Comparison of Marginal Structural Models to a missing data approach illustrated by data on breast cancer chemotherapies

Christine Gall, Angelika Caputo & Martin Schumacher

Universität Freiburg i. Br.

FDM-Preprint Nr. 103 August 2010

Zentrum für Datenanalyse und Modellbildung

Universität Freiburg

Eckerstraße 1

D–79104 Freiburg im Breisgau

und

Institut für Medizinische Biometrie und Medizinische Informatik

Universitätsklinikum Freiburg

Stefan-Meier-Straße 26

D–79104 Freiburg im Breisgau

christine.gall@fdm.uni-freiburg.de

angelika.caputo@novartis.com

ms@imbi.uni-freiburg.de

Abstract

One of the main objectives in clinical epidemiology is to detect a relation between treatment and outcome. We address data where treatment is applied repeatedly in time and the dose given at a specific time-point may be modified due to actual measurements on disease parameters. If such measurements are subsequently affected by the treatment, they might act as time-dependent confounders. Standard statistical methods cannot adequately address such confounders, but Marginal Structural Models (MSMs) proposed by Robins cope with them. However, these models are still controversely discussed because they are defined within the counterfactual framework.

We illustrate Robins' approach as an extension of a common approach developed for the handling of missing outcomes which does not explicitely use counterfactuals. We address two questions on breast cancer chemotherapy schemes given in repeated cycles. First, we examine the therapy effect and compare two different chemotherapy schemes by the outcome after the fully applied chemotherapy regimen. We account for confounding due to early stopping by Inverse-Probability-of-Censoring-Weighting. Secondly, we investigate the dose effect of one chemotherapy, i.e. the influence of the number of given cycles on the outcome which is modeled by a MSM. Now, the effect is defined by counterfactual variables and time-dependent confounders are accounted for by estimating the parameters of the MSM via Inverse-Probability-of-Treatment-Weighting. We illustrate the concepts of MSMs by showing parallels to the first analysis and pointing out the differences.

Keywords: Causal Inference, Counterfactuals, Intermediate Variables, Inverse-Probability-Weighting, Marginal Structural Model, Time-dependent Confounding

1 Introduction

In observational studies on the effect of time-varying treatments, the administered dose is often based on actual measurements of disease parameters. Also, treatment might be discontinued due to side effects. Then, estimating the therapy or dose effect is complicated due to time-dependent confounders. Such confounders affect the outcome and are simultaneously associated with the treatment in a time-dependent way: subsequent application of the treatment is influenced by the confounder which itself may be influenced by previous applications. Marginal Structural Models (MSMs) proposed by Robins [1, 2] cope with this problem. They are defined within the counterfactual framework which is controversially discussed because it explicitely makes assumptions whose validity is untestable. For some statisticians this is unacceptable [3]. Others [4] appreciate that this makes aware of the limitations of empirical research on causal effects and offers the opportunity to modify experimental design or evaluation techniques towards plausible assumptions. In the field of missing data, including the method of Inverse-Probability-of-Censoring-Weighting (IPCW) [5, 6, 7], one also deals with untestable assumptions about the reasons for missingness, but the necessity is largely accepted.

To provide insight into Robins' approach, we compare it to this missing data approach. We consider two treatment arms where treatment is applied in various cycles. First, we compare both arms with respect to the fully applied treatment, i.e. by the outcome after the application of all planned cycles. We refer to the difference in outcome as the therapy effect. Here, we consider the outcome as missing, if another number of cycles were applied. Secondly, we address the effect of a treatment regime comprising a fixed number n of application cycles. Here, the difference to the outcome after the application of $n-1$ cycles is regarded as the dose effect. The MSM is a parametric model to evaluate these differences for all possible numbers n.

Both analyses use a reweighting scheme to account for time-dependent confounding which requires similar untestable assumptions on the confounding mechanism. We illustrate the concepts of MSMs by showing parallels of the estimation steps and emphasizing the differences.

For illustration, we use data from a randomised clinical trial in breast cancer on preoperatively applied chemotherapy schemes which are given in repeated cycles. Clinical parameters such as palpation result measured before each cycle and side effects may lead to stop chemotherapy prior to the last planned cycle.

2 Data example

The GEPARDUO study [8] is a randomised controlled clinical trial run by the German Breast Group to compare two chemotherapy schemes which are applied preoperatively. There is a short and a long treatment arm which involve the application of a chemotherapy scheme of four and eight cycles, respectively.

The primary endpoint is pathological complete remission (pCR) in the breast and axillary nodes. It was measured by the resected breast specimen and axillary lymph nodes at subsequent surgery. Reasons for early stopping were partly foreseen in the study design and consisted amongst others of toxicity and progress of disease diagnosed by palpation carried out before each cycle. If chemotherapy was discontinued, immediate surgery was performed. Therefore, response assessment was possible for every patient which allows to investigate the effect of a certain number of cycles within one treatment arm. As we focus on the binary outcome complete remission, we do not consider, that after early stopping, the endpoint was observed prior to the end of study.

To maintain the beneficial effects of randomisation, the therapy effect was analysed in [8] by the Intention-To-Treat (ITT) principle. This means, patients are analysed as be-

longing to their randomised arm, regardless of whether they discontinued chemotherapy. This analysis is typically conservative, i.e. it tends to underestimate the effect. Now, by the first approach, we aim to reduce the possible bias introduced by ignoring early stopping. Per treatment arm, we estimate the outcome after all planned cycles adjusted for early stopping and then compare both treatment arms. The second approach addresses a different problem. We only consider one treatment arm and regard the dose effect by comparing the outcome after differential numbers of cycles.

3 Characteristics of time-dependent confounders

In our setting, a time-dependent confounder is a prognostic factor for tumor status which is characterised as a possible reason to stop treatment and furthermore as being affected by previously given cycles. One of the possible time-dependent confounders in our data set is increased leucocytes measured in WHO toxicity grades 1 to 4. Part of the confounding situation is shown in Figure 1. Application of treatment cycle n is given by the status variable Δ_n. If cycle n is given, Δ_n equals 1, otherwise it is set to 0. The covariates measured just before cycle n are denoted by X_n. Baseline covariates are included in X_1. The outcome is determined by Y. To simplify the graph, the figure only shows two cycles and does not contain the impact of baseline covariates. We see, that the number of leucocytes influences the doctor's decision to continue or stop treatment, i.e. $X_{n-1} \rightarrow \Delta_{n-1}$ and $X_n \rightarrow \Delta_n$. But also, by $\Delta_{n-1} \rightarrow X_n$, there is a relation the other way round, as depending on whether cycle $n-1$ is given, the number of leucocytes is affected. As this pathway carries on to Y, the value of X_n partially resembles a treatment effect. Therefore, usual adjustment for time-dependent confounders is unsuitable.

4 First approach: Therapy effect

The therapy effect refers to the difference of both treatment arms with respect to the outcome after the application of all planned cycles. For those patients who stopped chemotherapy early, this outcome is not observed and considered as censored. As both treatment arms are randomised subgroups of the population sample and are thus expected to be similar with respect to baseline characteristics, we adjust for censoring seperately per treatment arm.

Due to time-dependent confounding, there is a non-random selection of patients who received all planned cycles. This is accounted for by the missing data approach of Inverse-Probability-of-Censoring-Weighting (IPCW) [5, 6, 7], which we use to estimate the expected outcome adjusted by confounders separately per treatment arm. The idea of IPCW is that censored patients are replaced by observations with similar covariate history up to the censoring time, here, the time until early stopping. For this purpose, uncensored patients are weighted by the probability of not being censored which, here, equals the probability of receiving all planned cycles. The weights are deduced from assumptions on the censoring mechanism in the underlying data structure.

4.1 Outcome after application of all planned cycles per treatment arm

In this section, we consider each treatment arm separately and focus on the outcome after the application of all planned cycles. To indicate that chemotherapy was fully applied, we define a new outcome variable Y_{all} which is equal to Y, if all cycles are given and missing otherwise. This means, if the outcome is measured after less than the planned number of cycles, it is considered as censored. Censoring is induced by the relations between X_n and Δ_n shown in Figure 1. We can interpret Δ_n as the censoring

indicator which manifests in time. To account for the time-dependent aspects and the chronological order of cause and effect, the censoring mechanism is usually characterised by sequentially defined conditional independence assumptions (denoted by $\perp\!\!\!\perp$) which require for all possible numbers of cycles n:

$$Y_{\text{all}} \perp\!\!\!\perp \Delta_n | \overline{X_n}, \overline{\Delta_{n-1}} \tag{1}$$

Here, we write $\overline{\Delta_n}$ for $(\Delta_1, ..., \Delta_n)$ where $\overline{\Delta_0}$ is defined to be 0 to simplify notation. $\overline{X_n} = (X_1, ..., X_n)$ is the covariate history prior to cycle n. Baseline covariates are included in X_1.

The assumptions in (1) can be interpreted in two ways. Firstly, by comparison of two patients just before cycle n who received chemotherapy so far and have the same covariate history $\overline{X_n}$. According to (1), knowing that for one of them cycle n is withheld, does not imply that complete remission, had all planned cycles been given, is more or less likely than for the other patient where cycle n is given. This means, the fact that cycle n is given or not does not improve the prediction on the outcome after all planned cycles based on the covariate history $\overline{X_n}$. Secondly, (1) implicitely ensures that all information on the disease status which influences the doctor's decision to stop chemotherapy is included in $\overline{X_n}$. For example, if the number of leucocytes, which has a prognostic effect on response, is not included in $\overline{X_n}$, a patient who discontinues chemotherapy after cycle $n-1$ due to a high number of leucocytes is more likely to have complete remission after all cycles had been given than a patient with a low number of leucocytes. Then, knowing $\Delta_n = 0$ does improve the prediction on the outcome.

Now, we use (1) to transform the parameter of interest, $E(Y_{\text{all}}) = P(Y_{\text{all}} = 1)$, into a term that only includes probabilities of uncensored outcomes. These probabilities can then be estimated by weighting the observed data. The weights are read off from the resulting term. The transformation is done by iterative multiplication by a factor which

is equal to 1. For the first cycle, (1) implies $P(Y_{\text{all}} = 1|X_1 = x_1) \cdot P(\Delta_1 = 1|X_1 = x_1) = P(Y_{\text{all}} = 1, \Delta_1 = 1|X_1 = x_1)$ which leads to the following:

$$\begin{aligned}
P(Y_{\text{all}} = 1) &= \sum_{x_1} P(X_1 = x_1) \cdot P(Y_{\text{all}} = 1|X_1 = x_1) = \\
&= \sum_{x_1} P(X_1 = x_1) \cdot P(Y_{\text{all}} = 1|X_1 = x_1) \cdot \overbrace{P(\Delta_1 = 1|X_1 = x_1)/P(\Delta_1 = 1|X_1 = x_1)}^{=1} \overset{(1)}{=} \\
&= \sum_{x_1} P(X_1 = x_1) \cdot P(Y_{\text{all}} = 1, \Delta_1 = 1|X_1 = x_1)/P(\Delta_1 = 1|X_1 = x_1) = \\
&= \sum_{x_1} P(Y_{\text{all}} = 1, \Delta_1 = 1, X_1 = x_1)/P(\Delta_1 = 1|X_1 = x_1)
\end{aligned}$$

Applying this procedure iteratively, we obtain with p the number of planned cycles

$$P(Y_{\text{all}} = 1) = \sum_{\overline{x_p}} P(Y_{\text{all}} = 1, \overline{\Delta_p} = 1_p, \overline{X_p} = \overline{x_p})/P^{(1_p, \overline{x_p})} \tag{2}$$

where we sum over all possible covariate vectors $\overline{x_p} = (x_1, ..., x_p)$ with

$$P^{(\overline{\delta_p}, \overline{x_p})} = \prod_{i=1}^{p} P(\Delta_i = \delta_i|\overline{\Delta_{i-1}} = \overline{\delta_{i-1}}, \overline{X_i} = \overline{x_i})$$

evaluated at $\overline{\delta_p} = (\delta_1, ..., \delta_p) = 1_p$, where 1_p is the vector of length p with all elements equal to 1.

The weights can be deduced from (2) as

$$w = 1/P^{(\overline{\delta_p}, \overline{x_p})}$$

for $\overline{\delta_p} = 1_p$. Stabilised weights are used to obtain more efficient estimates [2]:

$$sw = \prod_{i=1}^{p} P(\Delta_i = \delta_i|\overline{\Delta_{i-1}} = \overline{\delta_{i-1}}) \Big/ P^{(\overline{\delta_p}, \overline{x_p})}. \tag{3}$$

The expected outcome per treatment arm, $E(Y_{\text{all}})$, is estimated by the mean of Y_{all} in the weighted uncensored subset. Assumption (1) makes sure that this estimate is unbiased.

4.2 Estimation of the weights

The weights are estimated by a statistical model appropriate for the mode of treatment application. In our case, skipping one chemotherapy cycle is not allowed. Thus, we regard the number of the last cycle applied, i.e. the failure time variable

$$N := \max\{n : \Delta_n = 1\}$$

and model the probability of having received a certain number n of cycles by a discrete proportional hazards model [9]. We estimate the denominator of (3) by adjusting for the time-dependent and baseline covariates X_n. The estimate of the nominator is just the relative frequency of cycles applied. Per treatment arm, a different model is fitted.

4.3 Therapy effect: comparison of both treatment arms

The therapy effect corresponds to the comparison of both treatment arms with respect to the outcomes after all planned cycles. As the outcome is binary, i.e. $E(Y_{\text{all}}|Z) = P(Y_{\text{all}} = 1|Z)$, with random variable Z that indicates the treatment arm, the therapy effect is quantified by the odds ratio:

$$\text{Therapy Effect} = \frac{P(Y_{\text{all}} = 1|Z = \text{long arm})/(1 - P(Y_{\text{all}} = 1|Z = \text{long arm}))}{P(Y_{\text{all}} = 1|Z = \text{short arm})/(1 - P(Y_{\text{all}} = 1|Z = \text{short arm}))} \qquad (4)$$

The expectations are estimated as shown above separately for both treatment arms. The odds ratio is then calculated from these estimates. For comparison with the MSM

below, we also write down a statistical model for it:

$$logitP(Y_{\text{all}} = 1|Z) = \alpha_0 + \alpha_1 \cdot Z \tag{5}$$

The parameters α_0 and α_1, where α_1 equals the odds ratio, can be estimated by weighted logistic regression fitted by the subset of observations who received all planned cycles using the weights in (3).

5 Second approach: Dose effect

In the previous section, we focussed on the therapy effect and compared two chemotherapy schemes. Now, we only regard data from one treatment arm as if it would have been a prospective observational study. We focus on the dose effect of one chemotherapy scheme by comparing the outcome after the application of differential numbers of cycles. The dose effect is quantified by the odds ratio of these outcomes. Primarily, we compare two groups where n and $n-1$ cycles were applied, respectively. This comparison is facilitated by the counterfactual framework due to its convenient definition of outcome variables. We proceed analogously to the missing data approach. First, we set up assumptions corresponding to the confounding mechanism in the data structure. These assumptions allow to rewrite probabilities of counterfactual variables in terms of probabilities of observable variables. Weights can be deduced from the resulting term which facilitate to estimate the expected outcomes after the application of n and $n-1$ cycles from the reweighted subsets of patients who actually received such numbers of cycles. Then, to regard these comparisons for all n simultaneously, we use a Marginal Structural Model which allows to analyse the question on different doses within one model by making a parametric assumption on the differences in outcome. Its parameters are estimated by Inverse-Probability-of-Treatment-Weighting.

5.1 Counterfactual model setup

To distinguish the outcome after a differential number of cycles by notation, we regard a whole vector $(Y_1, Y_2, ..., Y_p)$ of outcome variables. Such a variable Y_n represents the outcome after the application of n cycles. As one patient can only be treated according to one chemotherapy scheme, there are more outcome variables than can be observed. They are linked to the observed outcome Y as follows. If actually $N = n$ cycles were given, the observed outcome Y is equal to Y_n. In other respects, if $N \neq n$, we do not know the outcome after n cycles and call Y_n counterfactual. Analogously to (1), we claim the following sequentially defined conditional independence assumptions for all $n \leq p$:

$$(Y_1, Y_2, ..., Y_p) \perp\!\!\!\perp \Delta_n | \overline{X_n}, \overline{\Delta_{n-1}} \tag{6}$$

It is called the assumption of no unmeasured confounders and used as identifying assumption to make estimation feasible. Additionally, estimating procedures are based on the positivity assumption

$$P(\overline{\Delta_{n-1}} = \overline{\delta_{n-1}}, \overline{X_n} = \overline{x_n}) > 0 \Rightarrow P(\Delta_n = \delta_n | \overline{\Delta_{n-1}} = \overline{\delta_{n-1}}, \overline{X_n} = \overline{x_n}) > 0 \tag{7}$$

for all possible $\overline{\delta_n}$ and $\overline{x_n}$. It is also called the assumption of experimental treatment assignment [10, 11]. It claims that at every level of the confounder history measured just before cycle n, there is a positive probability of receiving the next cycle n and stopping after cycle $n-1$, respectively.

5.2 Dose effect: comparison of two groups

The dose effect is quantified by the odds ratio of the outcome after application of n versus $n-1$ cycles:

$$\text{Dose Effect} = \frac{P(Y_n = 1)/(1 - P(Y_n = 1))}{P(Y_{n-1} = 1)/(1 - P(Y_{n-1} = 1))} \tag{8}$$

The estimation of $P(Y_n = 1) = E(Y_n)$ is done analogously to the estimation of $E(Y_{\text{all}})$ for the evaluation of the therapy effect in section 4. However, now we address Y_n and the subset of observations where actually n cycles were applied. For patients, who did not receive n cycles, Y_n is counterfactual, i.e. not observed. Using (6), $P(Y_n = 1)$ can be deviated in terms of observables analogously to (2). Now, the sequence $\overline{\Delta_p}$ we are interested in depends on the respective n and the stabilised weights in (3) are deduced for the sequence $\overline{\delta_p}$ where only the first n cycles are given, i.e. $\delta_i = 1$ for $i \leq n$ and $\delta_i = 0$ otherwise. Thus, $P(Y_n = 1) = E(Y_n)$ is estimated as the mean of Y_n in the reweighted subset of patients who received n cycles. This reweighting of counterfactual variables is usually called Inverse-Probability-of-Treatment-Weighting (IPTW). It implicates the exchangeability of the treatment groups not only with respect to baseline characteristics. Also, the remaining differences in time-dependent covariates only reflect the effect of treatment. The reweighted population mimics the adequate trial to test for the dose effect where patients are sequentially randomised before each cycle either to receive another cycle or to stop treatment.

If various doses are compared and there is information on the relation of the odds ratios for different n, the dose effect can be analysed by fitting a Marginal Structural Model as shown in the next section.

5.3 Dose effect: comparison of all groups by a Marginal Structural Model

For our data situation with binary outcome, a simple model which links the marginal distributions of the counterfactual variables Y_n is given by the logistic model:

$$logitP(Y_n = 1) = \beta_0 + \beta_1 \cdot \mathbf{n}. \tag{9}$$

It assumes that the odds ratios in (8) for different n are constant and equal to β_1. This is a Marginal Structural Model as proposed by Robins [1, 2]. The parameter n is not a random variable but indicates the counterfactual outcome. The estimation of the parameters β_0 and β_1 is done in two steps. First, stabilised weights as in (3) are assigned to every observation according to the respective $\overline{\Delta_p}$, which corresponds to the observed N, and the covariate history $\overline{X_N}$. Then, the model is fitted by weighted regression using all observations. This can be done by standard statistical software.

6 Comparison of both approaches

In this section, we compare the modelling assumptions and estimating procedures of the approaches described above. Both approaches measure a population effect. This means, they estimate a marginal effect which depends on the characteristics of the overall population. The missing data approach allows inference on an optimal study conduct where everybody received the planned treatment regime and no early stopping, i.e. censoring, occurs. In contrast, the MSM approach simultaneously predicts the marginal effect of all potential outcomes which are coexistent and of equal interest. Instead of defining Y_{all} as the outcome variable of interest, we regard a whole vector $(Y_1, Y_2, ..., Y_p)$ of outcome variables. These outcome variables allow the definition of the effects without relation

to the way on how patients are actually assigned to treatment regimes. These relations belonging to the data structure are considered by the assumptions (1) and (6), respectively. They describe the mechanism that controls which outcome variable is observable. They are set up analogously. As $Y_p = Y_{\text{all}}$, (6) contains (1).

For the missing data approach, assumptions on the censoring mechanism are needed which are similarily structured as the assumptions made by the MSM approach. In both situations, they are not verifiable by the data but can only be made plausible by sensitivity analyses [12].

The weights, deduced from these assumptions, are derived in the same way and their estimates use the information on covariates X_n and on Δ_n of all observations in both analyses. However, the analyses differ in the second step. Estimating the therapy effect, only outcomes are used for observations with all planned cycles given, whereas with respect to the dose effect, all outcomes are considered and weighted differently according to the observed number of given cycles.

The MSM is a structural model which defines the relation between the marginal probabilities of the counterfactual variables Y_n. The dependent variable n is not random in contrast to a conventional associational model as in (5) where Z is a random variable and the outcome variable is a conditional probability. Note, that in both cases, time-dependent confounding is accounted for by reweighting, such that neither (5) nor (9) contain any covariates for adjustment.

7 Results

7.1 The data

We investigate the full-efficacy population of the GEPARDUO study chosen in [8] for the primary endpoint analysis. This data set involves 855 patients of which 441 were randomised to the long treatment arm consisting of eight cycles. We use a subset of the data shown in Table 1 where the outcome was documented and covariate information used to model the weights is not missing. As possible time-independent confounders between early stopping and the primary endpoint pCR at surgery, we consider the predictors of pCR identified as significant from multivariate logistic regression analysis in [8]: tumor grade (1 and 2 vs 3) and hormone receptor status (HR+ vs HR-). As possible time-dependent confounders, we regard palpation result and WHO toxicities where at least 5% of the patients showed grade 3 or 4 and less than 30% of the patients have missing values. These are alopecia, fatigue and increased leucocytes.

The original analysis in [8] yields a pCR rate of 14.3% and 7.0% for the long and the short arm, respectively. Within the multivariate analysis, the odds ratio of the long versus the short arm results to 2.42. These results are based on an ITT-analysis where no adjustment for early stopping is performed.

7.2 Estimation of the weights

The stabilised weights in (3) are estimated in two steps. To obtain the nominator, we calculate the relative frequencies of cycles applied. For the denominator, we fit a discrete proportional hazards model for the time until the last cycle applied for each treatment arm. With respect to the time-dependent variables, we first perform model selection by univariate analyses. According to the Akaike criterion [13], we include only

those time-dependent covariates where the unadjusted p-value of the two-sided Wald test is $p \leq 0.157$. Patients with missing values are excluded in those analyses where the missing covariate is considered. Estimates per treatment arm are given in Table 3. As recommended in [11], we check that the stabilised weights have a mean of one (see Table 2), which is a necessary condition for correct model specification.

7.3 Therapy effect

The therapy effect after the application of all planned cycles quantified by (4) results to an odds ratio of 2.62 with confidence interval $[1.48; 4.64]$. It is calculated from the estimates of $E(Y_{\text{all}})$ of the long and the short treatment arm which equal the pCR rates. They result to 16.0% (Std Error 0.0235) and 6.8% (Std Error 0.0146), respectively. In comparison to the original analysis, the pCR rate of the long arm is higher whereas it is almost the same for the short arm, resulting in a higher odds ratio. This confirms our expectations as in the long arm 21% of the patients discontinue therapy which diminishes the therapy effect. In the short arm, only 7% do not receive all planned cycles which hardly affects the estimates.

7.4 Dose effect

The trial was not designed to assess the dose effect and most of the patients received all planned cycles. So, its analysis can only be used for illustrating the method. We restrict the data to the long treatment arm with eight cycles, as there is more variation with respect to early stopping. For estimating the parameters of the MSM by the weighted logistic model (9), we use the SAS program Proc Genmod. To obtain robust variances, we follow [2] and choose the option "repeated" and specify an independence working correlation matrix.

Figure 2 shows the probability of pCR with respect to the number of cycles applied. The dots show the values of $P(Y_n = 1) = E(Y_n)$ estimated by the weighted means in the subset of patients who received n cycles. The line indicates the estimate of the MSM which assumes that the logistic model (9) holds. N_{obs} indicates the number of patients who actually received n cycles. We see, that the MSM mainly adjusts for the value at $n = 8$ where the data gives most of the information.

8 Discussion

In this paper, we address the handling of time-dependent confounders in observational studies. MSMs cope with such confounders but are still controversely discussed as they are defined within the counterfactual framework where untestable assumptions are needed for identification. The counterfactual framework facilitates the definition of the treatment effect because counterfactual outcomes are defined separately from any probability model about the way in which observations are assigned to treatment groups. To get more familiar with this approach, we illustrate it by comparison to a common approach developed for the handling of missing outcomes which is based on related unverifiable assumptions. By presenting similarities and discrepancies, we give insight into the structural model approach.

We exemplify our statements by data on breast cancer chemotherapy schemes. Here, the results for the dose effect can only be used for illustration, as there are not enough observations with early stopping to obtain reliable estimates. As this analysis is not meant to gain knowledge about breast cancer and its rehabilitation, the choice of prognostic factors and time-dependent confounders, as well as the assumptions made by the model for the weights, was not agreed with medical experts. Furthermore, we did not respect the change in chemotherapeutic decomposition and used a simple structural

model. If one is in doubt whether the model for the treatment assignment mechanism or the model for the counterfactual data might be misspecified, one might turn to doubly robust estimators [14].

MSMs estimate a population based effect, i.e. the marginal effect in the recruited patient collective. It is estimated from observational data where certain rules are used to decide on further application doses with respect to actual measurements of disease parameters, i.e. where so called dynamic treatment regimes are applied. However, the estimated MSM effect applies to the effect of a static treatment regime where the application doses are determined at baseline and do not change in response to past history of the individual patient. Effects of dynamic treatment regimes are challenging even with respect to their definition and interpretation. They are increasingly addressed in the literature [15, 16].

MSMs are very flexible as all kinds of regression models for the weights and the relation of the treatment with the outcome can be chosen according to the data. In case of a survival outcome, time-dependent weights are used [17]. For most of the regression models, standard statistical software packages provide a weighted fit.

MSMs make the assumption of experimental treatment assignment shown in (7) which claims that at every time-point every patient has a chance to get further treatment. As sometimes treatment must be stopped due to severe adverse events such as toxicity of high grade, this is not absolutely satisfied in our data example. In [18], van der Laan and Petersen describe how to relax this assumption and propose causal effect models which only consider possible treatment options. To apply such models to the data of the GEPARDUO study and explore their properties is subject of further considerations.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation), Research Unit FOR 534.

References

1. Robins JM. Marginal structural models. *ASA Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, 1998; 1–10.

2. Robins JM, Hernán MA, Brumback BA. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.

3. Dawid A. Causal Inference without Counterfactuals. *Journal of the American Statistical Association* 2000; **95**(450):407–424.

4. Greenland S, Morgenstern H. Confounding in health research. *Annual Review of Public Health* 2001; **22**:189–212.

5. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**(427):846–866.

6. Robins J, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology — methodological issues*, Jewell N, Dietz K, Farewell V (eds.), Birkhauser, Boston, 1992; 24–33.

7. JM R, Rotnitzky A. Inverse probability weighted estimation in survival analysis. Second edn., John Wiley & Sons: New York, 2005; 2619–2625.

8. von Minckwitz G, Raab G, Caputo A, Schütte M, Hilfrich J, Blohmer J, Gerber B, Costa S, Merkle E, Eidtmann H, *et al.*. Doxorubicin with cyclophosphamide followed

by docetaxel every 21 days compared with doxorubicin and docetaxel every 14 days as preoperative treatment in operable breast cancer: the geparduo study of the german breast group. *Journal of Clinical Oncology* 2005; **23**(12):2676–85.

9. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. Second edn., Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 2002.

10. Mortimer KM, Neugebauer R, van der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology* 2005; **162**(4):382–388.

11. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**(6):656–664, doi:DOI: 10.1093/aje/kwn164.

12. Brumback BA, Hernán MA, Haneuse SJ, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine* 2004; **23**:749–767, doi:10.1002/sim.1657.

13. Teräsvirta T, Mellin I. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics* 1986; **13**:159–171.

14. Robins JM, Bang H. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**:962–972.

15. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics* 2010; **6**(2, Article 18), doi: 10.2202/1557-4679.1212. Available at: http://www.bepress.com/ijb/vol6/iss2/18.

16. Orellana L, Rotnitzky A, Robins JM. Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content. *The International Journal of Biostatistics* 2010; **6**(2, Article 8), doi: 10.2202/1557-4679.1200. Available at: http://www.bepress.com/ijb/vol6/iss2/8.

17. Hernán MA, Brumback BA, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* 2000; **11**(5):561–570.

18. van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics* 2007; **3**(1, Article 3), doi:10.2202/1557-4679.1200. Available at: http://www.bepress.com/ijb/vol3/iss1/3.

Tables

Table 1: **Analysis set per treatment arm: number of patients by number of cycles received**

	number of cycles received							
	1	2	3	4	5	6	7	8
Short arm (319 Obs)	2	7	14	296	-	-	-	-
Long arm (311 Obs)	2	3	2	13	23	11	12	245

Table 2: **Summary of stabilised weights per treatment arm**

	stabilised weights			
	Mean	Max	Min	Std Dev
Short arm (4 cycles, 319 Obs)	1.017	5.680	0.093	0.476
Long arm (8 cycles, 311 Obs)	1.008	4.179	0.268	0.355

Table 3: **Results of multivariate discrete proportional hazards model for denominator of the stabilised weights**

	Short treatment arm			Long treatment arm		
Characteristics	Hazard ratio	95% CI	p-value	Hazard ratio	95% CI	p-value
Baseline characteristics						
Tumor grade (1 and 2 vs 3)	0.29	0.09;0.89	0.03	0.71	0.41;1.26	0.24
Hormone receptor (HR+ vs HR-)	1.28	0.41;3.96	0.67	0.802	0.44;1.45	0.47
Time-dependent characteristics						
Palpation result	not included			not included		
WHO toxicities						
Alopecia	not included			0.81	0.50;1.32	0.40
Fatigue	1.64	1.05;2.54	0.03	1.71	1.31;2.24	≤ 0.001
Increased leucocytes	1.78	1.28;2.47	≤ 0.001	1.18	0.96;1.44	0.12

CI: confidence interval, p-value from two-sided Wald test

Figures

Figure 1: **Time-dependent confounding: relations between time-dependent covariates X_n, application of treatment Δ_n and outcome Y**

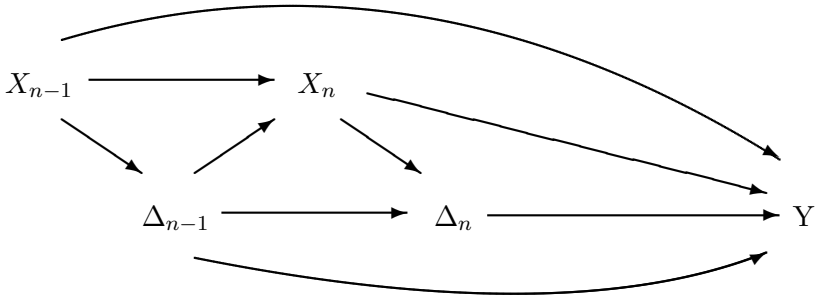

Figure 2: **Probability of pCR by number of given cycles n estimated by MSM (line) and by weighted means (dots) with $N_{\mathbf{obs}}$ number of patients who received n cycles**