biased by the partial correlation that is assumed between the potential outcomes (Gelman, Carlin, Stern, & Rubin, 2004, p. 219).

Other assumptions that are more general to estimating a treatment effect also hold for MI of unobserved potential outcomes. An obvious but important assumption is that the models correctly represent the relationship between the potential outcomes and covariates. Another requirement is that the active and control treatment group should have a substantial overlap in the distribution of the covariates. Otherwise, extrapolation is required to estimate the unobserved potential outcomes, which strongly relies on various assumptions. See King & Zeng (2006) for an elaborate clarification on the dangers of extrapolation. Since this paper is restricted to completely randomized groups, no further assumptions about the model of the distribution of the treatment assignment (T_i) are required.

4 Simulation studies

In this section, the conducted research plus its result are discussed. The conducted research consists of two parts, which are both performed using R64 2.10.1(R Development Core Team). In part one, MI of unobserved potential outcomes will be evaluated by comparing the performance of the novel method to the Student's t-test and ANCOVA for a dataset which is as natural as possible. Several research settings are used in order to mimic a variety of possible situations for when MI of unobserved might perform differently compared to the other methods. In part two, it is investigated with which set of covariates the novel method works best. Also, it is checked if a violation of the assumption $\rho_{Y_1 Y_0 \cdot X} = 0$ influences performance.

In this paper, the MI described above is carried out using Multivariate Imputation by Chained Equations (MICE) (van Buuren & Oudshoorn, 2000), available as R package. The number of imputations used is $m = 20$, since the rate of missing information is likely to be high (Rubin, 1987, p. 132). The number of iterations is set at 10 because with predictive mean matching it is especially important to check whether the estimate has converged (Little, 1988; van Buuren & Oudshoorn, 2000).

Part I: Comparative evaluation of the novel method

To evaluate multiply imputing unobserved potential outcomes, three simulation studies all using an empirical synthetic population are conducted. In this way, the true treatment effect in the population will be known, and the problem that naturally occurring populations rarely conform to assumptions of a simple parametric model is avoided. First, the whole natural dataset will be used to compare the three methods, where the structure of the covariates meets the requirements of ANCOVA. Next, a trimmed version of the natural dataset using only a subset of the covariates is used, where the relation of the covariates with the potential outcomes is artificially altered to compare performance when the dataset does not meet the ANCOVA requirements. On top of this, two aspects which directly influence efficiency and power are varied: sample size and effect size.

The empirical data on which the synthetic populations are based is the Social Medical Survey of Children attending Child Health Clinics (SMOCC) (Herngreen, Reerink, van Noord-Zaadstra, Verloove-Vanhorick & Reys, 1992; Herngreen, van Buuren, Wieringen, Reerink, Verloove-Vanhorick, & Ruys, 1994). The SMOCC is a longitudinal study project concerning a representative Dutch sample of a birth cohort including various perinatal and longitudinal measures of the child and various maternal and other parental characteristics. For the simulation study, the effect of breastfeeding on weight of the child at an average age of two months is utilized as the treatment effect of interest. The synthetic populations are created using the steps outlined in Appendix A.1.

The empirical synthetic populations

All three populations created consist of one million cases. In the synthetic populations, the longitudinal design, item nonresponse or selective dropout of the SMOCC data is not mimicked because it would detract from the main issue at hand. The selective dropout of this data is highlighted by van Buuren (2010). For each of the one million cases in each dataset, the potential outcomes and various covariates are available, which are all described next.

The outcome measure weight at an average age of two months is measured in grams. Each baby has two potential outcomes: weight when exclusively breastfed (*EBF*, denoted by Y_{i1}) and weight when not exclusively breastfed (*no EBF*, denoted by Y_{i0}). For each baby, both potential outcomes are simulated for all datasets. The value of the treatment assignment indicator (T_i) reflects which of the two outcomes are used as observed outcome in the simulation. Table 2 shows the simulated outcomes for a small subsample of babies from the base empirical synthetic population, where D_i denotes the unit-level treatment effect for each baby. Also, the mean and standard deviation for each outcome is included, which is similar for all three populations. The mean unit-level treatment effect for the populations represents ATE_1. The effect of breastfeeding on weight at 2 months is small to medium, with Cohen's measure of effect size d= .26.

Table 2: Illustration of the empirical synthetic population dataset: Treatment assignment (T_i), potential outcomes (Y_{i1}) and (Y_{i0}), and unit-level treatment effect (D_i) for a small subset of babies plus mean (M) and standard deviation (SD) in population

Baby i	T_i	Y_{i1}	Y_{i0}	D_i
1	1	5008	4637*	371*
2	1	4126	4922*	- 796*
3	0	4498*	5198	- 700*
4	1	5456	5348*	108*
.	.	.	.	.
.	.	.	.	.
1.000.000	0	5908*	5759	149*
M in population	0.5	5207.9*	5047.0*	160.9*
SD in population	0.5	649.6*	588.3*	514.4*

Note. T_i = breastfeeding behavior (1 = baby is exclusively breastfed, 0 = baby is not exclusively breastfed); Y_{i1} = weight at 2 months when exclusively breastfed; Y_{i0} = weight at 2 moths when not exclusively breastfed. * Values are known to the researcher but are not used in the analyses since they are unobserved.

Besides the potential outcomes for weight, various covariates of the SMOCC data are included in the synthetic datasets. The most crucial covariate is *birthweight* (*birthw*), since it is highly correlated to the outcome measure (r = .626) and so an effective predictor of weight at 2 months. All covariates used in this research plus their mean, standard deviation and relation to both potential outcomes are presented in Table 3. The decision to include these variables is based on the study by Herngreen, van Buuren, van Wieringen, Reerink, Verloove-Vanhorick and Ruys (1994) on the relationship between background characteristics and length and weight for children followed-up from birth to the age of two years. In the base empirical synthetic population, all described covariates are used. In both trimmed synthetic populations, the covariates presented in bold in Table 3 are used.

For the base empirical synthetic population, the structure of the covariates meets the classical ANCOVA assumptions of linearity and parallel slopes. For the trimmed synthetic populations with altered covariates, all means, standard deviations and correlations presented in Table 2 and 3 hold unless specified otherwise. For the first trimmed synthetic population, the correlations of the covariates with the active and control treatment are altered in such a way that the assumption of equal slopes for both treatment groups is violated. Details are provided in Appendix A.1. The correlations do not only differ between active and control treatment outcome, but also in the strength of their relation with the potential outcomes. This is because the more different the relations with the potential outcomes are for the covariates, the more one single model to describe the relation between the covariates and potential outcomes is a misfit. For the second trimmed synthetic population, the alteration is applied in the kind of relationship between the covariates and potential outcomes. For the covariates *birthw* and *days2* this is made curvilinear instead of linear. For both covariates, a deviation from the mean covariate value results in an higher outcome *weight at 2 moths*. For both trimmed datasets, not the most simplest model with one covariate was used since MI of unobserved potential outcomes needs a certain amount of explained variance by the covariates for each potential outcome to perform well.

Table 3: Description, mean, standard deviation and relation to potential outcomes of covariates in synthetic populations. All presented covariates are included in the base empirical synthetic population, only the covariates presented in boldface are included in the trimmed synthetic populations

Name	*Description*	*M*	*SD*	ρ_{Y_1}	ρ_{Y_0}
Birthw	**Weight in grams at birth**	**3613.8**	**506.9**	**.60**	**.64**
Days2	**Age of child in days on measurement 2**	**58.9**	**6.2**	**.19**	**.28**
Sex	**Sex of the child**	**0.5**	**0.5**	**- .34**	**- .30**
SES	Highest attained formal educational level of the mother (1 = low, 2 = mid-low, 3 = mid-high, 4 = high)	1.7	0.8	.09	.01
AgeM	Age of mother at delivery in completed years	29.2	4.5	.05	.01
Parity	Number of live and still-births after a gestation of 23 weeks or more	0.9	0.9	.15	.13
Dutch	Parents Dutch (1 = Dutch, 0 = otherwise)	0.9	0.3	- .07	- .04
Med	Parents Mediterranean (1 = Mediterranean, 0 0 = otherwise)	0.0	0.2	.08	.03
HightM	Height mother in completed centimeters	168.1	6.8	.19	.19
HightF	Height of father in completed centimeters	178.0	7.5	.12	.11
GA	Gestational age in completed weeks	39.6	1.6	.21	.23

Design

The simulations are performed as follows:

1. Randomly, an active treatment sample of size $n_1 = 100$ and a control treatment sample of size $n_0 = 100$ are drawn without replacement from the base empirical synthetic population, creating a total sample size of $N = 200$. The original effect size of .26 is used.

2. The MI of unobserved potential outcomes, the Student's t-test and ANCOVA is carried out for the sample to estimate the treatment effect. The covariates used are *birthw*, *sex*, *age of mother*, *gestational age*, *height of mother*, *height of father* and *days2*. Only main effects are used. This simplified model is used in order to create an extra dimension of reality, for the reason that in applied research not all variables and relations of the mechanism behind the outcomes are available or even known.

3. The outcome measures are as follows: $\overline{ATE}_1$ for MI of unobserved potential outcomes, $\widehat{ATE}_2$ for the Student's t-test and $\widehat{ATE}_{2,adj}$ for ANCOVA, the SE of all estimated treatment effects, bias of the treatment effects obtained, inclusion of true treatment effect in 95 per cent coverage interval and significance of treatment effect at $\alpha = .05$. The ATE's, standard deviations and biases are all in grams.

4. Step 1 to 3 are repeated 10000 times, for every 1000^{th} sample it is checked if convergence is reached.

5. The outcome measures that are needed for each cell in the design are obtained: mean $\overline{ATE}_1$, mean $\widehat{ATE}_2$, mean $\widehat{ATE}_{2,adj}$, their mean SE in order to assess the precision of the obtained estimates, the RMSE to assess bias of and variability between the

obtained estimates and percentage of significant treatment effects at $\alpha = .05$ to solely assess power. Besides the ATE's and outcomes to asses efficiency and power, percentage coverage of 95 per cent CI and mean absolute bias is obtained to evaluate the performance of the methods. The mean $\widehat{ATE}$'s, mean SE's, mean biases and RMSE's are all in grams.

6. Steps 1 to 5 are repeated using an effect size of .00 and .50. To obtain the different required effect sizes, a constant of 160.9 grams is subtracted and a constant of 149.0 grams is added to all outcomes of the active treatment group (Y_{i1}). The values of the constants are based on the value required for the empirical synthetic population to reach the wanted effect size.

7. Step 1 to 6 are repeated using a total sample size of 50.

8. Step 1 to 7 are repeating using the trimmed empirical synthetic populations, where the relation of the covariates differs for the active and control treatment outcome for the first dataset, and the covariates have a curvilinear relation with the active and control outcome for the second dataset.

Results Part I

For MI of potential outcomes, inspection of every 1000^{th} sample of the parameter estimates against iteration number indicates that convergence is reached for all variations in the simulation. The values of the mean and standard deviation of the imputations vary with iteration number and dataset number m and do not show any definite trend with iteration number.

Both bias and difference in performance of the methods will be more easily statistically significant than it may be practically relevant. Therefore other criteria than significance are used. To evaluate bias, the rule of thumb that the magnitude of the bias in an estimate should not exceed 40 per cent of its SE is used. According to Collins, Schafer and Kam (2001) this is the value when the bias starts to impair the performance of CI's and hypothesis tests. To assess the difference in performance of the methods, the differences between the outcomes will be evaluated while accounting for simulation error. For the outcomes which are scaled in grams, a simulation error of 2 grams is used, for outcomes scaled in percentages, a simulation error of 1 per cent is used.

General performance Tables 4 and 5 summarize the results of the simulations for the base empirical synthetic dataset and dataset where the slopes of the covariates vary for the treatment groups, for all used methods, sample sizes and effect sizes. As can be seen in Table 4 and 5, in general the novel method performs up to standards. The observed biases are small compared to the mean SE's, and percentages of 95 per cent CI's covering the true population values are above 90 per cent. In comparison with the commonly used methods, the coverage is approximatly 2.5 per cent lower. These results do not vary for the applied alterations in the study.

For considerations of space, a table with results for the dataset where covariates have a curvilinear relation with the potential outcomes is not provided in the text. Details of these results are available from the authors upon request. Although CI coverage and bias are up to standards for this dataset, the results clearly show that a more complicated model is needed when relations are non-linear for the novel method to perform well in terms of efficiency and power. Therefore, the results for this dataset will not be further discussed.

Table 4: Performance of two commonly used methods with MI of unobserved potential outcomes for the base empirical synthetic population: mean difference (ATE^*), mean SE of difference ($S.E.^*$), bias ($Bias$), RMSE ($RMSE$), percent coverage of 95 per cent CI's (Cov) and percentage of significant treatment effect at $\alpha = .05$ ($Sign$)

	n		$ATE^*(SE^*)$	$Bias$	$RMSE$	Cov	$Sign$
$d = 0.26$							
($ATE = 160.9$)							
		t-test	157.7 (173.4)	3.2	177.4	94.8	13.8
	50	ANCOVA	160.4 (132.6)	0.5	136.6	94.5	22.9
		MI	160.7 (131.9)	0.3	145.6	92.7	24.0
		t-test	159.0 (87.4)	2.0	89.3	95.1	43.7
	200	ANCOVA	161.4 (62.2)	0.4	62.9	95.3	73.1
		MI	161.4 (62.1)	0.4	67.8	92.9	71.5
$d = 0.00$							
($ATE = 0.0$)							
		t-test	3.2 (173.4)	3.2	177.4	94.8	5.1
	50	ANCOVA	0.5 (132.6)	0.5	136.6	94.5	5.5
		MI	0.3 (131.9)	0.2	145.7	92.4	7.6
		t-test	2.0 (87.4)	2.0	89.3	95.1	4.9
	200	ANCOVA	0.4 (62.2)	0.4	62.9	95.3	4.7
		MI	0.5 (62.2)	0.4	67.7	92.8	7.2
$d = 0.50$							
($ATE = 309.9$)							
		t-test	306.6 (173.4)	3.2	177.4	94.8	40.3
	50	ANCOVA	309.3 (132.6)	0.5	136.6	94.5	62.2
		MI	309.6 (131.9)	0.2	146.0	92.3	61.4
		t-test	307.9 (87.4)	2.0	89.3	95.1	93.9
	200	ANCOVA	310.3 (62.2)	0.4	62.9	95.3	99.9
		MI	310.1 (62.1)	0.2	67.6	92.7	99.7

Note: ATE^*, $S.E.^*$, $Bias$ and $RMSE$ are given in grams.

Usage of a dataset that meets the ANCOVA requirements or a dataset where the slopes of the covariates vary for the treatment groups, sample size and effect size influences differences in power and efficiency between the assessed methods in the following manner.

Power For the base empirical synthetic population there are no differences in power between ANCOVA and the novel method which exceed the simulation error (4). For the dataset where the slopes of the covariates vary for the treatment groups, MI of unobserved potential outcomes yields on average the most precise estimate of the ATE per sample with the smallest mean SE at a sample size of 50 (5). This obviously results in a higher percentage of significant results. Both differences with ANCOVA are however not very substantive and disappear at a sample size of 200. The Student's t-test preforms worst with a substantive difference in mean SE and percentage of significant results compared to the other methods. The above presented differences between the methods do not differ when effect size varies.

Efficiency The value of the RMSE is smaller for ANCOVA for all variations in the

Table 5: Performance of two commonly used methods with MI of unobserved potential outcomes for a dataset where covariates have unequal slopes for active and control treatment group: mean difference (ATE^*), mean SE of difference ($S.E.^*$), bias ($Bias$), RMSE ($RMSE$), percent coverage of 95 per cent CI's (Cov) and percentage of significant treatment effect at $\alpha = .05$ ($Sign$)

	n		$ATE^*(SE^*)$	$Bias$	$RMSE$	Cov	$Sign$
$d = 0.26$							
$(ATE = 160.9)$							
		t-test	159.5 (174.2)	1.4	177.3	94.8	14.4
	50	ANCOVA	161.4 (142.9)	0.5	144.8	94.6	19.7
		MI	159.9 (138.7)	1.0	154.2	92.0	22.4
		t-test	162.5 (87.6)	1.6	89.1	95.1	45.2
	200	ANCOVA	162.6 (70.0)	1.7	72.3	94.7	63.5
		MI	162.3 (69.6)	1.3	76.2	93.1	62.3
$d = 0.00$							
$(ATE = 0.00)$							
		t-test	1.4 (174.2)	1.4	177.3	94.8	5.2
	50	ANCOVA	0.5 (142.9)	0.5	144.8	94.6	4.9
		MI	1.0 (138.6)	1.0	153.2	92.5	7.6
		t-test	1.6 (87.6)	1.6	89.1	95.1	4.9
	200	ANCOVA	1.7 (70.0)	1.7	72.3	94.7	5.3
		MI	1.2 (69.5)	0.7	76.0	92.9	7.1
$d = 0.50$							
$(ATE = 309.7)$							
		t-test	308.3 (174.2)	1.4	177.3	94.8	40.1
	50	ANCOVA	310.2 (142.9)	0.5	144.8	94.6	56.2
		MI	309.0 (138.5)	0.6	153.4	92.1	57.6
		t-test	311.3 (87.6)	1.6	89.1	95.1	94.1
	200	ANCOVA	311.4 (70.0)	1.7	72.3	94.7	99.4
		MI	310.7 (69.6)	1.0	75.7	93.0	99.0

Note: ATE^*, $S.E.^*$, $Bias$ and $RMSE$ are given in grams.

simulations compared to MI of unobserved potential outcomes. Since the biases are similar, the differences in RMSE can be attributed to a difference in variability between the estimated treatment effects. ANCOVA is approximately 10 per cent more efficient than MI of unobserved potential outcomes, where the relative efficiency is calculated as RMSE of ANCOVA divided by RMSE of MI unobserved potential outcomes. The Student's t-test preforms worst with a substantive difference in RMSE compared to the other methods. The above presented differences in efficiency do not differ when sample size or effect size varies.

Part II: Dataset characteristics that influence performance

Because the imputed values of the unobserved potential outcomes get more precise when more and better-predicting covariates are used, there is probably a trade-of between number and quality of covariates used for the imputation and the advantage in terms of efficiency and power of the novel method. Also, it is possible that the performance of the

method will differ with different partial correlations between the potential outcomes given the covariate. To investigate the influence of number of covariates and of the correlations and partial correlations of the variables on the performance of MI of unobserved potential outcomes, four parametric datasets are created each consisting of one million cases. A detailed description of the creation of the four parametric datasets is given in the Appendix A.2.

The simulations are executed in similar fashion as described in Part I. The simulation is only executed for MI of unobserved potential outcomes, and sample size and effect size are not varied. Used sample size is $n = 200$. Used levels of the varied number of covariates, strength of correlation between covariate and potential outcomes and partial correlation between the potential outcomes given the covariate are provided in Table 6.

Results Part II

For MI of potential outcomes, inspection of every 1000^{th} sample of the parameter estimates against iteration number indicates that convergence is reached for all variations in the simulation. To assess the difference in performance of the methods, the differences between the outcomes are evaluated while accounting for simulation error. For the outcomes which are scaled according to the original measurement, a simulation error of 0.01 is used, for outcomes scaled in percentages, a simulation error of 1 per cent is used.

Table 6 summarizes the results of the simulations described in the previous subsection for all four parametric datasets. Usage of different partial correlation between potential

Table 6: Performance of MI of potential outcomes for different number of used covariates $X_{ik}, k = 1, ..., K$, correlations between potential outcomes and covariate ρ_{XY} and partial correlation between potential outcomes given the covariate(s) $\rho_{Y_1Y_0 \cdot X}$: mean difference (ATE^*), mean SE of difference ($S.E.^*$), bias ($Bias$), RMSE ($RMSE$), percent coverage of 95 per cent CI (Cov) and percentage of significant treatment effect ($Sign$)

			MI of potential outcomes				
	ρ_{XY}	$\rho_{Y_1Y_0 \cdot X}$	$\overline{ATE}_1^*(S.E.^*)$	$Bias$	$RMSE$	Cov	$Sign$
$K = 1$							
	.10	.29	1.00 (0.45)	- 0.01	0.48	92.80	63.45
	.10	.60	1.00 (0.45)	0.00	0.47	93.21	64.86
	.30	.23	1.00 (0.30)	0.00	0.30	94.05	90.74
	.30	.56	1.00 (0.30)	0.00	0.30	93.80	90.96
	.60 *	- .09	1.00 (0.23)	0.00	0.25	92.38	98.66
	.60	.38	1.00 (0.23)	0.00	0.25	92.34	98.66
$K = 5^{**}$							
	.50 & .36*	- .09	1.00 (0.23)	0.00	0.25	92.41	98.46
$K = 10^{**}$							
	.50 & .33*	- .09	1.00 (0.23)	0.00	0.25	92.89	98.13

Note: * These models only differ in the number of covariates used, since the explained variance ($R^2 = .36$) for these models is equal. The effect of using a different number of covariates to obtain the model and predicted outcomes for Y_{i1} and Y_i0 can be inferred from these lines. ** The correlation between the covariates is set at .30.

outcomes given covariate and usage of a different number of covariates do not influence the outcomes, since all differences are below the set value of the simulation error.

Usage of different correlations between covariates and the potential outcomes only influences the outcome measures used to assess efficiency and power. The precision of the estimated ATE and the value of the RMSE decrease with larger correlations. Because the bias is zero, it can be said that the decrease in the RMSE is caused by the decreased variability between the estimated effect sizes. The percentage of significant results increases when the used correlation increases, although the difference in percentage of significant results decreases with high correlations.

5 Application to a real dataset

In this section, the approach to multiply impute unobserved potential outcomes is illustrated using data from a clinical trial of the effects of chemotherapy on epileptic seizures (Thall & Vail, 1990). The study compared the effect of the anti-epileptic drug progabide with a placebo on 59 patients suffering from simple or partial seizures. Both treatments are administered on top of standard chemotherapy, since progabide works as an adjuvant for chemotherapy. The number of seizures was counted over four two-week periods. Measured covariates were *baseline seizure rate* (BSR) based on an 8-week pre-randomization seizure count and age of the patient. The slope of the covariate BSR is significantly different for the active and control treatment group, $p < .01$. The covariate *age* does meet the ANCOVA requirement of parallel slopes. The relationship between the covariates and outcome measure is linear.

In this paper, the number of seizures counted over the four two-week periods are transformed to a sum score and both covariates will be used. The sum-score of the observed seizure rate in a four week period ranges from 0 to 302.

When the Student's t-test is applied to the observed data using Equation 3, an estimated average treatment effect of -3.31 ($SE = 11.89$) is obtained. Fitting the ANCOVA model

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 BSR_i + \beta_3 age_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{12}$$

to the observed data gives

$$Y_i = -31.92 + -2.57 T_i + 1.45 BSR_i + 0.74 age_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, 25.72^2). \tag{13}$$

An estimated average treatment effect of -2.57 ($SE = 6.80$) is obtained.

Applying MI of unobserved potential outcomes, the following models for the active and control treatment are obtained:

$$Y_{i1} = -55.54 + 1.78 BSR_i + 1.12 age_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, 26.16^2). \tag{14}$$

$$Y_{i0} = -26.32 + 1.05 BSR_i + 0.97 age_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, 21.92^2). \tag{15}$$

The MI is again performed using the R package MICE (van Buuren & Oudshoorn, 2000), using the FCS model combined with PMM. The number of imputations is $m = 20$, the number of used iterations is 10. The resulting estimated average treatment effect is -3.69 ($SE = 6.13$).

All estimated average treatment effects result in an nonsignificant result and the estimated treatment effect of the three models agree closely. However, the SE of the Student's

t-test is approximately twice as high as the SE's of the MI of unobserved outcomes and ANCOVA. The SE's of last two methods do not differ much. These results are consistent with the simulation studies.

6 Discussion

The goal of this paper was to evaluate MI of unobserved potential outcomes and compare its performance to two commonly used methods. Simulations show that MI of unobserved potential outcomes performs up to standards: the bias is very small compared to the SE and the coverage levels of the 95 per cent CI are above 90 per cent. With a value of around 92,5 per cent the coverage is however not perfect, since the chance of making a type I error is about 1.5 times higher than expected.

Compared to the frequently used Student's t-test, the MI of unobserved outcomes is indeed more powerful and more efficient when the relation between the covariates and potential outcomes is linear. Compared to ANCOVA, the novel method performs approximately equally well. When the assumptions of parallel slopes are violated, the power of the novel method is somewhat higher for small sample size. The differences are however not substantial and disappear at a sample size of 200. Furthermore, for all covariate structures the novel method is somewhat less efficient than the classical ANCOVA, since the RMSE of the novel method is somewhat higher.

The imputation of quadratic relations imposes some extra difficulties (von Hippel, 2009). At least, non-linear relations should be included in the imputation model even when using PMM.

On the question under which circumstances MI of potential outcomes works best, the only tested aspect that seems to matter is the amount of correlation between the covariate and potential outcomes. The performance of the method increases in terms of efficiency and power when the correlation between the covariate and potential outcomes increases. When resulting in the same amount of explained variance, it makes little difference how many covariates are used. The strength of the partial correlation between the potential outcomes given the covariate has no influence on the outcomes.

This first evaluation of the novel way to obtain the treatment effects in case of a completely randomized experiment shows positive results. The performance of the method can possibly be further improved by some fine-tuning. With the current use and data, there is no gain in efficiency compared to the commonly used ANCOVA. It is once again proved that when used for completely randomized groups ANCOVA is an efficient and powerful method despite its negative reputation in some research areas, and is quite robust to violations of its assumptions (Little, An, Hohanns & Giordani, 2000; Porter & Raudenbush). For now, the advantage of MI of potential outcomes in comparison to ANCOVA remains a principal one, for the reason that it makes the existence of both potential outcomes and the definition of the treatment effect explicit to the user.

The first issue that could use some further improvement is that the imputations of the unobserved potential outcomes induce too much variance since they are not precise enough to really outperform ANCOVA for these datasets. A closer inspection at the imputed datasets confirms this. Not only does the mean estimated treatment effect per simulation $\overline{\widehat{ATE}}_1$ vary substantially, also the estimated treatment effect per dataset $\widehat{\widehat{ATE}}_{1j}$ varies to a large degree. Because of the uncertainty in estimating ATE, the RMSE increases and the pooled SE is a lot higher than it could be, making the method less powerful as it possibly

could be. Second, one could wonder whether the SE's are slightly underestimated, because the low coverage is probably caused by too narrow CI's. Even when mean SE of the novel method is similar compared to the other methods, coverage is still lower. Because the value of the estimated ATE fluctuates more using the novel method, the value of the SE probably needs to be larger than the SE of the ANCOVA to reach good coverage.

To better understand how the slight undercoverage and relatively high RMSE of the novel method could possibly be improved, a brief additional experiment was performed. One possibility is that the coverage and RMSE could be improved by using Bayesian linear regression as imputation method instead of PMM. Since estimated values around a regression equation are used instead of observed values, the range of imputed values increase and is less strongly dependent on the outcome values in the sample. A short simulation study similar to the previous simulations is conducted. The coverage does increase to 96.5 per cent, but RMSE is not influenced by using Bayesian linear regression. Concluding, using Bayesian linear regression instead of PMM does not solve the problem of not precise enough imputations. Another option is the use of auxiliary variables. An auxiliary variable is one which is not part of the intended analysis, but can improve imputation by providing extra information about the incomplete variables. Since the estimation of the treatment effect and the imputation of the unobserved potential outcomes are performed in two separate steps, covariates that do not relate to the theoretical model but do improve imputation can be incorporated. The possibility to use auxiliary variables is an advantage over ANCOVA.

This paper focusses only on completely randomized groups, but these can be hard to obtain in practice because of selective drop-out and noncompliance. Also, well-designed observational studies provide valuable information in addition to randomized controlled trials (Concato, Shah & Horwitz, 2000). A whole new advantage emerges when applying MI to non-random groups. Since the treatment effect is obtained at the individual level and then aggregated to a group effect, it is suggested that the outcome will not be biased by non-random groups. It will be interesting to further look into these matters.

Another interesting extension to this research is concerning the partial correlation between the potential outcomes given the covariates. The partial correlation between Y_{i0} and Y_{i1} given the covariates is unknown. Although this does not bias the estimation of the treatment effect (Gelman, Carlin, Stern, & Rubin, 2004, p. 219), it could improve the uncertainty in estimating the average treatment effect. When the relation between potential outcomes is introduced to the models used to estimate the potential outcomes. A possible method to obtain information about the unobserved relation between potential outcomes could be repeated measurements, like Steyer (2005) proposes. If for example more pretest measurements are available, it would be possible to predict for a unit placed in the active treatment group what its outcome would be when it was placed in the control treatment group. This would also permit the construction of a full joint model, instead of using fully conditional specification. Predicting the unknown outcome in this way however does require extra assumptions, in particular the correct extrapolation to the unobserved treatment outcome.

References

Buuren, S. van. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219-242.

Buuren, S. van. (2010). Effects of selective dropout on infant growth standards. In Lucas, A., Makrides, M., Ziegler, E. E. (eds): Importance of growth for health and development. *Proceedings of Nestl Nutr Inst Workshop Ser Pediatr Program*, *65*, 167-179.

Buuren, S. van, Brand, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049 - 1064.

Buuren, S. van, & Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: Mice v1.0 user's manual* (Tech. Rep.). Leiden: TNO Preventie en Gezondheid.

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

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*, 1887 - 1892.

Dominici, F., Zeger, S. L., Parmigiani, G., Katz, J., & Christian, P. (2006). Estimating percentile-specific treatment effects in counterfactual models: a case -study of micronutrient supplementation, birth weight and infant mortality. *Applied Statistics*, *55*, 261-280.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Baysian data analysis*. London: Chapman and Hall.

Herngreen, W. P., Buuren, S. van, Wieringen, J. C. van, Reerink, J. D., Verloove-Vanhorick, S. P., & Ruys, J. H. (1994). Growth in length and weight from birth to 2 years of a representative sample of netherlands children (born in 1988-89) related to socioeconomic status and other background characteristics. *Annals of Human Biology*, *21*, 449-463.

Herngreen, W. P., Reerink, J. D., Noord-Zaadstra, B. M. van, Verloover-Vanhorick, S. P., & Ruys, J. H. (1992). Smocc: Design of a representative cohortstudy of live-born infants in the netherlands. *The European Journal of Public Health*, *2*, 117-122.

Hippel, P. T. von. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265 - 291.

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Reserach Methodology*, *5*, 28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945 - 970.

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

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*, 131 - 159.

Little, R. J. (1986). Missing data in cencus bureau surveys. In *Proceedings of the second annual census bureau research conference.*

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*, 287 - 296.

Little, R. J., An, H., Hohanns, J., & Giordani, B. (2000). A comparison of subset selection and Analysis of Covariance for the adjustment of confounders. *Psychological Methods*, *5*, 459 - 476.

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, *21*, 121-145.

Miller, I., & Miller, M. (2004). *John E. Freund's mathematical statistics, 7^{th} edition.* New Yersey: Prentice Hall.

Neyman, J. (1923). On the application of probability theory to agricultural experiment. Essay on principles. Section 9. *Roczniki nauk rolniczych tom x [in Polish]; translated in Statistical Sciences*, *5*, 465 - 480.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, *34*, 383 - 392.

R Development Core Team. (2009). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Rubin, D. B. (1974a). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, *69*, 467 - 474.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrica*, *63*, 581 - 592.

Rubin, D. B. (1978). Baysian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*, 34 - 58.

Rubin, D. B. (1980). Discussion of 'randomization analysis of experimantal data in the fisher randomization test," by D. Basu. *Journal of the American Statistical Association*, *75*, 591 - 593.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley and Sons.

Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472 - 480.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *The Scandinavian Journal of Statistics*, *31*, 161-170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322-331.

Rubin, D. B. (2006). Conceptual, computational and inferential benefits of the missing data perspective in applied and theoretical statistical problems. *Allgemeines Statistisches Archiv*, *90*, 5001-513.

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

Schafer, L. S., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279 - 313.

Steyer, R. (2005). Analyzing individaul and average cuasal effects via structural equation modeling. *Methodology*, *1*, 39 - 54.

Thall, P. F., & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, *46*, 657 - 671.

A Appendix

A.1 How the empirical synthetic populations are generated

The base synthetic empirical population and the two trimmed empirical synthetic populations where the relation of the covariate with the potential outcomes are artificially altered are created in the following manner.

Base empirical synthetic population The procedure used to create the empirical synthetic population, is similar to how Schafer and Kang (2008) built their population in order to review different strategies for estimating the average causal effect for non-random groups. Distributions used to create the synthetic population are estimated from the SMOCC data, using cases for which at least the dependent variable *weight at 2 months*, exclusively breastfeeding (*EBF*) on measurement 2 and *birthweight* are observed. This leaves 1540 babies in the empirical dataset. With the data on these 1540 babies the empirical synthetic population of 1 million cases large is constructed as follows.

1. One million values for the possible combinations of the variables *sex*, *ethnicity* and *social economic status* (SES) are randomly sampled from a discrete multivariate distribution. To obtain the correct marginal, pairwise relations and three-way interactions for the variables, the unweighted proportions of the *sex* (2) $\times$ *ethnicity* (3) $\times$ *SES* (4) contingency table of the SMOCC dataset are used as sampling probabilities.
2. The values for the ordinal variable *parity* are obtained by randomly sampling a discrete value between 0 and 4, based on the unweighted proportions for each discrete value of *parity*. These proportions differ with the values observed for the previously sampled variables *ethnicity* and *SES*. The used proportions are derived from the contingency table *ethnicity* (3) $\times$ *SES* (2) of the SMOCC dataset. The lowest and highest two levels of *SES* are aggregated and the covariate *sex* is not used to avoid a low number of values within the cells.
3. The values for the covariate *age of mother* are obtained using a regression model estimated from the SMOCC data which includes the main effects for the previously obtained covariates in the sequence.
4. To the just simulated values of the covariate an error term with a marginal distribution similar to that in the SMOCC dataset is added.
5. Step four and five are repeated for subsequently the covariates *birthweight*, *gestational age*, *height of mother*, *height of father* and *age of child on measurement 2*.
6. Using the simulated covariates, the active treatment potential outcomes *EBF* (Y_{i1}) are obtained using an elaborate regression model estimated from the SMOCC dataset. All main effects of the covariates are included, plus a quadratic relation for *age of child on measurement 2* and an interaction term for *sex* $\times$ *birhtweight*, *sex* $\times$ *gestational age*, *sex* $\times$ *parents Mediterranean* and *low SES* $\times$ *parents Mediterranean*. The added quadratic relation and interaction terms are theory driven, not data driven.

7. The control treatment potential outcomes *no EBF* (Y_{i0}) are obtained in similar fashion as step 6, but with it's own intercept and slopes estimated from the SMOCC dataset.
8. To the potential outcomes a correlated random residual is added, which marginal distribution matches the empirical distribution of the residuals in the SMOCC data. The residual for the potential outcomes are correlated since it is not realistic that the covariates account for all observed differences between the potential outcomes. The correlation between the potential outcomes is set at an value of .30. This is an educated guess, because the correlation between the potential outcomes cannot be estimated from the SMOCC dataset.
9. One million values for the treatment indicator are randomly sampled with $Pr(T) = 0.5$, where 1 denotes active treatment and 0 control treatment. For each case, the value of the treatment indicator indicates which of the potential outcomes is treated as observed in the dataset.

The simulated covariates are very similar to the covariates in the SMOCC data regarding the marginal distribution, pairwise relations and most three-way interactions, just as the simulated potential outcomes.

Covariates with unequal slopes for treatment groups To simplify interpretation, the variance-covariance matrixes are provided in standardized fashion as correlation matrixes. The standard deviations for the potential outcomes Y_{i1} and Y_{i0}, *birthweight*, *age of child on measurement 2* and *sex* are similar to the obtained standard deviations in the base empirical synthetic population: 649.6, 588.3, 506.9, 6.2 and 0.5.

One million sets of observations ($Y_{1\ 1}$, $Y_{1\ 0}$, $birthweight_1$, $age\ of\ child\ on\ measurement\ 2_1$, sex_1), ..., ($Y_{1000000\ 1}$, $Y_{1000000\ 0}$, $birthweight_{1000000}$, $age\ of\ child\ on\ measurement\ 2_{1000000}$, $sex_{1000000}$) are drawn from

$$N\left(\begin{pmatrix} 5207.9 \\ 5046.9 \\ 3613.8 \\ 58.9 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & .30 & .60 & .20 & -.10 \\ .30 & 1 & .20 & .60 & -.50 \\ .60 & .20 & 1 & -.04 & -.15 \\ .20 & .60 & -.04 & 1 & .04 \\ -.10 & -.50 & -.15 & .04 & 1 \end{pmatrix}\right).$$

Each case is assigned a treatment indicator, with $Pr(T) = 0.5$.

Covariates have quadratic relation with potential outcomes The distributions used to create the population are estimated from the base empirical synthetic population instead of the SMOCC data, in order to obtain the same means and variance-covariance structure for the quadratic data as the base empirical population.

1. The simulated covariates *birthweight* and *age of child on measurement 2* of the base empirical synthetic population are centered.
2. Using these centered covariates plus the covariate *sex* of the base empirical synthetic population, the active treatment potential outcomes *EBF* (Y_{i1}) are obtained using a regression model estimated from the base empirical synthetic population plus added quadratic term for *birthweight* and *age of child on measurement 2*. The slope of the first quadratic term is 0.001, the slope of the

second quadratic term is 2.5. This results in a strong curvature for *birthweight* and a light curvature for *age of child on measurement 2*. The intercept is slightly adjusted to obtain the same mean *EBF* as in the base empirical population while including the quadratic terms.

3. The control treatment potential outcomes *no EBF* (Y_{i0}) are obtained in similar fashion as step 2, but with it's own intercept and slopes estimated from the base empirical synthetic population. The used values for the quadratic slopes are equal.
4. To the potential outcomes a random residual is added, which are correlated with $\rho = .30$. The SE of the marginal distributions of the residuals of the base empirical population are slightly adjusted to obtain the same SE for the potential outcomes as in the base empirical population while including the quadratic terms.
5. Each case is assigned a treatment indicator, with $PR(T) = 0.5$.

A.2 How the parametric synthetic populations are generated

All datasets are created by randomly drawing the values of the variables using the R package 'mvtnorm' (REF), using the settings specified below. To simplify interpretation, the variance-covariance matrixes are provided in standardized fashion as correlation matrixes. In all datasets, the standard deviation of both potential outcomes Y_{i1} and Y_{i0} is 2 and the standard deviation of all covariates X_{ik} is 5. As can be seen below, the only difference between the datasets is the number of covariates and correlation matrix used. The four dataset are next denoted in terms of means and correlation matrix used.

Dataset 1 One million sets of observations ($Y_{1\ 1}$, $Y_{1\ 0}$, $X_{1\ 1}$, $X_{1\ 2}$, $X_{1\ 3}$), ..., ($Y_{1000000\ 1}$, $Y_{1000000\ 0}$, $X_{1000000\ 1}$, $X_{1000000\ 2}$, $X_{1000000\ 3}$) are drawn from

$$N\left(\begin{pmatrix} 9 \\ 8 \\ 25 \\ 25 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & .30 & .10 & .30 & .60 \\ .30 & 1 & .10 & .30 & .60 \\ .10 & .10 & 1 & .30 & .30 \\ .30 & .30 & .30 & 1 & .30 \\ .60 & .60 & .30 & .30 & 1 \end{pmatrix}\right).$$

The resulting partial correlations between the potential outcomes given each of the covariates are .29, .23 and -.09. Each case is assigned a treatment indicator, with $Pr(T) = 0.5$. In the simulations, one covariate X_k is used at a time.

Dataset 2 One million sets of observations ($Y_{1\ 1}$, $Y_{1\ 0}$, $X_{1\ 1}$, $X_{1\ 2}$, $X_{1\ 3}$), ..., ($Y_{1000000\ 1}$, $Y_{1000000\ 0}$, $X_{1000000\ 1}$, $X_{1000000\ 2}$, $X_{1000000\ 3}$) are drawn from

$$N\left(\begin{pmatrix} 9 \\ 8 \\ 25 \\ 25 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & .60 & .10 & .30 & .60 \\ .60 & 1 & .10 & .30 & .60 \\ .10 & .10 & 1 & .30 & .30 \\ .30 & .30 & .30 & 1 & .30 \\ .60 & .60 & .30 & .30 & 1 \end{pmatrix}\right).$$

The resulting partial correlations between the potential outcomes given each of the covariates are .60, .56 and .36. Each case is assigned a treatment indicator, with $Pr(T) = 0.5$. In the simulations, one covariate X_k is used at a time.

Dataset 3 One million sets of observations ($Y_{1\ 1}$, $Y_{1\ 0}$, $X_{1\ 1}$, $X_{1\ 2}$, $X_{1\ 3}$, $X_{1\ 4}$, $X_{1\ 5}$), ..., ($Y_{1000000\ 1}$, $Y_{1000000\ 0}$, $X_{1000000\ 1}$, $X_{1000000\ 2}$, $X_{1000000\ 3}$, $X_{1000000\ 4}$, $X_{1000000\ 5}$) are drawn from

$$N\left(\begin{pmatrix} 9 \\ 8 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & .30 & .50 & .36 & .36 & .36 & .36 \\ .30 & 1 & .50 & .36 & .36 & .36 & .36 \\ .50 & .50 & 1 & .30 & .30 & .30 & .30 \\ .36 & .36 & .30 & 1 & .30 & .30 & .30 \\ .36 & .36 & .30 & .30 & 1 & .30 & .30 \\ .36 & .36 & .30 & .30 & .30 & 1 & .30 \\ .36 & .36 & .30 & .30 & .30 & .30 & 1 \end{pmatrix}\right).$$

The resulting partial correlation between the potential outcomes given all covariates is -.09. Each case is assigned a treatment indicator, with $Pr(T) = 0.5$. All five covariates are used simultaneously in the simulation.

Dataset 4 One million sets of observations ($Y_{1\ 1}$, $Y_{1\ 0}$, $X_{1\ 1}$, $X_{1\ 2}$, $X_{1\ 3}$, $X_{1\ 4}$, $X_{1\ 5}$, $X_{1\ 6}$, $X_{1\ 7}$, $X_{1\ 8}$, $X_{1\ 9}$, $X_{1\ 10}$), ..., ($Y_{1000000\ 1}$, $Y_{1000000\ 0}$, $X_{1000000\ 1}$, $X_{1000000\ 2}$, $X_{1000000\ 3}$, $X_{1000000\ 4}$, $X_{1000000\ 5}$, $X_{1000000\ 6}$, $X_{1000000\ 7}$, $X_{1000000\ 8}$, $X_{1000000\ 9}$, $X_{1000000\ 10}$) are drawn from

$$N\left(\begin{pmatrix} 9 \\ 8 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & .30 & .50 & .33 & .33 & .33 & .33 & .33 & .33 & .33 & .33 & .33 \\ .30 & 1 & .50 & .33 & .33 & .33 & .33 & .33 & .33 & .33 & .33 & .33 \\ .50 & .50 & 1 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & 1 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & 1 & .30 & .30 & .30 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & 1 & .30 & .30 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & 1 & .30 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & .30 & 1 & .30 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & .30 & .30 & 1 & .30 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & 1 & .30 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & 1 & .30 \\ .33 & .33 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & .30 & 1 \end{pmatrix}\right).$$

The resulting partial correlation between the potential outcomes given all covariates is -.09. Each case is assigned a treatment indicator, with $Pr(T) = 0.5$. All ten covariates are used simultaneously in the simulation.