# Extending balance assessment for the generalized propensity score under multiple imputation

Anna-Simone Josefine Frank[1,*] David S. Matteson[2,3]
Hiroko K. Solvang[4] Angela Lupattelli[5] Hedvig Nordeng[5,6]

[*]Corresponding author: anna-simone.frank@uib.no
[1]Comp. Biol. Unit (CBU), Dept. of Informatics, Univ. of Bergen, Norway
[2]Dept. of Statistical Science, Cornell Univ., Ithaca, New York, USA
[3]Dept. of Biol. Statstics and Comp. Biology, Cornell Univ., Ithaca, New York, USA
[4]Institute of Marine Research, Bergen Norway
[5]Dept. of Pharmacy, Univ. of Oslo, Oslo Norway
[6]Dept. of Child Health and Development, Nasjonalt folkehelseinstitutt, Oslo Norway

Abstract

This manuscript extends the definition of the Absolute Standardized Mean Difference (ASMD) for binary exposure ($M = 2$) to cases for $M > 2$ on multiple imputed data sets. The Maximal Maximized Standardized Difference (MMSD) and the Maximal Averaged Standardized Difference (MASD) were proposed. For different percentages, missing data were introduced in covariates in the simulated data based on the missing at random (MAR) assumption. We then investigate the performance of these two metric definitions using simulated data of full and imputed data sets. The performance of the MASD and the MMSD were validated by relating the balance metrics to estimation bias. The results show that there is an association between the balance metrics and bias. The proposed balance diagnostics seem therefore appropriate to assess balance for the generalized propensity score (GPS) under multiple imputation.

# 1 Introduction

It is impossible in observational studies to control how subjects are assigned into treatment groups, and this may potentially result in biased effect estimates caused by confounding (Rubin (2004), Hernán, Alonso, Logan, Grodstein, Michels, Stampfer, Willett, Manson, and Robins (2008)). The application of weights derived from the propensity score (the conditional probability of receiving treatment given observed covariates) aims to balance characteristics between treatment groups (Rubin (2004)). When this is achieved, the result is reduced bias effect in the estimates (Austin (2011)). For binary exposures, a common approach to checking whether balance between treatment and control group has been achieved is to calculate the Absolute Standardized Mean Difference (ASMD, $d$). For a continuous covariate $x$, the ASMD is defined as

$$d = \frac{|\bar{x}_{treatment} - \bar{x}_{control}|}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}},$$

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ denote respectively, the sample mean of the covariate in the treated and control groups, while $s^2_{treatment}$ and $s^2_{control}$ represent the sample variance in the respective treatment groups (Austin (2018), McCaffrey, Griffin, Almirall, Slaughter, Ramchand, and Burgette (2013)). When defining $d$ for categorical variables, mean values ($\bar{x}$) are replaced with proportions and variances ($s^2$) become functions of proportions (Austin (2018)).

If balance has been achieved, the ASMD value should be below a pre-defined threshold after propensity score analysis has been performed (Austin (2018), McCaffrey et al. (2013)).

There is however inconsistency in balance thresholds across the literature, where values vary between 0.1 and 0.25 for $d$ (Stuart, Lee, and Leacy (2013), Nguyen, Collins, Spence, Daurès, Devereaux, Landais, and Le Manach (2017), Austin (2011)). Given these differences, a recent study investigated the effectiveness of pre-defined thresholds for propensity score matching with binary exposure (Nguyen et al. (2017)). The authors concluded that if balance was below 0.1 on most variables, and if variables with balance above 0.1 are included in the outcome model as adjustment variables, bias was small. Yet, a specific threshold for the generalized propensity score (GPS) has not been tested.

The GPS is the generalization of the propensity score for binary exposure. The GPS makes it possible to estimate the effect of multiple treatment exposure on the outcome (Lechner (2001), Imai and Van Dyk (2004)). The GPS has been applied in various studies on observational data, for example, (Spreeuwenberg, Bartak, Croon, Hagenaars, Busschbach, Andrea, Twisk, and Stijnen (2010), Feng, Zhou, Zou, Fan, and Li (2012), Jiang and Foster (2013), Sugihara (2010)).

A common problem of observational data is missing information due to non-response, especially where data collection is performed via questionnaires (Pandis (2014)).

Multiple imputation techniques are often applied to fill the information gaps (Miri, Hassanzadeh, Rajaeefard, Mirmohammadkhani, and Angali (2016)). For binary exposure, several studies have combined imputation techniques with propensity score analysis (Eulenburg, Suling, Neuser, Reuss, Canzler, Fehm, Luyten, Hellriegel, Woelber, and Mahner (2016), Rosenbaum and Rubin (1984), Lavori, Dawson, and Shera (1995), Mitra and Reiter (2016), Kupzyk and Beal (2017),

Doidge (2018), Hsu and Yu (2018), Qu and Lipkovich (2009), Hill (2004), Hayes and Groner (2008)). A systematic review by Malla, Perera-Salazar, McFadden, Ogero, Stepniewska, and English (2018) summarizes how missing data are combined with propensity score analysis on actual patient data. The authors showed that the majority of reviewed articles performed propensity score analysis in combination with complete case data (Malla et al. (2018)).

Unlike for propensity score analysis, only one recent study, by De Vries, Van Smeden, and Groenwold (2018), considered the combination of missing data and the GPS (De Vries et al. (2018)). The authors found that multiple imputation of data, followed by propensity score estimation using classification and regression trees resulted in least biased estimates (De Vries et al. (2018)). However, no previous study has assessed balance for the GPS under multiple imputations, partly due to the computational burden involved (De Vries et al. (2018)).

This paper proposes an approach for balance assessment of GPS under multiple imputation. It uses simulated data to evaluate the proposed diagnostics. The article is organized in the following way. Section 2, the methodology section, defines the quantitative concepts applied in this study, such as the GPS, its estimation, a multiple treatment definition and multiple imputation. Existing balance diagnostics are reviewed for binary propensity score models and finally extended to the multiple treatment case under multiple imputation. Section 3, the simulation study section, describes the data-generation process and details about the implementation. The results of simulated data example are presented in Section 4, followed by the discussion of the results in Section 5 and finally the conclusions in Section 6.

# 2 Methodology

In this section we describe, the methodological concepts and ideas that are later applied to simulated data examples.

## 2.1 Generalized propensity score

The aim of propensity score analysis is to reduce the dimension of observed pre-treatment variables $X$ and bias, due to confounding, by re-weighting them.

For multiple treatments, this can be achieved by using the GPS (Imbens (2000)) In the present study, we will estimate GPS using generalized boosted models (GBM), and combine it with inverse probability of treatment weights (IPTW) to obtain the average treatment effect (ATE) (McCaffrey, Ridgeway, and Morral (2004), McCaffrey et al. (2013)).

We applied IPTW based on the results by Nian, Yu, Ding, Wu, Dupont, Brunwasser, Gebretsadik, Hartert, and Wu (2019), who showed that this approach had preferred performance compared to other approaches, such as matching, stratification or GPS adjustment. In addition propensity score methods may differ in the population where an overall treatment estimate shall be calculated Kurth, Walker, Glynn, Chan, Gaziano, Berger, and Robins (2005). The results in Kurth et al. (2005) have shown that IPTW is well suited to calculate the treatment effect of the total population.

Following, Imbens (2000) and Feng et al. (2012), for the subject index $i = 1, \ldots N$, let $\mathscr{T} = \{1, 2, \ldots, M\}$ denote the set of $M$ multiple treatments, and $Y_i(m)$ denote the potential outcome of subject $i$, if subject $i$ has been assigned to treatment $m \in \mathscr{T}$ (Imbens (2000), Feng et al. (2012)). Let $T_i$ be the treatment that subject $i$ received and the indicator that subject $i$ receives treatment $m$ is

$$I_m(T_i) \quad = \quad I(T_i = m). \tag{1}$$

**Definition 2.1.** (Generalized Propensity Scores (GPS) (Imbens (2000), McCaffrey et al. (2013))) Let $X_i$ be the set of observed pre-treatment variables of subject $i$. The GPS $r(m, X_i)$ is the conditional probability of receiving a particular level of the treatment $m$ given $X_i$:

$$r(m, X_i) \quad = \quad Pr(T_i = m | X_i) = \mathbb{E}[I_m(T_i) | X_i]. \tag{2}$$

Hence, for the set $\mathscr{T}$ of $M$ different treatment groups, we obtain $M$ GPSs, for all subjects $i$.

Given $r(m, X_i)$, the empirical expected outcome $\hat{\mathbb{E}}(Y(m))$ is estimated by the weighted mean (McCaffrey et al. (2013)), i.e.

$$\hat{\mathbb{E}}(Y(m)) \quad = \quad \frac{\sum_{i=1}^{N} I_m(T_i) Y_i(m) w_i(m)}{\sum_{i=1}^{N} I_m(T_i) w_i(m)}, \tag{3}$$

$$\text{where} \quad T_i = m, \quad w_i(m) = \frac{1}{r(m, X_i)}. \tag{4}$$

Then, the average treatment effect (ATE) is estimated, comparing treatment $t \in \mathscr{T}$ versus treatment $\ell \in \mathscr{T}$ ($t \neq \ell$) across the subjects

$$\widehat{\text{ATE}}_{t\ell} \quad = \quad \hat{\mathbb{E}}[Y(t)] - \hat{\mathbb{E}}[Y(\ell)], \tag{5}$$

under the condition that treatment ($T_i$) is independent of the outcome $Y_i(T_i)$, and that the compared groups are representatives of the population (Feng et al. (2012)). The latter condition does not hold in general, but can be achieved by the key assumption that treatment assignment is *weakly unconfounded* given observed covariates $X$. For multi-valued treatments, without missing data, the *weak unconfoundedness* assumption was defined by Imbens (2000), see Definition S1.1 in the Supplementary material section S1. This definition means that treatment and potential outcome are independent given the observed covariate (Imbens (2000)). It was shown in Leyrat, Seaman, White, Douglas, Smeeth, Kim, Resche-Rigon, Carpenter, and Williamson (2019a), that this assumption holds for a binary exposure under multiple imputation. However, to be able to define the ATE for the present situation, the following assumptions have to hold for the GPS after multiple imputation:

For a proof of these assumptions, see the Supplementary material sub-section S1.1.

Previous studies estimated GPS with multinomial and ordinal logistic regression models (Feng et al. (2012), Bray, Dziak, Patrick, and Lanza (2018)). However, these parametric approaches have been shown to lead to less robust ATE estimates than non-parametric approaches,

such as generalized boosted models, which we describe briefly below (McCaffrey et al. (2004, 2013)).

### 2.1.1 Generalized boosted models

Firstly, we describe how GBM are applied to estimate the propensity score for binary treatment, i.e. $\mathscr{T} \in \{0,1\}$ (McCaffrey et al. (2004, 2013)). Then the GBM algorithm is generalized to the multiple treatment case analogous to McCaffrey et al. (2004, 2013).

Let $X$ be the set of observed pre-treatment variables and let $X_i$ be the set of observed pre-treatment variables for subject $i$. For the binary treatment case, the treatment indicators for subject $i$ in equation (1) is given by $T_i = 1$, simplified with $I_1(T_i) = 1 - I_0(T_i)$ and $r(1,X_i) = 1 - r(0,X_i)$. Instead of directly estimating the propensity scores, the algorithm finds the maximum-likelihood estimate of the function $g(X)$, which is the log-odds of treatment assignment, i.e. $g(X) = \log(r(1,X)/(1-r(1,X)))$. Therefore, the GBM algorithm iteratively adds regression trees together to fit a non-linear logistic regression model to treatment indicator $I_1(T_i)$ (McCaffrey et al. (2004)). Initially, the algorithm sets $g_0(X) := \log(\bar{I}/(1-\bar{I}))$, where $\bar{I}$ is the average treatment assignment indicator for the whole sample. Then to improve the propensity score fit to the data, at each new iteration $j$ ($j = 1,\ldots J$), a new regression tree $h(X)$ is added to the current model $g_{j-1}(X)$, if it is the best fit to the residuals $I_1(T_i) - g_{j-1}(X_i)$ and provides the greatest increase in the log-likelihood for the data. When the regression trees are combined a shrinkage coefficient $\alpha$ is introduced to improve smoothness of the resulting piecewise constant model. In order to avoid overfitting of the data, GBM selects a number of trees in order to minimize imbalance on pre-treatment covariates across "treatment" and "control" groups. More details about the boosting algorithm can be found in McCaffrey et al. (2004).

### 2.1.2 GBM extension to more than two treatment groups ($M > 2$)

The approach in Section 2.1.1 can be extended to more than two treatment groups in the following way (McCaffrey et al. (2004, 2013)): Firstly, we define indicator functions for each of the $M$ treatment groups as in equation (1). The GBM algorithm is then applied respectively to all $M$ treatment indicator functions $I_m(T_i)$, resulting in $M$ propensity scores $r(m,X_i) = Pr(T_i = m|X_i)$. The respective IPTW $w_i(m) = \frac{1}{r(m,X_i)}$ are then applied in equation (3).

As mentioned previously, missing data are common in observational data. Therefore, techniques, which fill information gaps have been invented, such as multiple imputation by chained equations, which we describe below (Van Buuren and Groothuis-Oudshoorn (2011)).

## 2.2 Multiple imputation by chained equations

Let $\mathscr{S}$ be a data set and let $\mathscr{Z}$ be the subset for the elements that are missing in $\mathscr{S}$. Multiple imputation by chained equations (MICE) consists of three main steps (Van Buuren and Groothuis-Oudshoorn (2011)): Firstly, there is imputation of the missing data component $\mathscr{Z}$, using chained equations. After an initial random filling of all missing data, a MCMC algorithm draws iteratively

from the conditional distributions based on a collection of observed and missing variables (Chen and Ip (2015), Van Buuren and Groothuis-Oudshoorn (2011).) This results in $Q$ multiple imputed data sets $\mathscr{S}^{(q)}$, where $q = 1, \ldots, Q$, that are identical on each observed element, but differ on the initially missing entries. On each of the $q$ imputed data sets, propensity score analysis is performed, and leads to $Q$ effect estimates of the propensity score analysis. Finally, all $Q$ effect estimates and their variance are pooled into one estimate. Following Murray et al. (2018), as many variables as possible, including the outcome variable, should be incorporated into the imputation model (Murray et al. (2018), Leyrat et al. (2019a)).

Initially, the literature recommended that $5 \leq Q \leq 10$ imputed sets are sufficient (Azur, Stuart, Frangakis, and Leaf (2011)). However more recent developments showed that in order to also detect small effect sizes a minimum of $Q = 40$ imputed sets are needed (this however depends on percentage of missing data), see Graham, Olchowski, and Gilreath (2007).

There are three approaches to combine multiple imputation and propensity score analysis (Leyrat, Seaman, White, Douglas, Smeeth, Kim, Resche-Rigon, Carpenter, and Williamson (2019b)): After multiple imputation, one approach averages the estimated propensity scores to generate one outcome analysis (Mitra and Reiter (2016), Leyrat et al. (2019b)). Another approach, calculates the propensity score based on the pooled covariates over all imputed data sets. A third approach averages the $Q$ treatment effects, which were estimated based on the propensity score on each imputed data set. Leyrat et al. (2019b) compared all three approaches and concluded that the third one leads to least biased results. This approach was therefore applied in the below illustrative example (see Section 3).

Next we review and extend the definition of balance.

## 2.3 Balance assessment

In order to assess if the propensity score analysis was able to balance treatment groups across observed pre-treatment variables $X$, balance is calculated. Previous studies extended balance diagnostics to GPS for continuous and multiple treatment for complete data (Zhu, Coffman, and Ghosh (2015), Fong, Hazlett, and Imai (2018), Austin (2018), Bray et al. (2018), McCaffrey et al. (2013)). We first review the definition of balance for binary and multiple treatment groups and finally extend this definition for the case under multiple imputation. In addition to previous notations, let $k = 1, \ldots, K$ denote the covariate index and let $C = \binom{M}{2}$ denote the total number of pairwise comparisons of $M$ treatments and let $c$ be an index over comparison pairs, i.e. $c = 1, \ldots, C$.

### 2.3.1 Balance assessment without multiple imputation

**Definition 2.2.** (Absolute Standardized Mean Difference (ASMD) (McCaffrey et al. (2013))) For each covariate $k$ and binary treatment $m$, the ASMD equals the absolute value of the difference between the weighted mean of the covariate in the treatment group ($m = 1$) minus the weighted mean of the covariate in the control group ($m = 0$), divided by the unweighted standard deviation

of the pooled population for the ATE (see equation (5)). Given covariate $k$,

$$ASMD_k \quad = \quad \frac{|\bar{X}_{k,1} - \bar{X}_{k,0}|}{\hat{\sigma}_k},$$

where $\bar{X}_{k,m}$ is the weighted mean for covariate $k$ and treatment ($m = 1$) or control ($m = 0$) group, and $\hat{\sigma}_k$ is the unweighted averaged (pooled) within standard deviation for all treatment groups.

For GPS, Burgette, Griffin, and McCaffrey (2017) suggest that, balance shall be assessed, via the ASMD, for all pairwise comparisons when determining the ATE.

Definition 2.3 extends Definition 2.2 to $M$ treatment groups $m = 1, \ldots, M$, which will result in $C$-pairwise comparisons (Burgette et al. (2017), McCaffrey et al. (2004)).

**Definition 2.3.** (Multiple ASMD (McCaffrey et al. (2013))) For $M$ multiple treatment groups, and given the set of $K$ covariates and treatment pairs $(t, \ell), t \neq \ell$,

$$ASMD_{k,c} \quad = \quad \frac{|\bar{X}_{k,t} - \bar{X}_{k,\ell}|}{\hat{\sigma}_k}.$$

$\bar{X}_{k,t}$ denotes the weighted mean for covariate $k$ for treatment group $t$, while $\bar{X}_{k,\ell}$ is the weighted mean for covariate $k$ for treatment group $\ell$. The denominator $\hat{\sigma}_k$ is the unweighted averaged (pooled) standard deviation of all treatment groups (same as in Definition 2.2).

A similar definition holds for categorical variables, where the mean is replaced by weighted proportions.

Except for McCaffrey et al. (2013), two recent articles (Li and Li (2019), Yang, Imbens, Cui, Faries, and Kadziola (2016)) included discussions on balance assessments for more than two treatment arms: The Multiple ASMD is a special case of the pairwise absolute standardized differences (ASD) defined by Li and Li (2019), with $w_i(m) = \frac{1}{r(m,X_i)}$ and tilting function $h(X) = 1$. In addition, from McCaffrey et al. (2013), Li and Li (2019) extended the population standardized difference (PDS) for varying weight functions ($w_i(m)$). Instead of comparing weighted covariate means pairwise between treatment groups, the PDS compares weighted covariate means between the treatment group and the target population. The balance metric in Yang et al. (2016) are defined for matching and stratification, therefore $w_i(m) = 1$. Another major difference is that covariate means for each treatment group $m$ are not compared pairwise, but with the covariate mean of all other treatment groups combined, except $m$ (i.e., $m^c = T$). Finally, Yang et al. (2016) also proposed a metric that allows the assessment of balance in covariate distributions, where the GPS are considered instead of the covariates directly.

In the next section, Definition 2.3 is extended to multi-treatment exposure under multiple imputation.

### 2.3.2 Balance assessment for GPS under multiple imputation

**Definition 2.4.** (Maximal Maximized Standardized Difference (MMSD)) For each covariate $k$, the MMSD is defined by,

$$MMSD_k = \max_c \max_{q=1,\ldots,Q} ASMD_{q,k,c}, \tag{6}$$

where $ASMD_{q,k,c}$ refers to the ASMD for covariate $k$ and pairwise comparison $c$ on the multiple imputed set $q$.

**Definition 2.5.** (Maximal Averaged Standardized Difference (MASD)) Similarly, for each covariate $k$, the MASD is defined by,

$$MASD_k = \max_c \frac{1}{Q} \sum_{q=1}^{Q} ASMD_{q,k,c}, \tag{7}$$

where $ASMD_{q,k,c}$ is defined as above in Definition 2.4.

It is assumed that balance is obtained when, after weighting, the MMSD or the MASD are below a pre-defined threshold. If all covariates are balanced under MMSD then they are automatically balanced under MASD. The reverse argument does not hold.

Pre-defined thresholds have been reported to vary between 0.1 and 0.25 for balance assessment in the literature (Stuart et al. (2013)). For validation of the balance diagnostic, we follow a similar approach as Franklin, Rassen, Ackermann, Bartels, and Schneeweiss (2014), which creates a summary score of balance diagnostics over all covariates. Therefore, for the balance summary score, we decided to consider a threshold of 0.1 appropriate.

It is also important to assess, if the *positivity assumption* is fulfilled. The *positivity assumption* states that each treatment group has a positive probability of receiving the treatment (Austin and Stuart (2015), McCaffrey et al. (2013, 2004)). Across all imputed data sets, we therefore select the minimum and maximal weights, in order to make sure that there are no extreme values. Occurrence of extreme values could be indicative of a violation of the *positivity assumption*. Hence in cases they occur, the weights are truncated or stabilized weights are used as alternatives (Austin and Stuart (2015)).

Besides checking that the positivity assumption is fulfilled, we also checked that the selected GPS model was well specified by quantifying the standard deviation of the weights (Austin and Stuart (2015), McCaffrey et al. (2013, 2004)).

# 3 Simulation study

The aim of this simulation study is to evaluate balance metrices, MASD and MMSD, in their ability to detect imbalances in confounders after IPTW based on the GPS on multiple imputed data sets. We will compare the imbalance after weighting with the biased estimates of a continuous outcome.

**Motivation of simulated data: Clinical problem**

The simulation example for this study is motivated by observational studies investigating the associations between prenatal exposure to medications on pregnancy outcomes (Nordeng, Van Gelder, Spigset, Koren, Einarson, and Eberhard-Gran (2012), Lupattelli, Wood, Lapane, Spigset, and Nordeng (2017), Nezvalová-Henriksen, Spigset, Brandlistuen, Ystrom, Koren, and Nordeng (2016)). Such studies are necessary as clinical studies are rarely ethical among pregnant women (Blehar, Spong, Grady, Goldkind, Sahin, and Clayton (2013)). As many diseases have three or more therapeutic options, we decided to focus on a multi-group comparison of three medication alternatives. For example, hyperthyroidism during pregnancy can be treated with methimazole/carbimazole (MMI/CMZ), propylthouracil (PTU) or left untreated (Moleti, Di Mauro, Sturniolo, Russo, and Vermiglio (2019), Alexander, Pearce, Brent, Brown, Chen, Dosiou, Grobman, Laurberg, Lazarus, Mandel et al. (2017)). Although, we motivated the generation of simulation data on the above mentioned clinical example, the main focus of this manuscript is of primary methodological nature and allows no clinical implications and conclusions. The manuscript uses simulated data and has therefore no ethical issues.

## 3.1 Data generation

Data sets of sample size $n = 1000$ were generated, in order to mirror an observational study comparing three treatment options ($M = 3$), $\mathbf{T} = (T_1, T_2, T_3)$, with $T_2$ as reference treatment, on a fully observed continuous outcome $Y$ (birth weight in gram) with five measured covariates $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$. The covariates represent, respectively Body Mass Index (BMI), maternal educational level and age, marital status and parity. The covariates $X_1, X_2$ and $X_3$ were categorical, and $X_4$ and $X_5$ binary. While the variables $X_1$ and $X_2$ were partially observed, the other covariates were fully observed. Simulation details on covariates are outlined in Table 1. The data generation process is summarized in a directed acyclic graph (DAG) (see Figure 1).

Table 1: Simulation details for covariates[1]

| Covariates | Categories | Properties (%) |
|---|---|---|
| $X_1$ | BMI[2] $\leq 18$, $19-29$, $\geq 30$ | 2.0%, 82.0%, 16% |
| $X_2$ | Education in years $< 9$, 9-12, 13-16, and $> 16$ | 1.4%, 25.2%, 42.5% and 30.9% |
| $X_3$ | Age in years $\leq 24$, 25-29, 30-34 and $\geq 35$ | 45.0%, 17.0%, 30.0% and 8.0% |
| $X_4$ | Married/Cohabiting vs Other | 94.6 vs 5.4% |
| $X_5$ | Primiparity vs Multiparity | 42.3 vs 57.7% |

[1] Numbers were inspired by real world data in Frank (2019), Frank, Lupattelli, Matteson, and Nordeng (2018)

[2] BMI categories ($kg/m^2$), $19 \leq BMI \leq 24$ (normal weight) and $25 \leq BMI < 29$ (overweight) are merged into one category in this dataset

Abbreviations: $\mu$, mean, $\sigma$, standard deviation

Figure 1: Diagram of data generation

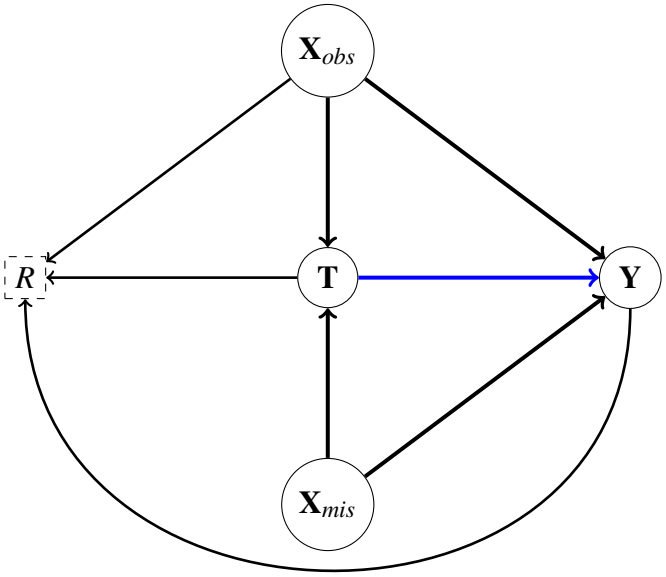



Figure 1 summarized the data generation in a directed acyclic graph (DAG). $T$ represents the treatment and $Y$ the outcome, while the blue arrow from $T$ to $Y$ represents the treatment effect. $X_{obs}$ and $X_{miss}$ represent, respectively, the completely and partially observed covariates, and $R$ the missing data indicator.

**Treatment assignment**

For each subject $i$, the probability of treatment assignment was determined from a multinomial logit function for $M > 2$ based on a linear combination of all five covariates $\mathbf{X}$ and with treatment reference $T_2 = 2$ via (Menard (2002), Borooah (2002)):

$$\ln\left(\frac{\mathbb{P}(T=m)}{\mathbb{P}(T=2)}\right) = \alpha_m + \sum_{c=1}^{C} \beta_{m,c} X_{i,c} = Z_{m,i}$$

With $M > 2$ treatment classes, there are $M-1$ predicted log-odds, one for each treatment category, with respect to the reference category. For our simulation example with $M = 3$, $Z_{1,i}$ and $Z_{3,i}$ were defined as follows:

$$\begin{aligned} Z_{1,i} &= log(0.2) + log(1.4)X_1 + log(2.0)X_2 + log(1.25)X_3 + log(1.3)X_4 + log(0.2)X_5, \\ Z_{3,i} &= log(0.5) + log(1.4)X_1 + log(1.2)X_2 + log(1.25)X_3 + log(1.3)X_4 + log(0.8)X_5 \end{aligned}$$

The coefficients in models $Z_{1,i}$ and $Z_{3,i}$ were chosen empirically, such that $p_m > 0, m = 2,..,M$ and $p_1 > 0$. Then, for $m \in \{1,3\}$, treatment assignment probabilities were calculated the following way:

$$p_m := \mathbb{P}(T=m) = \frac{\exp(Z_{m,i})}{1 + \sum_{m\in\{1,3\}} \exp(Z_{m,i})}$$

The reference group probability was calculated, as

$$\begin{aligned} p_2 := \mathbb{P}(T=2) = \frac{1}{1 + \sum_{m\in\{1,3\}} \exp(Z_{m,i})} &= 1 - \sum_{m\in\{1,3\}} p_m, \\ \text{since } p_2 + \sum_{m\in\{1,3\}} p_m &= 1. \end{aligned}$$

Treatment was assigned to all subjects by simulating random categorical treatment groups using the probabilities $p_m$, $m \in \{1,2,3\}$.

**Continuous outcome**

The continuous outcome was simulated via a linear model based on the five covariates and the treatment,

$$Y_i = \beta_T T_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \tag{8}$$

with $\varepsilon \sim \mathcal{N}(\mu = 3000, \sigma = 250)$, coefficients $\beta = (\log(2.4), \log(4), \log(1.25), \log(1.3), log(1.1))$ and with treatment effect of $\beta_T = 200$. This means that we modeled a linear mean effect of treatment. The above models define the full data situation. Next we will generate missing data with difference missing percentages in covariates the $X_1$ and $X_2$.

**Missing data mechanism**

In this simulation study, we assume that data are Missing At Random (MAR). Hence, the missing information in the variables, $X_1$ and $X_2$, depends on the fully observed covariates $X_3$ and $X_4$ (but not $X_5$, the treatment assignment variable $T$, as well as the outcome $Y$. To introduce missing information, we defined missing indicators $R$, respectively, for $X_1$ and $X_2$ via logistic models:

$$logit(p(R_1 = 0|T, X_3, Y)) = \gamma_0 + \gamma 1T + \gamma_2 X_3 + \gamma_4 Y \quad (9)$$

$$logit(p(R_2 = 0|T, X_3, X_4, Y)) = \eta_0 + \eta_1 T + \eta 2X_3 + \eta_3 X_4 + \eta_4 Y \quad (10)$$

The values were set to *NA* when the missing indicators were equal to zero. Parameters describing the missing indicators are presented in Table 2. We varied several factors *at-a-time*, in order to generate several missing percentages.

Table 2: Missing percentages and parameter values for equ. (9) and (10)

| Missing (%) | $\gamma_0, \eta_0$ | $\gamma_1, \eta_1$ | $\gamma_2, \eta_2$ | $\gamma_3, \eta_3$ | $\gamma_4, \eta_4$ |
|---|---|---|---|---|---|
| 14.9 | $\log(0.2), \log(0.3)$ | $\log(1.75), \log(1.5)$ | $\log(4.0), \log(2.0)$ | $-, \log(0.8)$ | $\log(1.0), \log(1.001)$ |
| 31.8 | $\log(0.2), \log(0.3)$ | $\log(1.25), \log(1.0)$ | $\log(3.0), \log(2.0)$ | $-, \log(0.8)$ | $\log(1.0), \log(1.001)$ |
| 57.5 | $\log(0.2), \log(0.2)$ | $\log(1.12), \log(1.0)$ | $\log(2.0), \log(1.1)$ | $-, \log(0.5)$ | $\log(1.0), \log(1.001)$ |
| 83.1 | $\log(0.2), \log(0.2)$ | $\log(1.12), \log(1.0)$ | $\log(1.1), \log(0.9)$ | $-, \log(0.5)$ | $\log(1.0), \log(1.001)$ |

Missing data percentages of 14.9%, 31.8% and 57.5% are well within empirically observed values, which vary between 2-65% (Karahalios, Baglietto, Carlin, English, and Simpson (2012), Marston, Carpenter, Walters, Morris, Nazareth, and Petersen (2010), Dong and Peng (2013)). Though the 83.1% level of missing data falls outside the empirical range, it was chosen to test ability of the methodology to accommodate the extreme values, because the literature documents evaluation of imputation techniques for 80% missing information (Lee and Huber Jr (2011)). For missing percentages larger than 10%, it is assumed that the analysis is potentially biased and imputation analysis are therefore recommended (Dong and Peng (2013)).

## 3.2 Estimates

We estimated the mean treatment effects $\hat{\beta}$, intercepts $\hat{\beta}_0$ and approximate 95% confidence intervals (CIs) on the full data set, and the imputed data sets. Over the Q imputed data sets, the effect estimates and CIs were averaged applying Rubin's rule (Murray et al. (2018), Leyrat et al. (2019b)).

## 3.3 Methods

For all missing scenarios, we imputed $Q = 10$ data sets via chained equations (using R package "MICE" from Van Buuren and Groothuis-Oudshoorn (2011)), due to the computational burden

involved when calculating the GPS (Azur et al. (2011)). The imputation model included the outcome, all covariates and the treatment (Murray et al. (2018), Frank (2019)). For each scenario, we then estimated the GPS based on all five covariates, and calculated the MMSD and MASD for each covariate separately, before and after IPTW. Then similar to the approach in Franklin et al. (2014), we created balance summary scores, that average the balance metrics over all covariates. To evaluate whether the balance metrics describe appropriately imbalance in covariates after IPTW, we plotted the estimated bias versus the balance summary score, after IPTW. Balance assessment was based on the function "bal.table" in R package "TWANG" (Burgette et al. (2017), Ridgeway, McCaffrey, Morral, Burgette, and Griffin (2017)). The GPS and IPTW were calculated based on GBM using the R function "mnps" in the R package "TWANG" (Burgette et al. (2017), Ridgeway et al. (2017)). With help of the "SURVEY" R package, IPTW was performed on each imputation set by using the "imputationList"-tool from the package "MITOOLS" (Lumley (2018, 2015)). The pooled effect estimates and CIs across the multiple imputed data sets were obtained by applying Rubin's rule via the function "MIcombine" from the "MITOOLS" R package (Lumley (2015)).

### 3.4 Performance measures

For each scenario, and the treatments, $T_{1,3}$, we estimated bias, coverage, and relative percent error in the model standard error (RelError), together with the Monte Carlo standard error (MCSE) for $\hat{\beta}$. These performance measures were calculated with help of the R package "rsimsum" (Gasparini and Lang (2018)). Based on conservative estimates from an initial small simulation run, we assumed that $\mathrm{SD}(\hat{\theta}) \leq 1676$, and that the MCSE of the bias should be lower than 52g (Morris, White, and Crowther (2019)). Using equations for the MCSE of the bias and coverage, as well as the maximized MCSE of the coverage, as provided by Morris et al. (2019), we are required to simulate $n_{sim} = 1053$ repetitive runs (see Figure S1 in Supplementary Material Section S2). The MCSE of the bias estimate of 52g is very conservative, as this was based on initial runs. It is therefore likely that the required $n_{sim}$ value may be higher in reality. Figure S1 was created in MATLAB (Version 9.4.0.813654 (R2018a)).

Given the computational burden involved in calculating the GPS (De Vries et al. (2018)), we were unable to perform analysis over all required Monte Carlo simulations ($n_{sim} = 1053$). We therefore iteratively reduced the number of $n_{sim}$, and used the number of simulated sets for which we had enough memory capacity (that was the case for $n_{sim} = 10$). These 10 sets were randomly selected out of the 1053 simulated data sets and statistical analysis was performed on them. Since, for each of the 10 selected data sets, we imputed 10 data sets for the missing data cases, we analyzed the full data set over $n_{sim} = 100$ Monte Carlo runs.

The results of the analysis and balance assessment are presented next.

## 4 Results of simulation study

For all simulation scenarios, Figure 2 shows the bias on the x-axis versus the balance summary score on the y-axis, respectively for the MASD and MMSD. For high imbalance, we would expect

Table 3: Balance summary scores

| Balance metric | Full data set | Imputed data sets[1] with missing values of | | | |
|---|---|---|---|---|---|
| | | 14.9% | 31.8% | 57.5% | 83.1% |
| Unweighted [MASD] | 0.096 | 0.108 | 0.107 | 0.113 | 0.115 |
| Weighted [MASD] | 0.035 | 0.032 | 0.037 | 0.034 | 0.039 |
| Unweighted [MMSD] | 0.122 | 0.138 | 0.141 | 0.165 | 0.149 |
| Weighted [MMSD] | 0.040 | 0.058 | 0.066 | 0.069 | 0.063 |

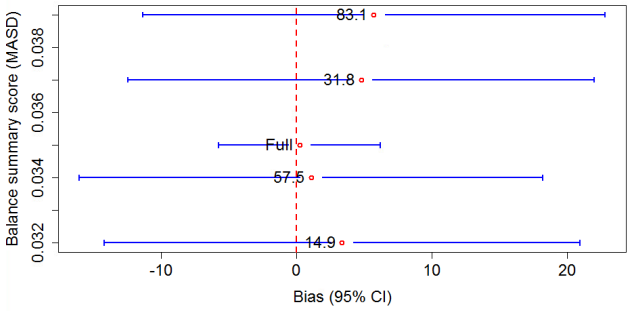
[1] The balance summary scores are the averaged MASD and MMSD over all covariate values.

Table 4: Performance measures per treatment group

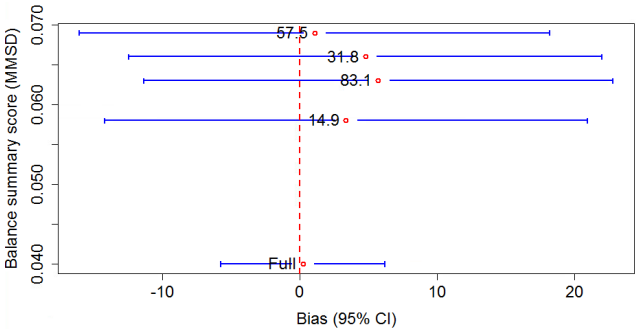
| Treatment | Full data set | Imputed data sets with missing values of | | | |
|---|---|---|---|---|---|
| $T_1$ | | 14.9% | 31.8% | 57.5% | 83.1% |
| Bias[1](MCSE) | 0.22 (3.05) | 3.35 (8.96) | 4.76 (8.79) | 1.07 (8.74) | 5.7 (8.71) |
| Coverage (MCSE) | 0.96 (0.02) | 0.90 (0.09) | 0.90 (0.09) | 0.90 (0.09) | 0.90 (0.09) |
| RelError (MCSE) | -2.32 (7.60) | 6.19 (25.69) | 8.96 (26.2) | 11.69 (27.04) | 8.71 (25.94) |
| $T_3$ | | 14.9% | 31.8% | 57.5% | 83.1% |
| Bias[1](MCSE) | -0.70 (2.75) | -11.01 (12.64) | -12.17 (13.21) | -8.97 (13.14) | -13.71 (13.52) |
| Coverage (MCSE) | 0.96 (0.02) | 0.90 (0.09) | 0.90 (0.09) | 0.90 (0.09) | 0.90 (0.09) |
| RelError (MCSE) | 9.68 (8.28) | -26.47 (17.66) | -29.45 (16.87) | -27.60 (17.40) | -31.60 (16.25) |

[1] Bias is measured in grams (g).
Abbreviations: MCSE, Monte Carlo standard error, ModSE, Model standard error, RElError, Relative % error in ModeSE

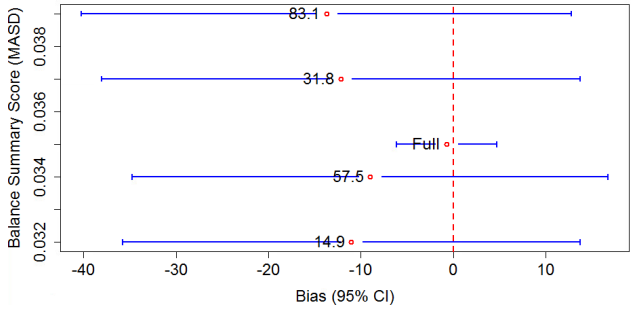
Figure 2: Bias vs Balance summary scores after IPTW.



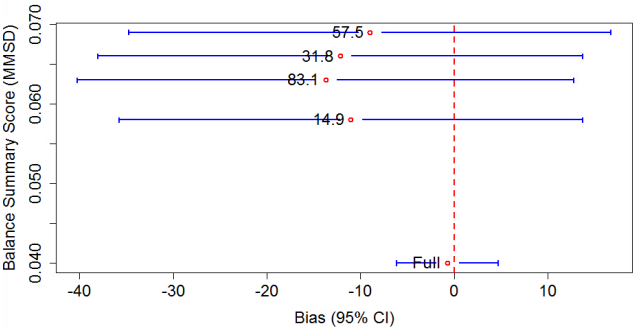

a. Bias vs Balance summary score (MASD) for $T_1$




b. Bias vs Balance summary score (MMSD) for $T_1$




c. Bias vs Balance summary score (MASD) for $T_3$




d. Bias vs Balance summary score (MMSD) for $T_3$

In the panels a-d, each line represents a missing data level setting. The line indicated with label "Full" represents the results obtained from the full data set, before missing values were introduced. Similarly, the lines notated with labels "14.9%", "31.8%", "57.5%" and "83.1%", represent results of the imputed data sets with missing values of respectively, 14.9%, 31.8%, 57.5% and 83.1%.

a high (i.e.,$\geq$ 0.1) balance summary score, and for low imbalance, a low (i.e.,$<$ 0.1) balance summary score. In Figure 2a-2d, the full data set had consistently low bias and balance summary score, for each treatment case ($T_1$ and $T_3$), when estimated over both, the MASD and MMSD. Compared to the imputed data scenarios, the confidence intervals of the full data set analysis were narrow. For the imputed data sets, we see higher bias together with larger balance summary scores (Figure 2a.-2d.), and larger variation in bias, as compared to the full data set. In total, we find a non-linear, but consistent association between the balance measures, and covariate imbalance, due to multiple imputation. For this illustrative example presented, the low balance summary scores (Table 3) reflected low imbalance in covariates. Table 3 also shows a clear reduction of balance summary scores from unweighted to the weighted sets.

The estimates of the full and imputed data sets are presented in Table S1 in Supplementary Material Section S2. Table 4 presents performance measures for all simulated scenarios. The mean coverage rate was above the 95% rate for the full data set, however slightly lower in the imputed case scenarios. For all scenarios, we see low MCSE of the coverage rate. The relative percentage error in the model standard error (Table 4) shows that the results are conditioned on adequate number of simulated Monte Carlo runs. However, even for 10 Monte Carlo simulations (due to computational burden of the GPS algorithm), we can see in Table 4 that the relative percentage error is still relatively low.

For all imputed sets, diagnostic tools, considering the distribution of the observed and imputed data, as well as convergence of the MICE algorithm for multiple imputations, indicated that the imputed values were plausible. Given that the multiple imputations seemed plausible, the GPS for each scenario was calculated and balance assessed. Furthermore, on each imputed data set, and for each missing case, the maximal standard deviation of the weights were small (i.e. $\leq 1.5$), indicating that the propensity score model was well specified (Xu, Ross, Raebel, Shetterly, Blanchette, and Smith (2010)).

# 5 Discussion

This study extended the definition of balance diagnostic to include multiple treatment exposure under multiple imputation. It has used simulations to investigate the performance of the proposed approach. Results from the analysis show that the mean differences after multiple imputation were unbiased for 14.9% to 83.1% missing data. The results were in accordance with the MASD and the MMSD balance summary scores after weighting, which indicate that covariates were balanced between groups.

The significance of this study can be appreciated when one considers the preponderance of data with missing covariates, especially in pharmacoepidemiological studies involving multiple treatment exposure (Bandoli, Kuo, Sugathan, Chambers, Rolland, and Palmsten (2018), Franklin, Shrank, Pakes, Sanfélix-Gimeno, Matlin, Brennan, and Choudhry (2013)). Specifically, the current approach allows incorporating external validation data, as suggested by Webb-Vargas, Rudolph, Lenis, Murakami, and Stuart (2017), and for multiple treatment exposure. A recent study by Frank, Lupattelli, Matteson, Meltzer, and Nordeng (2019), combined the incorporation of validation data and multiple treatment exposure after multiple imputation. For balance assessment, the authors

applied the MASD balance metric. In contrast to the present study, where the balance summary score of the metrics was calculated for validation purposes, the authors in Frank et al. (2019) assessed the balance of each covariate separately. This approach seems appropriate when the balance metrics are applied to real world data. When analyzing and interpreting balance metrics for each covariate separately, it should be taken into account that propensity score methods cannot balance covariates with small sample sizes (Rubin (1997)). Hence, when there are small sample sizes within treatment groups among a specific covariate, the balance metrics might fail to reduce balance for that covariate.

Each propensity score method (i.e., weighting, stratification, matching or adjustment), has its own merits and limitations. Our choice of IPTW in combination of GPS has been confirmed in the recent published article by Nian et al. (2019). However, the simulation results by Yang et al. (2016), identified IPTW with multiple exposures as the worst performing approach. Although, more research is needed to identify under which condition, a GPS approach performed best, such a methodological comparison exceeds the scope of this current study.

With the illustrative example presented, we could validate the proposed balance metric for multiple treatment groups under multiple imputation. However, we did not see a linear relationship (as proportional to the missing data percentage) between the summary scores and bias of imputed data sets. One explanation could be that we were not able to perform analysis on the required number of Monte Carlo runs, and therefore the result might be more prone to noise due to multiple imputations. Although, we cannot rule out some influence of chance, it should be noted that low imbalance and low balance summary scores were consistent for all presented data cases. The results were also in accordance with the data-generation process. Nevertheless, the conclusions made, based on the simulation data example presented, should be considered taking into account the drawbacks of the data generation, as well as the low number of Monte-Carlo runs. The latter limitation should however be attributed to the method applied. In the present study the MASD and MMSD diagnostics resulted in the same conclusion. It seems however that the MMSD diagnostic maintains the balance characteristics better than the MASD for small percentages of missing data (i.e, $< 30\%$). Therefore, when MASD should be applied versus MMSD needs further investigation. A recently published article considered the use of GPS under multiple imputation of missing covariates (De Vries et al. (2018)). The authors found that bias was introduced when the Classification and Regression Trees (CART) algorithm was applied directly for multiple imputation. However the authors concluded that the CART approach worked fine, when multiple imputation was applied before the estimation of the propensity score. Results presented in this paper confirm the finding.

The results reported come with some limitations. The determination of GPS and balance assessment are based on algorithms by McCaffrey et al. (2013), Burgette et al. (2017), McCaffrey et al. (2004). Therefore, limitations and strength of their approach will also hold for the present study. One such limitation is the improper confidence intervals, while a strength of the determination of GPS via boosted regression models is the relatively small weights. Another limitation, is that our proposed balance diagnostic does not assess interaction and higher order terms (De Vries et al. (2018)). Based on our results, this however did not seem to be a major problem, as long as the imputation and propensity score models are well-specified.

Furthermore, in this illustrative simulation example, we assumed that for missing data the missing at random (MAR) assumption holds. However, future studies need to investigate how balance diagnostics perform under the missing completely at random and missing not at random assumptions, as well as when the missing data process is complex, i.e. linear and non-linear dependents on more than two observed variables. In addition, our example assumed that there are no unmeasured confounders. For real data, this assumption is unrealistic. Therefore, one could evaluate the effect of bias due to unmeasured confounders with trimming methods for the IPTW, as presented in Yoshida, Solomon, Haneuse, Kim, Patorno, Tedeschi, Lyu, Franklin, Stürmer, Hernández-Díaz et al. (2018). There was not time varying confounding present in the illustrative example presented. However, if this should be the case, methods, such as proposed by Jackson (2016), Imai and Ratkovic (2015) chould be implemented in the analysis.

In this study, we choose to propose and evaluate the diagnostic based on the standardized difference, as this is most commonly used in the pharmacoepidemiology literature, although other diagnostics exist, such as the Kolmogorov-Smirnov test statistic or the C-statistic (Austin and Stuart (2015), Franklin et al. (2014)). As other balance measures (e.g., C-statistic), showed promising results in assessing imbalance in covariates (Franklin et al. (2014)), the MASD and MMSD could be alternated based on these balance measures. Such alternations however require separate validation. The different scenarios of missing data percentages investigated can be considered as a strength.

Different methods exist, which combine propensity scores and multiple imputation for a binary exposure as presented in Leyrat et al. (2019a). Such approaches should also be applied to the GPS in future studies, in order to better understand the effect of balance and missing data on causal effect estimates.

Given that, to our knowledge, no previous study has proposed or analyzed balance diagnostics for the GPS under multiple imputed data, this study is the first of its kind and will hopefully influence future research.

# 6 Conclusion

Based on simulation data, the proposed balance diagnostics seemed appropriate for balance assessment of the GPS after multiple imputation. Further research is needed to validate the results of the present study under different assumptions and conditions, and to apply the proposed diagnostics to real world data.

# References

Alexander, E. K., E. N. Pearce, G. A. Brent, R. S. Brown, H. Chen, C. Dosiou, W. A. Grobman, P. Laurberg, J. H. Lazarus, S. J. Mandel, et al. (2017): "2017 guidelines of the american thyroid association for the diagnosis and management of thyroid disease during pregnancy and the postpartum," Thyroid, 27, 315–389.

Austin, P. C. (2011): "An introduction to propensity score methods for reducing the effects of confounding in observational studies," Multivariate Behavioral Research, 46, 399–424.

Austin, P. C. (2018): "Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures," Statistical Methods in Medical Research, 0962280218756159.

Austin, P. C. and E. A. Stuart (2015): "Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies," Statistics in Medicine, 34, 3661–3679.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011): "Multiple imputation by chained equations: what is it and how does it work?" International Journal of Methods in Psychiatric Research, 20, 40–49.

Bandoli, G., G. M. Kuo, R. Sugathan, C. D. Chambers, M. Rolland, and K. Palmsten (2018): "Longitudinal trajectories of antidepressant use in pregnancy and the postnatal period," Archives of Women's Mental Health, 21, 411–419.

Bia, M. et al. (2007): "The propensity score method in public policy evaluation: a survey," POLIS Working Paper Series.

Billingsley, P. (2008): Probability and measure, John Wiley & Sons.

Blehar, M. C., C. Spong, C. Grady, S. F. Goldkind, L. Sahin, and J. A. Clayton (2013): "Enrolling pregnant women: issues in clinical research," Women's Health Issues, 23, e39–e45.

Borooah, V. K. (2002): Logit and probit: Ordered and multinomial models, 138, Sage.

Bray, B. C., J. J. Dziak, M. E. Patrick, and S. T. Lanza (2018): "Inverse propensity score weighting with a latent class exposure: estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later," Prevention Science, 1–13.

Burgette, L., B. A. Griffin, and D. McCaffrey (2017): "Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package," R package. Rand Corporation, (Accessed July 2018).

Chen, S.-H. and E. H. Ip (2015): "Behaviour of the gibbs sampler when conditional distributions are potentially incompatible," Journal of Statistical Computation and Simulation, 85, 3266–3275.

De Vries, B. B. P., M. Van Smeden, and R. H. Groenwold (2018): "Propensity score estimation using classification and regression trees in the presence of missing covariate data," Epidemiologic Methods.

Doidge, J. C. (2018): "Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random," Statistical Methods in Medical Research, 27, 352–363.

Dong, Y. and C.-Y. J. Peng (2013): "Principled missing data methods for researchers," SpringerPlus, 2, 222.

Eulenburg, C., A. Suling, P. Neuser, A. Reuss, U. Canzler, T. Fehm, A. Luyten, M. Hellriegel, L. Woelber, and S. Mahner (2016): "Propensity scoring after multiple imputation in a retrospective study on adjuvant radiation therapy in lymph-node positive vulvar cancer," PloS One, 11, e0165705.

Feng, P., X.-H. Zhou, Q.-M. Zou, M.-Y. Fan, and X.-S. Li (2012): "Generalized propensity score

for estimating the average treatment effect of multiple treatments," Statistics in Medicine, 31, 681–697.

Fong, C., C. Hazlett, and K. Imai (2018): "Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements," The Annals of Applied Statistics, 12, 156–177.

Frank, A. S., A. Lupattelli, D. S. Matteson, H. M. Meltzer, and H. Nordeng (2019): "Thyroid hormone replacement therapy patterns in pregnant women and perinatal outcomes in the offspring," Pharmacoepidemiology and Drug Safety.

Frank, A. S., A. Lupattelli, D. S. Matteson, and H. Nordeng (2018): "Maternal use of thyroid hormone replacement therapy before, during, and after pregnancy: agreement between self-report and prescription records and group-based trajectory modeling of prescription patterns," Clinical Epidemiology, 10, 1801–1816.

Frank, A.-S. J. (2019): "Thyroid hormone replacement therapy during pregnancy–quantifying medication patterns and associated outcomes in the offspring," Series of dissertations submitted to the Faculty of Mathematics and Natural Sciences, University of Oslo., 1–251, URL `http://urn.nb.no/URN:NBN:no-73653`.

Franklin, J. M., J. A. Rassen, D. Ackermann, D. B. Bartels, and S. Schneeweiss (2014): "Metrics for covariate balance in cohort studies of causal effects," Statistics in Medicine, 33, 1685–1699.

Franklin, J. M., W. H. Shrank, J. Pakes, G. Sanfélix-Gimeno, O. S. Matlin, T. A. Brennan, and N. K. Choudhry (2013): "Group-based trajectory models: a new approach to classifying and predicting long-term medication adherence," Medical Care, 51, 789–796.

Gasparini, A. and M. Lang (2018): "rsimsum: Summarise results from monte carlo simulation studies." Journal of Open Source Software, 3, 739.

Graham, J. W., A. E. Olchowski, and T. D. Gilreath (2007): "How many imputations are really needed? some practical clarifications of multiple imputation theory," Prevention Science, 8, 206–213.

Hayes, J. R. and J. I. Groner (2008): "Using multiple imputation and propensity scores to test the effect of car seats and seat belt usage on injury severity from trauma registry data," Journal of Pediatric Surgery, 43, 924–927.

Hernán, M. A., A. Alonso, R. Logan, F. Grodstein, K. B. Michels, M. J. Stampfer, W. C. Willett, J. E. Manson, and J. M. Robins (2008): "Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease," Epidemiology (Cambridge, Mass.), 19, 766–779.

Hill, J. (2004): "Reducing bias in treatment effect estimation in observational studies suffering from missing data," Report no.04-01, Columbia University, US, January 2004.

Hsu, C.-H. and M. Yu (2018): "Cox regression analysis with missing covariates via nonparametric multiple imputation," Statistical Methods in Medical Research, 0962280218772592.

Imai, K. and M. Ratkovic (2015): "Robust estimation of inverse probability weights for marginal structural models," Journal of the American Statistical Association, 110, 1013–1023.

Imai, K. and D. A. Van Dyk (2004): "Causal inference with general treatment regimes: Generalizing the propensity score," Journal of the American Statistical Association, 99, 854–866.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," Biometrika, 87, 706–710.

Jackson, J. W. (2016): "Diagnostics for confounding of time-varying and other joint exposures," Epidemiology (Cambridge, Mass.), 27, 859.

Jiang, M. and E. M. Foster (2013): "Duration of breastfeeding and childhood obesity: a generalized propensity score approach," Health Services Research, 48, 628–651.

Karahalios, A., L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson (2012): "A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures," BMC Medical Research Methodology, 12, 96.

Kupzyk, K. A. and S. J. Beal (2017): "Advanced issues in propensity scores: Longitudinal and missing data," The Journal of Early Adolescence, 37, 59–84.

Kurth, T., A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, K. Berger, and J. M. Robins (2005): "Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect," American journal of epidemiology, 163, 262–270.

Lavori, P. W., R. Dawson, and D. Shera (1995): "A multiple imputation strategy for clinical trials with truncation of patient data," Statistics in Medicine, 14, 1913–1925.

Lechner, M. (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in Econometric Evaluation of Labour Market Policies, Springer, 43–58.

Lee, J. H. and J. Huber Jr (2011): "Multiple imputation with large proportions of missing data: How much is too much?" in United Kingdom Stata Users' Group Meetings 2011, No. 23, Stata Users Group.

Leyrat, C., S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson (2019a): "Propensity score analysis with partially observed covariates: How should multiple imputation be used?" Statistical Methods in Medical Research, 28, 3–19.

Leyrat, C., S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson (2019b): "Propensity score analysis with partially observed covariates: How should multiple imputation be used?" Statistical Methods in Medical Research, 28, 3–19.

Li, F. and F. Li (2019): "Propensity score weighting for causal inference with multiple treatments," The Annals of Applied Statistics, 13, 2389–2415.

Lumley, T. (2015): "mitools: Tools for multiple imputation of missing data," `https://cran.r-project.org/web/packages/mitools/mitools.pdf`, (accessed July 2018).

Lumley, T. (2018): "survey: Analysis of complex survey samples," `http://r-survey.r-forge.r-project.org/survey/`, (accessed July 2018).

Lupattelli, A., M. Wood, K. Lapane, O. Spigset, and H. Nordeng (2017): "Risk of preeclampsia after gestational exposure to selective serotonin reuptake inhibitors and other antidepressants: A study from the norwegian mother and child cohort study," Pharmacoepidemiology and Drug Safety, 26, 1266–1276.

Malla, L., R. Perera-Salazar, E. McFadden, M. Ogero, K. Stepniewska, and M. English (2018):

"Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review," Journal of Comparative Effectiveness Research, 7, 271–279.

Marston, L., J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, and I. Petersen (2010): "Issues in multiple imputation of missing data for large general practice clinical databases," Pharmacoepidemiology and Drug Safety, 19, 618–626.

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette (2013): "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," Statistics in Medicine, 32, 3388–3414.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004): "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," Psychological Methods, 9, 403–425.

Menard, S. (2002): Applied logistic regression analysis, volume 106, Sage.

Miri, H. H., J. Hassanzadeh, A. Rajaeefard, M. Mirmohammadkhani, and K. A. Angali (2016): "Multiple imputation to correct for nonresponse bias: Application in non-communicable disease risk factors survey," Global Journal Health Science, 8, 133–158.

Mitra, R. and J. P. Reiter (2016): "A comparison of two methods of estimating propensity scores after multiple imputation," Statistical Methods in Medical Research, 25, 188–204.

Moleti, M., M. Di Mauro, G. Sturniolo, M. Russo, and F. Vermiglio (2019): "Hyperthyroidism in the pregnant woman: Maternal and fetal aspects," Journal of Clinical & Translational Endocrinology, 100190.

Morris, T. P., I. R. White, and M. J. Crowther (2019): "Using simulation studies to evaluate statistical methods," Statistics in Medicine, 38, 2074–2102.

Murray, J. S. et al. (2018): "Multiple imputation: a review of practical and theoretical findings," Statistical Science, 33, 142–159.

Nezvalová-Henriksen, K., O. Spigset, R. E. Brandlistuen, E. Ystrom, G. Koren, and H. Nordeng (2016): "Effect of prenatal selective serotonin reuptake inhibitor (ssri) exposure on birthweight and gestational age: a sibling-controlled cohort study," International Journal of Epidemiology, 45, 2018–2029.

Nguyen, T.-L., G. S. Collins, J. Spence, J.-P. Daurès, P. Devereaux, P. Landais, and Y. Le Manach (2017): "Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance," BMC Medical Research Methodology, 17, 78.

Nian, H., C. Yu, J. Ding, H. Wu, W. D. Dupont, S. Brunwasser, T. Gebretsadik, T. V. Hartert, and P. Wu (2019): "Performance evaluation of propensity score methods for estimating average treatment effects with multi-level treatments," Journal of Applied Statistics, 46, 853–873.

Nordeng, H., M. M. Van Gelder, O. Spigset, G. Koren, A. Einarson, and M. Eberhard-Gran (2012): "Pregnancy outcome after exposure to antidepressants and the role of maternal depression: results from the norwegian mother and child cohort study," Journal of Clinical Psychopharmacology, 32, 186–194.

Pandis, N. (2014): "Bias in observational studies," American Journal of Orthodontics and Dentofacial Orthopedics, 145, 542–543.

Qu, Y. and I. Lipkovich (2009): "Propensity score estimation with missing values using a multiple imputation missingness pattern (mimp) approach," Statistics in Medicine, 28, 1402–1414.

Reardon, S. F. and S. W. Raudenbush (2009): "Assumptions of value-added models for estimating school effects," Education Finance and Policy, 4, 492–519.

Ridgeway, G., D. McCaffrey, A. Morral, L. Burgette, and B. A. Griffin (2017): "twang: Toolkit for weighting and analysis of nonequivalent groups." `https://cran.r-project.org/web/packages/twang/twang.pdf`, (Accessed July 2018).

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

Rosenbaum, P. R. and D. B. Rubin (1984): "Reducing bias in observational studies using subclassification on the propensity score," Journal of the American Statistical Association, 79, 516–524.

Rubin, D. B. (1986): "Comment: Which ifs have causal answers," Journal of the American Statistical Association, 81, 961–962.

Rubin, D. B. (1997): "Estimating causal effects from large data sets using propensity scores," American Journal of Epidemiology, 127, 757–763.

Rubin, D. B. (2004): "On principles for modeling propensity scores in medical research," Pharmacoepidemiology and Drug Safety, 13, 855–857.

Spreeuwenberg, M. D., A. Bartak, M. A. Croon, J. A. Hagenaars, J. J. Busschbach, H. Andrea, J. Twisk, and T. Stijnen (2010): "The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health," Medical Care, 48, 166–174.

Stuart, E. A., B. K. Lee, and F. P. Leacy (2013): "Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research," Journal of Clinical Epidemiology, 66, S84–S90.e1.

Sugihara, M. (2010): "Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score," Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry, 9, 21–34.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011): "mice: Multivariate imputation by chained equations in r," Journal of Statistical Software, 1–68. Available at https://www.jstatsoft.org/v45/i03/. Accessed July 20, 2018, URL `https://www.jstatsoft.org/v45/i03/`.

Webb-Vargas, Y., K. E. Rudolph, D. Lenis, P. Murakami, and E. A. Stuart (2017): "An imputation-based solution to using mismeasured covariates in propensity score analysis," Statistical Methods in Medical Research, 26, 1824–1837.

Xu, S., C. Ross, M. A. Raebel, S. Shetterly, C. Blanchette, and D. Smith (2010): "Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals," Value in Health, 13, 273–277.

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola (2016): "Propensity score matching and subclassification in observational studies with multi-level treatments," Biometrics, 72, 1055–1065.

Yoshida, K., D. H. Solomon, S. Haneuse, S. C. Kim, E. Patorno, S. K. Tedeschi, H. Lyu, J. M. Franklin, T. Stürmer, S. Hernández-Díaz, et al. (2018): "Multinomial extension of propensity score trimming methods: a simulation study," American Journal of Epidemiology, 188, 609–616.

Zhu, Y., D. L. Coffman, and D. Ghosh (2015): "A boosting algorithm for estimating generalized propensity scores with continuous treatments," Journal of Causal Inference, 3, 25–40.

# Supplementary material

# S1 Proof of weak unconfoundedness and balance property for the GPS under Multiple Imputation

The proofs and notations were adapted from Leyrat et al. (2019b), Bia et al. (2007).

## Notations

We adopt the following notations, for the subject index $i = 1, ..., N$.

- Let $\mathscr{T} = \{1, 2, ..., M\}$ denote the set of M multiple treatments.
- Let $T_i$ be the treatment that subject $i$ received and the indicator that subject $i$ received treatment $m$ is

$$I_m(T_i) = I(T_i = m). \tag{11}$$

- Let $Y_i(T_i = m)$ denote the potential outcome of subject $i$, if subject $i$ has been assigned to treatment $m \in \mathscr{T}$. Only when the factual treatment happens to be $m$, then equivalence of $Y_i(T_i = m)$ and $Y_i^{obs}$ can be considered true.
- $X_i$, the vector of covariates is split into observed $X_{obs,i}$, and 'partially missing' $X_{miss,i}$ components, $X_i = (X_{obs,i}, X_{miss,i})$.
- $X_{miss,i}^q$ be the imputed value for $X_{miss,i}$ in the $q^{th}$ imputed data set ($q = 1, ..., Q$),
- $\alpha^{(q)}$ be the true propensity score (PS) parameters in the $q^{th}$ imputed data set (with $\alpha^{(q)} = \alpha$, as the overall true propensity score (PS) parameters, $q = 1, ..., Q$).

## Assumptions

We also make the following assumptions (Rosenbaum and Rubin (1983), Rubin (1986), Reardon and Raudenbush (2009), Leyrat et al. (2019b), Li and Li (2019)):

- Positivity, i.e., $0 < r(m, X_{obs,i}, X_{miss,i}^{(q)}) < 1$
- Manipulability assumption, i.e.,
  - the potential outcome of each subject $i$ is assigned to each of the M treatment groups, ensuring that each subject $i$ has at least one potential outcome per treatment group (i.e., $Y_i(1), ..., Y_i(M)$)

- Stable Unit Treatment Value Assumption (SUTVA) (a.k.a consistency assumption)
  - SUTVA assumption implies that each subject $i$ possesses one and only one potential outcome in each treatment group. Hence the observed outcome for subject $i$ ($Y_i^{obs}$) is the potential outcome corresponding to the treatment received, i.e., $Y_i(T_i = m) \equiv Y_i^{obs}(T_i = m) = Y_{i,m}^{obs}$.
- Ignorable treatment assignment (SITA) (a.k.a exchangeability assumption):
  - There are no unmeasured confounders, and
  - For full data sets, Definition S1.1 holds.
    **Definition S1.1** (Weak unconfoundedness assumption (Imbens (2000)))**.** Assignment to treatment $T_i$ is weakly unconfounded, given the pre-treatment variables $X_i$, if

$$Y_i(m) \perp T_i | X_i \tag{12}$$

- Missing At Random (MAR) assumption
  - Missing information in covariates ($X_{miss,i}$) depends on observed variables ($T_i, Y_i^{obs}, X_{obs,i}$)
- The imputation model is correctly specified.
- Temporal aspect with three stages: 1) Pre-treatment variables are measured, 2) treatment is assigned or selected, and 3) the outcomes is measured. This implies that a sequence

$$(T_1, X_1, Y_1(1), ..., Y_1(M)), ..., (T_N, X_N, Y_N(1), ..., Y_N(M))$$

  is i.i.d, so that the sequence of realized values

$$(T_1, X_1, Y_1^{obs}(1), ..., Y_1^{obs}(M)), ..., (T_N, X_N, Y_N^{obs}(1), ..., Y_N^{obs}(M))$$

  is also i.i.d.

## Basic definitions

The basic definitions are taken from Billingsley (2008) and they are applied in the proofs of subsection S1.1.

**Definition S1.2** (Bayes Theorem)**.** Suppose $E, F \subset \Omega$ with $\mathbb{P}(E), \mathbb{P}(F) > 0$. Then

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F)} \tag{13}$$

**Extension of Definition S1.2 (Ex. Def. S1.2):** If $E_1, ..., E_k$ are disjoint with $\mathbb{P}(E_i) > 0$ for $i = 1, ..., k$ and form a partition of $F \subset \Omega$, then

$$\mathbb{P}(E_i|F) = \frac{\mathbb{P}(F|E_i)\mathbb{P}(E_i)}{\sum_{j=1}^{k} \mathbb{P}(F|E_j)\mathbb{P}(E_j)} \tag{14}$$

**Definition S1.3** (Theorem of total probability)**.** Let $E_1, ..., E_k$ be a finite partition of $\Omega$, and let $F \subseteq \Omega$, then

$$\mathbb{P}(F|E_i) = \sum_{i=1}^{k} \mathbb{P}(F|E_i)\mathbb{P}(E_i) \tag{15}$$

## S1.1 Derivation of Proofs

With help of the above notations, assumptions and definitions, we show that

1. the *weak unconfoundedness* assumption holds, in each imputed data set, i.e.,
   (a) $r(m, X_{obs,i}, X^{(q)}_{miss,i}) = \mathbb{E}[I_m(T_i)|X_{obs,i}, X^{(q)}_{miss,i}]$ , and
   (b) $Y_i(m) \perp I_m(T_i)|X_{obs,i}, X^{(q)}_{miss,i}, \forall\, T_i = m,\ m \in \mathscr{T}$

   Then, we show that
2. the GPS balances covariates on the complete and imputed data sets (balance property), in each imputed data set $q$:
   (a) $X_{obs,i} \perp I_m(T_i)|r(m, X_{obs,i}, X^{(q)}_{miss,i})$,
   (b) $X^{(q)}_{miss,i} \perp I_m(T_i)|r(m, X_{obs,i}, X^{(q)}_{miss,i})$

For the following proofs, we will neglect the subject indices $i$, to enhance readability.

Given the assumptions, $Y^{obs} = Y(T)$, hence $Y^{obs}_t = Y(T = t), t \in \mathscr{T}$. Let $Y$ be a continuous outcome.

### 1 (a): $\mathbb{E}[I_m(T)|X_{obs}, X^{(q)}_{miss}] = r(m, X_{obs}, X^{(q)}_{miss}; \alpha^{(q)})$

*Proof.* Due to the extension of Definition S1.2, it holds, that

$$\begin{aligned} \mathbb{E}[I_m(T)|X_{obs}, X^{(q)}_{miss}] &= \mathbb{P}(T = m|X_{obs}, X^{(q)}_{miss}) \\ &\overset{Ex.Def.1.2}{=} \frac{\mathbb{P}(X^{(q)}_{miss}|T = m, X_{obs})\mathbb{P}(T = m|X_{obs})}{\sum_{t=1}^{M} \mathbb{P}(X^{(q)}_{miss}|T = t, X_{obs})\mathbb{P}(T = t|X_{obs})}. \end{aligned} \tag{16}$$

Next express $\mathbb{P}(X^{(q)}_{miss}|T = t, X_{obs})$ in terms of (unobserved) missing values $X_{miss}$:

$$\begin{aligned} \mathbb{P}(X^{(q)}_{miss}|T = t, X_{obs}) &\overset{Def.1.3}{=} \int_{\mathbb{R}} f_{X^q_{miss}|T=t, X_{Obs}=x, Y^{obs}_t=y}(X^q_{miss} = x|t, x, y) \cdot f_{Y^{obs}_t|T=t, X_{obs}=x}(y|t, x)dy \\ &\overset{MAR}{=} \int_{\mathbb{R}} f_{X_{miss}|T=t, X_{Obs}=x, Y^{obs}_t=y}(X_{miss} = x|t, x, y) \cdot f_{Y^{obs}_t|T=t, X_{obs}=x}(y|t, x)dy \\ &= \mathbb{P}(X_{miss}|T = t, X_{obs}), \end{aligned} \tag{17}$$

if values are imputed from the true distribution. Substituting this back into equ (16), yields

$$\begin{aligned}
\mathbb{E}[I_m(T)|X_{obs}, X_{miss}^{(q)} = x] &= \mathbb{P}(T = m|X_{obs}, X_{miss}^{(q)} = x) \\
&= \frac{\mathbb{P}(X_{miss}^{(q)} = x|T = m, X_{obs})\mathbb{P}(T = m|X_{obs})}{\sum_{t=1}^{M}\mathbb{P}(X_{miss}^{(q)} = x|T = t, X_{obs})\mathbb{P}(T = t|X_{obs})} \\
&\overset{Subst.\ eq.\ (17)}{=} \frac{\mathbb{P}(X_{miss} = x|T = m, X_{obs})\mathbb{P}(T = m|X_{obs})}{\sum_{t=1}^{M}\mathbb{P}(X_{miss} = x|T = t, X_{obs})\mathbb{P}(T = t|X_{obs})} \\
&\overset{MI\ b.o.\ \alpha}{=} \mathbb{P}(T = m|X_{obs}, X_{miss} = x; \alpha) \\
&\overset{Def.\ PS}{=} r(m, X_{obs}, x; \alpha) \\
&\overset{\alpha^{(q)}}{=} r(m, X_{obs}, X_{miss}^{(q)} = x; \alpha^{(q)}).
\end{aligned} \tag{18}$$

□

Equality 1(a) shows that $r(m, X_{obs}, X_{miss}^{(q)})$ is a generalized propensity score in each imputed data set.

## 1 (b): $Y_m^{obs} \perp I_m(T)|X_{obs}, X_{miss}^{(q)}$

*Proof.*

$$\begin{aligned}
\mathbb{P}(T = m|Y_m^{obs} = y, X_{obs}, X_{miss}^{(q)}) &\overset{Bayes\ thm.}{=} \frac{\mathbb{P}(X_{miss}^{(q)}|T = m, Y_m^{obs} = y, X_{obs})\mathbb{P}(T = m|Y_m^{obs} = y, X_{obs})}{\mathbb{P}(X_{miss}^{(q)}|Y_m^{obs} = y, X_{obs})} \\
&\overset{MAR}{=} \frac{\mathbb{P}(X_{miss}|T = m, Y_m^{obs} = y, X_{obs})\mathbb{P}(T = m|Y_m^{obs} = y, X_{obs})}{\mathbb{P}(X_{miss}|Y_m^{obs} = y, X_{obs})} \\
&\overset{Y_m=Y\ when\ T=m}{=} \frac{\mathbb{P}(X_{miss}|T = m, Y_m^{obs} = y, X_{obs})\mathbb{P}(T = m|Y_m^{obs} = y, X_{obs})}{\mathbb{P}(X_{miss}|Y_m = y, X_{obs})} \\
&= \mathbb{P}(T = m|Y_m^{obs} = y, X_{obs}, X_{miss}) \\
&\overset{Y_m^{obs}\perp I_m(T)|X}{=} \mathbb{P}(T = m|X_{obs}, X_{miss}) \\
&\overset{eq.\ (17)}{=} \mathbb{P}(T = m|X_{obs}, X_{miss}^{(q)})
\end{aligned} \tag{19}$$

□

1(b) shows independence of the potential outcome and the treatment at the level of interest, given observed and imputed covariate values.

Together, the proofs of 1(a) and 1(b) show that the *weak unconfoundedness* assumption holds for imputed data.

Next it needs to be shown that the true PS, $r(m, X_{obs}, X_{miss}^{(q)})$, is a balancing score in each imputed data set.

## 2 (a): $X_{obs} \perp I_m(T) | r(m, X_{obs}, X^{(q)}_{miss})$

*Proof.* Needs to be shown that

$$\mathbb{P}(T = m, X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) = \mathbb{P}(T = m | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})).$$

It holds that

$$\begin{aligned} \mathbb{P}(T = m, X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) &= \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(T = m | X_{obs}, r(m, X_{obs}, X^{(q)}_{miss})) \\ &= \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(T = m | X_{obs}), \end{aligned} \tag{20}$$

because $r(m, X_{obs}, X^{(q)}_{miss})$ is a function of $X_{obs}$, conditioning on $X_{obs}$ is equivalent to conditioning on $X_{obs} r(m, X_{obs}, X^{(q)}_{miss})$. Rewrite

$$\begin{aligned} \mathbb{P}(T = m | X_{obs}) &= \mathbb{P}(T = m | X_{obs}, Y^{obs}_m = y) \mathbb{P}(Y^{obs}_m = y | X_{obs}) \\ &\overset{Y^{obs}_m \perp I_m(T) | X_{obs}, X^{(q)}_{miss}}{=} \mathbb{P}(T = m | X_{obs}, X^{(q)}_{miss}, Y^{obs}_m = y) \mathbb{P}(Y^{obs}_m = y | X_{obs}, X^{(q)}_{miss}) \\ &= \mathbb{P}(T = m | X_{obs}, X^{(q)}_{miss}) \end{aligned} \tag{21}$$

Substituting eq. (22) in eq. (21), yields

$$\mathbb{P}(T = m, X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) = \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(T = m | X_{obs}, X^{(q)}_{miss}). \tag{22}$$

It was previously shown that $\mathbb{P}(T = m | X_{obs}, X^{(q)}_{miss}) = \mathbb{E}[I_m(T) | X_{obs}, X^{(q)}_{miss}] = r(m, X_{obs}, X^{(q)}_{miss})$. Then

$$\begin{aligned} \mathbb{P}(T = m | r(m, X_{obs}, X^{(q)}_{miss})) &= \mathbb{E}[I_m(T) | r(m, X_{obs}, X^{(q)}_{miss})] \\ &= \mathbb{E}[\mathbb{E}[I_m(T) | X_{obs}, X^{(q)}_{miss}] | r(m, X_{obs}, X^{(q)}_{miss})] \\ &= \mathbb{E}[r(m, X_{obs}, X^{(q)}_{miss}) | r(m, X_{obs}, X^{(q)}_{miss})] \\ &= r(m, X_{obs}, X^{(q)}_{miss}). \end{aligned} \tag{23}$$

Hence, eq. (21) becomes

$$\begin{aligned} \mathbb{P}(T = m, X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) &= \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(T = m | X_{obs}, X^{(q)}_{miss}) \\ &\overset{Def\ PS}{=} \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) r(m, X_{obs}, X^{(q)}_{miss}) \\ &\overset{eq.(23)}{=} \mathbb{P}(X_{obs} | r(m, X_{obs}, X^{(q)}_{miss})) \mathbb{P}(T = m | r(m, X_{obs}, X^{(q)}_{miss})). \end{aligned} \tag{24}$$

□

**2 (b)** $X^{(q)}_{miss} \perp I_m(T) | r(m, X_{obs}, X^{(q)}_{miss})$, can be shown in a similar manner as **2 (a)**.

Independences 2 (a) and (b) show that in each imputed data set, the propensity score balances respectively the observed and imputed values of the covariates across treatment groups.

# S2 Additional information

Figure S1: Simulation run plots of performance measures

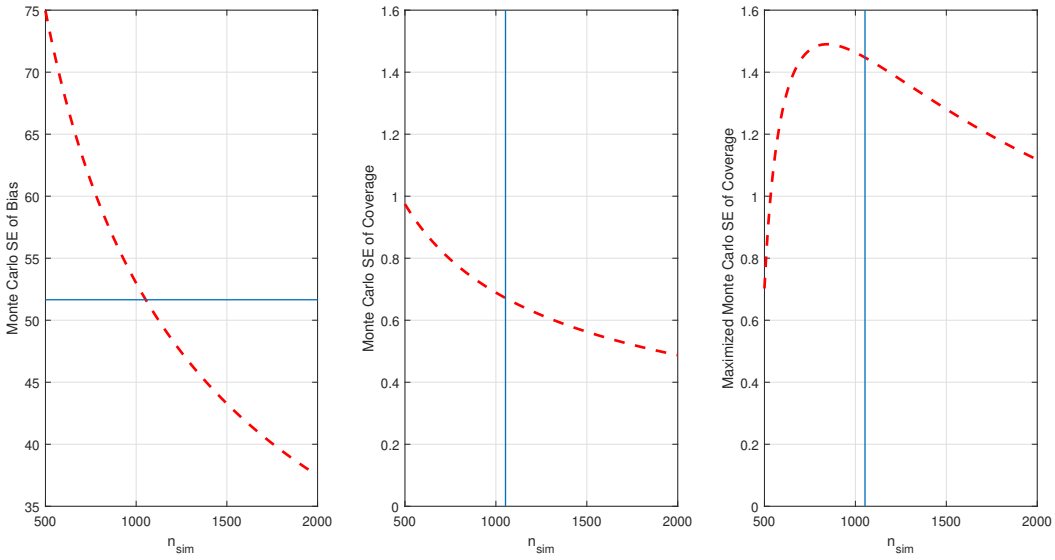



Figure S1 plots the estimated MCSE of the bias and coverage, as well as the maximized MCSE of the coverage, against the number of repetitive simulation runs, $n_{sim}$. Given that bias is the key measure of performance, we choose $n_{sim} = 1053$, such that the MCSE of the bias estimate is below 52, which is indicated by the blue vertical line in the left graph. In the middle and right graph, the blue horizontal lines indicate $n_{sim} = 1053$.

Table S1: Estimates of full data and imputed data sets

| Treatment groups | Full data set | Imputed data sets with missing values of | | | |
|---|---|---|---|---|---|
| | | 14.9% | 31.8% | 57.5% | 83.1% |
| | $\hat{\beta}$, 95% CI | $\hat{\beta}$, 95% CI | $\hat{\beta}$, 95% CI | $\hat{\beta}$, 95% CI | $\hat{\beta}$, 95% CI |
| $T_1$ | -205 (-262,-149) | -207 (-263,-150) | -205 (-261,-148) | -204 (-261,-146) | -201 (-257,-145) |
| $T_2$[1]($\hat{\beta}_0$) | 3413 (3363,3463) | 3418 (3367,3468) | 3416 (3365,3466) | 3416 (3364,3467) | 3413 (3362,3463) |
| $T_3$ | 194 (137,251) | 185 (128,243) | 188 (129,246) | 188 (129,247) | 190 (133,248) |

[1] Treatment group $T_2$ is the reference group, with respect to which the other groups are compared to. Its estimate is the intercept $\hat{\beta}_0$.

Abbreviations: CI, confidence intervals, IPTW, inverse probability of treatment weighting