# Estimating efficacy in the presence of non-ignorable non-trial interventions in the Helsinki Psychotherapy Study

Tommi Härkänen*
National Institute for Health and Welfare
Helsinki, Finland

Elja Arjas
University of Helsinki, and
National Institute for Health and Welfare
Helsinki, Finland

Maarit A. Laaksonen
National Institute for Health and Welfare
Helsinki, Finland

Olavi Lindfors
National Institute for Health and Welfare
Helsinki, Finland

Jari Haukka
University of Helsinki
Helsinki, Finland

Paul Knekt
National Institute for Health and Welfare, and
Social Insurance Institution
Helsinki, Finland

May 31, 2012

**Abstract**

In a randomised clinical trial with a longitudinal outcome, analyses of the efficacy of the study treatments may be complicated by both non-trial interventions, which have not been administered by the researcher,

*Corresponding author: Tommi Härkänen, THL/VETO, P.O.Box 30, FI-00271 Helsinki, Finland
Email tommi.harkanen@thl.fi, Telephone +358 29 524 8719.

and sparsely measured outcome values. The delay between the change in outcome and the starting of the non-trial intervention may be much shorter than the time intervals between the actual measurements.

We propose a model that accounts for the possible dynamic interdependence between the longitudinal outcome and time-to-event data. The model is based on discretizing the time into short intervals. This results in a missing data problem, which we tackle using Bayesian inference and data augmentation. The method is based on the assumption that decisions to initiate non-trial interventions are not confounded by unobservable factors.

The Helsinki Psychotherapy Study data are used as an illustration. Different psychotherapies were compared, and possible episodes of psychotropic medication were viewed as non-trial interventions.

Simulation studies suggest that our method provides reasonable estimates of the effects of both the study treatment and the non-trial intervention also showing some robustness against possible latent background factors. An application of marginal structural modeling, however, appeared to underestimate the differences between the treatments.

Keywords: Bayesian inference; Clinical trial; Data augmentation; Depression; Markov chain Monte Carlo; Missing data; Non-compliance; Repeated measurements.

# 1 Introduction

In a randomised trial the comparison of the efficacy of the study treatments may be complicated by non-trial interventions, such as auxiliary treatments [1]. For example, a comparison of the efficacy of different psychotherapies can be complicated by psychotropic medication, which is not a part of the study design and is not administered by the researcher but may affect the outcome. The decision of starting an auxiliary treatment (medication) is likely to depend not only on the study treatment group and on the baseline characteristics of the patient, but also on his or her outcome level (symptoms) during the follow-up.

Another frequent complication is the sparse measurement of the longitudinal outcome during the follow-up, which prevents it from being recorded in real time. An increase in symptoms can cause an auxiliary treatment to start almost immediately, which in turn may decrease the symptoms. In this case, the dependency between the outcome and the auxiliary treatment often cannot be fully explained by using the observed history alone, containing baseline covariates, repeated measurements of longitudinal outcomes and time-to-event data.

These complications have not been fully addressed in the literature so far. Eerola *et al.* [2] introduced a joint dynamical model for ear infections, immunity and risk of disease. The antibody levels were measured sparsely during the follow-up, and they were in turn assumed to influence without delay the risk of ear infection. From a modelling perspective, measured antibody levels in their model have a similar role as the longitudinal outcomes in our context, with ear infections as the auxiliary treatments. The drift, i.e. the expected change in the

antibody level between consecutive time points, was assumed to be a (negative) constant. A corresponding assumption may not be realistic in our case, because the drift can depend on the past history of event times. For example, the drift can have large values soon after the start of a treatment but small values later on. Therefore, in this work we generalise the model[2] by estimating the effects of the study treatments and the auxiliary treatment with time-dependent drifts. Bayesian modeling and inference were applied systematically, and the OpenBUGS software[3] based on MCMC methods was used in the numerical work.

Marginal structural models (MSM), estimated by applying the method of inverse probability weighting (IPW)[4,5], have been commonly applied in analysing follow-up data with time-dependent confounding. For example, in[6] the IPW method was applied for analysing observational data with dynamic treatment regimes, in which patients may stop following the regimes of interest. We compare our new approach to the IPW method in a simulation study, and show that a straightforward application of the IPW method does not necessarily yield unbiased results. Moreover, sparsely placed measurement times and a weighting model based only on the observed history may in fact further increase such bias.

Our method relies on the assumption that the likelihood of a non-trial intervention on a patient at a given time depends, in addition to measured patient characteristics, on the value of the concurrent outcome variable. Since the latter variables are not observed except at the times of the actual measurement and are therefore here treated as data-augmented latent variables, we can only say that the "no unobserved confounders" (NUC) postulate can serve as a basis of a reasonable approximation. NUC is commonly assumed in statistical analyses of observational data, and only some methods based on randomisation[7,8] can avoid the NUC assumption effectively. These methods are, unfortunately, rather restrictive, and it seems that they have not been generalised to empirical applications such as the one presented in this paper.

The structure of the paper is as follows. In Section 2 we introduce briefly the Helsinki Psychotherapy Study (HPS)[9] data set, which gave the original motivation for this study. Section 3 describes the model and Section 4 the methods of inference. A common concern in modelling is the possible confounding effect of latent factors on the results. We therefore perform sensitivity analyses using simulation experiments based on different assumptions concerning the latent baseline factors in Section 5, and the compare our results with the MSM method. As an illustration of our method when applied to a real data set, we return in Section 6 to the HPS presenting briefly the results that were obtained. Section 7 concludes the paper with a discussion.

## 2 The Helsinki Psychotherapy Study (HPS)

As a background of our method, the present study is based on a subpopulation of the 276 patients from the HPS[9], which conforms to the Helsinki Declaration of the World Medical Association[10]. These patients were diagnosed with

depressive disorder at the baseline and randomly assigned to solution-focused therapy (SFT), short-term psychodynamic psychotherapy (SPP), or long-term psychodynamic psychotherapy (LPP).

It has been suggested[11] that treatment effects can continue after the therapy has ended, because recovery unfolds sequentially and is associated with patient's overall satisfaction. Especially, this perceived satisfaction can create spillover effects, which cannot be separated from later effects of the study therapies.

The average delay from randomisation to the start of the treatment was 65 days, and the 5 and 95 percentile points were 26 and 127 days, respectively. The waiting times varied due to various reasons, e.g. therapists' schedules, other external non-patient factors, and unknown patient factors.

The assessments were completed at the time of randomisation, and during the follow-up at 3, 7, 9, 12, 18, 24 and 36 months after the treatment was started. The measurement times were defined in the protocol before the recruitment of patients and, accordingly, were not influenced by the outcome. However, it cannot be excluded that some patients chose the exact day for measurement depending on their concurrent outcome state. For this reason we adjusted for the difference between the theoretical measurement day based on the protocol and the actual day of measurement. This adjustment did not attenuate the results based on intention-to-treat models[9]. Thus, we conclude that the measurement times were most likely not significantly influenced by the level of the outcome variable. The outcome measure was depressive symptoms, which was assessed by the Beck Depression Inventory (BDI)[12].

The regular use of psychotropic medication (antidepressant, anxiolytic, neuroleptic or psychiatric combination medication) was considered to form the auxiliary treatment in this empirical example. The times at which such medication episodes has been started were self-reported using questionnaires and based on data from a nationwide health register on medication purchases. Patients were asked whether they were using medication regularly at the time of filling in the questionnaire, and if so, the date on which they had started such use. In the event that the information that was provided was not sufficient for the determination of the exact start or end date of the episodes, simple midpoint approximations were used. Moreover, two alternative approximations, described more closely in Section 6, were used to study the sensitivity of the results of our analysis to these particular choices.

# 3 Model

## 3.1 Notation

Assume that the observed data are based on a clinical trial and consist of $n$ individuals indexed by $i = 1, 2, \ldots, n$. The follow-up time is $T_{\max}$. Time is discretised and indexed by $t = 0, 1, \ldots, T_{\max}$. All event times are rounded to the nearest integer. The measurements on individual $i$ are made at time $t_{i1} := 0$ and during the follow-up at measurement times $t_{i2}, \ldots, t_{im_i}$. The

measurement times are assumed to be independent of outcomes. We call the time interval between consecutive measurements a measurement interval. The study treatment (S) starts at time $a_i^{\mathrm{S}} \geq 0$ and ends at time $b_i^{\mathrm{S}} \geq a_i^{\mathrm{S}}$. Time $a_i^{\mathrm{S}}$ is the waiting time before the starting of the treatment.

The partially observed longitudinal outcome of individual $i$ at time point $t = 0, 1, 2, \ldots, T_{\max}$ is denoted by $Y_{it}$, and the corresponding values observed at times $t_{ik}$ are denoted by $y_{it_{ik}}$. The observed horizontal vector of the baseline covariates $X_i$ is measured at time 0. The observed times of starting (resp. stopping) times of the auxiliary treatment (Aux) are denoted by $\left(a_{ij}^{\mathrm{Aux}}\right)_{j=1}^{n_i}$ (resp. $\left(b_{ij}^{\mathrm{Aux}}\right)_{j=1}^{n_i}$), where $n_i$ is the number of observed auxiliary treatment episodes for individual $i$. If individual $i$ receives no auxiliary treatment, then $a_{i1}^{\mathrm{Aux}} := \infty$. If the last auxiliary treatment episode does not end before the end of the follow-up at $t_{im_i}$, then $b_{in_i}^{\mathrm{Aux}} := \infty$. Define indicator variable $Z_{it} := 0$ if individual $i$ does not take auxiliary treatment at time $t$, and $Z_{it} := 1$ if he/she does. Note that $(Z_{it})_{t\geq 0}$ contains the same information as $\left(a_{ij}^{\mathrm{Aux}}\right)_{j=1}^{n_i}$ and $\left(b_{ij}^{\mathrm{Aux}}\right)_{j=1}^{n_i}$.

## 3.2 Structural assumptions and distributions

The model for the longitudinal outcome is based on drift $\Delta_{it}$ (expected changes of the outcome between time points $t-1$ and $t$) and residual variation $\epsilon_{it} \sim \mathrm{N}(0,\, \sigma_\epsilon^2)$:

$$Y_{it} := Y_{i,t-1} + (\Delta_{it} + \epsilon_{it}) = Y_{i0} + \sum_{s=1}^{t} (\Delta_{is} + \epsilon_{is}),\ t > 0, \tag{1}$$

where $Y_{i0} \sim \mathrm{N}(\mu_0,\, \sigma_0^2)$. Here we assume conditional independence of the outcome value at time $t$ and outcome values before $t-1$ given the outcome value at time $t-1$, drift $\Delta_{it}$ and model parameters $\theta$.

[Figure 1 about here.]

Drift $\Delta_{it}$ is determined by the waiting time before starting of the study treatment ($a_i^{\mathrm{S}}$), and the starting and/or stopping of the possible auxiliary treatment in the past. Let variable $v_i^{\mathrm{S}} := v_i^{\mathrm{S}}(t) := \max(t - a_i^{\mathrm{S}},\, 0)$ indicate how much time has elapsed since the starting of the study treatment. Let similarly $v_{ij}^{\mathrm{Aux},a} := v_{ij}^{\mathrm{Aux},a}(t) := \max(t - a_{ij}^{\mathrm{Aux}},\, 0)$ and $v_{ij}^{\mathrm{Aux},b} := v_{ij}^{\mathrm{Aux},b}(t) := \max(t - b_{ij}^{\mathrm{Aux}},\, 0)$ indicate how much time has elapsed since the $j^{\mathrm{th}}$ auxiliary treatment episode was started or stopped, respectively. These variables have value zero before the corresponding events occurred. The drift is now defined as

$$\Delta_{it} := \omega(t)\mathbf{1}\{t \leq a_i^{\mathrm{S}}\} + \vartheta(v_i^{\mathrm{S}},\, X_{i,\mathrm{S}}) + \sum_{j:\, a_{ij}^{\mathrm{Aux}} \leq t} \left(\zeta(v_{ij}^{\mathrm{Aux},a}) + \eta(v_{ij}^{\mathrm{Aux},b})\right), \tag{2}$$

where indicator function $\mathbf{1}\{A\}$ has value one if $A$ is true and zero otherwise, and covariate $X_{i,\mathrm{S}}$ denotes the treatment assignment group. Parameters $\omega(\cdot)$,

$\vartheta(\cdot)$, $\zeta(\cdot)$ and $\eta(\cdot)$ correspond, respectively, to the mean profile in symptoms during the waiting time before the study treatment, the mean profiles of the considered study treatments, and the effects of the starting and stopping an auxiliary treatment. These parameters controlling drift $\Delta_{it}$ are assumed to have a random walk structure *a priori* due to the large amount of missing data in the longitudinal outcome values. A random walk structure also smoothes out variation. For all $\rho \in \{\omega, \vartheta, \zeta, \eta\}$, let $\rho(t) \sim \mathrm{N}\big(\rho(t-1), \sigma_\rho^2\big)$ for all $t > 0$ defining $\rho(0) := 0$.

The model of drift $\Delta_{it}$ in (2) can easily be extended using, for example, baseline covariates. Note that $\Delta_{it}$ depends on the histories of both the study treatment and the auxiliary treatment. Note also that (2) does not depend on $b_i^{\mathrm{S}}$, so we have assumed that the mean profile of the study treatment will remain the same regardless of when the treatment is stopped. This assumption could be relaxed by defining $v_i^{\mathrm{S},b}$ and $\vartheta^b(\cdot)$ similarly as above, and then incorporating them into (2). If, for example, a patient decides to stop the study treatment prematurely, it is possible that the study treatment has been unsuccessful or that some latent confounding factors might have influenced this decision. In the case of HPS, there were relatively few cases of non-compliance with the study treatment, thus we did not incorporate $b_i^{\mathrm{S}}$ into (2). A similar remark concerns the times of stopping an auxiliary treatment. Figure 1 illustrates the notation and a possible history of individual $i$.

We assume that the auxiliary treatments can depend on observed baseline covariates, on the concurrent, often unobserved outcome value $Y_{it}$ and on the indicator $\mathbf{1}\{b_i^{\mathrm{S}} \geq t\}$ describing whether the study treatment was ongoing at time $t$ or not. All these covariates are contained in horizontal vector $X_{it}^{\mathrm{Aux}}$ and depend on two corresponding vertical vectors of regression parameters $\beta^{\mathrm{Aux},j}$, $j \in \{a, b\}$ for starting and stopping an auxiliary treatment, respectively. Logistic regression models are applied for the time-to-event data defined by the auxiliary treatments:

$$
\begin{aligned}
Z_{it} \sim \mathrm{Bernoulli}\Big( \quad & \mathbf{1}\{Z_{i,t-1} = 0\} \frac{\exp\{X_{it}^{\mathrm{Aux}} \beta^{\mathrm{Aux},a}\}}{1 + \exp\{X_{it}^{\mathrm{Aux}} \beta^{\mathrm{Aux},a}\}} + \\
& \mathbf{1}\{Z_{i,t-1} = 1\} \frac{1}{1 + \exp\{X_{it}^{\mathrm{Aux}} \beta^{\mathrm{Aux},b}\}} \quad \Big).
\end{aligned}
\tag{3}
$$

Here we assume that the covariate vectors $X_{it}^{\mathrm{Aux}}$ are the same for both starting and stopping the auxiliary treatment, but it is straightforward to generalise this.

A key requirement to being able to estimate and compare the effects of different study treatments is that the auxiliary treatments do not severely confound the inferences that can be drawn from the data. In the case of the HPS this would mean that all information that would be relevant for predicting the outcome, that the psychiatrist uses when he/she decides whether to prescribe medication initiating an auxiliary treatment of individual $i$ at time $t$, or to stop an ongoing one, is contained in $X_{it}^{\mathrm{Aux}}$, and that such values would be contained in the available data. Here it would be natural to assume that such $X_{it}^{\mathrm{Aux}}$ consisted of the measured baseline covariates and of the concurrent outcomes

$Y_{it}$.

Another potential source of confounding is due to the role of the waiting time $a_i^{\mathrm{S}}$ before commencing the study treatment. In (2) we assumed that there can be a drift $\omega(t)$ in the outcome values during this waiting time, but after $a_i^{\mathrm{S}}$ it no longer has an effect on the outcome, and the drift $\Delta_{it}$ depends on the study treatment and the possible auxiliary treatments. The individual variation in the psychological symptoms during the waiting time was modelled using $\epsilon_{it}$ in addition to $\omega(t)$. We included the possible change in symptoms from randomisation to the start of the study treatments expressed by $\sum_{t=1}^{a_i^{\mathrm{S}}} \omega(t)$, as there has been speculation that patients' expectations might decrease symptoms before the study treatment starts. These expectations were assumed to play no further role after the study treatment started.

A possible generalisation of model (3) would be to condition also on the observed longitudinal outcome values at the measurement points $\{t_{i1}, \ldots, t_{ik};\ t_{ik} < t\}$ as in[13]. Another extension might be to incorporate a random effect or vector $V_i$ in (2) and (3) as in, for example,[14].

## 3.3 Prior distributions

Here the prior distributions were chosen to correspond to our empirical example. The prior distribution for the mean and variance parameters of the outcome distribution at baseline are $\mu_0 \sim \mathrm{N}(0, 100)$ and $\sigma_0^2 \sim \mathrm{InverseGamma}(2, 1)$, respectively. The residual variance $\sigma_\epsilon^2$ is defined using the precision parameter $\tau_\epsilon := 1/\sigma_\epsilon^2$ with the prior distribution $\tau_\epsilon - 0.01 \sim \mathrm{Gamma}(2, 1)$, which constrains $\sigma_\epsilon^2 \in (0, 100)$ to achieve numerical stability. The constant 0.01 was chosen so that the upper limit (100) for the variance is higher than the total empirical variance estimate of the longitudinal outcome (83.3).

For the parameters controlling drift $\Delta_{it}$, $\rho \in \{\omega, \vartheta\}$ we did not want to set any fixed value for the variance parameter, which controls the fluctuation of the random walk, thus we let $\sigma_\rho^2 \sim \mathrm{InverseGamma}(4, 1)$. For the auxiliary treatment parameters $\rho \in \{\zeta, \eta\}$ the variance parameters had to be fixed, thus we let $\sigma_\rho^2 := 0.1$ for $s < 12$, and for $s \geq 12$ we let $\sigma_\zeta^2 := 0.01$ and $\sigma_\eta^2 := 0$.

The elements of the regression parameter vectors $\beta^{\mathrm{Aux},\cdot}$ are assumed to be normally and independently distributed *a priori* with mean zero and variance 10, except the intercept terms, which have variances equal to 100.

Because of the sparse nature of the observed data, we chose prior distributions to be rather informative in order to smooth out random variation in the results. We also conducted a sensitivity analysis on the choice of the prior distributions. There, the variances for the prior distributions of the regression coefficients $\beta^{\mathrm{Aux},\cdot}$ were multiplied by factor 100, the prior distribution $\sigma_\rho^2$ was set to $\mathrm{InverseGamma}(2, 1)$ ($\rho \in \{\omega, \vartheta\}$), and the variance parameters controlling the fluctuations of $\zeta(\cdot)$ and $\eta(\cdot)$ were multiplied by factor 10.

### 3.4 Direct effect and efficacy

We define the direct effect of study treatment $k$ at time $t$ in our study as the cumulative sum of drift parameters corresponding to the study treatment:

$$\mathcal{T}_{kt} := \sum_{s=1}^{t} \vartheta(s,\, k). \tag{4}$$

This corresponds to a hypothetical situation where the entire possibility of auxiliary treatments is withdrawn from all individuals. Using the concepts and terminology introduced by[15,16], we estimate the direct treatment effect in a situation in which the total effect can be mediated by possible auxiliary treatments.

The efficacy is defined as the difference of direct effects (4) of two study treatments $k_1 \neq k_2 \in \{\text{SPP, LPP, SFT}\}$ at time $t$:

$$\mathcal{E}_t(k_1,\, k_2) := \mathcal{T}_{k_2,t} - \mathcal{T}_{k_1,t}. \tag{5}$$

## 4 Estimation

Data management and the analyses based on the IPW method were performed using the R software packages[17]. OpenBugs software[3] was applied in the Bayesian analyses. The convergence of the MCMC was assessed using autocorrelations of the MCMC chain.

## 5 Simulation study

### 5.1 Parameter values

We conducted simulation studies based on three scenarios. For simplicity, we assumed in all scenarios no waiting time, a follow-up time of $T_{\max} := 60$ time units with $K := 10$ measurement points at times 0, 3, 7, 9, 12, 18, 24, 36, 48 and 60, two treatment groups labelled A and B, treatment lengths 12 time units and $n := 300$ individuals.

[Table 1 about here.]

In the first scenario no latent factors were assumed, and the simulated data were based on the model described in Section 3. The parameter values were $\mu_0 = 18$, $\sigma_0^2 = 25$, $\sigma_\epsilon^2 = 1$. The effects of the study treatments were as follows. In treatment A $\vartheta(t,\, A)$ equalled -8/3 for $0 \leq t \leq 3$, -2/9 for $4 \leq t \leq 12$ and 0 for $t \geq 12$. In treatment B $\vartheta(t,\, B)$ equalled -3, -2/9 and 5/48, respectively. The effect of starting an auxiliary treatment was $\zeta(t) := -1$ for $t = 1, 2, \ldots, 10$ and zero for $t > 10$. The effect of stopping an auxiliary treatment was $\eta(t) := 0.5$ for $t = 1, 2, \ldots, 4$ and zero for $t > 4$. The times of starting and stopping of auxiliary treatment episodes were assumed to depend on the treatment group encoded as

two dummy $\{0,1\}$-valued variables, with group A chosen as the reference class (abbreviated by ST=A), current outcome value $Y_{it}$, and the status of the study treatment $\mathbf{1}\{b_i^{\mathrm{S}} \geq t\}$, i.e. whether it was ongoing or not. In addition to these covariates, vector $X_i^{\mathrm{Aux}}$ of equation (3) contains 1 as the intercept. The true parameter values are presented in Table 1. The prevalences of the auxiliary treatment in the case of the first scenario are presented in Table 2. In this scenario 19% of the individuals in treatment group A and 63% in treatment group B had at least one auxiliary treatment episode.

[Table 2 about here.]

In the second scenario, a random effects model for starting auxiliary treatments was assumed. In addition to (at least partially) observed quantities, such as the outcome, the risk of starting auxiliary treatments depended on individual latent factors. The linear predictor (3) of starting an auxiliary treatment was in this case $X_{it}^{\mathrm{Aux}}\beta^{\mathrm{Aux},a} + U_{i1} + U_{i2}$, where the random effects $U_{i1} \sim \mathrm{N}(0, 2^2)$ and $U_{i2} \sim \mathrm{N}(-0.0625 \times U_{i1},\ 0.125^2)$ represented the latent factor. Note that there was a negative correlation between $(U_{i1},\ U_{i2})$ for preventing an increasing variance of the outcome during the follow-up. Random effect $U_{i1}$ also had an influence on the baseline value of outcome $Y_{i0} \sim \mathrm{N}(\mu_0 - 5 \times U_{i1},\ \sigma_0^2)$. The other random effect $U_{i2}$ had an influence on the individual trend in the outcome values. The drift (2) was redefined in this scenario as $\Delta_{it}^* := \Delta_{it} + U_{i2}$. Individuals who had higher values of $U_{i1}$, and thus had a higher tendency to start auxiliary treatment, had lower outcome values at baseline in this scenario, but they also more often had an increasing trend in the outcome. Other parameter values were as in the first scenario. The distributions of random effects $(U_{i1},\ U_{i2})$ in different treatment groups were equal, which is a realistic assumption in the case of a randomised trial. In observational studies, however, latent confounders may have different distributions in the treatment groups, and therefore an assumption of equal means and variances might not be realistic.

The third scenario was based on the first scenario, except that individuals took other auxiliary treatments (denoted by Aux2), which were similar to the auxiliary treatment in the first scenario (Aux). Aux2 was assumed to be unobserved, and thereby introducing latent time-dependent confounding into the observed outcome data, whereas Aux was observed and included in the analysis. Aux2 was assumed to have the same effects on the outcome and also the parameters controlling starting and stopping Aux2 were the same as those of Aux. If Aux (respectively Aux2) was ongoing at time $t$, then the probability of starting Aux2 (Aux) was assumed to be smaller; log OR of starting Aux2 (Aux) was assumed to be -1.

The Bayesian model presented in Section 3 was compared to two other methods: observed means, which represent the intention-to-treat estimates as if there were no confounding baseline factors, and the inverse probability weighting (IPW) method, which has been applied for analysing observational data with dynamic treatment regimes[6], where individuals may stop following a regime of interest. Here we defined the regimes of interest to be the two treatments A

and B without auxiliary treatments. At the time at which the auxiliary treatment would have started the individual was censored by setting the outcome values missing after that time. The inverse probability weights were based on the treatment group, time and the outcome value at the previous measurement point as the covariates and using the R-package `ipw` (version 1.0-6). Note that the assumptions made here are even stronger than our assumptions, because the contemporary symptoms are likely to be the most important predictor of initiating the auxiliary treatment. Recall that the actual measurements were far apart, thus the most recent measurement was made a long time ago at most time points, which deteriorated the predictive power of the logistic model. The analyses were performed using generalised estimating equations implemented in the `geeglm` function in the `geepack` of the R software[17]. We report the IPW averages using the sparse data based on the ten measurement points, but we also analysed the full data on all 61 time points on which the simulated data were generated.

We generated 100 simulated data sets based on the model and parameter values described above for each scenario. The point estimates (observed means, estimates based on generalised estimating equations and posterior expectations) were stored, and the different methods were compared using medians of the corresponding point estimates together with 90% reference ranges (RefR), which were based on the 5% and 95% quantiles of the point estimates.

## 5.2 Results

Even in the case of no latent baseline factors (scenario one), the observed means of the outcome and a standard IPW method failed to produce unbiased results. There were few differences in the observed means between the treatment groups (Table 2). The means were too low compared to the true values, especially in treatment group B after three time units of follow-up.

The results obtained using the MSM method did not contain systematic bias in treatment group A, but in treatment group B this method seemed to underestimate the treatment effect. There were more auxiliary treatment episodes in treatment group B than in A. The MSM method also underestimated the efficacy the treatments at the end of the follow-up (0.7), where the true difference equalled 4. The corresponding difference of the medians of the MSM point estimates using the data on all 61 time points was 2.05, which suggests that the MSM method was sensitive to sparse measurement times.

The joint analysis of the longitudinal outcome and time-to-event data with Bayesian inference performed better, and the medians were close to the true values in all scenarios. In scenarios two and three the RefRs seemed to become wider at the end of the follow-up. There was some instability in analysing some of the 100 simulated data sets, possibly due to the large amount of missing data. The RefRs of efficacies in scenario two contained the true efficacy at all measurement points, for example at time point 60 the RefR was (2.49, 6.44), median 4.56 and the true difference was 4.

In scenario three the corresponding RefR was (1.15, 4.18) and the median

2.61, but at time points 3, 7 and 9 the efficacy was overestimated, for example at time 9 the RefR was (1.04, 2.23) and the median the 1.74 when the true efficacy was 1. The medians were in general below the true values especially in treatment group B, because most Aux2's were taken in this group. As the Aux2's were not observed, their effects on the outcome cannot be adjusted for.

# 6 Results of The HPS

[Table 3 about here.]

In the following, we assess the statistical significance of the results by considering the 95% credible interval of a parameter. If the credible interval does not contain zero (or one in the case of odds ratios), then we consider the result statistically significant.

At baseline before the randomisation, the means did not differ, but thereafter symptoms seemed to decrease faster in the short-term therapy groups, SPP and SFT, than in the LPP group (Table 3,[18]). In the long run the LPP seemed to yield better results, but as the duration of this therapy was approximately 36 months, this study does not provide information on the stability of the effectiveness of LPP after the end of the therapy. The prevalence of medication was almost the same in the three treatment groups during the first 24 months of the follow-up. At 36 months the short-term therapy groups appeared to have a higher prevalence of medication, but the difference was not statistically significant.

[Table 4 about here.]

A clear decrease in the posterior expectation of the effect was demonstrated in all treatment groups during the follow-up (Table 4). In the short-term therapy groups the posterior expectations of the effects were slightly lower during the early stages of the follow-up than in the LPP group. At the end of the follow-up the LPP group appeared to have lower posterior expectations of the effect than the short-term therapy groups. These differences were, however, not statistically significant, as the credible intervals of the efficacies were wide in all cases. Only the efficacy of SPP versus LPP at 36 months appeared to be close to statistically significant. When these results were compared with the observed means (Table 3), it is interesting to note that the posterior expectations of the short-term therapies at 36 months were slightly higher than the observed means, whereas the posterior expectation of LPP remained almost on the same level as the observed mean.

Based on model (3), individuals in the LPP and SFT groups seemed to have lower odds of starting medication than did individuals in the SPP group ($\mathbb{E}[\mathrm{OR} \mid \mathrm{data}] = 0.661$ and 0.654, respectively) (Table 5). The severity of the symptoms increased the probability of starting medication ($\mathbb{E}[\mathrm{OR} \mid \mathrm{data}] = 1.06$). Individuals who had started medication in the SFT group were less prone to stop medication than those in the SPP group ($\mathbb{E}[\mathrm{OR} \mid \mathrm{data}] = 0.541$).

[Table 5 about here.]

The effect of starting medication on the outcome was not statistically significant during the first two months after the start of a medication episode: thereafter, medication caused the outcome to decrease on average by 2.1 BDI units (Figure 2). The effect remained statistically significant from 4 to 7 months. After that the pointwise credible interval widened rapidly due to the short follow-up time of 36 months. The effect of stopping medication was not statistically significant.

For performing sensitivity analyses, we considered two alternative approximations for determining the duration of a period during which psychotropic medication was taken. For forming these approximations of a given period, consider the time interval from the last measurement point at which a patient still reported having taken such medication, to the next measurement point at which the patient reported no medication. In the approximation, the duration of the episode in which taking medication was continued was then set to be 10% (resp. 90%) of the length of this interval. The sensitivity analyses altered the results only little (data not shown). In the 90% scenario stopping of taking medication appeared to have a slightly stronger effect, but the other results remained virtually unchanged.

The sensitivity analyses based on less informative prior distributions on the regression coefficients and the variance parameters controlling the drift parameters, had only little influence on the results (data not shown).

[Figure 2 about here.]

# 7 Discussion

This study presents a new dynamical Bayesian model for estimating the direct effects and efficacies of study treatments in a randomised clinical trial with repeated measurements on a longitudinal outcome. This model provides easily interpretable results on both the direct effect and the efficacy of the study treatments, as well as on the effect of the considered auxiliary treatment. In addition, it provides results on the relationship between the outcome and the probability of starting or stopping the auxiliary treatment. We have presented extensions to both the repeated measurements methods and the methods for analysing non-compliance, and assessed our method using both simulated data sets and the HPS data set.

Most publications on repeated measurements data have been based only on the actual, possibly sparsely spaced measurement points. This is not fully satisfactory in the case of simultaneous modelling of several time-dependent processes, which may not have been measured even at these time points or which correspond to event times such as starting or stopping an auxiliary treatment. In our work such processes have been embedded into a dense grid of time points used in the analyses. The random walk structure of the prior distributions of the drift parameters and the outcome values allowed us to estimate the effects

of the study treatments and the auxiliary treatment without making strong parametric assumptions such as linearity.

Our model allows for an interdependency between the outcome process and the auxiliary treatments. We modelled the average changes of the outcome between consecutive time points, using not only the outcome value itself and auxiliary treatment status at the previous time point, but also the time since starting or stopping the study treatment, or times from the start of the study treatment and of a possibly ongoing auxiliary treatment (or as the case may be, from the time since an auxiliary treatment was stopped), thereby applying non-homogenous Markov model. It would be tempting to incorporate also lagged values of the outcome as covariates in the logistic regression models controlling the starting and stopping of the auxiliary treatment. This, however, is likely to require a dense grid of observations on the outcome process or, for example, separate randomised measurement times for each individual. In this case the data set would provide also short intervals between measurements during various stages of the follow-up, which would allow both an identification of the effect of different lags and a more accurate estimation of the residual variance. Incorporating the observed outcome history into our model is another possible, and more plausible, future development.

The results based on our model can provide valuable insights for assessing model assumptions. Based on our simulation studies, our method appears to provide reasonable estimates also in cases where latent background factors might have influenced both the probabilities of starting the auxiliary treatment and/or the outcome values themselves. In the case of the HPS, the results are also well justified. The more severe the symptoms, the more likely the patient was to start medication. The effect of medication on the outcome is in accordance with earlier research on the effectiveness of psychotropic medication[19]. However, since most individuals in our data set stopped medication in consensus with the psychiatrist, it cannot be concluded from this result that medication could be stopped at any time and/or at any condition without an increase in the symptoms.

This method, like statistical methods in general, is not insensitive to modelling assumptions. Perhaps the most influential assumption here, and in general for drawing causal conclusions based on observational data and using statistical inferential methods, is that there is no unmeasured confounding. This assumption is problematic because it can never be tested empirically in a situation in which all information in the available data is already effectively used in forming model covariates. One such possibility of introducing additional conditioning into the model, which we have not reported here, would be to account in the model for the waiting time from randomisation to the start of the study treatment as an additional covariate in estimating the mean effect of the study treatment. Such additional conditioning would seem to be relevant e.g. in studies relating to malignant tumours or in AIDS studies[20], where the natural disease progression continuing beyond the initiation of the treatment would then be accounted for in the definition of the treatment effect. In the case of psychiatric disorders, however, treatment initiation has a very different character. For

example, a patient's knowledge of having been assigned to a treatment group, already before the actual treatment has started, can have an effect on the measured symptoms[21]. We also note that no systematic association between the length of waiting time and treatment effect was found in an analysis that was performed (data not shown). As a result, we decided not to elaborate on this issue further.

Note that our model has similarities with hidden Markov (HMM) and state space models, which have been popular in analysing parallel processes, of which some are (partially) unobserved. An important special case of state space models are the dynamic linear models (DLM)[22,23]. The DLM's have been applied not only in the cases of linear, Gaussian processes[24,25] but also in cases of generalized linear models[26]. A direct application of these models in situations such as the present one does not seem to be straightforward, however. This is because the HMMs and DLMs do not generally incorporate in their model specifications (functions of) lagged values of the considered processes, apart from the lag one value of the hidden process. In contrast to this, we have considered such functions of lagged values, which were in our case the time elapsed from starting or stopping a medication episode. This is important not only in our application, but potentially also in other research areas involving time-to-event data and sample-continuous processes.

We compared our method with the popular MSM method[4,5] in a simulation study. The MSM method did not produce unbiased results, especially in the cases of sparse measurement times or a high prevalence of auxiliary treatments. Our method was fully informed about the data-generating model in scenario 1, in which case our method gave correct results, and thus the comparison with the MSM method was not completely fair. In scenarios 2 and 3, however, the data-generating model differed from our analysis model.

So-called doubly robust (DR) methods[27,28] have gained popularity recently. These methods join a regression model to estimate the relationship between covariates and the outcome, and a model to estimate the probability of missingness. The (inverses of) estimated probabilities are then used as weights in the estimation. The benefit is that if at least one of these models is correctly specified, then the DR method estimates the average treatment effect consistently. If neither of these models is correct, then the results based on the DR methods are not necessarily better than those of traditional methods, however[29]. If the missingness cannot be predicted well, then the weighting might introduce some additional randomness in the results instead of removing the bias created by a misspecified outcome model. We assessed the possibility of predicting the start of a medication episode, based on the past observed history, using the scaled generalised $R^2$[30]. Different models based on the past observed outcome values, treatment group and measurement time yielded $R^2 = 0.06$ at best. This is likely to be again due to the sparse measurement times, which deteriorated the predictive power of the observed outcome history. A successful application of DR methods could therefore depend on the correctness of the outcome model.

Structural models have been popular in causal inference, and traditionally they have been based on assumed linear dependencies between the quantities

involved. In our case this restriction, variables that are not normally distributed, and large amounts of missing data, would make implementation of standard structural modelling methods problematic.

An auxiliary treatment can be viewed as a mediating factor.[31] applied principle stratification by[8], which has similarities with Rubin's model, to adjust for the effect of a post-randomisation factor such as an auxiliary treatment. The approach has been applied also in a repeated measurements study design[32], and their use of shared random effects to model possible latent factors is a potential extension to our model.

The times of starting and stopping of the auxiliary treatment may depend on the observed outcome values as well as on the intervening unobserved values, handled, in our application, by Bayesian data augmentation. Thus the auxiliary treatment can be considered to constitute non-ignorable non-compliance. Our work can be viewed as a generalisation of the approach of[13], who also constructed a model with a dense grid of time points, but conditioned the distribution of the outcome only on the observed history, whereas our model is based on the augmented concurrent outcome value. An alternative approach for modelling and analysing longitudinal outcome and time-to-event data was presented by[33], who used the framework of counting processes in real time. However, their method still needs further development in order to be suitable for practical applications comparable to the present one. Additional work of direct relevance to ours has been done by[16], whose concept of post-intervention distribution of the outcome is similar to our definition of efficacy, and by[34], who presented several definitions for a direct effect.

In future work, simulation studies could be conducted to assess the effect of different distributions of possible latent factors, such as confounders or effect modifiers, and model assessment using predictive distributions and cross-validation by leaving one individual out at a time. There can be interactions between the treatment effects and the baseline symptoms or other baseline factors, but an assessment of these interactions can be complicated by the limited size of our data set.

We applied the OpenBUGS software, which is commonly applied in Bayesian inference. This choice allowed fast implementation of our model, but general-purpose programs such as OpenBUGS are generally not optimal in all research problems. There are several more efficient methods to simulate the latent states to analyse DLM's or HMM's. Also there are Kalman filter based imputation methods, which can handle time non-homogeneity and missing data. In our case the execution times were only few hours, which we consider reasonably fast.

Our method is not restricted to randomised clinical trials, but can also be applied in observational studies, in which the assumption of equal distributions of latent baseline factors in the comparison groups does not generally hold. If several measurements of the most important factors relating to the outcome are recorded and the follow-up period is long, then the assumptions of conditional independence between the outcome, auxiliary treatment and possible latent background factors given the baseline observations might be realistic.

## Funding

The work is supported by the Academy of Finland [grant numbers 138876, 253088].
The financial support of Oskar Öflunds Stiftelse for participating in the ISCB conference in Montpellier 2010 is gratefully acknowledged.

## Conflict of Interest Statement

The Authors declares that there is no conflict of interest.

## References

1 White IR, Carpenter J, Pocock SJ and Henderson RA. Adjusting treatment comparisons to account for non-randomized interventions: an example from an angina trial. *Stat. Medicine* 2003; **22**(5): 781–793.

2 Eerola M, Gasbarra D, Mkel P, Linden H and Andreev A. Joint modelling of recurrent infections and antibody response by bayesian data augmentation. *Scand. J. Stat.* 2003; **30**(4): 677–698.

3 Thomas A, O'Hara B, Ligges U and Sturtz S. Making BUGS open. *R News.* 2006; **6**(1): 12–17.

4 Robins JM, Hernan MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000; **11**: 550–560.

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

6 Hernan MA, Lanoy E, Costagliola D and Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic &Clin. Pharmacol. &Toxicol.* 2006; **98**: 237–242.

7 Imbens GW and Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals Stat.* 1997; **25**(1): 305–327.

8 Frangakis CE and Rubin DB. Principal stratification in causal inference. *Biom.* 2002; **58**(1): 21–29.

9 Knekt P, Lindfors O, Härkänen T, Välikoski M, Virtala E, Laaksonen M, Marttunen M, Kaipainen M, Renlund C and Helsinki Psychotherapy Study Group. Randomized trial on the effectiveness of long-and short-term psychodynamic psychotherapy and solution-focused therapy on psychiatric symptoms during a 3-year follow-up. *Psychol Med.* 2008; **38**: 689–703.

10 Knekt P and Lindfors O. *A randomized trial of four forms of the psychotherapy on depressive and anxiety disorders.* No. 77 in Studies in social security and health. The Social Insurance Institution, Finland, 2004.

11 Perry J and Bond M. The sequence of recovery in long-term dynamic psychotherapy. *J Nerv Ment Dis.* 2009; **197**(12): 930–937.

12 Beck AT, Ward CH, Mendelson M, Mock J and Erbaugh J. An inventory for measuring depression. *Arch. Gen. Psychiatry.* 1961; **4**: 561–571.

13 Tsiatis AA and Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat. Sinica.* 2004; **14**: 809–834.

14 Hogan JW and Daniels MJ. A hierarchical modelling approach to analysing longitudinal data with drop-out and non-compliance, with application to an equivalence trial in paediatric acquired immune deficiency syndrome. *Appl. Stat.* 2002; **51**(1): 1–21.

15 Pearl J. *Causality: models, reasoning, and inference.* Cambridge University Press, 2000.

16 Pearl J. An introduction to causal inference. *The Int. J. Biostat.* 2010; **6**(2): Iss. 2, Article 7.

17 R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

18 Knekt P, Laaksonen MA, Härkänen T, Maljanen T, Heinonen E, Virtala E and Lindfors O. Helsinki Psychotherapy Study: effectiveness, sufficiency, and suitability of short- and long-term psychotherapy. In: Levy R, Ablon JS and Kächele H (eds.) *Evidence-Based Practice and Practice-Based Evidence Series*, Humana Press, New York. 2012; .

19 Salminen JK, Karlsson H, Hietala J, Kajander J, Aalto S, Markkula J, Rasi-Hakala H and Toikka T. Short-term psychodynamic psychotherapy and fluoxetine in major depressive disorder - a randomized comparative study. *Psychother Psychosom.* 2008; **77**: 351–357.

20 Robins JM, Orellana L and Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Medicine.* 2008; **27**(23): 4678–4721.

21 Posternak MA and Miller I. Untreated short-term course of major depression: a meta-analysis of outcomes from studies using wait-list control groups. *J Affect. Disord.* 2001; **66**: 139–146.

22 West M and Harrison J. *Bayesian Forecasting and Dynamic Models.* Springer, 1997.

23 Harvey AC. *Forecasting, structural time series models and the Kalman filter.* Cambridge University Press, 1989.

24 Carter CK and Kohn R. On gibbs sampling for state space models. *Biom.* 1994; **81**: 541–553.

25 Frühwirth-Schnatter S. Data augmentation and dynamic linear models. *J. Time Ser. Analysis.* 1994; .

26 Durbin J and Koopman SJ. A simple and efficient simulation smoother for state space time series analysis. *Biom.* 2002; **89**: 603–616.

27 Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biom.* 2005; **61**(4): 962–972.

28 Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY and Heagerty P (eds.) *Proceedings of the Second Seattle Symposium on Biostatistics*, Springer. 2004; .

29 Kang JDY and Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci.* 2007; **22**(4): 523–539.

30 Nagelkerke N. A note on a general definition of the coefficient of determination. *Biom.* 1991; **78**(3): 691–692.

31 Gallop R, Small DS, Lin JY, Elliott MR, Joffe M and Ten Have TR. Mediation analysis with principal stratification. *Stat. Medicine.* 2009; **28**(7): 1108–1130.

32 Lin JY, Have TRT and Elliott MR. Nested markov compliance class model in the presence of time-varying noncompliance. *Biom.* 2009; **65**: 505–513.

33 Commenges D and Gégout-Petit A. A general dynamical statistical model with causal interpretation. *J. Of The Royal Stat. Soc. Ser. B.* 2009; **71**(3): 719–736.

34 Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J. Of The Royal Stat. Soc. Ser. B.* 2007; **69**(2): 199–215.

# List of Figures

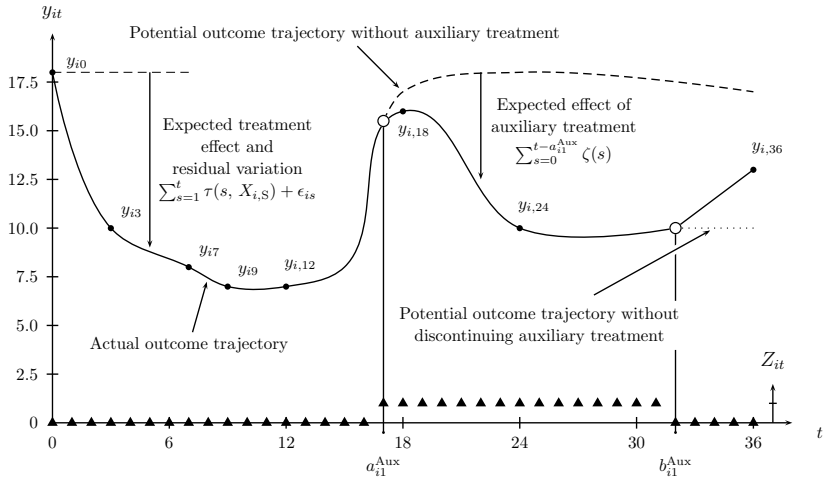


Figure 1: An illustration of a possible outcome history of an individual with one episode of auxiliary treatment. The horizontal axis corresponds to the follow-up time $t$ and the vertical axis to the outcome value. The solid curve denotes the actual outcome trajectory $Y_{it}$, and the black dots correspond to the observed values $y_{it_{ik}}$. The thick dashed line corresponds to the potential outcome values if there is no auxiliary treatment, and the dotted line to the potential outcome values if the auxiliary treatment is continued to the end of the observation interval. The times of starting and stopping the first auxiliary treatment episode are $a_{i1}^{\mathrm{Aux}}$ and $b_{i1}^{\mathrm{Aux}}$, respectively. The black triangles correspond to the auxiliary treatment indicator $Z_{it}$.

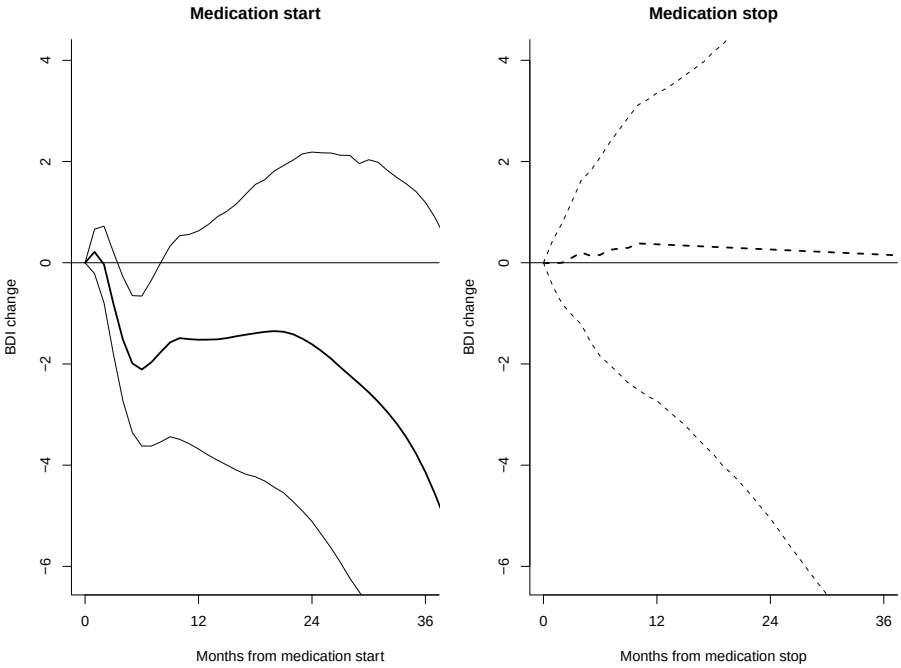


Figure 2: The effect of starting and stopping medication on the outcome defined as the cumulative sum $\sum_{s=0}^{t} \zeta(s)$ and $\sum_{s=0}^{t} \eta(s)$. Time $t$ is from the start of the medication episode in the left-hand side panel and from the end of the episode in the right-hand side panel.

# List of Tables

Table 1: The regression coefficients of the logistic regression models controlling probabilities of starting and stopping the auxiliary treatment. Medians (and the 5% and 95% quantile points) of the posterior expectations based on the 100 simulated data sets. True values are listed on the title lines.

| Scenario | $\beta^{\mathrm{Aux},a}_{\mathrm{Intercept}} = -8.5$ | $\beta^{\mathrm{Aux},a}_{\mathrm{ST=B}} = 2.0$ | $\beta^{\mathrm{Aux},a}_{\mathrm{Symptoms}} = 0.2$ | $\beta^{\mathrm{Aux},a}_{\mathrm{STOngoing}} = 1.0$ |
|---|---|---|---|---|
| 1 | -8.0 (-9.01, -6.35) | 1.88 (1.35, 2.58) | 0.17 (0.0 , 0.22) | 1.12 (0.52, 2.03) |
| 2 | -6.01 (-6.28, -5.72) | 1.11 (0.81, 1.4) | 0.0 (-0.08, 0.08) | 1.97 (1.73, 2.28) |
| 3 | -6.13 (-6.43, -5.88) | 1.06 (0.71, 1.41) | 0.01 (-0.07, 0.09) | 2.0 (1.79, 2.31) |
| | $\beta^{\mathrm{Aux},b}_{\mathrm{Intercept}} = -2.0$ | $\beta^{\mathrm{Aux},b}_{\mathrm{ST=B}} = 0$ | $\beta^{\mathrm{Aux},b}_{\mathrm{Symptoms}} = -0.2$ | $\beta^{\mathrm{Aux},b}_{\mathrm{STOngoing}} = 0$ |
| 1 | -2.6 (-3.77, -1.57) | 0.15 (-0.37, 0.99) | -0.13 (-0.23, 0.0) | 0.12 (-0.54, 0.68) |
| 2 | -2.89 (-3.18, -2.62) | 0.12 (-0.18, 0.53) | -0.01 (-0.1 , 0.08) | 0.06 (-0.25, 0.37) |
| 3 | -2.72 (-3.05, -2.34) | 0.26 (-0.14, 0.64) | -0.01 (-0.1 , 0.08) | -0.08 (-0.41, 0.26) |

Table 2: The differences of the true values and the medians of the average outcome values expressed by observed means, maximum likelihood estimates (frequentist inference) and posterior expectations (Bayesian inference) of the analyses of 100 simulated data sets. The widths of the 90% reference ranges based on 5% and 95% quantile points are in parentheses. The prevalences (%) of the auxiliary treatment are presented in the case of the first scenario.

| | | Scenario 1 | | | | | | | Scenario 2 | | Scenario 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Treatment | Time | AuxT prevalence | Observed means | | IPW | | Bayesian inference | | Bayesian inference | | Bayesian inference | |
| A | 0 | 0.0 | -0.1 | (1.3) | 0.0 | (1.1) | 0.0 | (0.9) | 0.0 | (0.9) | 0.0 | (0.8) |
| A | 3 | 5.3 | -0.2 | (1.4) | -0.1 | (1.5) | 0.0 | (1.6) | -0.1 | (0.9) | -0.2 | (1.0) |
| A | 7 | 7.3 | -0.5 | (1.4) | 0.8 | (1.7) | 0.1 | (1.4) | 0.0 | (1.1) | -0.3 | (1.0) |
| A | 9 | 7.8 | -0.6 | (1.3) | 0.0 | (1.6) | 0.1 | (1.4) | -0.2 | (1.2) | -0.5 | (1.1) |
| A | 12 | 7.4 | -0.8 | (1.4) | 0.0 | (1.7) | 0.1 | (1.5) | -0.1 | (1.2) | -0.7 | (1.3) |
| A | 18 | 6.0 | -0.9 | (1.4) | 0.1 | (1.8) | 0.0 | (1.8) | -0.2 | (1.5) | -0.7 | (1.5) |
| A | 24 | 5.6 | -1.0 | (1.4) | 0.7 | (2.2) | 0.1 | (2.2) | -0.3 | (1.5) | -0.7 | (1.5) |
| A | 36 | 5.4 | -1.2 | (1.6) | -0.1 | (2.4) | 0.2 | (2.9) | -0.5 | (1.5) | -0.7 | (1.8) |
| A | 48 | 5.9 | -1.4 | (2.0) | -0.2 | (3.0) | 0.2 | (3.9) | -0.4 | (2.1) | -0.9 | (2.1) |
| A | 60 | 6.9 | -1.7 | (1.9) | -0.1 | (3.2) | 0.2 | (8.4) | -0.6 | (2.0) | -0.9 | (2.0) |
| B | 0 | 0.0 | 0.0 | (1.2) | 0.0 | (1.2) | 0.0 | (0.9) | 0.0 | (0.9) | 0.0 | (0.8) |
| B | 3 | 26.0 | -0.7 | (1.3) | -0.1 | (1.6) | 0.0 | (1.7) | -0.1 | (1.0) | -0.4 | (0.9) |
| B | 7 | 29.2 | -1.9 | (1.6) | -0.3 | (2.1) | 0.1 | (1.3) | -0.1 | (1.2) | -0.9 | (0.5) |
| B | 12 | 24.4 | -3.3 | (1.8) | -0.6 | (2.7) | 0.0 | (1.6) | -0.3 | (1.4) | -1.3 | (1.4) |
| B | 18 | 17.6 | -3.7 | (1.9) | -0.5 | (2.8) | 0.1 | (1.5) | -0.3 | (1.6) | -1.4 | (1.5) |
| B | 24 | 15.0 | -3.8 | (1.8) | -0.8 | (2.3) | 0.2 | (2.3) | -0.3 | (1.9) | -1.5 | (2.0) |
| B | 36 | 15.6 | -4.5 | (2.0) | -1.6 | (3.5) | 0.1 | (3.5) | -0.3 | (2.2) | -1.7 | (2.2) |
| B | 48 | 17.7 | -5.2 | (1.9) | -2.6 | (5.2) | 0.0 | (4.4) | -0.3 | (2.9) | -2.0 | (2.7) |
| B | 60 | 20.3 | -6.0 | (2.3) | -3.4 | (5.4) | 0.2 | (12.6) | 0.0 | (3.5) | -2.2 | (2.5) |

Table 3: The number of observations (N), observed means and standard deviations (SD) of the outcome (BDI), and the prevalence of medication (Med%) in different treatment groups and measurement points.

| | SPP (group size 79) | | | | LPP (group size 113) | | | | SFT (group size 84) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | N | Mean | SD | Med% | N | Mean | SD | Med% | N | Mean | SD | Med% |
| 0 | 79 | 19.0 | 7.3 | 23 | 111 | 19.8 | 7.8 | 16 | 82 | 18.7 | 7.8 | 30 |
| 3 | 72 | 13.2 | 7.7 | 16 | 89 | 15.5 | 8.2 | 19 | 74 | 12.7 | 9.1 | 24 |
| 7 | 66 | 11.2 | 7.9 | 22 | 87 | 14.8 | 8.3 | 19 | 69 | 10.5 | 8.4 | 27 |
| 9 | 66 | 9.5 | 7.1 | 18 | 85 | 13.1 | 8.4 | 22 | 70 | 10.7 | 8.7 | 26 |
| 12 | 68 | 9.8 | 8.5 | 20 | 95 | 13.3 | 9.0 | 19 | 67 | 10.5 | 10.2 | 26 |
| 18 | 62 | 8.8 | 7.7 | 22 | 84 | 9.9 | 9.7 | 25 | 63 | 10.5 | 10.5 | 27 |
| 24 | 66 | 9.9 | 9.0 | 23 | 87 | 10.6 | 10.7 | 25 | 56 | 10.2 | 9.2 | 26 |
| 36 | 63 | 10.8 | 10.4 | 28 | 88 | 7.8 | 7.6 | 21 | 55 | 10.1 | 10.0 | 26 |

Table 4: Posterior expectations and 95% credible intervals of the sum of the baseline mean and the direct treatment effects $\mu_0 + \mathcal{T}_{kt}$ and the efficacies $\mathcal{E}_t(k_1, k_2)$, $k_1 \neq k_2 \in \{\text{SPP, LPP, SFT}\}$ at different measurement points. The standard deviations are in square brackets and the credible intervals within braces below the posterior expectations of the corresponding parameters.

| Month | SPP | LPP | SFT | SPP-LPP | SPP-SFT | LPP-SFT |
|---|---|---|---|---|---|---|
| 0 | 19.2 | 19.2 | 19.2 | | | |
| | [0.451] | [0.451] | [0.451] | | | |
| 3 | 15.2 | 15.6 | 15.0 | -0.456 | 0.149 | 0.606 |
| | [1.36] | [1.31] | [1.35] | (-2.49, 1.58) | (-1.97, 2.27) | (-1.35, 2.64) |
| 7 | 12.9 | 15.0 | 13.0 | -2.08 | -0.094 | 1.98 |
| | [1.6] | [1.52] | [1.6] | (-5.12, 0.995) | (-3.27, 3.05) | (-1.09, 5.09) |
| 9 | 11.8 | 13.6 | 12.7 | -1.84 | -0.962 | 0.881 |
| | [1.73] | [1.61] | [1.69] | (-5.28, 1.47) | (-4.53, 2.54) | (-2.57, 4.25) |
| 12 | 11.8 | 13.2 | 12.9 | -1.47 | -1.17 | 0.297 |
| | [1.89] | [1.7] | [1.84] | (-5.35, 2.27) | (-5.27, 2.8) | (-3.51, 4.05) |
| 18 | 11.1 | 10.7 | 13.3 | 0.476 | -2.18 | -2.66 |
| | [2.19] | [1.96] | [2.13] | (-4.32, 5.1) | (-7.21, 2.78) | (-7.27, 1.77) |
| 24 | 11.8 | 10.6 | 13.8 | 1.16 | -2.01 | -3.17 |
| | [2.44] | [2.18] | [2.44] | (-4.1, 6.49) | (-7.75, 3.76) | (-8.73, 2.14) |
| 36 | 14.2 | 8.4 | 13.0 | 5.82 | 1.21 | -4.62 |
| | [3.06] | [2.64] | [3.02] | (-1.14, 12.6) | (-6.18, 8.49) | (-11.4, 1.92) |

Table 5: Posterior medians (md), standard deviations, Monte Carlo errors representing the numerical accuracy of the posterior expectation and 95% credible intervals of the odds ratios of starting and stopping medication (OR = $\exp\{\beta_{\cdot}^{\text{medication},\cdot}\}$), and the effects of starting and stopping medication on the outcome.

| OR | md(OR) | SD | MC error | 2.5% | 97.5% |
|---|---|---|---|---|---|
| **Starting medication:** | | | | | |
| Intercept | 0.00823 | 0.00175 | 0.000067 | 0.00536 | 0.0122 |
| LPP vs. SPP | 0.643 | 0.147 | 0.00459 | 0.419 | 0.993 |
| SFT vs. SPP | 0.636 | 0.149 | 0.00357 | 0.416 | 1.0 |
| Study treatment ongoing $\mathbf{1}\{b_i^{\text{S}} \geq t\}$ | 1.39 | 0.286 | 0.00712 | 0.949 | 2.06 |
| Outcome $Y_{it}$ | 1.06 | 0.00872 | 0.000409 | 1.04 | 1.08 |
| **Stopping medication:** | | | | | |
| Intercept | 0.048 | 0.0111 | 0.000431 | 0.0303 | 0.0743 |
| LPP vs. SPP | 0.793 | 0.226 | 0.00404 | 0.464 | 1.34 |
| SFT vs. SPP | 0.522 | 0.144 | 0.00271 | 0.31 | 0.87 |
| Study treatment ongoing $\mathbf{1}\{b_i^{\text{S}} \geq t\}$ | 1.22 | 0.325 | 0.00522 | 0.732 | 2.02 |
| Outcome $Y_{it}$ | 1.0 | 0.0113 | 0.000667 | 0.979 | 1.02 |