Tackling non-ignorable dropout in the presence of time varying confounding

Tackling non-ignorable dropout in the presence of time-varying confounding

Marco Doretti†

University of Perugia - Department of Economics, Perugia, Italy

Sara Geneletti

London School of Economics - Department of Statistics, London, United Kingdom

Elena Stanghellini

University of Perugia - Department of Economics, Perugia, Italy

Abstract. In this paper we explore the sensitivity of time-varying confounding adjusted estimates to different dropout mechanisms. We extend the Heckman correction to two time points and explore selection models to investigate situations where the dropout process is driven by unobserved variables and the outcome respectively. The analysis is embedded in the Bayesian framework which provides a number of advantages. These include fitting a hierarchical structure to processes that repeat over time and avoiding exclusion restrictions in the case of the Heckman correction. We adopt the Decision Theoretic approach to causal inference which makes explicit the *No regime dropout dependence* (NRD) assumption. We apply our methods to data from the Counterweight Programme pilot, a UK protocol to address obesity in primary care. A simulation study is also implemented.

Keywords: Causal inference, Heckman correction, non-ignorable dropout, selection models, time-varying confounding.

1. Introduction

We are often interested in evaluating the causal effect of a treatment strategy implemented over successive time periods on a final response. This is the case in the Counterweight Programme pilot (Laws et al., 2004) (henceforth CWP), a UK based study aimed at evaluating the effect of different lifestyle interventions administered over time on weight loss. Lifestyle changes such as dieting and exercise have been linked to weight loss (Curioni

†*Address for correspondence:* Marco Doretti, University of Perugia - Department of Economics, Via A. Pascoli 1, 06123 Perugia. Email: doretti@stat.unipg.it

and Lourenco, 2005) and to a lesser extent so has dietary counselling (Dansinger et al., 2007). The CWP combined both dietary counselling and exercises and found that 43% of patients that participated in the study for 12 months (termed compliers in the paper) lost on average 5% of their body weight (Laws et al., 2004). However, no adjustment was made for the participants who dropped out. CWP suffered from high levels of dropout in excess of 50% at every measurement occasion. Further, as the dropout was likely to be non-ignorable, standard methods (Daniel et al., 2013) could potentially lead to biased effect estimates. The focus in this paper is on describing a method to assess how sensitive the effect estimates are to modelling assumptions that encode different dropout generating mechanisms.

When, as in our case, data involve *time-varying confounders* a form of recursive standardisation termed the g-computation formula or algorithm (Robins, 1986; Daniel et al., 2013) is often implemented. Examples include the effect of anti-retroviral medication on CD4 counts in HIV positive patients (Arjas and Saarela, 2010) and the effect of antiglycaemic drugs on blood glucose level for patients affected by type II diabetes (Daniel et al., 2013). While most of the literature in this area (with notable exceptions (Scharfstein et al., 1999; Rotnitzky et al., 1998)) assumes that when there is dropout over time, this is unrelated with either the outcome or any other unobserved variables conditional on the observed covariates (termed *missing at random* (MAR)), our paper looks at a study where MAR assumptions are not tenable in a similar spirit as Washbrook et al. (2014).

Specifically we use directed acyclic graphs (DAGs) to describe three mechanisms that could be responsible for dropout in the CWP. We then link the DAGs to models for handling non-ignorable dropout as characterised by Little and Rubin (2002). The dropout mechanisms we consider are as follows: a) MAR holds, b) dropout depends on unobserved factors and c) dropout is outcome dependent. These structures are easily described by DAGs and naturally lead to a wavewise complete case (WCC) analysis, a Heckman correction (Heckman, 1979) (HC) and a selection model (SM) approach respectively (Hogan et al., 2004). We plug our dropout adjusted equations into the g-formula in order to deal with both non-ignorable dropout and time-varying confounding. By comparing results from different models and simulation studies we can assess the sensitivity of estimates to the structural assumptions embodied in the DAGs.

Our sensitivity analyses revealed that provided the assumptions we made were correct and the models were not misspecified, the WCC and HC models gave similar results with patients losing at least 4% of their BMI regardless of the treatment strategy followed while

the SM analysis led to a lower 1.5% BMI loss. As the SM is the natural analysis if the dropout is outcome dependent, this fits with the story that only individuals who were losing weight remained in the study and explained the inflated results of the wavewise complete case analysis.

Our analysis is embedded in the Bayesian paradigm which is becoming more common in sensitivity analyses (Greenland, 2009; Geneletti et al., 2013) and has been implemented in the context of causal analyses of longitudinal data (Arjas and Parner, 2004; Arjas and Saarela, 2010). Furthermore, in the current context the Bayesian approach means exclusion restrictions (strong untestable assumptions needed to ensure model identification) can be avoided in the implementation of the Heckman correction (Puhani, 2000). This is an advantage for us as there are no clear exclusion restrictions in our application.

Further we adopt the Decision Theoretic approach to causal inference (Dawid and Didelez, 2010; Dawid and Constantinou, 2013). This allows us to state that there is *No regime dropout dependence* (NRD), which makes explicit the idea that dropout is independent of whether the study is observational or experimental conditional on subjects' personal information. As a consequence we can in principle make causal inference from these data even in the presence of dropout.

The paper is arranged as follows: Section 2 introduces the Counterweight Programme pilot, our substantive application. In Section 3 we set the notation and basic concepts about DAGs. Moreover, in this section we introduce the Decision Theoretic framework describing the standard approach to estimating causal effects of treatment regimes in the absence of missing data. Section 4 describes the dropout mechanisms we propose and the assumptions required to make inference about treatment strategies when dropout is present. We apply our sensitivity analysis method to the real data in Section 5. A simulation is described in Section 6. We discuss advantages and drawbacks of our approach and make concluding remarks in Section 7.

2. The Counterweight Programme pilot

We now introduce the application in this paper and embed in it further methodological issues. The Counterweight Programme pilot was a UK based non-randomized study designed to assess a range of primary care interventions to tackle obesity in general practice. The data we have cover the years 2001-2005. The aim was to evaluate whether a sequence of four treatments resulted in a reduction of the body mass index (BMI) of clinically overweight (BMI>25) and obese (BMI>30) patients by at least 5%. The body mass index of

	Occasion		
Variable	Baseline ($t = 0$)	Second ($t = 1$)	Third ($t = 2$)
BMI	37.03	36.94	37.11
ΔBMI (%)	–	-3.43	-4.85
Age	49.05	52.27	53.39
Gender (1=male)	0.23	0.27	0.27
Soft treatment	1080	450	–
Hard treatment	766	333	–
Total (dropout%)	1846(–)	783 (58%)	457 (42%)

Table 1: Table of mean values of the explanatory and outcome variables for patients who remained in the study as well as the numbers remaining and percentage dropout at baseline and the following two measurement occasions.

an individual is defined as their weight divided by the square of their height and therefore is measured in kg/m^2. Due to mismatches in the protocol implementation as well as very high dropout rates (over 70%) in the final two occasions, we only considered the first three measurements (the baseline period and the next two). Sample sizes at each measurement occasion and dropout rates are shown in the last line of Table 1 where it is evident that dropout is a very serious concern with rates of 58 and 42% between the baseline and second and the second and third measurement respectively.

When a patient entered the study, several indicators of their clinical status and lifestyle were recorded. These included sex, age, depression scores, history of heart conditions and diabetes as well as smoking, alcohol consumption and physical activity. Height and weight were also collected and from these the BMI was calculated. These variables form the baseline set. In our analysis we only included the most relevant (age, gender and BMI) as an analysis using the extended set of variables listed above did not lead to substantially different results. Mean values taken by these variables at each measurement occasion can be seen in Table 1. The average BMI at baseline was high at 37.03 which was in accordance with the study protocol which aimed at recruiting patients who were severely overweight. Notice that at the third occasion ($t = 2$) the mean value of BMI is slightly higher than the baseline value (37.11 kg/m^2) although the average percentage change in BMI with respect to the baseline is -4.85%. The men formed approximately 25% of the sample througout.

This is less than seen in Hospital Episode Statistics where men made up approximately 40% of admissions with a primary or secondary diagnosis of obesity (Eastwood, 2012) between 2001 and 2004. The average age at baseline was approximately 49. The age of patients who remained in the study increased slightly over the course of the study to 53.

After an initial meeting with the practice staff a treatment was assigned and further meetings were scheduled at three month intervals for the second and third, longer for subsequent meetings. At every subsequent meeting the observed percentage change in BMI was determined and a new treatment based on the set of baseline variables and the change in BMI assigned. The variable of causal interest is the final percentage change in BMI. For some individuals additional measurements (for example blood pressure) were taken after the baseline measurement. It is in principle possible that these measurements were considered when GP staff assigned new treatments. However only a small percentage of patients had these data and even for these patients measurements were not taken consistently. As a consequence we did not take them into account in our analysis. The models we propose in Section 4 extend when such measurements are available.

Emphasis in CWP was on lifestyle interventions rather than drug therapies. There were seven possible treatments and we chose (somewhat arbitrarily) to compare the effect of 'hard' (h) lifestyle changes – gym and diet – versus 'soft' (s) actions like goal setting and group meetings. We were therefore able to frame the problem in terms of a sequence of *binary* treatments. The number of patients administered each of these treatments at each time point is shown in Table 1. More soft treatments than hard treatments were administered at both time points.

The targets of inference in this context are the effects of the four possible static strategies (treatment plans) $\{(s,s),(s,h),(h,s),(h,h)\}$ and one dynamic strategy: "apply the hard treatment until a 5% loss in BMI is achieved" which we denote by (d). Under (d) the hard treatment is assigned to everyone at the baseline occasion and only to those who did not manage to lose at least 5% of their BMI at the second. From these effects we can determine if there was an overall effect for any strategy and whether some strategies were better than others. Notice that without the aggregation of the seven original treatments there would be a substantial number of strategies and their comparison would become cumbersome. Furthermore, as some of these treatments are rarely assigned, the estimates of their effects would likely be unstable.

In many applications, and specifically in our case, at every occasion (termed *wave* in Washbrook et al. (2014)) the sub-sample of complete cases is likely to be systematically

different from the one including those who dropped out of the study. This means that the MAR assumption is violated and an analysis conducted on the complete cases of every occasion (wavewise complete cases) without further adjustments is likely to produce biased results. We consider in our paper two dropout mechanisms that potentially lead to biased results in a wavewise complete case analysis. The first comes about when patients decide whether to attend a session based on personal characteristics that are unobserved by the general practice (GP) staff who administer the treatments. For example a patient might have a history of unsuccessful dieting which makes them demotivated and more likely to dropout. This variable is not recorded and is not known by the practice staff. A second plausible mechanism is where some patients drop out if they do not manage to lose a sufficient amount of weight. In both cases the dropout is termed *non-ignorable* (Little and Rubin, 2002) because it is associated with the outcome.

3. Background

We embed our subsequent discussions and analyses in the *Decision Theoretic* (DT) framework for causal inference. We offer a somewhat simplified description here, for a complete account of the formal details see Dawid and Didelez (2010); Dawid and Constantinou (2013). An analogous set-up based on potential responses can be found in Robins (1986); Daniel et al. (2013) and citations therein.

Fundamental to the DT framework is the concept of conditional independence. We say that two variables A and B are independent conditional on another variable C when $p(A, B|C) = p(A|C)p(B|C)$ and we write $A \perp\!\!\!\perp B|C$ (Dawid, 1979). Directed acyclic graphs (DAGs) are used to formally encode conditional independences via the moralisation criterion (Lauritzen, 1996). See Section 1 of the supporting materials for a brief overview of moralisation. While moralisation is necessary to derive conditional independences from DAGs it is not essential to understand the power of DAGs to visualise relationships between variables. For the purposes of this paper it is sufficient to view the DAGs as *influence diagrams* (Dawid, 2002) with directed edges representing influence.

3.1. Dynamic Treatment regimes

We are interested in evaluating the effect of a sequence of interventions over successive periods of time indexed by $t = 0, 1, \ldots, T + 1$. At each time point t we can record two types of information: the sequence of observed covariates (typically multi-valued) $(\boldsymbol{L}_0, \boldsymbol{L}_1, \ldots, \boldsymbol{L}_t) = \bar{\boldsymbol{L}}_t$ and the sequence of actions $(A_0, A_1, \ldots, A_t) = \bar{\boldsymbol{A}}_t$ taken. We drop

the individual index for simplicity. As is common in observational data, there are potential confounders $(U_0, U_1, \ldots, U_t) = \bar{\boldsymbol{U}}_t$ which we do not observe. Following convention, the collection $(\bar{\boldsymbol{L}}_t, \bar{\boldsymbol{A}}_t, \bar{\boldsymbol{U}}_t)$ is termed the *partial history* and $(\bar{\boldsymbol{l}}_t, \bar{\boldsymbol{a}}_t, \bar{\boldsymbol{u}}_t)$ is a realisation of this partial history. Note that we often refer to the observed partial history $(\bar{\boldsymbol{L}}_t, \bar{\boldsymbol{A}}_t) = \bar{\boldsymbol{X}}_t$.

In many settings, and in our motivating example in particular, *baseline* variables (the clinical and lifestyle indicators as well as initial BMI) are collected at the beginning of the study ($t = 0$) and are denoted by $\boldsymbol{L}_0$. During successive periods a single variable V_t (percentage change in BMI) is recorded. As baseline information is likely to play a role at advanced stages of the study, without loss of generality we can set $\boldsymbol{L}_t = (\boldsymbol{L}_0, V_t)$. At each occasion t therefore, some information $\boldsymbol{L}_t$ is collected and used to assign a binary treatment A_t (soft or hard lifestyle interventions). A single outcome Y (total percentage change in BMI) is measured only at the final period: in many applications, including our own, this will coincide with the V_t measured at the final point, *i.e.* $Y = V_{T+1}$.

In the DT framework causality is explicitly dealt with by introducing decision (non-random) variables termed regime indicators or simply *regimes* (Dawid, 2002). Specifically we define σ to be the regime indicator taking on values $\sigma = \{o, \mathcal{S}^\star\}$ where o is the observational regime and $\mathcal{S}^\star$ is a set of interventional *strategies* (Dawid and Didelez, 2010). Therefore $\sigma = o$ means that the data are observational while $\sigma = e$ with $e \in \mathcal{S}^\star$ means the data arise under a particular experimental setting. In practice a strategy e is a decision algorithm that determines, based on a partial history, the value of the next action. A strategy can be *static* or *dynamic*. The former is when each patient is administered the same sequence of treatments irrespective of the value of their partial history. The latter is when the next treatment is some (potentially probabilistic) known function of the value of the individual observed partial history. Thus for example a patient might be administered the hard treatment if they have lost no weight and the soft treatment if they have lost weight. We note that in the potential responses literature the term regimes is often used interchangeably with strategies. A necessary assumption in DT which formalises the distinction between regimes states that $e \in S^\star$ are *control strategies* (Dawid and Didelez, 2010). This means that when actions $\bar{\boldsymbol{a}}_T$ are set by intervention within an experiment ($\sigma = e$) their value depends **only** on the strategy e. This is in contrast to observational data ($\sigma = o$), where actions potentially depend probabilistically on both observed and unobserved covariates in an unknown fashion. The strategies described in Section 2 for our motivational example are control strategies.

Given that our target is the *causal* effect of a number of treatment strategies we would

ideally like to perform experiments representing static and dynamic strategies of interest. Such experiments would enable us to obtain unconfounded estimates of the effects of the aforementioned strategies of the form $E(Y|\bar{\boldsymbol{a}}_T; e)$ (termed $E(Y|do(\bar{\boldsymbol{a}}_T))$ by Pearl and Robins (1995)). In our context this would be the expected percentage change in BMI for the experimental strategy e. It is usually not possible to perform the necessary experiments and thus the data at our disposal are typically (and specifically in our case where the treatment assignment was not randomised) observational. Without further assumptions we can at best estimate $E(Y|\bar{\boldsymbol{a}}_T; o)$ (termed $E(Y|\bar{\boldsymbol{a}}_T)$ by Pearl and Robins (1995)) provided we have indeed observed the particular sequence of actions $\bar{\boldsymbol{a}}_T$. As there is no guarantee that $E(Y|\bar{\boldsymbol{a}}_T; e)$ and $E(Y|\bar{\boldsymbol{a}}_T; o)$ are going to be the same due to the presence of confounding and no way, other than performing all the experiments, to test this, we need to make some assumptions that allow us to relate the two quantities. These assumptions are most easily expressed using conditional independence statements.

3.1.1. Assumptions

The first assumption we make is that of *Extended stability* (ES):

$$(U_t, \boldsymbol{L}_t) \perp\!\!\!\perp \sigma|(\bar{\boldsymbol{L}}_{t-1}, \bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{U}}_{t-1}) \quad t = 0, \ldots, T+1. \tag{1}$$

In words this assumption states that conditional on all the past, both observed and unobserved, the current values of U_t and $\boldsymbol{L}_t$ do not depend on how the data were generated, whether from an experiment or from an observational study. In our context this translates to assuming that, conditional on the past, the values of the current level of motivation or weight loss do not depend on whether the study is experimental (with randomised treatments) or observational. While ES is a plausible (if untestable) assumption, it does not directly help us to estimate the target quantity $E(Y|\bar{\boldsymbol{a}}_T; e)$ as it involves the unobserved potential confounders $\bar{\boldsymbol{U}}_t$. In order to make some headway, we must make an additional assumption. In many contexts, and indeed for our motivating example, it makes sense to assume

$$A_t \perp\!\!\!\perp \bar{\boldsymbol{U}}_t|(\bar{\boldsymbol{L}}_t, \bar{\boldsymbol{A}}_{t-1}; \sigma) \quad t = 0, \ldots, T. \tag{2}$$

This assumption is akin to the conditional exchangeability or no unobserved confounders assumption in the potential responses literature (Daniel et al., 2013). In words, assumption (2) states that if we know the values of past observables and past actions taken, then the present action is independent of present and past unobserved factors. The reason this assumption makes sense for our application is that the GPs and nurses assigning

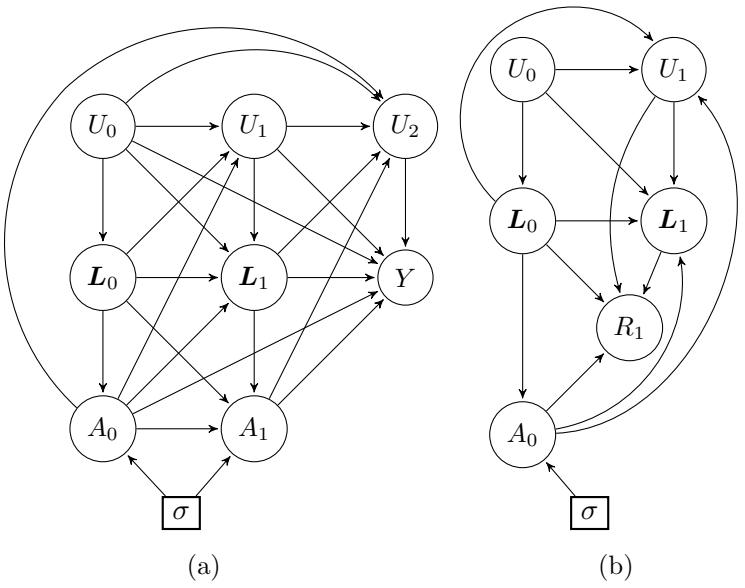

Figure 1: DAG (a) represents sequential randomisation. DAG (b) represents the NRD as well as SR.

the weight loss interventions have at their disposal a large number of health and lifestyle variables and thus unobserved variables are unlikely to enter into the treatment decision, even if they influence the outcome. The combination of ES and assumption (2) has been termed *Sequential randomization* (SR) by Dawid and Didelez (2010). Using the moralisation criterion it is easy to see that both DAGs in Figure 1 embody SR. DAG (a) shows the situation for three measurement occasions whereas DAG (b) refers to only the first two. We include DAG (b) for comparison with later DAGs in Figure 2 which describe the dropout mechanisms we consider.

In addition to SR we must make a further assumption: positivity. In broad terms this requires that all the strategies we want to estimate in the experimental setting are also observed in the observational regime. For details on all the assumptions and a formal treatment of dynamic treatment regimes in the DT framework see Dawid and Didelez (2010); Dawid and Constantinou (2013).

3.2. Time-varying confounding

Another problem we face in the context of evaluating the effect of treatment strategies is that $E(Y|\bar{\boldsymbol{a}}_T; e)$ cannot be written as a single regression equation due to the problem of *time-varying* confounding, specifically $\boldsymbol{L}_t$ or a component of it such as V_t is a time-varying confounder. This is best explained by looking at DAG (a) in Figure 1: the node

$\boldsymbol{L}_1$ (more precisely its component V_1) plays a double role, being a confounder for the pair (A_1, Y) and an intermediate variable on the pathway from A_0 to Y (termed a mediator for the pair (A_0, Y)). On the one hand, fitting a regression of Y upon $(\boldsymbol{L}_0, A_0, A_1)$ only (*i.e.* excluding V_1) yields a confounded effect for A_1 (in causal DAG terminology: the back-door path $A_1 \leftarrow \boldsymbol{L}_1 \rightarrow Y$ is not blocked). On the other hand including V_1 in the regression model blocks the causal (front-door) path $A_0 \rightarrow \boldsymbol{L}_1 \rightarrow Y$ and induces a spurious marginal association between the pairs (A_0, U_0) and (A_0, U_1) which also contributes to the non-identification of the target causal effect. This is known as *selection* or *collider stratification* bias and comes about by conditioning on a common child (Geneletti et al., 2009; Daniel et al., 2013). In our example, A_0, U_0 and U_1 are parents of $\boldsymbol{L}_1$ thus conditioning on it generates the spurious association between the two pairs (A_0, U_0) and (A_0, U_1).

3.3. The g-computation algorithm

One solution to this problem developed by Robins (1986) is to use a recursive approach. If SR holds the target quantity for a continuous outcome can be written as follows:

$$\begin{aligned} E(Y|\bar{\boldsymbol{a}}_T; e) = & \int_{\bar{\boldsymbol{l}}_t \in \bar{\boldsymbol{\mathcal{L}}}_T} \big[E(Y|\bar{\boldsymbol{A}}_T = \bar{\boldsymbol{a}}_T, \bar{\boldsymbol{L}}_T = \bar{\boldsymbol{l}}_T; e) \\ & \times \prod_{t=0}^{T} f_{\boldsymbol{L}_t|\bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{L}}_{t-1}}(\boldsymbol{l}_t|\bar{\boldsymbol{a}}_{t-1}, \bar{\boldsymbol{l}}_{t-1}; e) \, \mathrm{d}\, \bar{\boldsymbol{l}}_T \, \big] \\ = & \int_{\bar{\boldsymbol{l}}_t \in \bar{\boldsymbol{\mathcal{L}}}_T} \big[E(Y|\bar{\boldsymbol{A}}_T = \bar{\boldsymbol{a}}_T, \bar{\boldsymbol{L}}_T = \bar{\boldsymbol{l}}_T; o) \\ & \times \prod_{t=0}^{T} f_{\boldsymbol{L}_t|\bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{L}}_{t-1}}(\boldsymbol{l}_t|\bar{\boldsymbol{a}}_{t-1}, \bar{\boldsymbol{l}}_{t-1}; o) \, \mathrm{d}\, \bar{\boldsymbol{l}}_T \, \big] \end{aligned} \tag{3}$$

with $\bar{\boldsymbol{\mathcal{L}}}_T$ being the set of possible values along the covariate history and $f_{\boldsymbol{L}_t|\bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{L}}_{t-1}}$ the conditional densities assumed for the measurements (Daniel et al., 2013; Dawid and Didelez, 2010). We can go from the first form, conditional on e, to the second form, conditional on o, because SR holds. Equation (3) is known as the g-computation formula or the g-formula. The g-formula is relatively straight-forward to implement in a Bayesian framework for estimating causal effects of sequential treatment plans (Arjas and Saarela, 2010; Saarela et al., 2015). This approach does however have a number of drawbacks (Robins, 1986; Daniel et al., 2013) which we discuss in Section 7. Alternative methods have been suggested to tackle the problem of time-varying confounding. However these have a number of limitations of their own (Robins et al., 2000; Daniel et al., 2013) and embedding them in a Bayesian framework is not trivial, though some work has been recently undertaken in

this area (Saarela et al., 2015).

Even in the simplest contexts the g-formula (3) is such that the integral cannot be computed analytically. Thus numerical methods must be brought to bear. As it is not necessary in practice to define a model for the baseline row vector $\boldsymbol{L}_0$, in our application we only need to estimate $f_{V_1|\boldsymbol{L}_0,A_0}(v_1|\boldsymbol{l}_0, a_0; o)$ and $E(Y|\bar{\boldsymbol{A}}_1 = \bar{\boldsymbol{a}}_1, \bar{\boldsymbol{L}}_1 = \bar{\boldsymbol{l}}_1; o)$ and plug them in the numerical algorithm computing (3). For a detailed discussion of such algorithm see for example Daniel et al. (2011).

4. Dropout in the presence of time-varying confounding

In this section we consider the issues involved in making inference about treatment strategies in the presence of dropout and describe in terms of conditional independences and DAGs the possible mechanisms that lead to participants dropping out. To this end we define the binary random variable R_t taking value 1 for subjects observed at time t and 0 for subjects who have dropped out at time t. As we consider only monotone dropout patterns, we always assume that $R_0 = 1$ (namely that we observe everyone at the beginning) and that if a subject drops out at time t then $R_s = 0$ for all $s = t, \ldots, T+1$.

Before describing the dropout mechanisms we need to make sure that the introduction of selection nodes R_t does not lead to SR failing. More specifically, as at every occasion t data availability implies conditioning on $R_t = 1$, we need to verify that SR still holds when the row vector $\boldsymbol{L}_t$ is extended to include R_t (*i.e.* $\boldsymbol{L}_0$ is like in Section 3.1 while $\boldsymbol{L}_t = (\boldsymbol{L}_0, V_t, R_t)$ for $t = 1, \ldots, T+1$). To this purpose we consider the most general case where R_t is simultaneously influenced by $(\bar{\boldsymbol{X}}_{t-1}, V_t, U_t)$. This situation is pictured in Figure 1 (b) for two measurement occasions. As typically the only child of R_t is the following selection node R_{t+1}, by means of the moralisation criterion it is easy to see that both ES and conditional independence (2) hold when $\boldsymbol{L}_t$ is extended as above. This is true also for each of the three dropout mechanisms we consider. As it will become clear in the following subsections, these mechanisms are indeed obtained by deleting some arrows from DAG 1 (b): basic rules of DAGs state that every conditional independence holding in a DAG also holds when one or more arrows are removed from it.

As a consequence of these facts we notice that the *No regime dropout dependence* (NRD) assumption

$$R_t \perp\!\!\!\perp \sigma|(\bar{\boldsymbol{L}}_{t-1}, \bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{U}}_{t-1}, \bar{\boldsymbol{R}}_{t-1}) \quad t = 1, \ldots, T+1 \tag{4}$$

holds. Though directly implied by the "augmented" ES, the NRD assumption has a non-trivial interpretation and it is worth discussing its role. It means that whether individuals

drop out of a study does not depend on whether the study is observational or experimental conditional on the partial history for all t. Although it is an observational study as the treatment assignments were not randomized, the CWP has a formal protocol so patients were recruited and followed up in much the same way as they would have been in a trial (especially during the first few measurements). Thus the NRD assumption seems plausible in our context. However this might not always be the case. In both cross-sectional and longitudinal settings it is reasonable to argue that subjects are more willing to participate if they have been formally enrolled in a clinical trial. Therefore when using "purely" observational data (*i.e.* not coming from a well-established programme) the NRD assumption (and thus the extended SR) might be problematic. From a technical perspective, we remark that NRD fails only when arrows from σ to R_t are included in the DAG. This is a notable point as in the DT framework σ is usually intended to influence only the actions A_t. As we are willing to assume NRD, we drop the regime indicator σ for the remainder of the subsequent discussion for simplicity.

We are now ready to introduce the three dropout mechanisms. First we postulate how dropout might occur (at random, driven by unobserved factors, driven by the outcome). Second we use DAGs to encode and visualise the mechanisms. As the DAGs represent different data generating structures we term our approach structural. The DAGs naturally lead to three factorisation of the joint distribution of the variables involved in the problem and encode different conditional independences. The induced factorisations correspond to three statistical models that have been used in the literature (models based on MAR, Heckman correction and selection models) to address bias due to dropout or selection. Our approach is complementary to that adopted in the dropout literature (Hogan et al., 2004; Little, 1995).

Robins et al. (2000) develop a selection bias g-formula. This formula highlights the aspects of the data generating mechanism that are non-parametrically non-identified because of dropout. In the same spirit we consider conditional independences that help identify the target quantities. In this section we give the basic ideas referring to supporting materials Section 2 for the formal statement of all the conditional independences involved. While we are not entirely successful (see Section 4.2), the simulation study in Section 6 shows that each model always produces results that are closest to the true values when the associated dropout generating process holds.

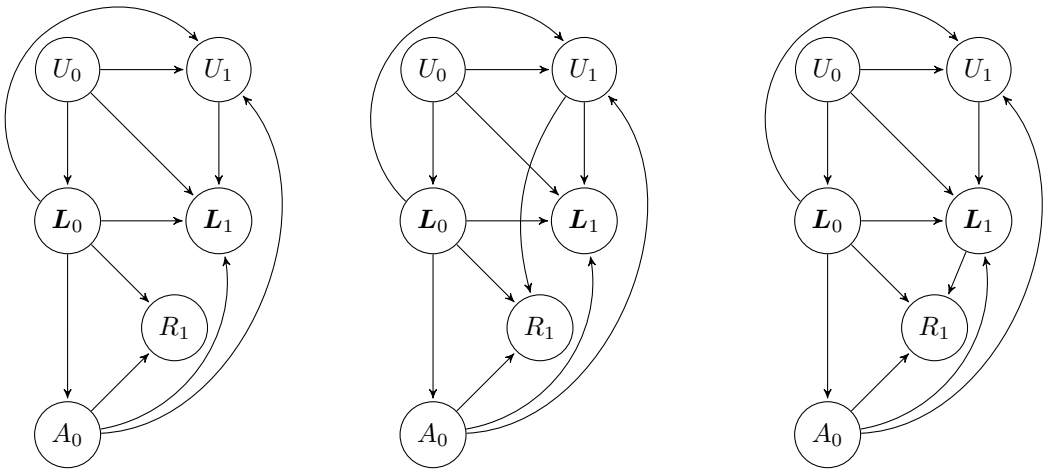

Figure 2: DAGs for dropout mechanisms over two measurement occasions. (a) represents the S-MAR scenario, (b) the *U-drop* scenario and (c) the *Y-drop* scenario.

4.1. Sequential missing at random

In the first mechanism we introduce data are missing at random. There are at least two possible MAR assumptions for longitudinal data (Hogan et al., 2004). Given that our interest is in estimating causal effects in the presence of time-varying confounding we make a *Sequential Missing At Random* (S-MAR) (Pearl and Mohan, 2013; Hogan et al., 2004; Daniel et al., 2013) assumption:

$$R_t \perp\!\!\!\perp (\bar{\boldsymbol{U}}_t, \boldsymbol{L}_t) | (\bar{\boldsymbol{L}}_{t-1}, \bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{R}}_{t-1}) \;\; t = 1, \ldots, T+1. \tag{5}$$

If S-MAR holds then we can say that dropout is *ignorable*. In our application S-MAR implies that the probability of one person attending one session is not influenced by his/her change in weight nor by any other unobserved covariates if we know the baseline variables and the history of weight loss. DAG (a) in Figure 2 shows the situation where S-MAR holds. Again, for simplicity we consider only the first two measurement occasions. When S-MAR holds the complete cases of each wave can be used to estimate the relevant quantities (wavewise complete case analysis). In order to identify the causal quantities of interest, in our application we can simply estimate $f_{V_1|\boldsymbol{L}_0,A_0,R_1}(v_1|\boldsymbol{l}_0, a_0, 1)$ and $E(Y|\bar{\boldsymbol{A}}_1 = \bar{\boldsymbol{a}}_1, \bar{\boldsymbol{L}}_1 = \bar{\boldsymbol{l}}_1, \bar{\boldsymbol{R}}_2 = \bar{\mathbf{1}}_2)$ (where $\bar{\mathbf{1}}_t$ is a sequence of t ones) and plug them in the g-formula (3) in place of $f_{V_1|\boldsymbol{L}_0,A_0}(v_1|\boldsymbol{l}_0, a_0)$ and $E(Y|\bar{\boldsymbol{A}}_1 = \bar{\boldsymbol{a}}_1, \bar{\boldsymbol{L}}_1 = \bar{\boldsymbol{l}}_1)$.

4.2. Dropout driven by unobserved factors

In the second scenario S-MAR (5) no longer holds as dropout at time t is driven also by an unobserved factor U_t. We therefore term this scenario *U-drop*. This is shown for the first two measurement occasions in DAG (b) in Figure 2. One approach to dealing with this type of dropout is based on the *Heckman correction* popular in the Econometric literature (Heckman, 1979). We describe this adjustment first for two measurement occasions and develop a novel extension to the third measurement occasion in Section 4.2.1.

While S-MAR no longer holds, we assume that

$$R_t \perp\!\!\!\perp (\boldsymbol{L}_t, \bar{\boldsymbol{U}}_{t-1}) | (\bar{\boldsymbol{L}}_{t-1}, \bar{\boldsymbol{A}}_{t-1}, U_t, \bar{\boldsymbol{R}}_{t-1}) \qquad t = 1, \ldots, T+1, \tag{6}$$

which prevents lagged unobserved factors from affecting dropout in the present. This is necessary as at time t the Heckman correction is designed to address the association between R_t and $\boldsymbol{L}_t$ induced by U_t and not by U_{t-1}. If for example U_0 and R_1 were directly associated (corresponding to an arrow $U_0 \to R_1$ in DAG 2 (b)) additional bias would be induced.

The basic idea of the Heckman correction is to partition the possibly biased expectation $E(V_1|\boldsymbol{X}_0, R_1 = 1)$, where $\boldsymbol{X}_0 = (\boldsymbol{L}_0, A_0)$, into the unbiased expectation $E(V_1|\boldsymbol{X}_0)$ plus a correction term that can be estimated from the data. Recall that V_1 is percentage change in BMI at the second measurement occasion and $\boldsymbol{L}_1 = (\boldsymbol{L}_0, V_1)$. The Heckman model assumes that the following underlying structure generates the data for the second occasion:

$$\begin{cases} R_1^\star = \boldsymbol{X}_0\boldsymbol{\alpha}_0 + U_1 \\ V_1 = \boldsymbol{X}_0\boldsymbol{\beta}_0 + f(\epsilon_1, U_1) \end{cases} \qquad R_1 = \begin{cases} 1 \text{ if } R_1^\star > 0 \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

In the context of our application we can think of $R_1^\star$ as a linear combination of the observed baseline covariates (age, gender, BMI) as well as an unobserved measure of motivation U_1. We can see that the structure assumed for the Heckman correction fits with DAG (b) in Figure 2 as an unobserved factor U_1 is influencing both the dropout indicator R_1 and the outcome of interest V_1. This type of adjustment for dropout is a type of *shared parameter model* in Hogan et al. (2004).

The vector $\boldsymbol{\alpha}_0$ is typically estimated by means of a generalised linear model for the binary indicator R_1 upon $\boldsymbol{X}_0$. In this regression, which is termed the *selection* equation, the link function depends on the distribution assumed for the error term U_1. The equation containing the vector of interest $\boldsymbol{\beta}_0$ is termed the *outcome* equation. As the outcome equation is typically fitted only on a self-selected sub-sample of subjects (namely those for

whom $R_1 = 1$) a simple linear least squares estimate of $\boldsymbol{\beta}_0$ is biased. Heckman's solution assumes that U_1 is standard normal and that

$$f(\epsilon_1, U_1) = \eta_1 = \tau_{11}^* U_1 + \epsilon_1$$

where $\tau_{11}^* = \text{Cov}(\eta_1, U_1)$ and ϵ_1 is a random variable independent of U_1, without any further distributional assumptions (Hutton and Stanghellini, 2010). It follows that the selection equation is characterized by the probit link

$$\Phi^{-1}(p(R_1 = 1)) = \boldsymbol{X}_0 \boldsymbol{\alpha}_0$$

while we have

$$E(V_1 | \boldsymbol{X}_0, R_1 = 1) = \boldsymbol{X}_0 \boldsymbol{\beta}_0 + \tau_{11}^* \lambda(k_1) \tag{8}$$

with $k_1 = \boldsymbol{X}_0 \boldsymbol{\alpha}_0$ and $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the probability density function and the cumulative density function of a standard normal random variable. The term $\lambda(\cdot)$ is known as an inverse Mills ratio (IMR). Equation (8) implies that in order to obtain an unbiased estimate of $\boldsymbol{\beta}_0$ using data from the non-random sub-sample it is necessary to add the covariate $\lambda(k_1)$ to the outcome equation. Heckman proposes a two-stage procedure that consists in estimating k_1 from the fitted values of the probit regression and using it to estimate the IMR, which is then included as a correction term in (8) based on the values for those units with $R_1 = 1$.

In finite samples the IMR is often almost perfectly correlated with the linear predictor $\boldsymbol{X}_0 \boldsymbol{\beta}_0$: this results in multicollinearity when fitting the adjusted outcome equation (8). The standard solution to the problem is the omission (termed *exclusion restrictions*) of one or more variables (termed *instruments*) from the model specification. Exclusion restrictions represent a pitfall as quite often the choice of instruments is arbitrary or one is forced to rule out some relevant information (Puhani, 2000; Washbrook et al., 2014). This problem exists in the multiple occasion framework as well. Heckman (1979) proposes a maximum likelihood approach to deal with this issue. However in practice at least one instrument is needed to obtain stable estimates or to reach convergence in the optimization algorithms (Genbäck et al., 2014; Washbrook et al., 2014). The Bayesian approach we adopt overcomes the exclusion restrictions in the same way as the maximum likelihood approach but does not present similar convergence problems.

4.2.1. Proposed extension to the Heckman model

To extend Heckman's framework to three measurement occasions we add another pair of equations so that model (7) becomes

$$\begin{cases} R_1^\star = \boldsymbol{X}_0\boldsymbol{\alpha}_0 + U_1 \\ V_1 = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \eta_1 \end{cases} \qquad \begin{cases} R_2^\star = \bar{\boldsymbol{X}}_1\boldsymbol{\alpha}_1 + U_2 \\ Y = \bar{\boldsymbol{X}}_1\boldsymbol{\beta}_1 + \eta_2 \end{cases} \qquad R_t = \begin{cases} 1 \text{ if } R_t^\star > 0 \text{ for } t = 1, 2 \\ 0 \text{ otherwise.} \end{cases} \tag{9}$$

As we are dealing with a monotone dropout mechanism we can insert as covariates those variables measured during the second visit (namely A_1 and V_1 that are contained in $\bar{\boldsymbol{X}}_1$). We assume a standard bivariate normal distribution for (U_1, U_2)

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & 1 \end{pmatrix} \right)$$

while no distributional assumptions are placed on the joint distribution of the error terms within the same temporal point. As before we let

$$\eta_1 = \tau_{11}^* U_1 + \epsilon_1$$
$$\eta_2 = \tau_{22}^* U_2 + \epsilon_2$$

where again $\tau_{22}^* = \mathrm{Cov}(\eta_2, U_2)$, ϵ_2 is independent of U_2 and $k_2 = \bar{\boldsymbol{X}}_1\boldsymbol{\alpha}_1$. Results for the bivariate truncated normal distribution (see Rosenbaum (1961) and Manjunath and Wilhelm (2010)) combined with some calculation permit us to write

$$E(Y|\bar{\boldsymbol{X}}_1, R_1 = 1, R_2 = 1) = \bar{\boldsymbol{X}}_1\boldsymbol{\beta}_1 + \tau_{22}^* C_2(k_1, \tilde{\rho}, k_2) \tag{10}$$

with

$$C_2(k_1, \tilde{\rho}, k_2) = \frac{\tilde{\rho}\phi(k_1)\left(1 - \Phi\left(\frac{\tilde{\rho}k_1 - k_2}{\sqrt{1-\tilde{\rho}^2}}\right)\right) + \phi(k_2)\left(1 - \Phi\left(\frac{\tilde{\rho}k_2 - k_1}{\sqrt{1-\tilde{\rho}^2}}\right)\right)}{p(U_1 > -k_1, U_2 > -k_2)}. \tag{11}$$

As for the two-occasion situation, Equation (11) provides the covariate term which it is necessary to adjust for in order to obtain an unbiased estimate of $\boldsymbol{\beta}_1$ in Equation (10) from the complete case sub-sample. See supporting materials Section 3 for further details.

For more than three measurement occasions, the mathematics of deriving the equivalent of C_2 based on a multivariate truncated normal distribution involves partial correlations (Tallis, 1961) and becomes intractable. Instead, we suggest using $C_t(k_{t-1}, \tilde{\rho}_{t-1,t}, k_t)$ where $\tilde{\rho}_{t-1,t} = \mathrm{Cor}(U_{t-1}, U_t)$ assuming $U_t \perp\!\!\!\perp \bar{\boldsymbol{U}}_{t-2}|U_{t-1}$ for $t = 3, \ldots, T+1$. Note also that this method is a correction for expectations and not for distributions while the g-formula requires that we sum over a probability distribution for the intermediate occasion ($t = 1$). This means that we are not able to fully identify the causal quantities of interest. However the simulation study in Section 6 shows that this method performs better than a WCC or

a SM analysis when dropout is of the *U-drop* kind. Therefore it is worth considering its application in those cases where unobserved factors rather than the outcome are likely to drive dropout.

4.2.2. The correlation parameter $\tilde{\rho}$

The parameter $\tilde{\rho}$ can be interpreted as the correlation between the unobserved variables that drive the dropout at the two time points. Thus in the special case where $\tilde{\rho} = 0$ we have $U_1 \perp\!\!\!\perp U_2$ and Equation (11) reduces to $\lambda(k_2)$: this is equivalent to performing two separate Heckman corrections as in Washbrook et al. (2014). If on the other hand $\tilde{\rho} = 1$ then the unobserved part of the dropout mechanisms is the same at both time points as U_2 is a perfect linear combination of U_1. In the general context of longitudinal data with non-ignorable dropout we can think of $\tilde{\rho}$ as taking on a high positive value as we expect similar forces to be responsible for the dropout at each time point.

It is important to note that the data carry no information about $\tilde{\rho}$ and thus it is an unidentified parameter. It is however essential as without it $C_2(k_1, \tilde{\rho}, k_2)$ is also unidentified and only independent corrections can be implemented over two time points. As they deal with only two occasions, Hutton and Stanghellini (2010) and Genbäck et al. (2014) propose a sensitivity analysis in which they investigate the effects of a range of possible values of τ_{11}^*. In the same vein, but from a Bayesian perspective, we handle the parameter $\tilde{\rho}$ placing a strongly informative prior on it. We discuss the choice of prior for our application and sensitivity of results to this prior in Section 5.2.

4.3. Outcome-driven dropout

Another plausible situation is when dropout is *outcome dependent* (Little, 1995). In our application this means that participation R_t is directly affected by the percentage change in BMI V_t. This situation which we term *Y-drop* is encoded in Figure 2 (c). Again S-MAR does not hold and therefore the participants will be systematically different from the non-participants at each occasion. Recalling that $\boldsymbol{X}_0 = (\boldsymbol{L}_0, A_0)$ and $\boldsymbol{L}_1 = (\boldsymbol{L}_0, V_1)$, it is easy to see that DAG 2 (c) naturally leads to the partition

$$p(V_1, R_1|\boldsymbol{X}_0) = p(V_1|\boldsymbol{X}_0)p(R_1|V_1, \boldsymbol{X}_0)$$

as V_1 depends only on $\boldsymbol{X}_0$ whilst R_1 depends on both V_1 and $\boldsymbol{X}_0$ (Hogan et al., 2004).

In line with Mason (2009), we define two equations for the two measurements. The first is the outcome equation, which relates the outcome to the relevant covariates. The

second models the dropout as a function of possibly the same covariates and the outcome. Thus we have:

$$\begin{cases} V_1 = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \eta_1 \\ logit(p(R_1 = 1)) = \boldsymbol{X}_0\boldsymbol{\gamma}_0 + \gamma_0^O V_1 \end{cases} \quad \begin{cases} Y = \bar{\boldsymbol{X}}_1\boldsymbol{\beta}_1 + \eta_2 \\ logit((p(R_2 = 1)) = \bar{\boldsymbol{X}}_1\boldsymbol{\gamma}_1 + \gamma_1^O Y. \end{cases} \tag{12}$$

Note that in order to identify the causal quantities $E(Y|\bar{\boldsymbol{a}}_1; e)$ for our treatment strategies, the outcome equations need to be fitted on the complete cases at baseline ($t = 0$) and at the intermediate occasion ($t = 1$) respectively (see Section 2.3 of supporting materials for details). As a consequence many units are missing information on the outcomes V_1 and Y. An advantage of the Bayesian approach in this context is that those units can still be included in the regressions as, given the observed covariates, the missing values are sampled within the Monte Carlo Markov Chain (MCMC) procedure (Glynn et al., 1993).

Of particular interest to us are the coefficients γ_0^O and γ_1^O of the outcomes V_1 and Y in the dropout equations in model (12). These parameters tell us how strongly the outcome is associated with the dropout. If the *Y-drop* mechanism is operating in a data-set we would expect these parameters to be significant.

4.4. Comparison of Heckman and selection models

It is well known that the Heckman and selection models are closely related. In particular, when the outcome equation is linear the Heckman model can be rewritten as a selection model. This is true for the two measurement scenario as well as for any extension with a general number of occasions. Washbrook et al. (2014) note that while the models are mathematically equivalent they are conceptually different. In the selection model there is a direct association (or in our case a causal relationship) between dropout and the outcome while in the Heckman model the effect of the outcome on dropout is simply contained in the selection error term like that of every other variable not included in the regression equation.

5. Sensitivity analysis

We now give details of the specific models we use to analyse the dropout in the data from CWP introduced in Section 2. Our aim is to provide some answers to the following two related questions: 1) Do the causal effect estimates change between the WCC analysis and the analysis which take into account dropout of the *U-drop* or *Y-drop* type? 2) Is

there any evidence of dropout of either type? We stress that these methods are intended to be used as part of a sensitivity analysis rather than a one-stop adjustment for dropout especially when dropout is high. Adjusted estimates should be compared to one another in the light of context specific information.

Prior to performing the sensitivity analysis we fit linear models to estimate the input quantities of the g-algorithm. We performed analyses of residuals as well as other diagnostic checks and investigated the presence of quadratic effects or interactions but these did not improve model fit. We therefore based our analyses on linear models throughout.

The sensitivity analysis we propose has two steps potentially. The first is to explore the dropout mechanism by investigating which of S-MAR, *U-drop* and *Y-drop* is driving participation. The second is an analysis of the sensitivity of results to choice of Bayesian priors especially on the poorly or unidentified parameters. We focus on the former with some discussion of the latter. Other sensitivity analyses based on the choice of priors can be found in other contexts in the literature (Scharfstein et al., 1999).

We implemented our Bayesian models using MCMC methods running on JAGS (Plummer et al., 2003). For each analysis the JAGS MCMC sampler was run for 2 chains for 20000 iterations of which 10000 were retained. Convergence was good overall. The means and 95% credible intervals were reported for each analysis. Below we report priors on the more important parameters and refer the reader to the supporting materials Section 2 for information on the priors on the remaining parameters. Notice that normal distributions are henceforth parametrized in terms of precisions rather than variances. Moreover, we code the soft treatment s as 0 and the hard treatment h as 1. The results of the frequentist analysis are shown in the supporting materials Section 4.

5.1. WCC analysis

In the wavewise complete case analysis we are assuming S-MAR as in DAG (a) in Figure 2. The models are given by

$$\begin{aligned} V_1 &\sim N(\mu_1, \varsigma_1) \\ \mu_1 &= \boldsymbol{L}_0\boldsymbol{\beta}_0^B + \beta_{0h}A_0 \qquad (13) \\ Y &\sim N(\mu_2, \varsigma_2) \\ \mu_2 &= \boldsymbol{L}_0\boldsymbol{\beta}_1^B + \beta_{1h}A_1 + \beta_{12}A_0 + \beta_{13}V_1. \qquad (14) \end{aligned}$$

$\boldsymbol{L}_0$ contains four values: the intercept, age, gender and initial BMI so that $\boldsymbol{\beta}_t^B$ are vectors of four parameters for $t = 0, 1$. We place a hierarchical structure on the parameters that

	WCC		HC		SM	
e	$E(Y\|\bar{\boldsymbol{a}}_1;e)$	95% CI	$E(Y\|\bar{\boldsymbol{a}}_1;e)$	95% CI	$E(Y\|\bar{\boldsymbol{a}}_1;e)$	95% CI
(ss)	-4.25	(-4.78,-3.74)	-4.17	(-5.35,-3.00)	-1.51	(-2.36,-0.63)
(sh)	-4.41	(-5.08,-3.58)	-4.26	(-5.53,-2.92)	-2.28	(-3.16,-1.40)
(hs)	-4.88	(-5.79,-4.09)	-4.93	(-6.25,-3.58)	-1.49	(-2.51,-0.41)
(hh)	-5.04	(-5.69,-4.38)	-5.02	(-6.17,-3.82)	-2.26	(-3.21,-1.35)
(d)	-4.98	(-5.63,-4.35)	-4.99	(-6.11,-3.79)	-1.99	(-2.91,-1.04)
parameters						
β_{0h}	-0.53	(-1.01,-0.07)	-0.51	(-1.00,-0.05)	-0.76	(-1.18,-0.34)
β_{1h}	-0.15	(-0.74,0.60)	-0.09	(-0.72,0.68)	-0.77	(-1.38,-0.15)
$\tilde{\rho}$	-	-	0.70	(0.10,0.99)	-	-
γ_0^O	-	-	-	-	-0.03	(-0.08,0.02)
γ_1^O	-	-	-	-	-0.77	(-0.98,-0.52)

Table 2: Results of the sensitivity analysis of the different structural assumptions.

correspond to the same processes over time. Thus

$$\beta_{tj}^B \sim N(\mu_j^B, \varsigma_j^B) \text{ for } t = 0, 1 \text{ and } j = 1, \ldots, 4$$

where μ_1^B, μ_2^B and μ_3^B and μ_4^B correspond to the intercept term, gender, age and the initial BMI. We also impose hierarchical priors on the two "direct" treatment effects:

$$\beta_{th} \sim N(\mu_h, \varsigma_h) \text{ for } t = 0, 1$$

with $\mu_h \sim N(-1, 1/2)$ to reflect our belief that the hard treatment results in modest additional loss in percentage of initial BMI with respect to the soft treatment. The precisions have distinct diffuse $G(0.001, 0.001)$ distributions. Finally β_{13} has a $N(1, 1/2)$ prior as we deem the association between the two changes in BMI, V_1 and Y, quite strong.

We explored a number of alternative prior structures including non-hierarchical priors for the regression coefficients as well as other specifications for the precisions. Results were not substantially different. See supporting materials Section 2.1 for details. The prior structure described here is maintained for the common parameters of other models.

The WCC columns in Table 2 show the results for the four static strategies, the dynamic strategy and the parameters of the treatments in the regressions (13) and (14) for the wavewise complete case analysis. Overall, there seems to be little added value between

strategies (s, s) and (h, h). Simply participating in the study leads to a loss in weight. More specifically, as V_1 and Y are percentage changes in BMI (see Section 3.1), the mean for the static strategy (s, s) is an average loss of 4.25% of initial BMI. This is not entirely unexpected as the soft interventions are still active treatments. The effect of the static strategies increases as the hard intervention is included so that the (h, h) strategy results in a loss of initial BMI of 5.04% on average. The dynamic strategy, indicated by (d), represents the situation where hard treatments are administered until 5% reduction in BMI is achieved and results in a loss of 4.98%. The parameters β_{0h} and β_{1h} are both negative, indicating that the direct effects of the treatment are negative, even though only β_{0h} is significant in this instance.

5.2. HC analysis

We now present the Bayesian version of the extended Heckman correction. Recalling model (9), we define

$$\begin{aligned} R_1 &\sim Bern(p_1) \\ \Phi^{-1}(p_1) &= \boldsymbol{L}_0\boldsymbol{\alpha}_0^B + \alpha_{0h}A_0 \\ V_1 &\sim N(\mu_1, \varsigma_1) \\ \mu_1 &= \boldsymbol{L}_0\boldsymbol{\beta}_0^B + \beta_{0h}A_0 + \tau_{11}^*\lambda(k_1) \qquad (15) \\ R_2 &\sim Bern(p_2) \\ \Phi^{-1}(p_2) &= \boldsymbol{L}_0\boldsymbol{\alpha}_1^B + \alpha_{12}A_0 + \alpha_{13}V_1 + \alpha_{1h}A_1 \\ Y &\sim N(\mu_2, \varsigma_2) \\ \mu_2 &= \boldsymbol{L}_0\boldsymbol{\beta}_1^B + \beta_{12}A_0 + \beta_{13}V_1 + \beta_{1h}A_1 + \tau_{22}^*C_2(k_1, \tilde{\rho}, k_2) \qquad (16) \end{aligned}$$

where $((\boldsymbol{\alpha}_0^B)^\top, \alpha_{0h})^\top = \boldsymbol{\alpha}_0$, $((\boldsymbol{\alpha}_1^B)^\top, \alpha_{12}, \alpha_{13}, \alpha_{1h})^\top = \boldsymbol{\alpha}_1$ and the quantities k_1, k_2 and $C_2(k_1, \tilde{\rho}, k_2)$ are defined as in Section 4.2.1. The priors for the outcome equation parameters are identical to those for the WCC analysis. Those for the selection equations are defined according to the same scheme with $\alpha_{tj}^B \sim N(\nu_j^B, \psi_j^B)$ for $t = 0, 1$, $j = 1, \ldots, 4$. Similarly for the treatment effects on participation we have $\alpha_{th} \sim N(\nu_h, \psi_h)$, $t = 0, 1$ with $\nu_h \sim N(0, 1/4)$. Again, precisions are given $G(0.001, 0.001)$ priors. Priors for ν_j^B and for α_{12} and α_{13} are reported in the supporting materials Section 2.2.

Additional parameters τ_{11}^*, τ_{22}^* have independent uniform priors on the interval $[-1, 1]$. These are strong priors as these parameters are poorly identified in the data and we needed to ensure good convergence. We chose this range as these parameters do not have a direct interpretation in terms of observable quantities and we did not want to impose negative or

positive values. Note that $\tau_{11}^* = 0 = \tau_{22}^*$ is equivalent to a S-MAR mechanism and thus a prior that included 0 as a possible value was important. We chose a uniform distribution on the interval $[0, 1]$ for $\tilde{\rho}$ to encode our belief that the correlation between the dropout processes over time will be positive. Section 2.2 of the supporting materials gives further details of the Bayesian priors including different choices of models for τ_{11}^*, τ_{22}^* and $\tilde{\rho}$, the model implementation as well as the JAGS code and the approximation used to calculate $C_2(k_1, \tilde{\rho}, k_2)$ based on Cox and Wermuth (1991).

The results for the Heckman correction are generally similar to those for the complete cases and are reported in the HC columns of Table 2. The (s, s) regime now as a slightly smaller expected loss in initial BMI, of 4.17%. Again, only the first treatment effect is significant.

5.3. SM analysis

We now describe the Bayesian selection model which handles dropout mechanisms of the type *Y-drop*. The models for the outcomes are the same as Equations (13) and (14) for the WCC analysis. However we also add

$$\begin{aligned} R_1 &\sim Bern(p_1) \\ logit(p_1) &= \boldsymbol{L}_0\boldsymbol{\gamma}_0^B + \gamma_{0h}A_0 + \gamma_0^O V_1 \qquad (17) \\ R_2 &\sim Bern(p_2) \\ logit(p_2) &= \boldsymbol{L}_0\boldsymbol{\gamma}_1^B + \gamma_{12}A_0 + \gamma_{13}V_1 + \gamma_{1h}A_1 + \gamma_1^O Y \qquad (18) \end{aligned}$$

where we can write $((\boldsymbol{\gamma}_0^B)^\top, \gamma_{0h})^\top = \boldsymbol{\gamma}_0$ and $((\boldsymbol{\gamma}_1^B)^\top, \gamma_{12}, \gamma_{13}, \gamma_{1h})^\top = \boldsymbol{\gamma}_1$ to be consistent with model (12). The prior structure is the same as for Heckman model for common parameters. As stated in Section 4.3 the parameters γ_0^O and γ_1^O are a measure of the association between participation and change in BMI in Equations (17) and (18). As the BMI variations V_1 and Y are partially missing in these equations, these parameters are poorly identified in the data. As a consequence, we place strong independent uniform priors on the interval $[-1, 1]$ on them. As with other parameters in the selection and outcome equations which are associated with relationships that repeat over time, we attempted to impose a single hyper prior on γ_0^O and γ_1^O as this would reflect our belief that they are correlated. However this resulted in poor convergence for these parameters although it did not change the values of the effects of the treatment strategies. As before, refer to supporting materials Section 2.3 for a detailed discussion of priors.

The SM columns of Table 2 contain the results for this model. These are different from

the results of the WCC and the HC analyses which are similar to one another. All the static strategies result in much smaller but still significant loss of BMI. This fits in with the soft treatment being an active treatment. Strategy (s, s) results in BMI loss of 1.51% and strategy (h, h) in 2.26% BMI loss. Another interesting feature is that the hard treatment is most effective if administered after the second measurement. The values of γ_0^O and γ_1^O are both negative indicating that lower (weight) BMI loss is associated with a higher chance of dropout (we recall that these coefficients are attached to covariates representing the percentage change in BMI). This is particularly true of the dropout between the second and third measurement.

Given our model specification is correct and the necessary assumptions hold, the sensitivity analysis we performed fits with context specific arguments suggesting that outcome driven dropout is the most plausible mechanism for these data. Conditional on baseline covariates, the change in (weight) BMI seems indeed an information patients are unlikely to ignore when deciding whether to attend the next scheduled meeting.

6. Simulation study

As pointed out in Section 5, we would like our models to adjust for non-ignorable dropout when estimating causal effects but more importantly to be reliable predictors of the underlying dropout mechanisms. To evaluate these properties for our estimators, we performed a simulation study. We consider a simplified context without baseline covariates and maintain the same notation of previous sections, thus the overall interpretation is unchanged.

The study has two parts. First we generate data for 500 units assuming simple linear models for the conditional expectations. As a consequence, obtaining the true causal effects for the four static strategies $\{(s, s), (s, h), (h, s), (h, h)\}$ is straight-forward, as shown in Havercroft and Didelez (2012). We are not going to consider the dynamic strategy here. At the second stage, monotone dropout patterns are simulated by constructing participation indicators R_1 and R_2. This step is based on model (9) for the *U-drop* case and on model (12) for the *Y-drop* case. The S-MAR mechanism is analogous to *Y-drop*, but the outcome-dependent terms in (12) are dropped. We reproduce two scenarios representing respectively low (25%) and high (50%) total participation with dropout rates roughly constant at each occasion. See Section 5 of supporting materials for a more detailed description of both data and dropout generating processes.

The WCC, HC and SM analysis are implemented for each mechanism so 9 models are fitted at every run. The whole procedure is then repeated for 500 runs. Relying on the

same arguments of Section 5.2, the $U(0,1)$ prior for $\tilde{\rho}$ is maintained. For the other dropout specific parameters $(\tau_{11}^*, \tau_{22}^*, \gamma_0^O, \gamma_1^O)$ we explored two different structures. We found that uniform priors lead to more robust estimates than normal priors and recommend their use. Again, details can be found in the supporting materials Section 5.

Figure 3 summarizes results for the low-participation scenario, whose dropout rates are close to the observed ones. The four horizontal lines represent the true expectations under each static strategy, *i.e.* the true causal effects. Note that for these strategies the notation E_e is used to denote $E(Y|\bar{\boldsymbol{a}}_1; e)$ without ambiguity (see the discussion about control strategies in Section 3.1). For every line, 9 boxplots (one for each model-mechanism combination) are drawn so different models within a dropout mechanism can be compared in terms of proximity of their boxplot to the horizontal line. For each boxplot the points beyond the whiskers are not depicted for clarity. Moreover, the empirical coverage rates are reported in the middle of the boxes in place of the usual line representing the median. As we are in the Bayesian framework, these numbers are the proportions of 95% credible intervals that contain the respective true values.

The plot shows that Heckman and selection models each outperform the other two in terms of proximity and coverage when the associated dropout mechanisms hold. Results for the SM analysis when *Y-drop* holds are better than those for the HC model in the *U-drop* case. This is not unexpected given the identification issues discussed in Section 4.2.1 which probably lead to these estimates being very variable especially for $E_{(ss)}$ and $E_{(sh)}$. It is encouraging that all models are able to detect a S-MAR situation, though we notice that the selection model tends to slightly underestimate the true values.

7. Conclusions and Caveats

This paper develops a sensitivity analysis to assess whether there is evidence of non-ignorable dropout in the context of evaluating the effect of some treatment strategies in a weight loss study. The dataset we analyse was gathered from the Counterweight Programme pilot, a study designed to determine the impact of lifestyle interventions on weight loss in overweight and obese patients in primary care in the UK. The methods we propose consider three different dropout mechanisms: sequential missing at random (S-MAR), dropout driven by unobserved factors (*U-drop*) and outcome-dependent dropout (*Y-drop*). We obtain causal effect estimates for static and dynamic strategies using three models that are associated to the three dropout mechanisms. The results of these analyses combined with subject matter knowledge and further evidence from the simulation

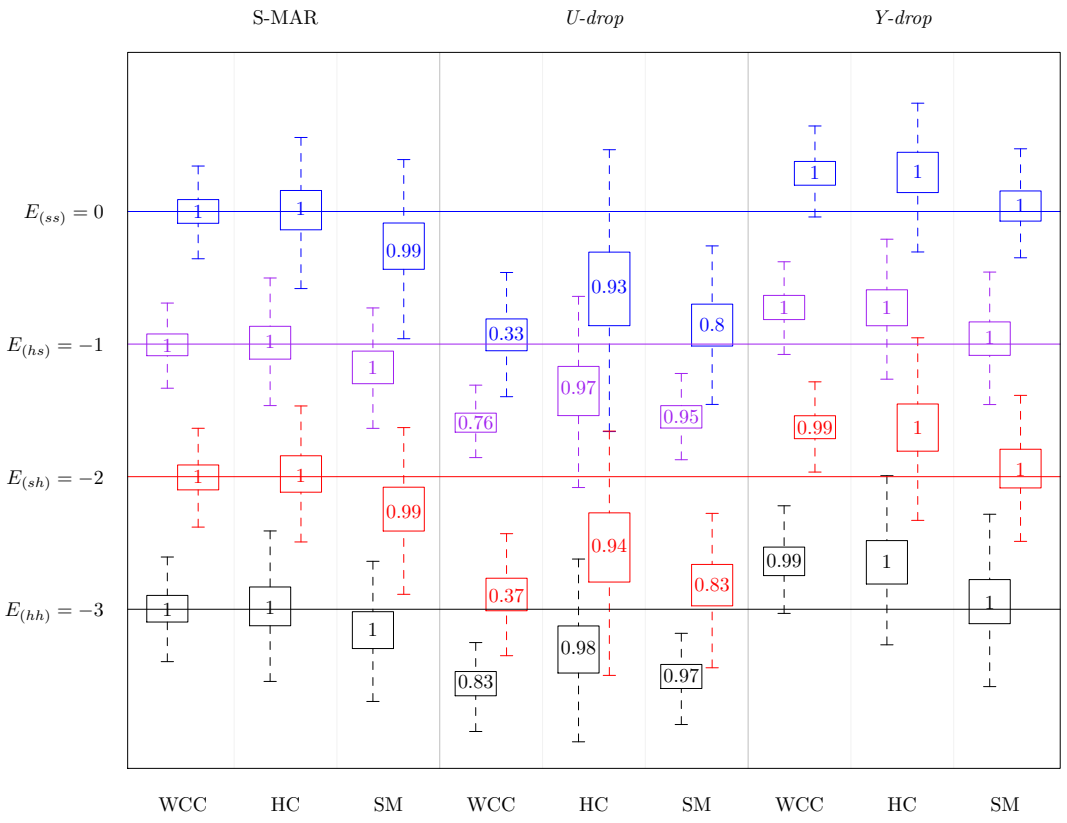

Figure 3: Simulation study results for the wavewise complete cases (WCC), Heckman correction (HC) and selection model (SM).

study lead to the conclusion that non-ignorable outcome dependent dropout is likely to characterize the data. Specifically it appears that individuals who did not lose weight tended to dropout. While the application and the simulation results are promising, we recall that this sensitivity analysis is based on a number of untestable assumptions. Thus the adjusted estimates we report must be viewed as part of a larger pool of context specific information.

The results were generally not sensitive to prior specification in the application. However there was some sensitivity to prior specification in the simulations, possibly due to the smaller sample sizes. This is an important point to bear in mind as the methods we propose might not be appropriate in situations where the sample size is small and the dropout rates high.

It is also worth bearing in mind that as we deal with a number of poorly or unidentified parameters, more complex models might impact negatively on the convergence of Bayesian MCMC procedures. However in the context of non-linear models some identification issues would disappear (Washbrook et al., 2014). Moreover, we arbitrarily chose to partition the seven possible treatments into soft and hard interventions. A different partition might have given different results.

The g-computation algorithm suffers from the *Null paradox*. Thus if we use regression models the effect parameters estimated using the g-formula will appear significant when their value is really zero. In our application there is no indication that the treatment effects are in fact zero in any of the scenarios, though this might be a problem with the simulations. Semi-parametric methods using inverse probability weighting such as marginal structural models (MSM) (Robins et al., 2000) have been put forward in the literature to deal with non-ignorable dropout. These have the advantage that they are not sensitive to model misspecification and they do not suffer from the Null paradox. An extension of this work could consider these alternative methods within the Bayesian framework as in Saarela et al. (2015). However, the Bayesian implementation of the g-formula we present here also has advantages. It allows us to place informative priors on poorly or unidentified parameters (Scharfstein et al., 1999) which is often simpler than assessing the sensitivity of results using a range of values (Genbäck et al., 2014; Rotnitzky et al., 1998). Furthermore it facilitates the identification of the causal effects in the *Y-drop* case and permits to overcome exclusion restrictions in the *U-drop* case.

In our application we deal with only three measurement occasions. However it is possible to deal with more time points. If the size of the history becomes too large it

is possible to make assumptions that reduce dependences between variables at any given time point to the previous one or two measurements only. These can be encoded in conditional independences. For the case of the Heckman correction this is necessary due to the difficulty in obtaining a correction term for three time measurements. We only consider monotone dropout patterns in our analysis as this is the standard in the field. Dealing with non-monotone patterns in this context would involve a number of novel challenges, especially in the *U-drop* case where adjustment terms like those in equations (8) and (10) can be defined in principle. The CWP data suffered from non-monotone dropout; however there were few patients (198) who attended the baseline and third measurement thus we feel that our monotone dropout assumption is justified overall.

Using the DT framework highlights that in order to make causal inference in the presence of dropout we must make the *No regime dropout dependence* assumption. Namely we have to assume that whether patients leave the study is independent of whether the study is experimental or observational conditional on the partial history of subjects. To the knowledge of the authors this assumption has not been made explicit elsewhere in the literature. Finally the approach we propose encourages careful exploration of the problem at hand. This ranges from attempting to understand how dropout is coming about to trying to formulate plausible priors on poorly identified parameters.

The authors would like to thank Prof. Gary Frost for giving us access to the Counterweight Programme Pilot dataset as well as the reviewers and Vanessa Didelez for useful discussion.

References

Arjas, E. and J. Parner (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics 31*(2), 171–187. 19th Nordic Conference on Mathematical Statistics, Stockholm, SWEDEN, JUN, 2002.

Arjas, E. and O. Saarela (2010). Optimal dynamic regimes: Presenting a case for predictive inference. *The International Journal of Biostatistics 6*(2), 1–21.

Cox, D. R. and N. Wermuth (1991). A simple approximation for bivariate and trivariate normal integrals. *International Statistical Review*, 263–269.

Curioni, C. C. and P. M. Lourenco (2005). Long-term weight loss after diet and exercise: a systematic review. *International Journal of Obesity 29*(10), 1168–1174.

Daniel, R. M., S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. A. C. Sterne (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine 32*(9), 1584–1618.

Daniel, R. M., B. L. De Stavola, and S. N. Cousens (2011). gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal 11*(4), 479.

Dansinger, M. L., A. Tatsioni, J. B. Wong, M. Chung, and E. M. Balk (2007). Meta-analysis: The effect of dietary counseling for weight loss. *Annals of Internal Medicine 147*(1), 41–50.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–31.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 70*(2), 161–189.

Dawid, A. P. and P. Constantinou (2013). A formal treatment of sequential ignorability. Technical report, University of Cambridge.

Dawid, P. A. and V. Didelez (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys 4*, 184–231.

Eastwood, P. (2012). Statistics on obesity, physical activity and diet: England, 2012. *The NHS Information Centre*.

Genbäck, M., E. Stanghellini, and X. de Luna (2014). Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. *Statistical Papers*, 1–19.

Geneletti, S., N. Best, M. B. Toledano, P. Elliott, and S. Richardson (2013). Uncovering selection bias in case-control studies using Bayesian post-stratification. *Statistics in Medicine 32*(15), 2555–2570.

Geneletti, S., S. Richardson, and N. Best (2009). Adjusting for selection bias in retrospective, case–control studies. *Biostatistics 10*(1), 17–31.

Glynn, R. J., N. M. Laird, and D. B. Rubin (1993). Multiple imputation in mixture-models for nonignorable nonresponse with follow-ups. *Journal of the Americal Statistical Association 88*(423), 984–993.

Greenland, S. (2009). Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International Journal of Epidemiology 38*(6), 1662–1673.

Havercroft, W. G. and V. Didelez (2012). Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine 31*(30), 4190–4206.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153–161.

Hogan, J. W., J. Roy, and C. Korkontzelou (2004). Tutorial in Biostatistics. Handling drop-out in longitudinal studies. *Statistics in Medicine 23*(9), 1455–1497.

Hutton, J. L. and E. Stanghellini (2010). Modelling bounded health scores with censored skew-normal distributions. *Statistics in Medicine*.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Laws, R. et al. (2004). A new evidence-based model for weight management in primary care: the Counterweight Programme. *Journal of Human Nutrition and Dietetics 17*(3), 191–208.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data, 2nd Edition*. Wiley.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association 90*(431), 1112–1121.

Manjunath, B. G. and S. Wilhelm (2010). Moments calculation for the double truncated multivariate normal density. *SSRN Working Paper Series*.

Mason, A. J. (2009). *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. Ph. D. thesis, Imperial College London. Available at www.bias-project.org.uk.

Pearl, J. and K. Mohan (2013). Recoverability and testability of missing data: Introduction and summary of results. *Available at SSRN 2343873*.

Pearl, J. and J. Robins (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 444–453. Morgan Kaufmann Publishers Inc.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, Volume 124, pp. 125. Technische Universit at Wien.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys 14*(1), 53–68.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9-12), 1393–1512.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94. Springer.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological) 23*(2), 405–408.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the Americal Statistical Association 93*(444), 1321–1339.

Saarela, O., E. Arjas, D. A. Stephens, and E. E. Moodie (2015). Predictive Bayesian inference and dynamic treatment regimes. *Biometrical Journal*.

Saarela, O., D. A. Stephens, E. E. Moodie, and M. B. Klein (2015). On Bayesian estimation of marginal structural models. *Biometrics*.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 223–229.

Washbrook, E., P. S. Clarke, and F. Steele (2014). Investigating non-ignorable dropout in panel studies of residential mobility. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(2), 239–266.