

PhD thesis

# **Innovative statistical methods for assessing post approval drug safety based on real world user data**

**Jeppe Ekstrand Halkjær Madsen**

Academic advisor: Thomas Scheike

Industrial advisors: Christian Pipper & Thomas Delvin

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on May 2, 2023.



# Preface

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen. The PhD was carried out at the Section of Biostatistics, University of Copenhagen and at the unit of Biostatistics at Leo Pharma A/S. It was funded by the Innovation Fund Denmark.

I would like to thank my supervisors, Thomas Scheike, Christian Phipper, and Thomas Delvin. Thomas Scheike for always being calm, patient and with valuable inputs. Christian Phipper for our uncountable number of discussions and for pushing me to put my thoughts and ideas into more words. And Thomas Delvin for spending time listening to us crazy biostatisticians and for trying to find the relevance in our ideas. And thanks to all of you for your flexibility while I was stuck on a Vietnamese island for over two months due to Covid restrictions.

I want to thank the Section of Biostatistics at University of Copenhagen. I knew a lot about models and math when I came, but little about epidemiology, survival analysis, causal inference and much more. I feel much more like a statistician now than when I came - you have had a significant effect on my knowledge about statistics. It is truly amazing that so much different, interesting research can come out of such a relaxed and friendly environment.

Thanks to Leo Pharma for giving me the chance to peek into the amazing world of dermatology, drug development, and how it is to work in the industry as a statistician. There has been ups and downs for Leo Pharma in the last few years, but it has remained a pleasant place to work. You have skin in the game, but your good deeds and intentions go beyond the skin. I hope to learn a lot more about all the important work you're all doing in the future.

A special thanks to Professor Jesper Hallas from SDU who helped us with data examples in Manuscripts I and III, and with very valuable comments for Manuscript I along with other good pieces of advice related to academic work.

I would also like to thank Dr. Lars Christian Lund and Data Manager Martin Thomsen Ernst from University of Southern Denmark for helping me with the data example in Manuscript III. Dr. Lars Christian Lund came with the idea for the data example that turned out to provide amazing results, and Martin Thomsen Ernst has been helping with running the code in the very last moment. Thanks!

And last but not least, thanks to my girlfriend Vi who has been incredibly supportive and with whom I can discuss everything statistical and non-statistical. You and I are together for a reason - you might even say that we have a causal relationship.



# Summary

The statistical science as well as the pharmaceutical industry are continuously evolving due to many factors including advances in theory, treatments, technology, ways of monitoring etc. This PhD was particularly motivated by the concepts of Real World Data (RWD) and Real World Evidence (RWE). These concepts have been scrutinized for many reasons, such as the enormous amount of data in existence now and especially in the future due to the development of medical devices that can record enormous amounts of data all the time. This dissertation aims at improving existing methods applied to such data. The general idea explored is to use designs where time-stable confounding is adjusted for implicitly, for example by comparing subjects to themselves. In Manuscript I, we highlight a concrete problem in the case-time-control design, and how to sample controls in order to alleviate it. In Manuscript III, we propose a novel way of adjusting for time-stable confounding without having to switch to a self-controlled design. This enables the use of standard types of models for time-to-event data while not having to fear bias due to confounding by time-stable confounding.

Developments in the causal inference literature, and an increased focus on what we want to estimate and what our research question really is, has also reached the pharmaceutical industry with the estimand framework.

The influence of the estimand framework on this PhD is clear in Manuscript II, where we show how a whole class of working models can lead to unbiased estimation of causal effect under arbitrary misspecification of the working model in crossover trials. This robustness, however, comes with the caveat that the variance has to be estimated from the so-called influence function in order to be completely robust towards misspecification. The estimand framework also shines through clearly in Manuscript III, where a causal effect is targeted despite unmeasured confounding.

All in all, this PhD has made contributions to the major developments in the industry along with research relevant both for RWD (Manuscripts I and III) and in clinical development (Manuscript II).

In this dissertation, I will introduce the currently used methods for the analysis of RWD and causal inference. This will give an overview of the literature and field to which the manuscripts from this PhD belong.



# Resumé

Statistik som videnskab såvel som farma industrien udvikler sig kontinuerligt af mange årsager såsom udvikling i teori, behandlinger, teknologi, monitorering osv. Denne PhD var særligt motiveret af koncepterne Real World Data (RWD) og Real World Evidence (RWE). Disse emner er blevet gransket af mange årsager, såsom den enorme mængde data, der eksisterer i dag og i særlig grad i fremtiden pga. udviklingen af medicinske apparater, der kan måle enorme mængder data hele tiden. Denne afhandling forsøger at forbedre eksisterende metoder til at analysere den slags data. Den overordnede idé, der er udforsket, er at bruge designs, hvor tidsstabil konfounding implicit er justeret for, f.eks. ved at sammenligne folk med dem selv. I det første manuskript belyser vi et konkret problem i case-time-control designet, samt hvordan man skal sample kontroller for at undgå det. I det tredje manuskript foreslår vi en metode til at justere for tidsstabil konfounding uden at være nødt til at skifte til et selv-kontrolleret design. Dette gør det muligt at bruge standard modeller til overlevelseseanalyser uden, at man bør frygte bias som følge af tidsstabil konfounding.

Udvikling i kausal inferens litteraturen og et øget fokus på, hvad vi estimerer samt hvad vores forsknings spørgsmål faktisk er nåede også til farmaindustrien med estimand frameworket.

Indflydelsen fra estimand frameworket på denne PhD er tydeligt i det andet manuskript, hvor vi viser, hvordan en hel klasse af modeller kan føre til unbiased estimation af kausale effekter i cross-over studier uanset, hvor misspecificeret modellen er. Denne robusthed kommer dog med det forbehold, at variansen skal estimeres ud fra den såkaldte influence funktion for at være helt robust overfor misspecification af modellen. Estimand frameworket skinder også tydeligt igennem i det tredje manuskript, hvor en kausal effekt bliver søgt til trods for unmeasured konfounding.

Alt i alt har denne PhD bidraget til de store udviklinger i industrien såvel som med forskning relevant for både RWD og i den kliniske udvikling.

I denne afhandling vil jeg introducere de gængse metoder for analysen af RWD og kausal inferens. Dette vil give et overblik over litteraturen og det felt som manuskripterne i denne PhD tilhører.

# Contents

Preface	i
Summary	iii
Resumé	v
<b>1 Industrial context for the PhD</b>	<b>1</b>
1.1 Real world data (RWD) and real world evidence (RWE)	1
1.2 The estimand framework	3
1.3 Causality and the estimand	4
1.4 Going from Real World Data to Real World Evidence via the estimand	5
<b>2 Epidemiological designs</b>	<b>9</b>
2.1 The cohort study	9
2.2 The case-control study design	11
2.2.1 Matching and the analysis of stratified data	13
2.3 Nested case-control	14
2.4 Poisson regression	15
2.5 Instrumental variables	16
2.6 What's the time?	18
<b>3 Self-controlled designs</b>	<b>19</b>
3.1 The case-crossover design	19
3.2 The case-time-control design	21
3.3 The Self-Controlled Case Series analysis (SCCS)	23
3.4 Crossover design	24
3.5 Modelling and semi-parametric efficiency theory	25
<b>4 Summary of manuscripts</b>	<b>29</b>
4.1 Manuscript I	29
4.2 Manuscript II	30
4.3 Manuscript III	31
<b>5 Perspectives</b>	<b>33</b>
<b>Bibliography</b>	<b>35</b>
<b>Manuscript I: Sampling in the case-time-control design among drug users when outcome prevents further treatment</b>	<b>41</b>
<b>Manuscript II: Unbiased and Efficient Estimation of Causal Treatment Effects in Cross-over Trials</b>	<b>57</b>

---

Manuscript III: Estimating causal effects while adjusting for unmeasured time-stable confounding	87
--	----



# 1 Industrial context for the PhD

This industrial PhD was motivated by current developments within the pharmaceutical industry and the challenges they pose for established statistical methodology. In this section, I will describe these developments and put them into a statistical framework to show the relevance and importance of the statistical problems addressed in this PhD, and how they relate to the problems of the pharmaceutical industry. The two main developments in the industry during my PhD have been:

- Real World Data (RWD) and Real World Evidence (RWE)
- The estimand framework

The main problem for the use of RWD and RWE from a statistical point of view, as we will see later in the dissertation, is the problem of confounding. This problem, along with standard ways to handle it and potential alternative ways of handling it, will be covered extensively in this dissertation. However, a disclaimer is warranted at this stage: no method or solution is perfect, but depends on specific assumptions that may be more or less realistic depending on the context. This is an active area of research, and more research will probably be necessary as long as technology and data sources develop and change. In particular, Manuscripts I and III of this dissertation are a part of this process of incrementally making the scientific community and the pharmaceutical sector better at handling this problem. The estimand framework is most directly related to the statistical field of causal inference. This field deals very directly with the research question, i.e. what we really want to estimate. This might also have been driven by technological and methodological developments in statistical modelling, and is indeed a very active area of research. Causal inference methodology is used extensively in Manuscripts II and III, but unfortunately we didn't succeed in applying the framework in a satisfactory way in Manuscript I for reasons covered later in the dissertation. In the following sections, RWD, RWE and the estimand framework will be described in detail, and the connection to statistical theory will be made explicit so that the reader will know exactly what this PhD is all about.

## 1.1 Real world data (RWD) and real world evidence (RWE)

The Randomized Controlled Trial (RCT) is considered the gold standard for regulatory drug approval, in part due to the fact that randomization ensures that treatment groups are comparable in terms of prognostic factors (Byar et al., 1976). However, there is an increasing amount of data, particularly from medical devices and administrative databases, that is not collected in the context of an RCT. These data can support authorities and pharmaceutical companies in regulatory decision-making. These developments have led to the concepts of RWD and RWE. The US Food and Drug Administration (FDA) defines RWD as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources", and RWE as

”the clinical evidence regarding the usage, and potential benefits or risks, of a medical product derived from analysis of RWD” (US Food and Drug Administration, 2017).

RWD and RWE is not just about using data because they exist. There are several use cases where these data can help the industry and regulators answer questions relating to the efficacy and safety of medicine, where the alternative would either be infeasible or worse. We will go through some use-cases in the following (Franklin et al., 2019).

### **External controls**

External controls, that is a control group from outside the trial, can be used if a new treatment is very promising, and no other satisfactory treatment exists. Especially if the disease is severe, in which case it is perceived as unethical to randomize subjects to placebo. Moreover, external controls could help if treatment is for a rare disease, in which case it might be hard to get enough power in an RCT (Franklin and Schneeweiss, 2017). External controls can help in these contexts (US Department of Health and Human Services (DHHS), US Food and Drug Administration (FDA), 2001). External controls could for example be patients with the same indication as the indication for the new drug, but who are or have been receiving either an older treatment or no treatment.

### **Indication expansion**

RCTs are often made in very specific populations with very specific, sometimes intermediate, outcomes. For example, we might be interested in the effect of a drug on the risk of stroke. In that context, the drug might be approved because it lowers blood pressure, which in turn should lower the risk of stroke. However, with RWD, we may actually be able to measure the effect directly on stroke. RWD could also lead to an estimate of the effect of treatment in a different population, such as patients in a different disease stage or in pediatrics, if the RCT was with adults (Franklin et al., 2019).

### **Post Authorization Safety Studies (PASS)**

A PASS study is “any study relating to an authorized medicinal product conducted with the aim of identifying, characterizing or quantifying a safety hazard, confirming the safety profile of the medicinal product, or of measuring the effectiveness of risk management measures.” (European Medicines Agency, 2017). That means, that PASS studies are conducted after the drug has been approved to ensure safety of the approved drug. PASS can be either clinical trials or non-interventional. A PASS is considered non-interventional if “the medicine is prescribed in the usual way in accordance with the terms of the marketing authorization”, if “the assignment of the patient to a particular therapeutic strategy is not decided in advance by a trial protocol but falls within current practice and the prescription of the medicine is clearly separated from the decision to include the patient in the study” and “no additional diagnostic or monitoring procedures are applied to the patients and epidemiological methods are used for the analysis of collected data” (European Medicines Agency, 2017). In short, the PASS has to use data that would exist even in the absence of the PASS. PASS can be imposed by the authorities or voluntarily (European Medicines Agency, 2017). A non-interventional PASS may be necessary, for example for investigating rare safety

signals, in which case the required sample size for an RCT would be infeasible in terms of resources (Franklin and Schneeweiss, 2017).

### RWD and RWE from a statistical point of view

The issues relating to RWD and RWE are not new from a statistical point of view. Arguably, they are classic problems in the field of pharmacoepidemiology. Many statisticians and epidemiologists would probably call RWD for *observational data*, although RWD arguably is a subset of observational data. Furthermore, RWE might simply be interpreted as the results of *observational studies*. Statisticians and epidemiologists are also well aware of the challenges related to the analysis of such data, in particular the problem of *confounding*. Therefore, it might be tempting for some to dismiss the ideas of RWD and RWE as industry hype. This is an understandable opinion with some truth to it. Nevertheless, I hope statisticians and epidemiologists will see the concepts of RWD and RWE as part of an industry development that makes statistical theory and knowledge more relevant for the real world. In my opinion, the industry needs statisticians and epidemiologists, and it would be incredibly valuable for the movement if statisticians and epidemiologists see themselves as a part of it. I hope this PhD can make the concepts more digestible for academics to the benefit of the industry, regulators and at the end of the day to the benefit of patients.

The concepts of RWD and RWE were the original motivation for this PhD. However, these are not the only hot topics in the pharmaceutical industry. The estimand framework came in the beginning of the PhD, and has had a major impact on how we have approached this project (International Council for Harmonisation, 2019).

## 1.2 The estimand framework

The causal inference literature has increased the focus on what we are estimating in our studies (Hernán, 2018; van der Laan and Rose, 2011). In the beginning of this PhD, this reached the pharmaceutical industry in the estimand framework (International Council for Harmonisation, 2019). The International Council for Harmonisation (2019) defines an estimand as a “precise description of the treatment effect reflecting the clinical question posed by a given clinical trial objective” with the following five attributes:

1. The **treatment** we want to quantify the effect of, and an alternative treatment (possibly no treatment) we want to compare to.
2. The **population** in which we want to estimate the effect.
3. The **variable/endpoint** obtained from each patient.
4. How to handle **intercurrent events**, which are “events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest”, such as rescue medication.
5. The **population-level summary measure** we want to use to quantify the effect of treatment.

An estimand addresses a trial objective. By clearly defining the five estimand attributes, we achieve not only a motivation for the chosen trial design, but also a transparent interpretation of trial data that aligns with the trial objective. The International Council for Harmonisation (2019) does not explicitly mention causal reasoning, but the treatment effects advocated herein are clearly causal by nature.

### 1.3 Causality and the estimand

In this section, I will put RWD, RWE, and the estimand framework into rigorous mathematical causal inference terms. This is necessary for two reasons. First, because causal inference isn't mentioned directly in the regulatory guidance documents related to RWD, RWE or the estimand framework (International Council for Harmonisation, 2019; US Food and Drug Administration, 2017). Second, because it makes the link to my research in this dissertation much more clear. This is not just an exercise in fitting some regulatory documents into a theoretical framework. Causal inference really seems to have played a big role in the thinking behind the estimand framework for example when explaining the effect of treatment as "how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e., had they not received the treatment, or had they received a different treatment)" (International Council for Harmonisation, 2019). Furthermore, the central problem for the use of RWD/RWE, namely confounding, is best described with this framework.

Suppose we are interested in the effect of treatment with the psoriasis medication Brodalumab (Encepp, 2021) on the risk of suicide. Then the *treatment* in the estimand framework is binary and could be denoted by  $A$ , where  $A = 1$  corresponds to treatment with Brodalumab, and  $A = 0$  could be treatment with a comparator drug. Admittedly, this is a strong simplification we make in the interest of illustration, since treatment in practice is complicated by the fact that subjects can get on and off treatment. The variable/endpoint would be suicide, which we can denote by  $Y$ , where  $Y = 1$  corresponds to suicide, and  $Y = 0$  corresponds to no suicide. In causal inference, we define *counterfactual outcomes*, also known as *potential outcomes*,  $Y^a$  corresponding to the outcome we would observe for the subject if they, possibly counter to fact, received treatment  $a$  (Hernán, M. A. and Robins, J. M., 2020). Usually we only observe subjects with one treatment, so the central problem in causal inference is a problem of missing data (see Figure 1.1).

In that case, it is possible to define a causal effect, for instance, the average treatment effect (ATE):

$$E(Y^1 - Y^0).$$

The ATE is the average outcome we would observe if all subjects received Brodalumab minus the average outcome we would observe if all subjects received the comparator treatment. Thus, the ATE indeed compares "how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment" (International Council for Harmonisation, 2019). The ATE could serve as the population-level summary measure in the estimand framework. The population we are interested in could, for example, be a population of psoriasis patients, arguably with a severe stage of psoriasis. The strategy for handling intercurrent events could, for example, be *the treatment policy strategy*, which simply ignores intercurrent events. On that note, intercurrent event strategies may align to several inference strategies, such as competing risk analysis (while on treatment strategy) or mediation analysis (hypothetical

## The data we want

	Treatment	Y1	Y0
1	0	2.5	1.3
2	0	0.5	0.3
3	1	2.5	1.2
4	1	-0.6	-1.1
5	1	0.5	-0.5
6	0	0.0	0.4

## The data we have

	Treatment	Y1	Y0
1	0	NA	1.3
2	0	NA	0.3
3	1	2.5	NA
4	1	-0.6	NA
5	1	0.5	NA
6	0	NA	0.4

Figure 1.1: If we had the choice, we would like to have the data on the left where we observe subjects in both treatment arms. Unfortunately, in real life, we only have the dataset on the right, where we only observe subjects in one treatment arm.

strategy) (VanderWeele, 2016). Now it is clear how to define an estimand theoretically. In the next section, I will go over how to use this in the data analysis.

## 1.4 Going from Real World Data to Real World Evidence via the estimand

In this chapter, we will cover popular ways of getting from estimand to estimate, which facilitates the use of RWD. Let's return to the example of Brodalumab and the risk of suicide. How would we provide an estimate of the treatment effect in this case? In theory, we could have an RCT where we randomize subjects with severe psoriasis to either Brodalumab or the comparator drug. Then it is easy to obtain a causal effect under the following three causal assumptions (Hernán, M. A. and Robins, J. M., 2020):

1.  $A = a \Rightarrow Y = Y^a$  (consistency).
2.  $Y^a \perp\!\!\!\perp A$ , for  $a = 0, 1$  (exchangeability).
3.  $0 < P(A = a)$ , for  $a = 0, 1$  (positivity).

The consistency assumption seems obviously true, and has the practical consequence that we actually observe some of the counterfactual outcomes of interest. This is partly illusory due to notation. As argued in VanderWeele (2009), we could define potential outcomes,  $Y^{a,k}$  which are the outcome the subject would get if treatment  $a$  was given through means  $k$ . Then the notation  $Y^a$  implies that we are assuming that the effect of the treatment doesn't depend on the means with which the treatment was given, i.e., we assume  $Y^a = Y^{a,k}$  for all  $k$  of interest. Assuming that, the consistency assumption states that  $A = a$  implies  $Y = Y^a = Y^{a,k}$ . For example, we assume it doesn't matter whether the subjects receive Brodalumab in an RCT or through other means. We assume that it is meaningful to actually talk about *the* treatment effect. The consistency assumption might for instance be violated if we are trying to estimate the effect of weight loss on life expectancy, since it probably matters whether weight

loss is through diet and exercise, or through surgery (Hernán, M. A. and Robins, J. M., 2020, p. 39). It could be resolved by specifying the means for the intervention so that weight loss through different means are considered different treatments. The bottom line is that the consistency assumption actually *is* a non-trivial assumption. We would usually not worry about this in an RCT, because the treatment is very well-defined, but it should be considered seriously in an actual data analysis with RWD. The exchangeability assumption is sometimes called no-unmeasured confounding and implies that the distribution of outcomes among those randomized to treatment is the same as what the distribution of outcomes would have been among the untreated, if they had instead been treated. The positivity assumption implies that we have subjects in both treatment arms, which is clearly necessary and known in any trial. Then it is possible to estimate the ATE simply by taking averages within groups, due to the following little calculation

$$\begin{aligned} E(Y^1 - Y^0) &\stackrel{2}{=} E(Y^1 | A = 1) - E(Y^0 | A = 0) \\ &\stackrel{1}{=} E(Y | A = 1) - E(Y | A = 0). \end{aligned}$$

The main problem with this trial in practice is that suicide, luckily, is a relatively rare outcome. Thus, the required sample size in order to have enough power in such a trial would be enormous. This would take an unacceptable amount of time, patients, and money to do in practice. There is indeed a need for RWD in this context.

Using RWD is potentially way faster and more cost-efficient, since we might already have all the data needed to answer our research question. The main problem with this approach, and with RWD in general, is the assumption of exchangeability. If, for example, we just compare whether patients on Brodalumab commit suicide more frequently than patients on the comparator drug, then we might have confounding. For example, if the comparator drug is used for subjects with less severe psoriasis, and if disease severity causes suicide, then we will see a higher proportion of suicides among Brodalumab users compared to those on the comparator drug even if there is no direct effect of Brodalumab on the risk of suicide. Researchers try to resolve this issue by replacing the exchangeability assumption with the assumption of conditional exchangeability

$$Y^a \perp\!\!\!\perp A | X, \quad (1.1)$$

for  $a = 0, 1$  for some set of confounders  $X$  that are sufficient for confounding adjustment. That is, we assume subjects with, for example, same sex, age, and disease severity, or whatever variables are included in  $X$ , are exchangeable (Hernán, M. A. and Robins, J. M., 2020). Furthermore, the assumption of positivity has to be changed to

$$P(A = a | X) > 0 \quad a.s.,$$

for  $a = 0, 1$  (Petersen et al., 2012). That is, we need to be able to observe subjects in both treatment arms for all the values of the covariates that we might observe. Then it is possible to estimate the ATE as

$$\begin{aligned} E(Y^1 - Y^0) &\stackrel{(1.1)}{=} E(E(Y^1 | X, A = 1)) - E(E(Y^0 | X, A = 0)) \\ &\stackrel{1}{=} E(E(Y | X, A = 1)) - E(E(Y | X, A = 0)), \quad (1.2) \end{aligned}$$

In practice this can be done in several ways, for example by fitting a model for the expected outcome  $\hat{E}(Y | X, A) = h(X, A, \beta)$ , where  $\beta$  is a vector of parameters, and

plugging into the g-formula (Robins, 1986)

$$\frac{1}{n} \sum_{i=1}^n h\left(X_i, 1, \hat{\beta}\right) - h\left(X_i, 0, \hat{\beta}\right). \quad (1.3)$$

The g-formula uses the empirical distribution of  $X$  for the distribution of  $X$  in (1.2). Alternatively, the ATE can be written as

$$E\left(\frac{I(A=1) \cdot Y}{P(A=1|X)} - \frac{I(A=0) \cdot Y}{P(A=0|X)}\right), \quad (1.4)$$

which motivates estimators where the treatment assignment is modelled resulting in an inverse probability of treatment weighted (IPTW) estimator (Hernán, M. A. and Robins, J. M., 2020)

$$\frac{1}{n} \sum_{i=1}^n \frac{I(A_i=1) \cdot Y_i}{\hat{P}(A=1|X_i)} - \frac{I(A_i=0) \cdot Y_i}{\hat{P}(A=0|X_i)}.$$

Equation (1.4) can be realized from the following calculation (Hernán, M. A. and Robins, J. M., 2020, p. 25):

$$\begin{aligned} E\left(\frac{I(A=a) \cdot Y}{P(A=a|X)}\right) &\stackrel{!}{=} E\left(\frac{I(A=a) \cdot Y^a}{P(A=a|X)}\right) \\ &= E\left[E\left(\frac{I(A=a) \cdot Y^a}{P(A=a|X)} \mid X\right)\right] \\ &\stackrel{(1.1)}{=} E(E(Y^a | X)) \\ &= E(Y^a). \end{aligned}$$

Intuitively, IPTW estimators create a pseudo population where there is exchangeability between treatment and outcome by up-weighting outcomes from subjects who are under-represented in each treatment arm and down-weighting those that are over-represented (Hejazi and van der Laan, 2023; Hernán, M. A. and Robins, J. M., 2020). The probabilities of treatment are also known as propensity scores and can be used in many ways due to the fact that it is the simplest transformation of  $X$  that implies conditional exchangeability, assuming conditional exchangeability given  $X$  (Rosenbaum and Rubin, 1983).

Of course, we never know for sure whether we have conditional exchangeability in practice, or whether there is some unmeasured confounder that affects both treatment and outcome. To return to the example with Brodalumab and suicide, if we compared Brodalumab users to the general population, psoriasis might be causing both use of Brodalumab and suicide. This is a clear example of what is known as confounding by indication (Salas et al., 1999), i.e., the indication for treatment, here psoriasis, is a confounder. This comparison would probably never be made, so a more interesting case is confounding by severity, where it is the severity of the indication, and not just the indication itself, which is the confounder (Salas et al., 1999). That would correspond to patients with more severe stages of psoriasis being more likely to take Brodalumab, as well as committing suicide. Confounding by severity is a special case of confounding by indication (Salas et al., 1999). Confounding by indication in general and confounding by severity specifically are hard to adjust for. One way to try to circumvent the problem

with unmeasured confounding is to use a so-called self-controlled design that implicitly adjusts for time-stable confounding. In the first manuscript, we show how to use such a design called the case-time-control design in a population of drug users (Madsen et al., 2022). An introduction to self-controlled designs, and in particular the case-crossover and the case-time-control designs, will be provided later in this dissertation. Since we never know for sure whether we have conditional exchangeability, it is tempting to blame any discrepancy between results from RCTs and RWE on confounding. However, there are other differences between randomized experiments and observational data than the randomization (Hernán et al., 2008). The effect of treatment is not necessarily the same for all subjects. Therefore, it makes sense to specify who we want an effect of treatment for, which is why an estimand includes the specification of the population in which we are targeting a treatment effect. This is clear in equation (1.2), where the expectation of  $Y$  under treatment may depend on other covariates,  $X$ . The ATE thereby depends on the distribution of  $X$ , i.e., the population we are estimating an effect in. If we find a different treatment effect in an observational study and an RCT, then it might simply be because we are estimating effects in different populations. An advantage with RWD is that we can estimate an effect in broader, and arguably more relevant, populations (Franklin and Schneeweiss, 2017). In Manuscript III, we consider a population of new-users. Generally, we are interested in an effect among those who will actually take the drug of interest, and a population of new users is exactly such a population (Ray, 2003).

All in all, the estimand framework addresses many of the issues pointed out by the causal inference literature, and the causal inference literature and general statistical theory provides the actual tools to be able to use the estimand framework in practice.

## 2 Epidemiological designs

In this chapter, we describe popular methods for analyzing RWD. Even though these designs suffer from bias when we have confounding, knowing these designs still makes it easier to understand what is going on in self-controlled designs since these are inspired by standard epidemiology designs in many cases. This chapter is also helpful for researchers who might want to develop theory for epidemiology or self-controlled designs themselves.

### 2.1 The cohort study

In a cohort study we follow a cohort over time, usually with the purpose of evaluating the effect of treatment on some time-to-event outcome, for example time until death. These types of data are often analyzed with methods from survival analysis due to the fact that outcomes often are *censored*, i.e., some subjects leave the cohort for some reason unrelated to the outcome. This could for example be that they move out of the country, haven't had the event of interest, leave the study etc. Additionally, we might have *late entries* in our data, that is, some subjects are only observed if their event time is after a certain value (see Figure 2.1).

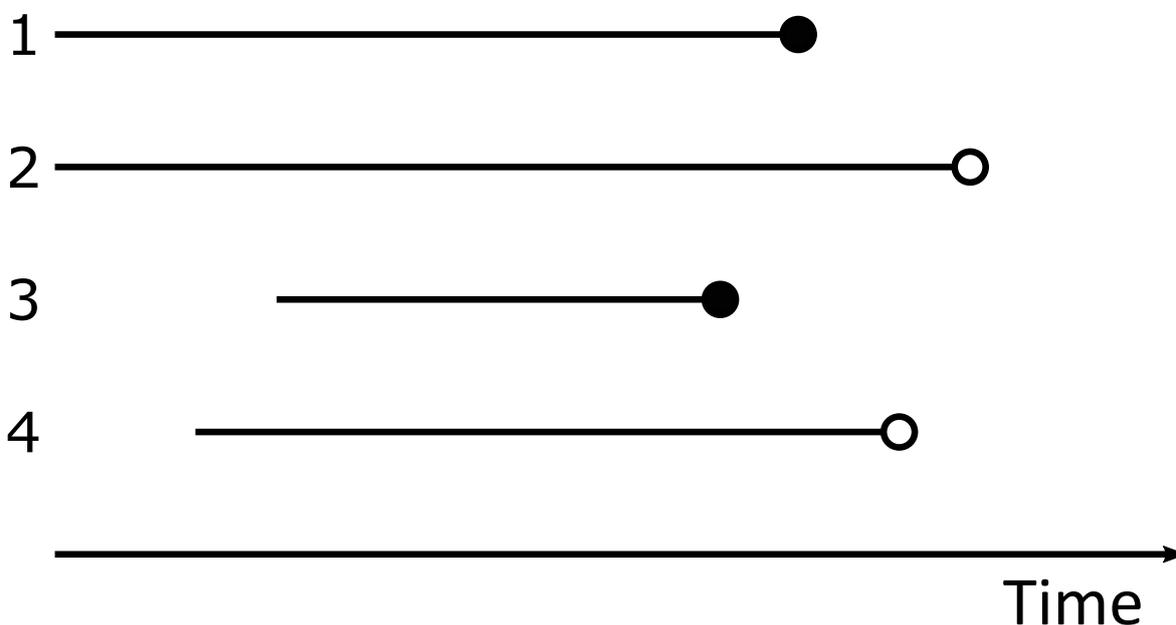


Figure 2.1: Filled dots indicate event and empty dots indicate censoring. The beginning of a line indicates the start of follow-up. Subject 1 is not censored and does not have late entry. Subject 2 is censored but doesn't have late entry. Subject 3 has late entry but is not censored, and subject 4 is censored and has late entry.

A crucial assumption in survival analysis is that censoring, and potential late entry, is uninformative, meaning that the censoring and late entry times are independent of

the event time, at least after conditioning on observed confounders (Martinussen and Scheike, 2006).

We denote the time of the event of interest by  $T^*$  and the censoring time by  $C$ . Then our observed data consists of the observed event time  $T := \min\{T^*, C\}$  and the event indicator  $\Delta := I(T^* \leq C)$ . In survival analysis, it is common to describe the distribution of  $T^*$  in terms of the *hazard function*, which is defined as

$$\alpha(t) = \frac{f(t)}{S(t-)} = \lim_{h \downarrow 0} \frac{P(t \leq T^* < t+h \mid T^* \geq t)}{h},$$

where  $f(t)$  is the density and  $S(t-) = P(T^* \geq t)$  is the survival function just before time  $t$ . There is a one-to-one correspondence between the hazard function and the density, thereby showing that the distribution indeed can be described in terms of the hazard function. The hazard function can be thought of as the instantaneous risk of an event at time  $t$  conditional on being at risk at time  $t$ .

The hazard function is very useful for studying the statistical properties of estimators in survival analysis due to the elegant counting process theory (Andersen et al., 1993). Let  $N(t) = I(T \leq t, \Delta = 1)$  be the counting process telling us whether the subject has been observed having the event at or before time  $t$ , and  $Y(t) = I(T \geq t)$  be the *at-risk indicator*, telling us whether the subject is still under observation with risk of event at time  $t$ . Let  $dN(t) = I(T \leq t, \Delta = 1) - I(T < t, \Delta = 1)$  denote the change in the counting process. The probability of having the event in the interval  $[t, t+dt)$  for some small number  $dt$  given the information until just before time  $t$ ,  $\mathcal{F}(t-)$ , is then  $E(dN(t) \mid \mathcal{F}(t-))$  and  $Y(t)\alpha(t)dt$  such that their difference equals zero. Thereby, heuristically,

$$M(t) = N(t) - \int_0^t Y(s)\alpha(s) ds$$

is shown to be a martingale (Andersen et al., 1993).

The hazard function is often modelled conditional on covariates,  $X$ , and treatment,  $A$ , in the Cox model (Cox, 1972):

$$\alpha(t \mid X, A) = \alpha_0(t) \cdot \exp(\beta^T X + \gamma A).$$

Nothing is assumed about the shape of the baseline hazard function,  $\alpha_0(t)$ , and it is typically estimated non-parametrically, which gives the model a great deal of flexibility. An important assumption in the Cox model is that the Hazard Ratio (HR) is independent of time and given by

$$HR(t) = \frac{\alpha(t \mid x, A = 1)}{\alpha(t \mid x, A = 0)} = \exp(\gamma).$$

Some self-controlled designs estimate quantities that under some conditions can be interpreted as HRs (Farrington et al., 2018; Marshall and Jackson, 1993; Vines and Farrington, 2001). Despite its widespread use, the HR has recently received criticism as a summary measure. One problem is if the true HR is not constant over time, for example if the risk is higher right after treatment than later. In that case, the estimated HR depends on the length of follow-up, or more generally, on the censoring distribution, which we usually don't care about (Hernán, 2010). The other issue is that an HR of, for example 2, after five years doesn't contrast comparable populations, since the population that survived five years of treatment may be healthier than the

population that survived five years without treatment (Hernán, 2010; Martinussen et al., 2020). In terms of the estimand framework, it is unclear that the HR ”summarises at a population level what the outcomes would be in the same patients under different treatment conditions being compared” (International Council for Harmonisation, 2019; Martinussen et al., 2020).

The counterfactual framework does exactly this. In a survival setting, an ATE could for example be risk at time  $\tau$ , corresponding to the outcome variable  $I(T \leq \tau)$  (Hernán, 2010). The ATE has the added advantage that it is easier to understand for most people. Additionally, the ATE is defined in terms of the variables in our data as opposed to the HR, which only makes sense if we happen to have proportional hazards. Generally, it has been argued that the population-level summary measure should be some function of the true data-generating mechanism, and not only well-defined if the model we use happens to be correctly specified (van der Laan and Rose, 2011).

Another important concept in epidemiology research is the concept of *competing risks*. Like censoring, competing risks mean that we stop observing the subject. However, unlike censoring, the subject experiencing the competing event may no longer be able to experience the event of interest. A very common example of a competing risk is death. You cannot get any event of interest after dying (Andersen et al., 2012).

Treating competing risks like censoring doesn’t jeopardize estimation of the hazard function, i.e., a Cox model or Nelson-Aalen estimator can still be applied, but the quantity they estimate is now the cause-specific hazard function. Let  $T$  denote the observed event time, and  $\epsilon$  denote what event we observe. Then the cause-specific hazard function for event  $k$  equals

$$\alpha_k(t) = \lim_{h \downarrow 0} \frac{P(t \leq T < t + h, \epsilon = k \mid T \geq t)}{h}.$$

These cause-specific hazard functions can be used to estimate the cumulative incidence for risk  $k$  by

$$P(T \leq t, \epsilon = k) = \int_0^t \alpha_k(s) \cdot \exp\left(-\sum_{l=1}^K \int_0^s \alpha_l(u) du\right) ds. \quad (2.1)$$

It is important to note that the survival function in the integral,  $\exp\left(-\sum_{l=1}^K \int_0^s \alpha_l(u) du\right)$ , includes the cause-specific hazard rates for all causes, i.e., it is the probability of not experiencing any competing risk before time  $t$ . If competing risks are treated as censoring when estimating the cumulative incidence, for example using a Kaplan-Meier estimator, the estimate of the cumulative incidence will be too high, since that corresponds to using  $\exp\left(-\int_0^s \alpha_k(u) du\right)$  for the survival function in (2.1).

In this PhD we have used survival analysis terminology and notation, among other things, to move from time periods in the case-crossover and the case-time-control designs to a continuous timescale. To our knowledge, competing risks had also not been considered in the case-time-control design until Manuscript I. Furthermore, Manuscript III exclusively considers a survival analysis setting.

## 2.2 The case-control study design

Intuitively, we should compare exposed to unexposed when we want to estimate the effect of treatment on the event of interest. In case-control studies, we do the opposite.

In case-control studies, we compare those with the event of interest, so-called cases, to those without the event of interest, so-called controls (Schulz and Grimes, 2002). When comparing exposed to unexposed, we would compare them in terms of the outcome. In case-control studies, we compare cases and controls in terms of exposure to see if their exposure distributions differ, thereby indicating that treatment and outcome are associated (Clayton and Hills, 1993, p. 153). When doing this, we don't need all the subjects in the background population to find the distribution of exposure among the controls. Therefore, the case-control study design is very cost-efficient and fast compared to using the full cohort in a standard cohort study. This comes at a fairly small price in terms of efficiency since every extra control after, say a few per case, will add a very small amount of extra precision to the estimation of the treatment effect (Clayton and Hills, 1993, p. 153). If there is no relationship between exposure and event, and if we have no confounding, then the distribution of exposures should be the same among the cases and the controls (Clayton and Hills, 1993, p. 153). The odds-ratio (OR) can be estimated from the two-by-two table in Table 2.1 as

$$\frac{D_1/H_1}{D_0/H_0} = \frac{D_1 \cdot H_0}{D_0 \cdot H_1}. \quad (2.2)$$

	Case	Control
Exposed	$D_1$	$H_1$
Unexposed	$D_0$	$H_0$

Table 2.1: Subjects from case-control studies can be divided into four groups in terms of exposure and outcome.

Comparing cases to controls in terms of exposure instead of exposed to unexposed in terms of outcome may seem backwards (Schulz and Grimes, 2002), but is completely legitimate because the OR is *symmetric* in the sense that if we, by accident, mixed up treatment and event status, then we would end up with an OR of

$$\frac{D_1/D_0}{H_1/H_0} = \frac{D_1 \cdot H_0}{D_0 \cdot H_1},$$

which is exactly the same as in (2.2). This explains why it is legitimate in the case-control design to compare cases and controls instead of exposed and unexposed. It simply doesn't matter as long as the focus is on the OR. The symmetry of the OR is also very useful in the case-crossover design, since it allows us analytically to focus on the odds of exposure instead of the odds of event. Moreover, the OR is approximately equal to the risk ratio when the event is rare, which is usually the case when conducting case-control studies, since the reason for choosing a case-control design often is that the event is rare. This follows straight from the fact that odds,  $p/(1-p)$ , are close to risks, or probabilities, when  $p$  is small (Clayton and Hills, 1993, p. 8). Measured confounders can be adjusted for in a logistic regression model. Denote sampled by  $S = 1$  and not sampled by  $S = 0$ . Denote the probability of outcome in the background population with treatment  $e$  by  $p_e$ . Then the probability in the sampled population is found with Bayes' theorem

$$P(Y = 1 | E = e, S = 1) = \frac{P(S = 1 | Y = 1) \cdot p_e}{P(S = 1)}, \quad (2.3)$$

and likewise for  $P(Y = 0 \mid E = e, S = 1)$ . Then the odds are

$$\frac{P(Y = 1 \mid E = e, S = 1)}{P(Y = 0 \mid E = e, S = 1)} = \frac{p_e}{1 - p_e} \cdot \frac{P(S = 1 \mid Y = 1)}{P(S = 1 \mid Y = 0)}.$$

The main point with the above calculation is to show that the odds in the sampled population are the same as in the background population multiplied by the ratio of sampled cases to sampled controls. Typically, all cases will be sampled such that  $P(S = 1 \mid Y = 1) = 1$ . Note that we have assumed sampling independent of treatment above, as we indeed should. The above calculation implies that if we fit a logistic regression to the outcome in the sampled population, then it is only the intercept that is influenced by the sampling (Clayton and Hills, 1993, p. 155).

### 2.2.1 Matching and the analysis of stratified data

A popular strategy to gain efficiency in case-control studies is matching. Matching means that controls are sampled such that a constant ratio between the number of cases and the number of controls is achieved within strata of the data. This could for instance be within each sex and age group. In the following, we use sex as an example of a matching variable to make the arguments easier to follow, but the argumentation would be exactly the same for any other matching variable. Intuitively, matching achieves more efficiency because it ensures that the distribution of men and women is the same among cases and controls, thus avoiding uncertainty due to a chance imbalance in the distribution of men and women. Unfortunately, some of the initial popularity of matching came from the mistaken idea that there is no need to adjust for a variable that has been used for matching. At first, this seems intuitive: the distribution of men and women is already the same among cases and controls, so why would we need to adjust for it? The reason is that the OR is non-collapsible, that is, the OR conditional on sex is not the same as when not conditioning on sex (Sjölander et al., 2016). In fact, the bias will predictably be towards an OR of one if we fail to adjust for sex (Clayton and Hills, 1993, p. 180). The necessity of adjusting for sex can also be realized from (2.3) by letting the sampling probabilities depend on sex, in which case the intercept in the logistic regression also depends on sex. Furthermore, the ability to interpret the effect of sex on outcome is lost when using it for matching, since the distribution of men and women is the same among cases and controls after matching (Clayton and Hills, 1993, p. 179). There is also the issue of overmatching which is when the matching variable is strongly related to exposure but not to event. In this case, matching may *reduce* precision since the amount of variability in exposure within groups of the matching variable will be reduced (Clayton and Hills, 1993, p. 181). For example, if sex strongly predicts exposure then we will have strata of men and women where most subjects within each stratum have the same exposure, thus leading to an estimate with less precision than we would have obtained if we did not match.

Estimation in matched case-control studies relies on logistic regression. This approach works well if the number of strata is small and the number of subjects in each stratum is big. However, that is not the case if for instance subjects are individually matched, that is, if we want each case and its matched controls to constitute a stratum. In that case, the data will typically be analyzed by use of the Mantel-Haenszel estimator or conditional logistic regression. To perform inference, we produce one two-by-two table as in Table 2.1 for each stratum. Denote the different strata by  $k$ , and the numbers in the corresponding two-by-two tables  $D_1^k, D_0^k, H_1^k, H_0^k$ , and the total number

of subjects in stratum  $k$  by  $n^k$ . Then the Mantel-Haenszel estimator of the OR equals (Clayton and Hills, 1993, p. 177)

$$\frac{\sum_{k=1}^K \frac{D_1^k H_0^k}{n^k}}{\sum_{k=1}^K \frac{H_1^k D_0^k}{n^k}}. \quad (2.4)$$

The Mantel-Haenszel estimator is a popular estimator of the OR and is close to the maximum likelihood estimator based on the hypergeometric distribution obtained by conditional logistic regression when the OR is close to one (Clayton and Hills, 1993, p. 177). Alternatively, matched case-control studies are analyzed with conditional logistic regression (Clayton and Hills, 1993, p. 234). Denote the binary outcome for subject  $i$  in stratum  $k$  by  $Y_{ik}$  and the covariates for the subject by  $X_{ik}$ . Then the conditional logistic regression models the probability of outcome by

$$\text{logit}[P(Y_{ik} = 1 \mid X_{ik} = x)] = \alpha_k + \beta^T x.$$

The intercepts are stratum specific and the rest of the regression looks like standard logistic regression. The likelihood contribution from each subject becomes the conditional probability of the subject being a case, given the total number of cases and controls in the stratum the subject is in. The stratum specific parameters  $\alpha_k$  cancel out in these calculations and therefore don't need to be estimated (Clayton and Hills, 1993, p. 293). Case-crossover designs can be seen as a special type of individually matched case-control studies, where control subjects very directly are replaced by control times and all the analysis methods are the same as for individually matched case-control studies. Individual matching is also used in the nested case-control design (Clayton and Hills, 1993, ch. 33), which we will describe in the next section.

### 2.3 Nested case-control

The nested case-control study design is an efficient alternative to a cohort study. In the nested case-control study, controls are sampled for each case among all subjects at risk at the time of event (see Figure 2.2) (Clayton and Hills, 1993, p. 330). Matching can be employed in the sampling just as in the case-control design. The data are analyzed with conditional logistic regression, where each case and its matched controls constitute a stratum (Clayton and Hills, 1993, p. 331). In that context, it seems natural to interpret the estimated treatment effect as an OR, but in fact it turns out the estimated treatment effect is an estimate of the HR (Borgan et al., 1995). The main advantage with a nested case-control study over cohort studies is that fewer subjects are needed, thereby making the analysis easier and more cost-efficient compared to using the full cohort, since data collection is only necessary for the cases and the sampled controls (Clayton and Hills, 1993, p. 329). This enables collection of expensive measurements such as biomarkers that would not be feasible for the entire cohort. Furthermore, the analysis is easier in case of time-dependent confounders since we only need the value of the confounders at the time of sampling for the analysis (Borgan et al., 1995). The main drawback is when we are interested in the effect of treatment on several outcomes. In that case, we need to sample and make separate analyses for each outcome. As a solution to this, it has been proposed to sample controls at the time of recruitment instead of at the time of event. This design has been termed the case-cohort design and has similar logistical advantages as the nested case-control study, but can handle

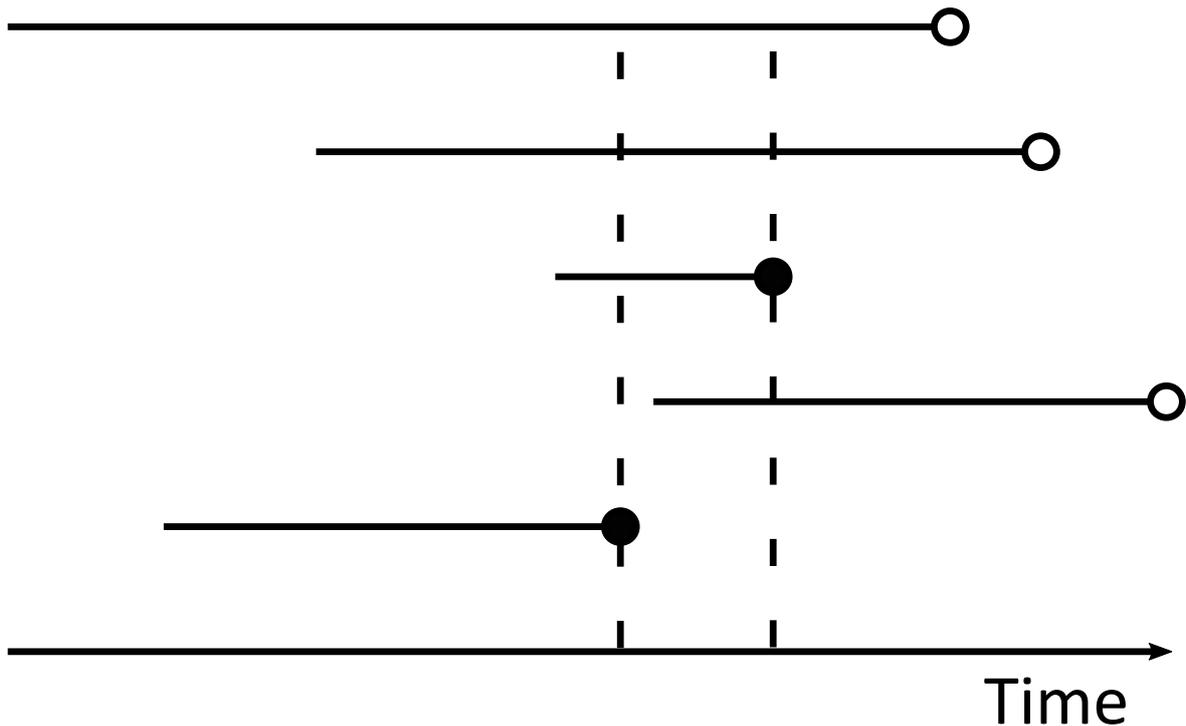


Figure 2.2: Filled dots indicate events, and empty dots indicate censoring. Controls are sampled among all subjects at risk at the time of event, including future cases.

several outcomes of interest (Clayton and Hills, 1993, p. 331). The statistical analysis of data from a case-cohort design is unfortunately quite challenging and will not be covered here (Clayton and Hills, 1993; Barlow et al., 1999).

## 2.4 Poisson regression

Poisson regression is a popular type of generalized linear model (glm) that is used for count data, i.e., data where the outcome  $Y$  can take the values  $0, 1, \dots$  (McCullagh and Nelder, 1998, ch. 6). The assumption is that this count comes from a Poisson distribution with a mean conditional on covariates  $X$  of the form

$$g(E(Y | X = x)) = \beta^T x, \quad (2.5)$$

where  $g$  is a *link function*, typically the logarithm, to ensure a positive expected number of counts (McCullagh and Nelder, 1998, ch. 6). In epidemiology, Poisson regression is particularly used to model data from contingency tables, such as the hypothetical Table 2.2 (Clayton and Hills, 1993, p. 227). Subjects can contribute with person-years to more than one age group if they happen to be between 30 and 40 years old in a part of the study and 41 to 50 years old later on (Clayton and Hills, 1993, p. 227). A Poisson regression model can be fitted to these types of data by using person-years as offset, for example by the following R code (R Core Team, 2022):

```
glm(cases ~ offset(log(personyears)) + Age + Exposure, family = "poisson")
```

Cases	Person-years	Age	Exposure
4	608	30-40	0
2	311	30-40	1
5	945	41-50	0
7	732	41-50	1
9	438	51-60	0
2	944	51-60	1

Table 2.2: Contingency table for Poisson regression. The data have three age groups and two treatment groups.

According to this model the expected number of cases, for example in row one in Table 2.2 is

$$\exp(\log(608) + \beta_0 + \beta_{Age_0} + \beta_{Exp_0}) = 608 \cdot \exp(\beta_0).$$

Treatment group 0 and age group 30-40 have been assumed to be reference levels in the calculation above. The point of the calculation is to show that the model states that the number of events in one person-year follows a Poisson distribution with a mean of  $\exp(\beta^T x)$  like in (2.5). The offset simply multiplies this expected number with the number of person-years. The estimates we get from such a model are rate ratios. For the effect of treatment, the rate ratio is the relative difference in the expected number of events in a given time frame among exposed compared to unexposed within age groups (Clayton and Hills, 1993). The likelihood approaches the likelihood from a Cox model when the age bands become smaller. Thus, the rate ratios will tend to be similar to the HR from a Cox model (Clayton and Hills, 1993, p. 299). Poisson regression seems to have been an inspiration behind the self-controlled case series analysis (SCCS), presented in the next chapter (Farrington et al., 2018).

## 2.5 Instrumental variables

Instrumental variables is an analysis method for observational data that does not rely on conditional exchangeability. Instead, treatment is initially replaced by a variable, known as an instrument, in the analysis. The intuition is that we approximate the effect of treatment on outcome with the association between an unconfounded proxy for treatment and outcome. The instrument needs to fulfill the following formal criteria (Baiocchi et al., 2014):

1. The association between instrument and outcome is unconfounded.
2. The instrument causes treatment. Thereby, the instrument outcome association reflects the treatment outcome association.
3. There is no direct effect of instrument on outcome. Such an effect would render the association between instrument and outcome a biased estimate of the treatment outcome association.

These three requirements can be summarized by the directed acyclic graph (DAG) in Figure 2.3.

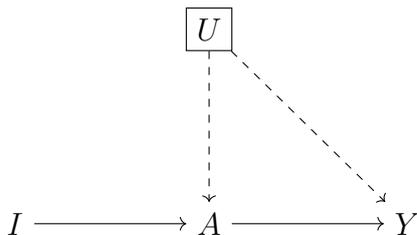


Figure 2.3: DAG for instrumental variable setup.  $I$  is the instrument,  $A$  is exposure,  $U$  is a set of unmeasured confounders, and  $Y$  is the outcome.

Note that these requirements make it a challenge to identify an instrumental variable in practice. However, nice examples of instrumental variables do exist. A particularly nice example is in randomized trials where some subjects don't take the treatment they were randomized to, also known as non-compliance (Baiocchi et al., 2014). If someone asked you whether we should use randomized treatment or actual treatment in our estimation, your first thought might be actual treatment. After all, this is the treatment the subjects actually received. However, it is clear that actual treatment might be confounded, whereas the randomized treatment is not. The randomized treatment often has no direct effect on outcome, either. Randomized treatment therefore often satisfies all the conditions for being an instrumental variable, and we might proceed by using randomized treatment, which in the pharma industry is known as an intent-to-treat analysis. Another result about instrumental variables might become clear from this example: we will tend to underestimate the true treatment effect by using randomized treatment instead of actual treatment, especially if we have a lot of non-compliance. This is adjusted for in an instrumental variable setting, for example by using the Wald formula

$$\frac{E(Y | I = 1) - E(Y | I = 0)}{P(A = 1 | I = 1) - P(A = 1 | I = 0)},$$

which estimates the effect among the compliers (Baiocchi et al., 2014). The numerator is the intent-to-treat estimator, and the denominator is the proportion of compliers. A bit of terminology might be necessary here: a complier is not just someone who takes the treatment they are randomized to. A complier is someone who would take the randomized treatment, no matter what it happened to be. The terminology is summarized in Table 2.3.

	$A^1$	$A^0$
Complier	1	0
Always-taker	1	1
Never-taker	0	0
Defier	0	1

Table 2.3:  $A^i$  denotes treatment when randomized to treatment  $i$ .

The intuition for why the estimate is an effect only for the compliers is also clear: we don't have any potential outcomes under treatment among the never-takers and no potential outcomes without treatment among the always-takers. Thus, we can't

extrapolate results to those subpopulations without further assumptions. The Wald estimator depends on an assumption of no-defiers in order for the denominator to be the proportion of compliers (Baiocchi et al., 2014).

Instrumental variables enable unconfounded estimation of a causal effect for observational data. The main downside is that good instruments are rare (Baiocchi et al., 2014).

## 2.6 What's the time?

Often different timescales will matter for the risk of outcome. Dependent on the specific study design, time can refer to calendar time, age, time since exposure etc. (Clayton and Hills, 1993, ch. 6). But what is the right timescale to use, and do you have to choose? Usually, age is important in epidemiological studies, since the risk of most outcomes vary substantially with age. Therefore, it makes sense to have age as your timescale, but it is not necessary since age can be included as a covariate in a regression model (Clayton and Hills, 1993, ch. 6). In fact, methods have been proposed, where it is not necessary to choose a specific timescale, but several timescales can be handled simultaneously in a flexible way, for example using splines (Iacobelli and Carstensen, 2013; Crowther and Lambert, 2014). This requires parametric modelling of the baseline hazard, but this might not be too big a problem considering the flexibility enabled by the methods, especially considering how similar results from Cox modelling and Poisson models tend to be (Clayton and Hills, 1993, p. 299). Nevertheless, the timescale is probably an underappreciated difference between RCTs and observational studies (Hernán et al., 2022). The difference of timescale has in some cases been the main reason for different results when comparing RCTs and observational studies in practice (Hernán et al., 2022). Specifically, time zero is more or less always the time of randomization, which is most likely the main motivation behind new-user designs (Ray, 2003). In this PhD, timescales have played a role in Manuscript I, where we consider calendar time as our timescale due to the fact that the time-trend we want to eliminate depends more on calendar time than age, although concerns about confounding by age could be alleviated by matching controls on age as well. In Manuscript III, time zero refers to the time of first treatment, which is a necessary requirement for the methodology of that paper to apply. Specifically, the choice of timescale made it possible to adjust for unmeasured time-stable confounding simply by adjusting for the number of treatment administrations.

## 3 Self-controlled designs

The basic idea in self-controlled designs is to compare subjects with the event of interest, also known as cases, to themselves at different time points. This has the added advantage that we only need cases in the analysis. The cases are their own controls (Maclure et al., 2012). It might seem challenging to say anything about the risk of outcome if we only observe subjects who have that outcome. However, our research question is not to determine the risk of outcome, but to determine whether treatment affects the risk of outcome. This is possible because we observe subjects over time, both under treatment exposure and without treatment exposure. Then it is possible to determine whether the risk of event is higher at times of exposure than at times without. It has the additional advantage that conditional exchangeability seems more reasonable since the subjects in active treatment are the same as the ones without treatment, although at different time points. However, it subtly answers a different research question than normally, namely “why now?” instead of “why me?” (Maclure, 2007). All the information in the estimation comes from subjects who have different exposure status at different times. Thereby, we get an effect in a population that changes treatment status. An extreme example would be in the case of sex, where the estimate would change from ‘the effect of being male’ to ‘the effect of having switched sex to male’, which is an entirely different research question (Maclure et al., 2012).

Unfortunately, even here conditional exchangeability is not something we can take for granted, since some confounders may change over time (Mittleman and Mostofsky, 2014). However, self-controlled designs effectively adjust for confounders that are stable in time. Perhaps the simplest form of a self-controlled design is the case-crossover design, which we will describe in the following along with the case-time-control design, which is an extension of the case-crossover design that handles time-trends in exposure, and which is the topic of Manuscript I.

### 3.1 The case-crossover design

The case-crossover design was developed by Maclure (1991). The idea is very simple: compare the exposure status at the event time to the exposure status at a number of time points before, known as reference times (see Figure 3.1). If subjects are exposed at the time of event more frequently than at the reference times, then it might be because treatment causes the event. The standard description of the design is in terms of time-periods. Here, I will describe the design in continuous time. Things like timescale, censoring and competing risks are more natural to discuss with a continuous timescale, and it reflects the practical use of the design with registry data, where we have, more or less, exact days of events and treatments, better.

The design is a perfect example of going from comparing subjects to times. It is the self-controlled design equivalent of a matched case-control study, where control subjects have been replaced by control times and each subject constitutes a stratum. In this light, it is very clear that the design is self-controlled, and every case quite

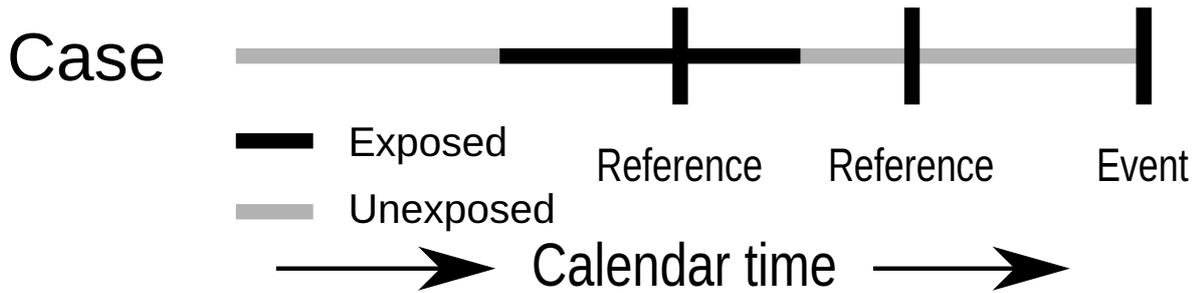


Figure 3.1: The exposure status at the time of event is compared to the exposure statuses at the two reference times here. Note that the Figure is in continuous time.

literally serves as their own control. The analysis methods are, not surprisingly, also the same as in individually matched case-control studies.

Denote the treatment status for subject  $i$  at time  $t$  by  $E_i(t)$ , the exposure at the time of event by  $E_i(t_\tau)$ , and the exposure status at the reference times by  $E_i(t_1), \dots, E_i(t_K)$ . Then the Mantel-Haenszel estimator in this setup equals

$$\frac{\sum_{i=1}^n \sum_{k=1}^K I(E_i(t_\tau) = 1, E_i(t_k) = 0)}{\sum_{i=1}^n \sum_{k=1}^K I(E_i(t_\tau) = 0, E_i(t_k) = 1)}.$$

The estimator is particularly simple in the case with only one reference time, where it becomes

$$\frac{\sum_{i=1}^n I(E_i(t_\tau) = 1, E_i(t_1) = 0)}{\sum_{i=1}^n I(E_i(t_\tau) = 0, E_i(t_1) = 1)}.$$

This is even easier to understand when looking at it in a two-by-two table as in Table 3.1. The Mantel-Haenszel estimator is simply the number of subjects exposed at event but not at reference divided by the number of subjects exposed at reference but not at event.

		Reference	
		Unexposed	Exposed
Event	Unexposed	a	b
	Exposed	c	d

Table 3.1: With one reference time point, the Mantel-Haenszel estimator simply becomes  $c/b$ .

The Mantel-Haenszel estimator has been shown to be unbiased for the HR under the following stratified Cox regression model

$$\lambda_i(t) = \lambda_{i0}(t) \cdot \exp(\beta \cdot E_i(t)),$$

when there is no time-trend in exposure and the outcome is rare (Vines and Farrington, 2001). The baseline hazard function  $\lambda_{i0}(t)$  is subject specific and represents the self-adjustment in the design. It is quite remarkable that nothing has to be assumed about the subject specific baseline hazard functions in order to estimate the HR with the case-crossover design. The requirement that there are no time-trends in exposure is very

intuitive: if the probability of being exposed increases over time, then the estimated OR will be greater than one simply because the event time happens to be later than the reference times, at least assuming there is not a strong protective effect of treatment. In the language of the case-control design, this can be seen as sampling controls from a different cohort than cases, namely a less exposed cohort. Time-trends essentially make time periods non-exchangeable (Mittleman and Mostofsky, 2014).

Traditionally, the case-crossover design has been described in terms of time periods, so that  $t_\tau$  and  $t_k$  are periods. In this framework, the HR can be given a causal interpretation (Shahn et al., 2022).

Conditional logistic regression has, not surprisingly given the parallel to case-control studies, also been suggested as an analysis method in the case-crossover design. The main advantage over the Mantel-Haenszel estimator is that maximum likelihood provides asymptotic optimality, thereby ensuring optimal use of data from the design (Vines and Farrington, 2001). Unfortunately, it is biased when the exposures within subjects are not globally exchangeable, unless we restrict the design to one control period, in which case the conditional logistic regression is unbiased for the HR as long as we don't have a time-trend in exposure. Global exchangeability implies no time-trend in exposure, but also that the dependence between exposure statuses at different time points are the same. This assumption is for example violated if exposure statuses at time-points closer to each other are more correlated than exposure statuses at more distant time points (Vines and Farrington, 2001). Similar results were discovered in the case-time-control design in simulation studies (Jensen et al., 2014). This is a big disadvantage compared to the Mantel-Haenszel estimator, which doesn't require global exchangeability in the case-crossover design (Vines and Farrington, 2001), and possibly also not in the case-time-control design, where it could be applied for cases and controls separately. The case-time-control estimate of the OR could then be obtained by dividing the case OR with the control OR. The bias of the conditional logistic regression can be alleviated through weighting methods that also enable adjustment for measured time-dependent confounders (Kubota et al., 2021).

A simple solution to the problem of time-trends in the case-crossover design is simply to have reference periods both before and after the event time, as is the case in the bidirectional case-crossover design (Navidi, 1998). The theoretical justification for this solution is that the usual case-crossover design can be seen as suffering from selection bias when there is a time-trend in exposure. That is, controls will systematically be less exposed than cases when sampling of control times only happens before the event time, but not after (Greenland, 1996). This bias is removed by allowing control times after the event time. However, this does not work if the event prevents subsequent treatment, such as if the event is terminal, unless we know what the treatment status would have been in the absence of the outcome. This is for example possible in studies where the exposure is pollution. Unfortunately, this is rarely the case in pharmacoepidemiology. The case-time-control design is an alternative solution that we will describe in the following.

## 3.2 The case-time-control design

The problem with time-trends in exposure may unfortunately often be a very real one. This could for example be the case if we want to test an effect of a drug in the period right after it was approved, in which case the usage must be expected to increase

over time. Therefore, Suissa (1995) proposed to sample controls, and compare their exposure status at the times of event and reference for their matched case (see Figure 3.2). Any relationship found among the controls must be attributed to the time-trend



Figure 3.2: Source: Madsen et al. (2022). Controls are compared at the event and reference times of their matched case in the case-time-control design.

in exposure, since the controls in fact didn't have the event of interest. The design was originally described for a situation with only one reference time, which we will use in the following, but it was extended to handle multiple reference times in Jensen et al. (2014).

Denote whether subject  $i$  is a case or a control by  $G_i$ . Then the data are modelled with a conditional logistic regression on the form

$$\begin{aligned}\text{logit}(P(E_i(t_1) = 1)) &= \alpha_i, \\ \text{logit}(P(E_i(t_\tau) = 1)) &= \alpha_i + \beta_1 + \beta_2 \cdot G_i.\end{aligned}$$

The self-adjustment in this design is reflected in the subject specific parameter  $\alpha_i$ . The OR for controls is  $\exp(\beta_1)$  and reflects the time-trend of exposure, whereas the OR for cases is  $\exp(\beta_1 + \beta_2)$ . This reflects the effect of time-trend through  $\exp(\beta_1)$ , and the effect of treatment on outcome, reflected by  $\exp(\beta_2)$ . The target of estimation in the case-time-control design is  $\exp(\beta_2)$ , which is the part of the OR for cases that can't be attributed to a time-trend in exposure. Unfortunately, we can't interpret this OR as an HR because the OR among controls depends on the joint distribution of treatment and the tail of the event time distribution, which we have no access to (Madsen et al., 2022). A more basic and important assumption of the design is that cases and controls have the same time-trend of exposure. It is argued in Suissa (1995) that this assumption is more realistic if we match controls on different covariates, such as sex and age. It implies that the unmeasured confounders, we hope to adjust for by using the design, don't have an effect on the time-trend of exposure. Arguably, this reintroduces the problem of lack of exchangeability to some extent (Greenland, 1996; Suissa, 1998). The case-case-time design has been proposed as a solution to this problem. In the case-case-time design, controls are sampled exclusively among future cases. This is supposed to achieve a similar time-trend among cases and controls, but there is a trade-off between having controls with an event close to the event time of the case in order to ensure a similar time-trend, and not having an event too close to the event time of the case to avoid the time-trend being due to the event of the control. Another problem is that the design is conditioning on the future (Hallas and Pottegård, 2014).

In Manuscript I, we consider the case-time-control design in a population of drug users. This has no effect on the OR for cases, since untreated don't contribute to the estimation. Nevertheless, it implies sampling controls from a different population that is hopefully more similar to the cases. We show how this results in a bias if controls are sampled traditionally. Furthermore, we show how to sample controls to avoid this bias in a setup that handles random censoring and competing risks in a continuous time setting.

### 3.3 The Self-Controlled Case Series analysis (SCCS)

The SCCS was developed to analyze the effect of vaccines in a self-controlled way (Farrington, 1995). Unlike the self-controlled designs covered so far, the SCCS analysis is not based on logic from the case-control design (Vines and Farrington, 2001). Instead, the design follows the logic from Poisson regression. More concretely, each subject has an observation period  $[a_i, b_i]$  in which a number of events happen according to a Poisson process with intensity rate function  $\lambda_i(t | x_i, y_i)$ , where  $t$  is time,  $x_i$  denotes exposure, and  $y_i$  denotes time-invariant confounders for subject  $i$  (Farrington et al., 2018). The design has since been extended in several ways, but the standard version of the design has specific age groups and treatment groups, just like Poisson regression. The intensity rate function when subject  $i$  is in age group  $j$  with exposure level  $k$  can be written as  $\lambda_{ijk}$ . Several parameterizations for  $\lambda_{ijk}$  exist, the most simple being (Farrington et al., 2018, p. 25)

$$\lambda_{ijk} = \phi_i \cdot \exp(\alpha_j + \beta_k).$$

How do we estimate the parameters from this model? The trick is to only consider cases. Then the likelihood contribution from each subject becomes the conditional probability of having the event times they had, given that they had at least one event. Denote the number of events subject  $i$  had in age group  $j$  under treatment  $k$  by  $n_{ijk}$  and the total amount of time spent in age group  $j$  under treatment  $k$  by  $e_{ijk}$ . Then the log-likelihood up to a constant becomes

$$\sum_{i=1}^n \sum_{j,k} n_{ijk} \cdot \log \left\{ \frac{\exp(\alpha_j + \beta_k) e_{ijk}}{\sum_{r,s} \exp(\alpha_r + \beta_s) e_{irs}} \right\},$$

the main point being that the subject specific parameters  $\phi_i$  are eliminated. Thereby, all confounding that is time-stable with a multiplicative effect on the intensity is implicitly adjusted for (Farrington et al., 2018). Several extensions exist, most importantly probably being the ability to include other covariates in the regression. There are four key assumptions in the SCCS design (Farrington et al., 2018, p. 18):

1. Events arise according to a Poisson process. Specifically, the number of events in non-overlapping time periods are independent. Furthermore, the design works if the outcome is binary assuming the event is rare, basically due to the fact that the binomial distribution and the Poisson distribution are similar when the probability of event is small.
2. Events do not influence the observation period. This assumption is for example violated if the event is stroke and stroke leads to death for some subjects, thus

ending their observation periods at the times of death (Farrington et al., 2018, p. 18).

3. Event does not influence exposure. This assumption is for example violated if the event is a contraindication to treatment. For example, some drugs are not allowed for clinical depression, so if we are interested in the effect of treatment on depression, this assumption is violated.
4. Cases are either the whole or a random subset of a well-defined population. This assumption is for example violated if cases are included in the study due to a suspicion that their event was caused by treatment (Farrington et al., 2018, p. 19). This assumption is not unique to the SCCS design, and reflects the focus on the population of interest in the estimand framework.

The main drawback with the SCCS design is that assumptions two and three above are violated when the event is terminal. Extensions of the design to handle such situations have been proposed, but rely on additional non-trivial assumptions (Farrington et al., 2018, ch. 7).

### 3.4 Crossover design

In crossover designs, subjects are randomized to one of several sequences of treatments. The simplest and most common example is with one treatment and one placebo. In that case, subjects are randomized to either get treatment first and then placebo or vice versa (Senn, 2002). The two periods have a washout period in between in order to ensure that the outcome in period two is unaffected by the treatment in period one (see Figure 3.3).

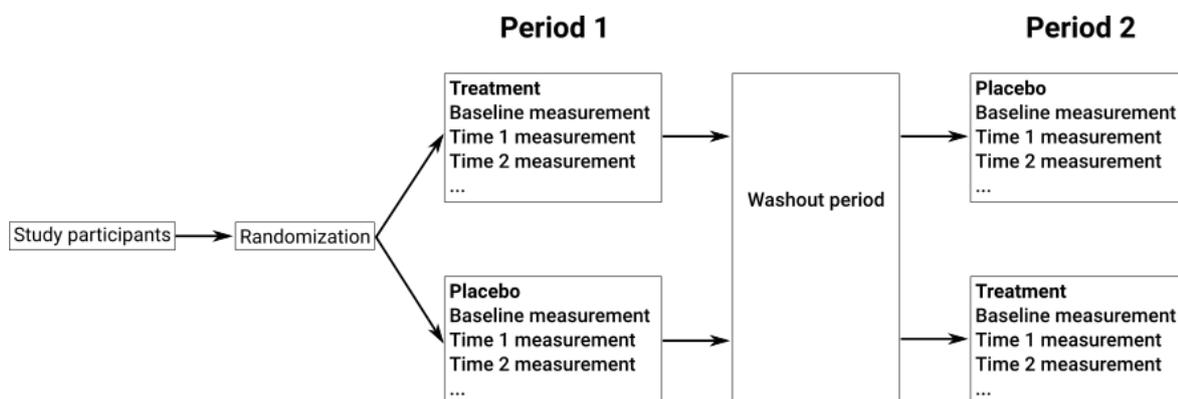


Figure 3.3: Source: Manuscript II.

Unlike the other self-controlled designs, the crossover design is made with randomized studies in mind. Thus, the point of crossover trials and comparing subjects to themselves is not to avoid bias due to lack of exchangeability. Instead, the point of crossover trials is to gain precision in order to save sample size, and consequently time and money as well (Senn, 2002, p. 7). They have the added advantage that they enable principal stratum analyses, i.e. estimation of treatment effects for those subjects who

can tolerate the drug of interest (Matthews et al., 2022; International Council for Harmonisation, 2019). Such an analysis is typically challenged by the fact that we don't know who in the comparison group would tolerate the drug of interest. All subjects get all treatments in crossover trials, and therefore we know whether they can tolerate the drug of interest or not.

Manuscript II deals with crossover designs and was motivated partly by our interest in designs with self-adjustment and partly by the lack of use of the estimand framework (International Council for Harmonisation, 2019) in Thorough QT (TQT) studies, which is a specific type of crossover trial conducted to ensure drug safety before drug approval.

Randomization opens up a whole new world of opportunities. With observational data, we might be biased even if we have all the right variables in our dataset in order to ensure conditional exchangeability. Bias may emerge simply if we misspecify the model or the functional form of the effects of covariates. In practice, we rarely know the true model or functional form of the effects of covariates. In randomized trials, such as the crossover design, it turns out we can estimate treatment effect unbiasedly even without adjusting for covariates, although such non-parametric estimators can often be seen as special cases of models with only treatment included as a covariate. If we do adjust for covariates in an outcome regression model, we will often estimate causal treatment effects unbiasedly even if crucial model assumptions are wrong (Rosenblum and Steingrimsson, 2016; Bartlett, 2018). In Manuscript II we extend some of these robustness results to a large class of working regression models in crossover designs. In this context, one important point is that even though unbiasedness is ensured by randomization, model based standard errors may still be wrong. Fortunately, robust standard errors can be estimated from the influence function. The theory behind influence functions will be described in the next section.

### 3.5 Modelling and semi-parametric efficiency theory

Anyone who has studied statistics has spent quite some time on modelling. But why do we model data? The ATE is not defined in terms of a model, after all. One answer might be to avoid confounding and be able to estimate the ATE from (1.2). This is indeed a very good reason for using a model. However, models are also used in RCTs where confounding is less of a problem due to the randomization. Missingness can break the randomization and motivate the use of models, but models also serve another purpose, which arguably is their main purpose in RCTs. Models help us gain efficiency, which leads to a smaller required sample size. Unadjusted estimators, typically averages within treatment arms, are known to be unbiased for causal effects due to randomization. However, it is not trivially true that we are unbiased if we fit a model to data and plug into (1.3). Somewhat surprisingly, this turns out to be the case in one-period trials when the working model is a generalized linear model (glm) with a canonical link function, no matter the true data-generating mechanism (Bartlett, 2018; Rosenblum and Steingrimsson, 2016; Wang et al., 2021). Manuscript II extends these results to cross-over trials with linear mixed models as working models.

In randomized trials, the working model becomes a tool to gain efficiency rather than an attempt at finding the true data-generating mechanism, since unbiased estimation would be easy to obtain even without a model. The robustness results of the

effect estimates unfortunately don't extend to the variance of the effect estimates. The inverse information may very well lead to a biased estimate of the variance under a misspecified model. This is particularly clear when the working model is a standard linear regression model in a cross-over trial. It follows from the main result in Manuscript II that linear regression is unbiased for the treatment effects of interest despite assuming all observations independent, including the different observations from the same subject on the same day. Fortunately, it is possible to estimate the asymptotic variance correctly and get asymptotic normality of our estimator. In order to obtain that, we need a bit of semi-parametric theory (Tsiatis, 2006, p. 23):

**Definition 1 (Asymptotically linear estimators)** *Let  $Z_1, \dots, Z_n$  be iid random vectors of data. We say that  $\hat{\beta}_n$  is an asymptotically linear estimator of the  $q$ -dimensional vector  $\beta_0$  if there exists a  $q$ -dimensional random function  $\varphi(Z)$  with  $E(\varphi(Z)) = 0$  and*

$$\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(Z_i) + o_p(1),$$

where  $o_p(1)$  is a term that converges to zero in probability. The function  $\varphi$  is called the influence function for  $\hat{\beta}_n$ .

Influence functions are almost surely unique, and examples of influence functions include (Tsiatis, 2006)

- $\varphi(Z_i) = (Z_i - \mu_0)$  for the sample average, where  $\mu_0$  is the true mean.
- $\varphi(Z_i) = I(\beta_0)^{-1} S(\beta_0)$  for the maximum likelihood estimator, where  $I(\cdot)$  is the information matrix and  $S(\cdot)$  is the score function.
- $\varphi(Z_i) = - \left[ E \left( \frac{\partial m(Z, \beta_0)}{\partial \beta^T} \right) \right]^{-1} m(Z, \beta_0)$  for  $m$ -estimators, i.e. estimators that solve the equation

$$\sum_{i=1}^n m(Z_i, \beta) = 0.$$

It follows from the central limit theorem and Slutsky's theorem that asymptotically linear estimators have the following asymptotic behavior:

$$\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left( 0, E \left( \varphi(Z) \varphi(Z)^T \right) \right).$$

It follows from this result that if we know the influence function of an estimator, then we can estimate its asymptotic variance by

$$\frac{1}{n^2} \sum_{i=1}^n \hat{\varphi}(Z_i) \hat{\varphi}(Z_i)^T,$$

where  $\hat{\varphi}(Z_i)$  is the estimated influence function. This estimator is (asymptotically) valid, no matter the true data-generating mechanism (Tsiatis, 2006). Unfortunately, the asymptotic variance from the influence function can be too low in small trials. Finite sample adjustments have been proposed for specific types of working models, but the simplest and most popular adjustment is to use

$$\frac{1}{n(n-1)} \sum_{i=1}^n \hat{\varphi}(Z_i) \hat{\varphi}(Z_i)^T$$

for the variance estimate, and then base the confidence intervals on the  $t$ -distribution with  $n - 1$  degrees of freedom instead of a standard normal distribution for confidence intervals and hypothesis tests (Colin Cameron and Miller, 2015). Admittedly, this solution is ad hoc and only has a firm theoretical justification for small samples in linear normal models. However, generally it has the right asymptotic properties since the difference between dividing by  $n$  and  $n - 1$  becomes negligible and the  $t$ -distribution converges towards the standard normal distribution as the sample size increases. We used this adjustment in Manuscript II, and it resulted in confidence intervals with correct coverage probabilities.

The influence function with the lowest possible variance is called the efficient influence function (Tsiatis, 2006). This semi-parametric theory was used in Bartlett (2018) to prove that the  $g$ -formula in combination with a correctly specified glm with canonical link yields the (asymptotically) best possible estimator in a one-period RCT in the sense that its influence function equals the efficient influence function.

As indicated, the RCT world and the observational data world require different tools and considerations. In the RCT world, we will be unbiased (almost) no matter what we do, whereas we risk being biased if any minor modelling assumption is violated in the observational data world. However, this is only a part of the picture. Semi-parametric efficiency theory has been used to derive doubly robust estimators for observational studies. As seen earlier, causal effect can be estimated by modelling the outcome and using the  $g$ -formula, or by modelling treatment in IPTW estimators. However, doubly robust estimators use both models. These estimators have the property that they are unbiased if either or both models are correctly specified, and they are efficient if both models are correctly specified (Hejazi and van der Laan, 2023). Denote covariates by  $X$ , treatment by  $A$ , the model for the expected outcome given covariates  $E(Y | X, A)$  by  $h(X, A)$ , and the propensity score model, i.e., the model for  $P(A = 1 | X)$ , by  $g(A | X)$ . Then the efficient influence function for  $\mu_0 := E(Y^1)$  is (Hejazi and van der Laan, 2023)

$$\frac{I(A = 1)}{g(1 | X)}(Y - h(X, 1)) + h(X, 1) - \mu_0. \quad (3.1)$$

The influence function for  $E(Y^0)$  is similar, thus enabling estimation of causal risk difference, causal risk ratio etc. The influence function has a term related to the  $g$ -formula estimator,  $h(X, 1)$ , and a residual term related to the propensity score. If we fit a propensity score model and an outcome model and plug into 3.1, and the outcome model is correctly specified, then the residual term will equal zero (on average) and the estimator is unbiased since the remaining part is  $g$ -formula with a correctly specified model. This is irrespective of whether the propensity score model is correctly specified or not. The influence function can also be rewritten to have a term related to the IPTW estimator and a residual term related to the outcome model (Hejazi and van der Laan, 2023). The implication of this rewriting is that we are unbiased if the propensity score model is correctly specified, even if the outcome model is misspecified. Thus, at least one model has to be correctly specified with RWD, but there are two chances to get it right to obtain unbiased estimates. However, the estimator is only completely efficient if both models are correctly specified (Hejazi and van der Laan, 2023)

The robustness results for RCTs are also often polluted by missing data that break the strict randomization in the theoretical results. Thus, the difference between RCT data and RWD might be smaller than it looks like from a theoretical point of view.



# 4 Summary of manuscripts

The manuscripts in this PhD contribute to the understanding and proper use of self-controlled designs, both in clinical development (Manuscript II) and observational studies for post approval safety (Manuscripts I and III).

## 4.1 Manuscript I

*Sampling in the case-time-control design among drug users when outcome prevents further treatment*

Manuscript I considers sampling of controls in the case-time-control design, applied to a population of drug users, when the outcome prevents subsequent treatment. For instance, if the outcome is terminal, we will not observe what future treatment a subject would have received in the absence of the outcome.

We show that a usual sampling leads to a bias in this setup. The bias can be understood from Figure 4.1. There are three cases, observations 1, 2, and 3 in the Figure. The case-time-control design requires that controls and cases have the same time-trend of exposure. This is not satisfied in the example if controls are sampled traditionally, i.e., among subject 4, 5, and 6 in the Figure. The problem is that observations 4 and 5 for example could be sampled as controls for case number 1, but if case number 1 had the treatment history of those observations, then it wouldn't be in our dataset to begin with, since there would be no treatment before the event time. The problem can to some extent be thought of as the problem of conditioning on the future (Lund et al., 2017). That is, we should not sample controls among subjects who are only in the dataset because they get exposure after the time of sampling (i.e. the event time of the case). As soon as this is understood, the solution of requiring controls to have had treatment before the event time of their matched case seems reasonable. This improved way of sampling is illustrated in the figure by tick marks and crosses, where tick marks indicate that a subject could be sampled for the corresponding case, and a cross indicates that the subject could not. The problem was not previously described in the literature, and the solution seemed rather hand-wavy, so we show in a mathematically rigorous framework that this sampling indeed solves the problem described in the Manuscript.

We illustrate the bias and the improved sampling in a simulation study and in a data example showing the effect of non-steroidal anti-inflammatory drugs on the risk of upper gastrointestinal bleeding. We find a meaningful difference in the results between the two ways of sampling controls, thereby highlighting the importance of the sampling method in practice. Furthermore, we extended the design to continuous time with competing risks. Originally, in line with the estimand framework and the result in Shahn et al. (2022), we had hoped to be able to give an interpretation of the target parameter in the case-time-control design. Unfortunately, the target parameter seems to depend on the relationship between the tail of the outcome distribution and the exposure distribution. We didn't want to make very strong assumptions about this and therefore had to settle with showing that if exposure increases risk of the

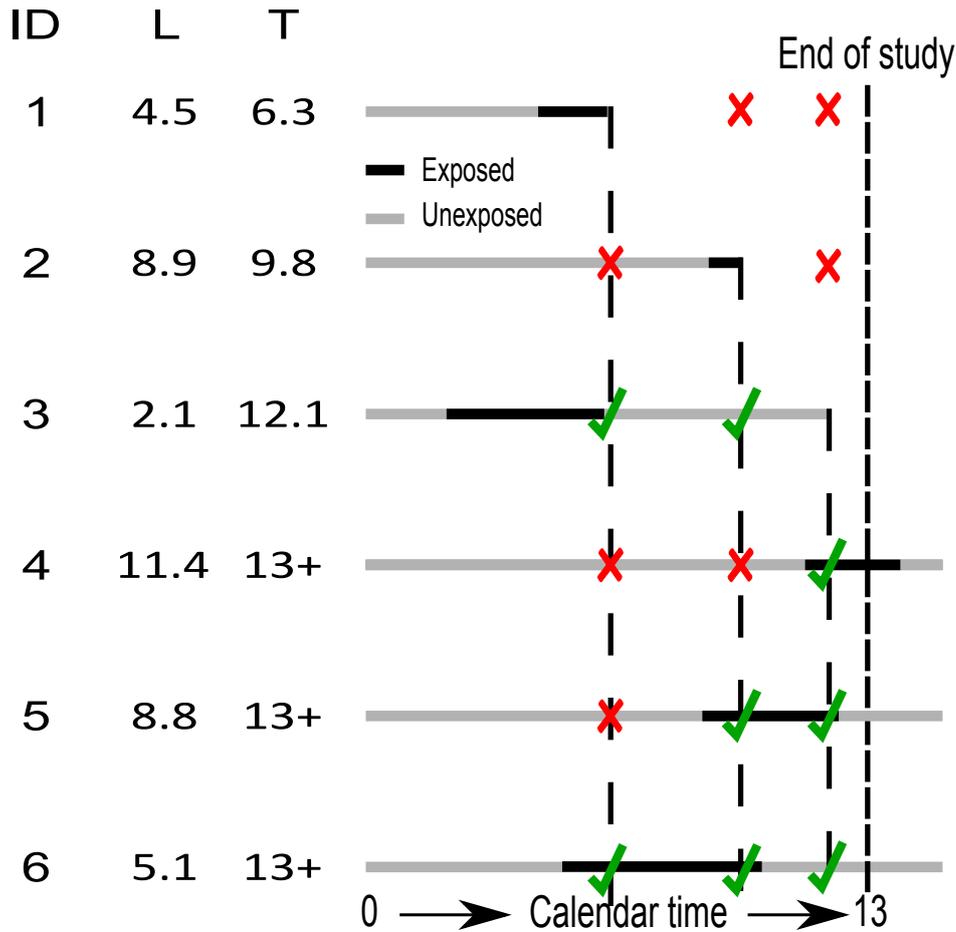


Figure 4.1: Source: Madsen et al. (2022). L denotes time of first treatment and T denotes time of event (13+ meaning no event in the study period).

event, then the target OR parameter is greater than one when the outcome is rare. It is possible that a causal target parameter could be obtained if the framework was formulated in terms of time-periods, as is usually done, and assuming that the logistic regression model is correctly specified.

All in all, the paper describes and solves a previously unknown problem while extending the design to continuous time and a competing risk setting.

## 4.2 Manuscript II

### *Unbiased and Efficient Estimation of Causal Treatment Effects in Cross-over Trials*

The combination of the estimand framework and our focus on self-controlled designs led us to the problem of cross-over trials, specifically motivated by Thorough QT (TQT) studies. The literature on TQT trials has mainly been focused on complex modelling of the data and use of baseline measurements, with little focus on what is really being estimated. In this paper, we wanted to go back to square one and figure out how to conduct cross-over trials from a causal inference perspective. This approach motivated a g-formula type estimator, and a semi-parametric estimator with the property of being unbiased no matter the choice of working model due to the randomization. We found, in fact to our surprise, the result that for a certain class of working models these two

estimators are equal, with the implication that also the g-formula estimator is unbiased for causal treatment effects in cross-over trials when a model from this class of working models is applied. This is of particular interest because the class of working models in the result includes a large class of mixed effects models that are routinely applied in TQT studies. The estimation in those models are based on assuming normally distributed outcomes and a specific mean and covariance structure. The result of the Manuscript ensures unbiased estimation of causal treatment effects, even if these assumptions are violated. We illustrate the flexibility that is facilitated by the result in a data example from a TQT study and in a simulation study. However, the robustness shown in the Manuscript does not extend to the variance estimates, which can be biased if the model is misspecified. Instead, we propose to estimate the variance from the influence function in order to be completely robust to misspecification of the applied working model. We derive the influence functions in the paper. Unfortunately, we discovered that TQT studies typically have such a small sample size that the asymptotic theory does not apply. Concretely, the influence function based variance estimates underestimate the uncertainty in the estimates. We applied a rather general, although heuristic adjustment, to the variance estimates and confidence intervals to resolve the problem, which seemed to work well in our simulations. However, more research could go into developing more satisfactory adjustments, and some research arguably already has (Pustejovsky and Tipton, 2018; Imbens and Kolesár, 2016; Colin Cameron and Miller, 2015). Another problem is that of efficiency. It would be advantageously to know, along the lines of Bartlett (2018); Rosenblum and Steingrimsson (2016), whether adjusting for covariates necessarily leads to lower asymptotic variance. We suspect this to be the case, but have so far not been able to show prove it. One potential way forward could be to derive the efficient influence function, but this is not an easy task since the assumption of having the same treatment effect in all periods necessarily has to be used in this derivation. Another drawback of the Manuscript is that the results rely on there being no missing data. An assumption that is usually violated in practice. Missing data breaks the randomization and essentially turns our data into observational data. There are methods in Tsiatis (2006) that involve modelling the missingness mechanism in order to gain some robustness towards misspecifying the outcome model. However, it could be interesting to explore this theory in the case of crossover trials in future research. A last suggestion for future research would be to generalize the result to more flexible working covariance matrices instead of the compound symmetry like one used in the Manuscript.

In any case, the manuscript shows that standard methods of analysis in crossover trials, and specifically in TQT trials, lead to unbiased estimates of causal contrasts that are easy to interpret without relying on the working model being correctly specified.

### 4.3 Manuscript III

*Estimating causal effects in the presence of unmeasured confounding when treatment duration is fixed*

In the third project, we wanted to explore to what extent adjustment for confounding by indication can be achieved without changing the whole design of the study and analysis method. We wanted to do this in a context relevant to treatment with for example biologics, i.e., antibody treatments targeted to severely sick subjects under very controlled circumstances, where among other things the time between consecutive

treatment administrations would be roughly constant.

In this context, we show how all unmeasured time-stable confounding can be adjusted for in any time-to-event model simply by adjusting for the number of treatment administrations. However, it is not possible to adjust directly for the number of treatment administrations if the outcome prevents subsequent treatment, such as when the outcome is terminal, since we don't know how many treatment administrations would have taken place if a subject had not had the outcome. We propose to solve this missing data problem by use of the EM algorithm. Even though an unconfounded estimate of the HR is possible to obtain in this setup, we propose to use the model as a stepping stone towards estimating a causal effect instead. The method is tested in a simulation study and applied to registry data to estimate the effect of treatment with antidepressant/anxiety treatment on the risk of poisoning from various medications.

However, getting a causal effect using measured covariates is challenged by the missingness since it makes non-parametric estimation of the covariate distribution in the g-formula difficult at best. Another weakness is the flexibility of the modelling. In particular, it is not possible to stratify on the number of treatment administrations in a Cox model due to lack of identifiability. Admittedly, this lack of flexibility does not seem to be bigger than for example what is required for frailty models (Balan and Putter, 2020).

All in all, the Manuscript allows adjusting for unmeasured confounding within whatever modelling framework is used simply by adjusting for the number of treatment administrations when the time between treatment administrations is fixed and time zero is defined to be the time of first treatment administration.

## 5 Perspectives

The research presented in this dissertation has contributed to important and active areas of research with importance to the pharmaceutical industry and researchers in the field of pharmacoepidemiology alike. The three Manuscripts use state-of-the-art theory from pharmacoepidemiology, causal inference, semi-parametric efficiency theory, and survival analysis, with relevance for the pharmaceutical industry in both clinical development and observational studies for post approval safety. Thus, the topics covered, and the techniques used in the Manuscripts are wide-ranging, but hopefully the introduction to this dissertation equips the unfamiliar reader with the necessary theoretical knowledge to understand them.

Although, the Manuscripts expand the amount of knowledge in these areas of research somewhat, there is still plenty to learn and research. We had hoped in the beginning of the PhD, that it would be possible to find a nicer, maybe even causal, interpretation of the target parameter in the case-time-control design. Unfortunately, the target parameter seemed to depend on the tail of the outcome distribution, which we have no access to in the data. Cutting some corners in the calculations would yield us a HR, but it was clear from the simulations that these corners were too big, and the results rarely would be very close to the HR, even if there actually were proportional hazards. This shortcoming was partly due to our insistence on formulating the design in continuous time, which we think was most appropriate. However, if the design is formulated in terms of periods, as is commonly the case, it is not impossible that some sort of causal interpretation could be obtained under some extra assumptions that would be beneficial to know. Admittedly, this causal target parameter could turn out to be a very unnatural one that would never come out of the causal roadmap (van der Laan and Rose, 2011). Manuscript II mainly has three shortcomings. First, it would be nice to generalize the results to more flexible variance structures that, for example, allows different variances for the different treatments. Second, efficiency seems within reach. The result from Bartlett (2018), unfortunately, does not extend since the influence function we derived is not the efficient influence function. Admittedly, the efficiency result does extend to situations, where the target parameter can be identified as a parameter in the model, due to the general result of asymptotic efficiency of the MLE. Maybe the semi-parametric theory used in Rosenblum and Steingrimsson (2016) could be used to show some efficiency gain from using a model. An alternative way forward could be to derive the efficient influence function, or at least an alternative estimator with the desirable property that the effect estimate equals a regression parameter for simpler models, that happens to be easier to show efficiency gains for compared to a simple non-parametric estimator. The last shortcoming is that the results rely on the assumption of no missing data. This is a hard problem to get around, since it essentially turns the data into observational, i.e. non-randomized, data. One way forward could be to use double robust estimators that model the missingness mechanism along the lines of Tsiatis (2006). The main shortcoming of Manuscript III is the assumption of no time-dependent confounding. This assumption is shared with other self-controlled designs and seems impossible to avoid without conditioning on the future (Hallas and Pottegård, 2014). Maybe the methodology of the paper can be adjusted for other

applications.

The Manuscripts have expanded the toolkit for pharmacoepidemiological research, which potentially can be a part of a movement towards a situation where self-controlled designs are understood and used sufficiently well to be more commonly taught and used instead of, or as a supplement to, analyses relying on an assumption of conditional exchangeability, which often happens to be known to be false a-priori.

As mentioned, this movement is partly driven by technology and new data sources along with regulatory requirements, but theory like the one presented in this dissertation needs to be able to catch up to these developments, and can help shape them as well to the benefit of statisticians, epidemiologists, regulators, industry, and most importantly, to the benefit of patients.

# Bibliography

- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.
- Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870.
- Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference: Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340.
- Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11):3424–3454.
- Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of Case-Cohort Designs. *Journal of Clinical Epidemiology*, 52(12):1165–1172.
- Bartlett, J. W. (2018). Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical statistics*, 17(5):648–666.
- Borgan, O., Goldstein, L., and Langholz, B. (1995). Methods for the Analysis of Sampled Cohort Data in the Cox Proportional Hazards Model. *The Annals of Statistics*, 23(5):1749–1778. Publisher: Institute of Mathematical Statistics.
- Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., and Ware, J. H. (1976). Randomized Clinical Trials: Perspectives on Some Recent Ideas. *New England Journal of Medicine*, 295(2):74–80.
- Clayton, D. and Hills, M. (1993). *Statistical models in epidemiology*. Oxford University Press, Oxford ; New York.
- Colin Cameron, A. and Miller, D. L. (2015). A Practitioner’s Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372.
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.
- Crowther, M. J. and Lambert, P. C. (2014). A general framework for parametric survival analysis. *Statistics in Medicine*, 33(30):5280–5297.
- Encepp (2021). The BRodalumab Assessment of Hazards: A Multinational Safety (BRAHMS) study in electronic healthcare databases. <https://www.encepp.eu/encepp/viewResource.htm?id=32635>. Accessed: November 18, 2022.
- European Medicines Agency (2017). Guideline on good pharmacovigilance practices (gvp): Module viii—post-authorisation safety studies (rev 3).

- Farrington, C. P. (1995). Relative Incidence Estimation from Case Series for Vaccine Safety Evaluation. *Biometrics*, 51(1):228–235. Publisher: [Wiley, International Biometric Society].
- Farrington, P., Whitaker, H., and Ghebremichael Weldeselassie, Y. (2018). *Self-controlled case series studies: a modelling guide with R*. Chapman & Hall/CRC biostatistics series. CRC Press, Taylor & Francis Group, Boca Raton.
- Franklin, J. M., Glynn, R. J., Martin, D., and Schneeweiss, S. (2019). Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. *Clinical Pharmacology & Therapeutics*, 105(4):867–877.
- Franklin, J. M. and Schneeweiss, S. (2017). When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?: Real world evidence and RCTs. *Clinical Pharmacology & Therapeutics*, 102(6):924–933.
- Greenland, S. (1996). Confounding and Exposure Trends in Case-Crossover and Case-Time-Control Designs. *Epidemiology*, 7(3):231–239.
- Hallas, J. and Pottegård, A. (2014). Use of self-controlled designs in pharmacoepidemiology. *Journal of Internal Medicine*, 275(6):581–589.
- Hejazi, N. S. and van der Laan, M. J. (2023). Revisiting the Propensity Score’s Central Role: Towards Bridging Balance and Efficiency in the Era of Causal Machine Learning. *Observational Studies*, 9(1):23–34.
- Hernán, M. A. (2010). The Hazards of Hazard Ratios. *Epidemiology*, 21(1):13–15.
- Hernán, M. A. (2018). The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health*, 108(5):616–619.
- Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. *Epidemiology*, 19(6):766–779.
- Hernán, M. A., Wang, W., and Leaf, D. E. (2022). Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*, 328(24):2446.
- Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Iacobelli, S. and Carstensen, B. (2013). Multiple time scales in multi-state models. *Statistics in Medicine*, 32(30):5315–5327.
- Imbens, G. W. and Kolesár, M. (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *Review of Economics and Statistics*, 98(4):701–712.
- International Council for Harmonisation (2019). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9 (R1).

- Jensen, A. K. G., Gerds, T. A., Weeke, P., Torp-Pedersen, C., and Andersen, P. K. (2014). Brief Report: On the Validity of the Case-Time-Control Design for Auto-correlated Exposure Histories. *Epidemiology*, 25(1):110–113.
- Kubota, K., Kelly, T.-L., Sato, T., Pratt, N., Roughead, E., and Yamaguchi, T. (2021). A novel weighting method to remove bias from within-subject exposure dependency in case-crossover studies. *BMC medical research methodology*, 21(1):214.
- Lund, J. L., Horváth-Puhó, E., Komjáthiné Szépligeti, S., Sørensen, H. T., Pedersen, L., Ehrenstein, V., and Stürmer, T. (2017). Conditioning on future exposure to define study cohorts can induce bias: the case of low-dose acetylsalicylic acid and risk of major bleeding. *Clinical Epidemiology*, 9:611–626.
- Maclure, M. (1991). The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology*, 133(2):144–153.
- Maclure, M. (2007). ‘Why me?’ versus ‘why now?’—differences between operational hypotheses in case-control versus case-crossover studies. *Pharmacoepidemiology and Drug Safety*, 16(8):850–853.
- Maclure, M., Fireman, B., Nelson, J. C., Hua, W., Shoaibi, A., Paredes, A., and Madigan, D. (2012). When should case-only designs be used for safety monitoring of medical products? *Pharmacoepidemiology and Drug Safety*, 21:50–61.
- Madsen, J. E. H., Hallas, J., Delvin, T., Scheike, T., and Pipper, C. (2022). Sampling in the case-time-control design among drug users when outcome prevents further treatment. *Pharmacoepidemiology and Drug Safety*, 31(4):404–410.
- Marshall, R. J. and Jackson, R. T. (1993). Analysis of case-crossover designs. *Statistics in Medicine*, 12(24):2333–2341.
- Martinussen, T. and Scheike, T. H. (2006). *Dynamic regression models for survival data*. Statistics for biology and health. Springer, New York, N.Y.
- Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26(4):833–855.
- Matthews, J., Bazakou, S., Henderson, R., and Sharples, L. D. (2022). Contrasting principal stratum and hypothetical strategy estimands in multi-period crossover trials with incomplete data. *Biometrics*, page biom.13777.
- McCullagh, P. and Nelder, J. A. (1998). *Generalized linear models*. Chapman & Hall/CRC, 2nd edition.
- Mittleman, M. A. and Mostofsky, E. (2014). Exchangeability in the case-crossover design. *International Journal of Epidemiology*, 43(5):1645–1655.
- Navidi, W. (1998). Bidirectional Case-Crossover Designs for Exposures with Time Trends. *Biometrics*, 54(2):596–605.
- Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

- Pustejovsky, J. E. and Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, 36(4):672–683.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ray, W. A. (2003). Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. *American Journal of Epidemiology*, 158(9):915–920.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rosenblum, M. and Steingrimsson, J. A. (2016). Matching the efficiency gains of the logistic regression estimator while avoiding its interpretability problems, in randomized trials. *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 281.
- Salas, M., Hotman, A., and Stricker, B. H. (1999). Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. *American Journal of Epidemiology*, 149(11):981–983.
- Schulz, K. F. and Grimes, D. A. (2002). Case-control studies: research in reverse. *The Lancet*, 359(9304):431–434.
- Senn, S. (2002). *Cross-over trials in clinical research*. Statistics in practice. J. Wiley, Chichester, Eng. ; New York, 2nd edition.
- Shahn, Z., Hernán, M. A., and Robins, J. M. (2022). A formal causal interpretation of the case-crossover design. *Biometrics*, page biom.13749.
- Sjölander, A., Dahlqvist, E., and Zetterqvist, J. (2016). A Note on the Noncollapsibility of Rate Differences and Rate Ratios:. *Epidemiology*, 27(3):356–359.
- Suissa, S. (1995). The Case-Time-Control Design. *Epidemiology*, 6(3):248–253.
- Suissa, S. (1998). The Case-Time-Control Design: Further Assumptions and Conditions. *Epidemiology*, 9(4):441–445.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer, New York.
- US Department of Health and Human Services (DHHS), US Food and Drug Administration (FDA) (2001). Guidance for industry: E10 choice of control group and related issues in clinical trials. <https://www.fda.gov/media/71349/download>.
- US Food and Drug Administration (2017). Use of real-world evidence to support regulatory decision-making for medical devices. *Guidance for industry and Food and Drug Administration staff*.
- van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. Springer.

- VanderWeele, T. J. (2009). Concerning the Consistency Assumption in Causal Inference. *Epidemiology*, 20(6):880–883.
- VanderWeele, T. J. (2016). Mediation Analysis: A Practitioner’s Guide. *Annual Review of Public Health*, 37(1):17–32.
- Vines, S. K. and Farrington, C. P. (2001). Within-subject exposure dependency in case-crossover studies. *Statistics in Medicine*, 20(20):3039–3049.
- Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2021). Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment. *Journal of the American Statistical Association*, pages 1–12.



# Manuscript I

## Sampling in the case-time-control design among drug users when outcome prevents further treatment

Jeppe Ekstrand Halkjær Madsen, Jesper Hallas, Thomas Delvin, Thomas Scheike & Christian Phipper

**Details:** Published in *Pharmacoepidemiology and Drug Safety* in 2022.

# Sampling in the case-time-control design among drug users when outcome prevents further treatment

Jeppe Ekstrand Halkjær Madsen<sup>1,2</sup>  | Jesper Hallas<sup>3</sup>  | Thomas Delvin<sup>1</sup> | Thomas Scheike<sup>2</sup> | Christian Phipper<sup>1</sup>

<sup>1</sup>Biostatistics and Pharmacoepidemiology, Medical Sciences, Leo Pharma A/S, Ballerup, Denmark

<sup>2</sup>Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen K, Denmark

<sup>3</sup>Clinical Pharmacology, Pharmacy and Environmental Medicine, Department of Public Health, University of Southern Denmark, Odense, Denmark

## Correspondence

Jeppe Ekstrand Halkjær Madsen, Section of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, 1014 Copenhagen K, Denmark.  
Email: jehm@sund.ku.dk

## Funding information

This study was funded by Innovation Fund Denmark.

## Abstract

**Purpose:** The objective of this article is to advocate a new way of sampling controls in the case-time-control design in a cohort of drug users when the studied outcome prevents further treatment.

**Methods:** Mathematically we demonstrate how a standard sampling of controls, where controls are sampled among all subjects without an event at end-of-study, leads to a biased effect estimate. We propose to add the requirement that controls initiate treatment before the calendar time of event of their matched case to circumvent this. The standard and proposed sampling methods are compared in a simulation study and in an empirical data example examining the effect of nonsteroidal anti-inflammatory drug usage on the risk of upper gastrointestinal bleeding.

**Results:** When the controls are sampled the standard way, the case-time-control design confers a bias because cases and controls have a different time-trend of exposure.

The bias has been upwards in all the scenarios we have investigated. The requirement we add to be a potential control ensures that cases and controls have the same time-trend of exposure when treatment and outcome are independent. The simulation study confirms that the proposed sampling method removes the bias between treatment and outcome. The proposed sampling method lowered the odds-ratio estimate from 3.72 to 3.26 in the data example.

**Conclusion:** The proposed sampling method makes it possible to use the case-time-control design in a cohort of subjects with registered use of a drug when outcome prevents further treatment.

## KEYWORDS

association, bias, case-time-control, sampling

## Key Points

- A standard sampling of controls leads to a biased estimate of the effect of treatment on outcome in the case-time-control design in a cohort of drug users when outcome prevents further treatment.
- Cases and controls have a different time-trend of exposure when controls are sampled in the standard way.

- A new way of sampling controls, where controls are required to have drug exposure before the event time of their matched case, ensures that cases and controls have the same time-trend of exposure when there is no effect of treatment on outcome.

## 1 | INTRODUCTION

Observational studies are often challenged by the bias incurred by confounding.<sup>1</sup> One solution to this problem is to compare cases, that is, subjects that get the event of interest, to themselves. This works because we can compare the exposure status of cases at the time of event to their exposure status at other time points. Consequently, we only need cases for the analysis. Accordingly, these designs are sometimes called case-only designs or self-controlled designs.<sup>2</sup> Their clear advantage is that time-invariant confounders are eliminated by design. However, research in these designs is ongoing, and guidelines on their use was only recently published by International Society for Pharmacoepidemiology.<sup>3</sup>

One example of a case-only design is the case-crossover design<sup>4</sup> where the exposure status at event is compared to the exposure status at a reference time, say a year, before the event time. A problem with this design arises when there is a time-trend in exposure. A time-trend could for example arise if there is an increase in drug usage in the period after drug approval or because of a change in treatment guidelines. In that case, the probability of exposure would be higher at the event time than at the reference time even if there is no effect of treatment on event. It has been suggested to remove this bias by including reference times both before and after event,<sup>5</sup> but this approach fails if the event affects future treatment, for example, if the event is severe or an absolute contraindication to further treatment. Moreover note that the self-controlled case series analysis is not a valid alternative as this design leads to biased statements when the event of interest affects subsequent treatment.<sup>6</sup> A popular solution is the case-time-control design,<sup>7</sup> an extension of the case-crossover design to handle time-trends in exposure. This is done by matched sampling of controls whose exposure status is compared at the calendar times of event and reference of their matched case.

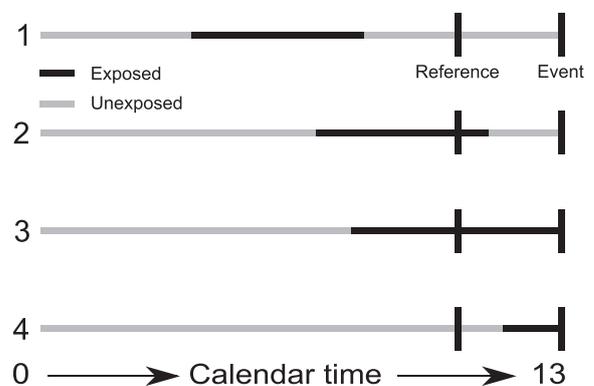
The case-time-control design enables studies of rare outcomes by use of data from large databases and eliminates commonly envisaged types of time-invariant confounding. However, we are not interested in targeting an effect in a very heterogenous population consisting of subjects who, for example, will never end up using the drug of interest. Instead, we wish to estimate an effect in a population of drug users. The idea of restricting focus to users is similar to the idea behind new-user designs<sup>8</sup>—namely to get a more homogenous population in terms of disease severity, which may be related to the event of interest. A more homogenous population would resemble a RCT more closely and lead to less bias resulting from confounding by indication.<sup>9</sup> The choice, unfortunately, also induces bias if the outcome of interest prevents subsequent drug use, for example, if the outcome is severe or an absolute contraindication to further treatment.

The purpose of this article is to consider the case-time-control design in this setup and propose a different way of sampling controls. The proposed sampling method is compared to a standard sampling method in a simulation study and an empirical data example.

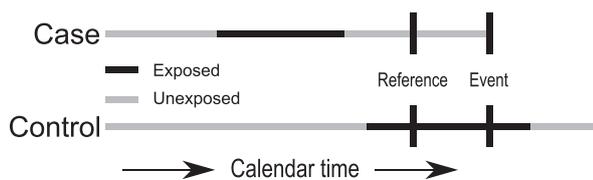
## 2 | THE CASE-TIME-CONTROL DESIGN

This section describes the case-time-control design as it is traditionally used. Subjects are observed in the study period  $[0, \tau]$  and have event times denoted by  $T$ . At this point it is important to note that time refers to time in the study period, that is, the underlying time scale is calendar time. This is the case because the time-trend we wish to eliminate is due to events that happen in calendar time, such as drug approval. Thus, time zero refers to the beginning of the study period. Note that the math is still valid if another time axis than calendar time is used. Cases are all the subjects with an event time in the study period. Exposure is compared at the event time,  $T$ , and a number of reference times before the event time. We consider a setup with exactly one reference time given by  $T - D_2$ , where  $D_2 > 0$  is a constant (see Figure 1).

If cases are exposed at their event time more frequently than at their reference time, this indicates either that the event is a side-effect of exposure or that there is a time-trend in exposure. The case-crossover design assumes no time-trend and therefore only includes cases. The odds-ratio (OR) among cases is given by the number of cases with exposure pattern 4 divided by the number of cases with exposure pattern 2 in Figure 1. This OR is the target parameter in the case-crossover design. To go from the case-crossover design to the case-time-control design, a fixed number of controls are sampled per case among subjects without an event. Exposure for controls is



**FIGURE 1** The four possible exposure patterns in the case-time-control design



**FIGURE 2** Exposures for controls are compared at the calendar time of event and reference of the case

compared at the calendar time of event and reference of their matched case (see Figure 2).

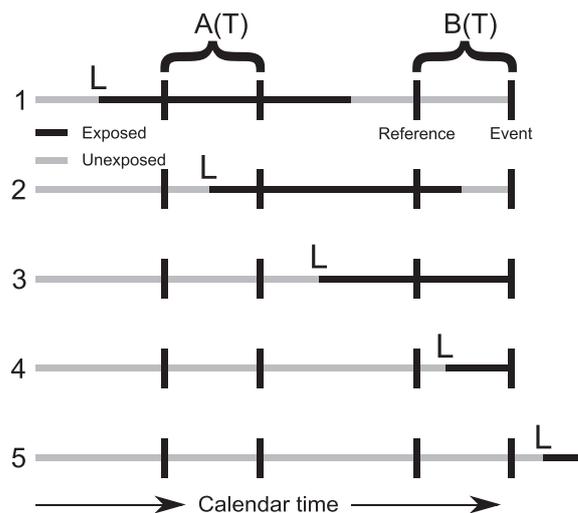
Since both of these times occur before the control has experienced a potential event, we assert that any difference between how often controls are exposed at the calendar times of event and reference of their matched case can be attributed to a time-trend in exposure. The *i*'th subject contributes with two observations to our dataset, one for the event time and one for the reference time. Let  $E_{i,ref}$  denote exposure for subject *i* at the reference time (1 = exposed, 0 = unexposed) and let  $E_{i,event}$  denote exposure for subject *i* at the event time. Let  $G_i$  be case status for subject *i* (1 = case, 0 = control). The estimation is done using the following conditional logistic regression model:

$$\begin{aligned} \text{logit}(P(E_{i,ref} = 1)) &= \alpha_i \\ \text{logit}(P(E_{i,event} = 1)) &= \alpha_i + \beta_1 + \beta_2 \cdot G_i. \end{aligned} \quad (1)$$

The subject specific intercept,  $\alpha_i$ , captures any difference in odds between subjects, which is not affected by time, and thus effectively controls for time-constant confounders. The OR for cases is  $\exp(\beta_1 + \beta_2)$  and reflects both the time-trend of exposure and the effect of exposure on event. The OR for controls is  $\exp(\beta_1)$  and reflects the calendar time-trend of exposure. Importantly, the odds ratio obtained as the ratio between OR for cases and OR for controls is identified as  $\exp(\beta_2)$ . It follows that this odds ratio informs us how much higher, if at all, the odds are of being exposed at the event time relative to the reference time for cases, after adjusting for time-trend in exposure in controls. Therefore, this OR is the OR of interest, and will be termed the trend-adjusted OR.

### 3 | ANALYSIS

We now characterize the bias that is induced when applying a standard case-time-control design in a cohort of drug users when the outcome of interest prevents further treatment. The resulting characterization prompts a simple alternative way of sampling controls that eliminates this problem. For ease of exposition, we consider a setup where each subject has exactly one treatment period initiated at a random time point *L* and a fixed duration  $D_1$ . In Appendix A, results and conclusions are extended to a general treatment profile with competing risks and random censoring. We are considering a cohort of



**FIGURE 3** The four possible exposure patterns are entirely determined by the time of treatment initiation in our setup

drug users which means that necessarily treatment initiation, *L*, will be in the study period  $[0, \tau]$ . Furthermore, since outcome prevents further treatment, we have treatment initiation before the event time for cases. The four exposure patterns in Figure 1 are entirely determined by the time of treatment initiation in this setup. A subject has exposure pattern 4 if it initiates treatment in the interval  $B(T)$ , which is defined as  $(T - D_2, T]$  when  $D_1 \geq D_2$ , and  $(T - D_1, T]$  when  $D_1 < D_2$ . A subject has exposure pattern 2 if it initiates treatment in the interval  $A(T)$ , which is defined as  $(T - D_1 - D_2, T - D_1]$  when  $D_1 \geq D_2$ , and  $(T - D_1 - D_2, T - D_2]$  when  $D_1 < D_2$ . The situation where  $D_1 \geq D_2$  is depicted in Figure 3. In that case,  $B(T)$  is simply the time between event and reference, and  $A(T)$  is the same interval but shifted by the treatment duration. Now, treatment initiation can be in five different regions (see Figure 3).

1. Before  $A(T)$ , where exposure has ended before the reference time.
2. In  $A(T)$ , where the subject will be exposed at the reference time but not at the event time.
3. Between  $A(T)$  and  $B(T)$ , where the subject will be exposed at both the reference time and the event time. If  $D_1 < D_2$  the subject will be exposed neither at event nor at reference. Not at event because exposure has ended before the event time, and not at reference because treatment is initiated later.
4. In  $B(T)$ , where the case will be exposed at the event time but not at the reference time.
5. After event, that is, after  $B(T)$ , where the subject is exposed neither at event nor at reference. Note that this is only possible for controls in our setup.

The OR for cases,  $\exp(\beta_1 + \beta_2)$  from (1), is identified as the probability, among cases, of initiating treatment in  $B(T)$  divided by the probability of initiating treatment in  $A(T)$ . The OR for controls is identified analogously. Consequently, for a randomly selected case *i* and control *j*, the trend-adjusted OR,  $\exp(\beta_2)$  from (1), is identified as

$$\frac{P(L_i \in B(T_i)T_i \leq \tau) / P(L_i \in A(T_i)T_i \leq \tau)}{P(L_j \in B(T_j)T_j \leq \tau, L_j \leq \tau, T_j > \tau) / P(L_j \in A(T_j)T_j \leq \tau, L_j \leq \tau, T_j > \tau)} \quad (2)$$

If L and T are independent, corresponding to no effect of exposure on event, the probabilities in the numerator of (2) are

$$P(L_i \in B(T_i)T_i \leq \tau) = \int_0^\tau P(L_i \in B(t)) \cdot P(T_i = t) dt$$

$$P(L_i \in A(T_i)T_i \leq \tau) = \int_0^\tau P(L_i \in A(t)) \cdot P(T_i = t) dt,$$

and the probabilities in the denominator are

$$P(L_j \in B(T_j)T_j \leq \tau, L_j \leq \tau, T_j > \tau) = p_{ij} \int_0^\tau P(L_j \in B(t)) P(T_j = t) \frac{P(L_j \leq t)}{P(L_j \leq \tau)} dt$$

$$P(L_j \in A(T_j)T_j \leq \tau, L_j \leq \tau, T_j > \tau) = p_{ij} \int_0^\tau P(L_j \in A(t)) P(T_j = t) \frac{P(L_j \leq t)}{P(L_j \leq \tau)} dt,$$

where  $p_{ij}$  is the probability that control  $j$  is sampled for case  $i$ . These probabilities cancel out in (2) and can thus be ignored.

The probabilities in the numerator for cases differ from the corresponding probabilities in the denominator for controls by the highlighted fractions within the control integrals. The cause of the problem is that treatment initiation must be before the event time for cases but can be anywhere in the study period for controls. Hence, the distribution of treatment initiation is different for cases and controls, and they will, as a result, have a different time-trend of exposure. In particular, controls can initiate treatment *after* the calendar time of event of their matched cases, which is not possible for cases, that is, controls can be in situation five from Figure 3.

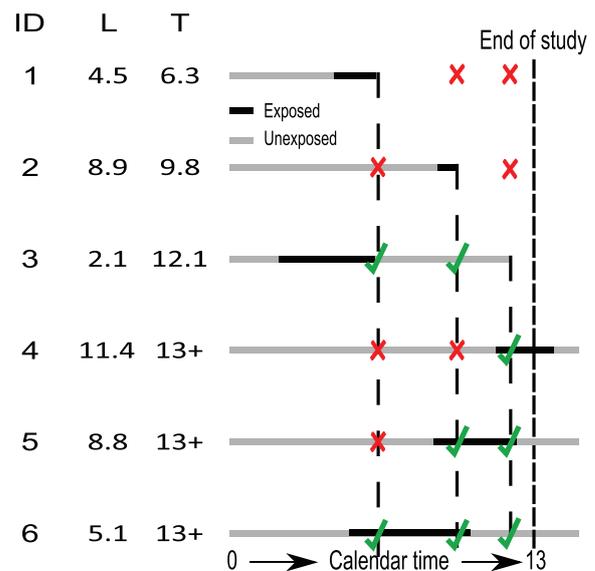
Our solution is straightforward: the controls are required to initiate treatment before the calendar time of event of their matched case. This ensures that the time-trend of exposure is the same for cases and controls when there is no effect of treatment on event. Note that  $T_j$  is nowhere in the calculations which implies that we can sample among all subjects at risk at the failure time of the case, including other future cases, as long as they initiate treatment before the calendar time of event. This is in line with the traditional risk-set sampling paradigm but unlike the standard sampling of controls in a case-non-case study, where controls are sampled among subjects without an event at end-of-study. This sampling is similar to the sampling in the nested case-control design and can be seen as a mixture of a classic case-time-control design and a case-case-time design.<sup>10</sup> Potentially this can lead to a lower bias when cases and controls have different treatment distributions since we would expect future cases to have a more similar treatment distribution than controls.<sup>11</sup> Furthermore, it makes it possible to use the design when we have random censoring as opposed to constant censoring at  $\tau$  like above (see Appendix A). From here on, sampling among all subjects at risk who have received treatment before the event time, will be termed “the proposed sampling method.”

### 4 | EXAMPLE

To illustrate the difference between the standard and the proposed sampling methods suppose you have the dataset in Figure 4. The first three subjects are cases and the rest are controls. Normally, we would just sample among all the controls. Which subjects can be sampled for case number 1 according to the proposed method in this article? The subject must be at risk at the calendar time of event, but initiate treatment before. All subjects are at risk at the calendar time of event of case number 1, but subjects 2, 4, and 5 initiate treatment after, and hence cannot be sampled. All subjects except subjects 1 and 4 can be sampled for case number 2. Subject 1 cannot since it is not at risk at the calendar time of event of case number 2 and subject 4 cannot since it initiates treatment after. All controls can be sampled for case number 3, but subjects 1 and 2 cannot since they are not at risk at the calendar time of event of case number 3. This means that there is no difference between the standard and the proposed ways of sampling controls for case number 3.

### 5 | SIMULATION

The standard and the proposed ways of sampling controls are compared in a simulation study. The data are simulated in the same setup as in the previous section with exactly one treatment period with a fixed duration,  $D_1$ , which is 6 years in this simulation study. Treatment initiation is simulated from a uniform distribution on the interval  $[0, 25]$ . It might seem surprising that we get a time-trend in exposure despite simulating from a uniform distribution. The reason is described in Appendix D. The survival times have been simulated from proportional hazards models with a baseline hazard rate of



**FIGURE 4** Event time of 13+ means subjects did not have an event in the study period. Crosses mean subject cannot be sampled for given case, tick marks mean subject can be sampled. ID, subject id, L, treatment initiation, T, event time

**TABLE 1** Geometric means of trend-adjusted OR estimates from simulation study

True HR	True trend-adjusted OR	Observed OR with proposed sampling	Observed OR with standard sampling
1.0	1.00	1.00	1.43
0.5	0.57	0.56	0.72
2.0	1.81	1.84	2.76
5.0	4.22	4.32	6.82

**TABLE 2** Exposure patterns among cases, controls sampled the standard way, and controls sampled in the proposed way in the simulation study

Cases			
		Reference	
		Unexposed	Exposed
Event	Unexposed	759	187
	Exposed	355	1992
Controls (standard)			
		Reference	
		Unexposed	Exposed
Event	Unexposed	2502	81
	Exposed	98	612
Controls (proposed)			
		Reference	
		Unexposed	Exposed
Event	Unexposed	767	190
	Exposed	360	1976

0.001 and hazard ratios of exposure given by 1, 0.5, 2, and 5. The time span between the event time and the reference time,  $D_2$ , is equal to 0.75, and the end of study,  $\tau$ , is 13. We carried out 10 000 simulations, each with 100 000 subjects, for each HR. The results are summarized in Table 1.

We clearly see the estimates are biased if we do not sample according to our proposal. In particular, we get an OR > 1 when treatment and event are simulated independently in the first row of Table 1. The proposed sampling method leads to estimates close to the true trend-adjusted OR.

One simulation like above with HR = 1, corresponding to no effect of treatment on outcome, and with 1 million subjects, has been run to further illustrate the problem with sampling in the standard way. The results are displayed in Table 2. Substantially more subjects are unexposed at both event and reference among controls than among cases when controls are sampled in the standard way. The results for cases and controls are very similar when controls are sampled in the proposed way.

## 6 | EMPIRICAL DATA EXAMPLE

The method is implemented on real world data to illustrate the importance of a correct sampling of controls in studies based on databases.

**TABLE 3** NSAID exposure patterns among upper gastrointestinal bleeding cases, controls sampled the standard way, and controls sampled in the proposed way when exposure is defined assuming DDD/d = 1

Cases			
		Reference	
		Unexposed	Exposed
Event	Unexposed	1633	148
	Exposed	525	363
Controls (standard)			
		Reference	
		Unexposed	Exposed
Event	Unexposed	2349	106
	Exposed	101	113
Controls (proposed)			
		Reference	
		Unexposed	Exposed
Event	Unexposed	2325	115
	Exposed	125	104

We used a case-control dataset which has been described in detail in previous publications.<sup>12</sup> In brief, we identified all patients admitted to a hospital in the Funen area in Denmark with severe upper gastrointestinal bleeding during 1995–2006. Hence, the year 1995 serves as time 0 in this case. All cases were manually validated by review of the discharge summary. Controls were sampled by a risk-set strategy, that is, for each case, 10 controls were recruited who were born in the same year, had the same sex, were residents of the Funen area and who had not yet been hospitalized with upper gastrointestinal bleeding. Controls were assigned an index date identical to the admission date of the corresponding case. Data on use of prescription drugs were retrieved from the OPED database.<sup>13</sup> We only included the subjects who had had at least one prescription for a nonsteroidal anti-inflammatory drug (NSAID) from the data to emulate the setup in this article. We were interested in the association between the use of NSAIDs and the occurrence of upper gastrointestinal bleeding. It should be noted that occurrence of upper gastrointestinal bleeding is an absolute contraindication to treatment with NSAID. Hence the dataset reflects the setup in this article. Subjects were considered exposed from the day of dispensing until the end of prescription duration. Prescription duration was calculated under the assumption of a daily intake of one defined daily dose (DDD/d). As a sensitivity analysis, we redid the analysis under the assumption of 0.5 DDD/d.

**TABLE 4** Regression results from data example

DDD/d	Sampling method	Controls OR		Trend-adjusted OR	
		OR	(95% CI)	OR	(95% CI)
1.0	Standard	0.95	(0.73, 1.25)	3.72	(2.69, 5.15)
	Proposed	1.09	(0.84, 1.40)	3.26	(2.38, 4.47)
0.5	Standard	1.12	(0.87, 1.44)	3.94	(2.88, 5.37)
	Proposed	1.18	(0.95, 1.47)	3.73	(2.79, 5.00)

The standard and the proposed sampling methods are implemented on these data. Choosing the interval between event and reference is a trade-off between on one hand having enough time to get enough subjects with a discordant exposure history for the design to work, and on the other hand not have too much time-trend in exposure to adjust for. We have chosen half a year for these data. The exposure patterns when assuming  $DDD/d = 1$  can be seen in Table 3. We notice that fewer controls are unexposed at both event and reference when sampling controls the proposed way compared to the standard way as expected, although the difference is small.

The results of the conditional logistic regression model (1) are shown in Table 4. We get a lower estimate of the trend-adjusted OR when sampling controls in the proposed way no matter what  $DDD/d$  we assume. This indicates that a part of the OR when sampling controls in the standard way is due to a different time-trend of exposure between cases and controls. Previous work on the association between NSAID and upper gastrointestinal bleeding has been using case-control studies, which could suffer from some confounding. The case-time-control design implicitly adjusts for time stable part of this confounding, but an incorrect sampling of controls could lead to additional bias due to an added difference in time-trend of exposure between cases and controls. Our proposal removes time-stable confounding by design and correctly adjusts for time-trend in exposure. It should also be noted that the reported odds ratio from a case-control study cannot be compared one-to-one with the odds ratio from the case-time-control design owing to the fact that the comparison made differs by design. That said, results from the case-time-control design are comparable to the results found in previous work despite the different approaches to sampling controls. Thus the conclusions are in essence not affected by how we sample controls.<sup>12</sup>

## 7 | DISCUSSION

By mathematical analysis and simulation, we have shown how the classical case-time-control design induces a biased estimate of the effect of treatment on event in a cohort of drug users when outcome prevents further treatment. Furthermore, we have shown that this bias disappears if we add the extra requirement that controls have initiated treatment before the calendar time of the event of their matched case. The change in sampling of controls is shown to matter, not just in theory, but also in an empirical data example. Moreover, the mathematical framework for analyzing the case-time-control design in this article paves the way for dealing with competing risks and random

censoring. Additionally, it makes it possible to analyze special setups for the design, like the one considered in this article, thus making the design more flexible.

This means that the design can still be used for outcomes that prevent further treatment in a cohort of drug users.

A downside of the suggested case-time-control design, and the case-time-control design in general, is that it is challenging to interpret the trend-adjusted OR if we in fact do find an association between treatment and event unless we make further assumptions. Moreover, the treatment distribution must be the same for cases and controls, although the problem of a different time-trend for cases and controls can be resolved to some extent by using the case-case-time design. Nevertheless, the design leads to a trend-adjusted OR of one when there is no effect of treatment on outcome and controls are sampled in the proposed way. Thus, the design can be used to answer the question of whether a given outcome is a side-effect of treatment while automatically adjusting for time-invariant confounders.

In summary, we have managed to solve a concrete problem for the case-time-control design and developed a new mathematical framework that potentially can be used to solve other problems in the case-time-control design, and other self-adjusted designs. The shortcomings are all related to the case-time-control design and are thus not specific for our study. More research is needed to resolve those issues.

### CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

### ETHICS STATEMENT

Registry-based studies do not need ethical board approval in Denmark.

### ORCID

Jeppe Ekstrand Halkjær Madsen  <https://orcid.org/0000-0002-7327-3224>

Jesper Hallas  <https://orcid.org/0000-0002-8097-8708>

### REFERENCES

- Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Lippincott Williams & Wilkins; 2008.
- Hallas J, Pottegård A. Use of self-controlled designs in pharmacoepidemiology. *J Intern Med*. 2014;275(6):581-589.
- Cadarette SM, Maclure M, Delaney JC, et al. Control yourself: ISPE-endorsed guidance in the application of self-controlled study designs

- in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf.* 2021;30(6): 671-684.
4. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol.* 1991;133(2): 144-153.
  5. Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics.* 1998;54:596-605.
  6. Farrington C. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics.* 1995;51:228-235.
  7. Suissa S. The case-time-control design. *Epidemiology.* 1995;6: 248-253.
  8. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol.* 2003;158(9):915-920.
  9. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol.* 2015;11(7):437-441.
  10. Borgan O, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann Stat.* 1995;23:1749-1778.
  11. Wang S, Linkletter C, Maclure M, et al. Future-cases as present controls to adjust for exposure-trend bias in case-only studies. *Epidemiol Camb Mass.* 2011;22(4):568.
  12. Dall M, De Muckadell OBS, Lassen AT, Hansen JM, Hallas J. An association between selective serotonin reuptake inhibitor use and serious upper gastrointestinal bleeding. *Clin Gastroenterol Hepatol.* 2009;7(12):1314-1321.
  13. Hallas J, Hellfritsch M, Rix M, Olesen M, Reilev M, Pottegård A. Odense pharmacoepidemiological database: a review of use and content. *Basic Clin Pharmacol Toxicol.* 2017;120(5):419-425.

#### SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Madsen JEH, Hallas J, Delvin T, Scheike T, Pipper C. Sampling in the case-time-control design among drug users when outcome prevents further treatment. *Pharmacoepidemiol Drug Saf.* 2022;31(4):404-410. doi:10.1002/pds.5410

## Appendix A: General case

The results in this paper are a special case of a more general result presented below. Let  $\{Z(t)\}$  be the treatment process ( $Z(t) = 1$  meaning that the patient is being treated at time  $t$ ,  $Z(t) = 0$  meaning that the patient is not being treated at time  $t$ ). Let  $L := \inf\{t \mid Z(t) = 1\}$  be the time of first treatment,  $T$  be the event time,  $\Delta$  be the time between event and reference,  $C$  be the censoring time, and  $X$  be a competing risk. Let  $C$  be independent of everything and  $X$  independent of everything except for  $T$ . We have the following odds-ratios

$$OR_{\text{case}} = \frac{P(Z_i(T_i) = 1, Z_i(T_i - \Delta) = 0 \mid L_i \leq T_i \leq \min\{C_i, X_i\})}{P(Z_i(T_i) = 0, Z_i(T_i - \Delta) = 1 \mid L_i \leq T_i \leq \min\{C_i, X_i\})},$$

$OR_{\text{ctrl}}$

$$= \frac{P(Z_j(T_i) = 1, Z_j(T_i - \Delta) = 0 \mid L_i \leq T_i \leq \min\{C_i, X_i\}, T_i \leq \min\{T_j, C_j, X_j\}, L_j \leq \min\{T_j, C_j, X_j\})}{P(Z_j(T_i) = 0, Z_j(T_i - \Delta) = 1 \mid L_i \leq T_i \leq \min\{C_i, X_i\}, T_i \leq \min\{T_j, C_j, X_j\}, L_j \leq \min\{T_j, C_j, X_j\})}.$$

In this notation the parameter  $\exp(\beta_2)$  is given by

$$\frac{OR_{\text{case}}}{OR_{\text{control}}}.$$

We need the following conditional probability to rewrite the probabilities above:

$$\begin{aligned} P(Z_i(T_i) = 1, Z_i(T_i - \Delta) = 0 \mid L_i \leq T_i = t \leq \min\{C_i, X_i\}) &= \\ \frac{P(Z_i(t) = 1, Z_i(t - \Delta) = 0, T_i = t \leq \min\{C_i, X_i\})}{P(T_i = t, L_i \leq t, X_i \geq t, C_i \geq t)} &= \\ \frac{P(Z_i(t) = 1, Z_i(t - \Delta) = 0, T_i = t, X_i \geq t)}{P(T_i = t, L_i \leq t, X_i \geq t)}. \end{aligned}$$

Assume independence between  $T$  and  $L$ , and between  $T$  and  $Z(t)$  for all  $t$ . Then the above probability becomes

$$P(Z_i(T_i) = 1, Z_i(T_i - \Delta) = 0 \mid L_i \leq T_i = t \leq \min\{C_i, X_i\}) = \frac{P(Z_i(t) = 1, Z_i(t - \Delta) = 0)}{P(L_i \leq t)}$$

which leads to

$$P(Z_i(T_i) = 1, Z_i(T_i - \Delta) = 0 \mid L_i \leq T_i \leq \min\{C_i, X_i\}) =$$

$$\int_0^{\infty} P(T_i = t, X_i \geq t)P(C_i \geq t) \cdot P(Z_i(t) = 1, Z_i(t - \Delta) = 0) dt$$

$P(Z_i(T_i) = 0, Z_i(T_i - \Delta) = 1 \mid L_i \leq T_i \leq \min\{C_i, X_i\})$  is found in a similar way and is given by

$$\int_0^{\infty} P(T_i = t, X_i \geq t)P(C_i \geq t) \cdot P(Z_i(t) = 0, Z_i(t - \Delta) = 1) dt.$$

The sampling of controls is done among all subjects at risk at the failure time of the case. The equations for controls are then:

$$\begin{aligned} &P(Z_j(T_i) = 1, Z_j(T_i - \Delta) = 0 \mid L_i \leq T_i = t, \min\{X_i, C_i\} \geq t, t \leq \min\{T_j, C_j, X_j\}, L_j \leq \min\{T_j, C_j, X_j\}) \\ &= p_{ij} \cdot \frac{P(Z_j(t) = 1, Z_j(t - \Delta) = 0, t \leq \min\{T_j, C_j, X_j\})}{P(t \leq \min\{T_j, C_j, X_j\}, L_j \leq \min\{T_j, C_j, X_j\})}, \end{aligned}$$

where  $p_{ij}$  is the probability of control  $j$  being sampled as a control for case  $i$ . Then

$$\begin{aligned} &P(Z_j(T_i) = 1, Z_j(T_i - \Delta) = 0 \mid L_j \leq \min\{T_j, C_j, X_j\}) = \\ &p_{ij} \cdot \int_0^{\infty} P(T_i = t, X_i \geq t)P(L_i \leq t)P(C_i \geq t) \cdot \frac{P(Z_j(t) = 1, Z_j(t - \Delta) = 0)}{P(L_j \leq \min\{T_j, C_j, X_j\})} dt \end{aligned}$$

The difference between the case and the control odds-ratios disappears if we require  $L_j \leq T_i$  and assume the same distribution of  $\{Z_i(t)\}$  and  $\{Z_j(t)\}$ , i.e. cases and controls have the same treatment distribution. Thus,  $\exp(\beta_2)$  is equal to one in this case when controls are sampled according to our proposal.

## Appendix B: Interpretation of trend-adjusted OR

We have demonstrated that the trend-adjusted OR is one in the case-time-control design

when there is no effect of treatment on event, and controls are sampled in the proposed way.

But what if we *do* have an association between treatment and event? The intuition behind the case-time-control design is the following: if more people are exposed at event than at reference, it is either because we have an association between exposure and event *or* because we have a time-trend in exposure. We remove the time-trend of exposure using controls and

expect the remaining association to reflect the pure effect from exposure on event. Hence, we would expect a trend-adjusted OR greater than one when treatment increases risk of event.

But is that certain to happen? Let  $\{Z_i(t)\}_{t \in [0, \tau]}$  be the exposure history of subject  $i$ , ( $Z_i(t) = 1$  meaning that subject  $i$  is exposed at time  $t$ ,  $Z_i(t) = 0$  meaning that subject  $i$  is unexposed at time  $t$ ). Let us assume the following frailty model:

$$\lambda_i(t | U_i, Z_i(t)) = \lambda(t | Z_i(t)) \cdot U_i. \quad (4)$$

The model says that the hazard rate at calendar time  $t$  only depends on  $t$ , the individual specific frailty,  $U_i$ , and treatment status at time  $t$ , but not on previous treatment history.

It might seem a bit unnatural that the hazard rate is a function of calendar time rather than age, but it is assumed that the effect of age, or rather birth year, is reflected by the frailty  $U_i$ , which reflects the confounders we want to adjust for implicitly by using a self-adjusted design. Assume without loss of generality that  $E(U_i) = 1$  for all  $i$ , and assume independence between  $\{Z_i(t)\}$  and  $U_i$ . As in Appendix A, let  $L := \inf\{l | Z(l) = 1\}$  be the time of first treatment. In addition, assume  $\lambda_i(t | Z_i(t), U_i)$  so small that  $S(t | Z_i(t), U_i) \approx 1$  in the study period (note that this implies  $P(T_i = t | Z_i(t), U_i) \approx \lambda_i(t | Z_i(t), U_i)$ ). We can calculate the probabilities in the OR for cases when we have an association between treatment and event, corresponding to  $\lambda(t | 1) > \lambda(t | 0)$  for all  $t$ , using the following conditional probability:

$$\begin{aligned} P(Z_i(t) = 1, Z_i(t - \Delta) = 0, T_i = t | U_i) &= \\ P(Z_i(t) = 1, Z_i(t - \Delta) = 0 | U_i) \cdot P(T_i = t | Z_i(t) = 1, Z_i(t - \Delta) = 0, U_i) &\approx \\ P(Z_i(t) = 1, Z_i(t - \Delta) = 0) \cdot \lambda(t | 1) \cdot U_i. \end{aligned}$$

We get  $P(Z_i(t) = 1, Z_i(t - \Delta) = 0, T_i = t)$  by integrating the frailty out:

$$P(Z_i(t) = 1, Z_i(t - \Delta) = 0, T_i = t) \approx P(Z_i(t) = 1, Z_i(t - \Delta) = 0) \cdot \lambda(t | 1)$$

The same procedure yields

$$P(Z_i(t) = 0, Z_i(t - \Delta) = 1, T_i = t) \approx P(Z_i(t) = 0, Z_i(t - \Delta) = 1) \cdot \lambda(t | 0).$$

This means that the OR of treatment among cases approximately is equal to

$$OR_{\text{case}} \approx \frac{\int_0^\tau \lambda(t | 1) \cdot P(Z_i(t) = 1, Z_i(t - \Delta) = 0) dt}{\int_0^\tau \lambda(t | 0) \cdot P(Z_i(t) = 0, Z_i(t - \Delta) = 1) dt}$$

Note that if we had proportional hazards,  $\lambda(t | Z(t)) = \lambda_0(t)e^{\gamma Z(t)}$ , and exchangeability of exposures between the event and the reference times,  $P(Z_i(t) = 1, Z_i(t - \Delta) = 0) = P(Z_i(t) = 0, Z_i(t - \Delta) = 1)$ , corresponding to no exposure time-trend, then the OR for cases would equal the hazard ratio,  $e^\gamma$ . Thus, the OR in the case-crossover design can be interpreted as a hazard ratio if the event is rare, there is no time-trend in exposure, and events arise according to a proportional hazards model. The OR of treatment among the controls is

$$OR_{\text{ctrl}} = \frac{\int_0^\tau \frac{P(T_i = t, L_i < t)}{P(T_j > t, L_j < t)} P(Z_j(t) = 1, Z_j(t - \Delta) = 0, T_j > t) dt}{\int_0^\tau \frac{P(T_i = t, L_i < t)}{P(T_j > t, L_j < t)} P(Z_j(t) = 0, Z_j(t - \Delta) = 1, T_j > t) dt}$$

This can be rewritten by calculating the probabilities in the integrals explicitly. The following conditional probability is needed for that:

$$\begin{aligned} P(Z_j(t) = 1, Z_j(t - \Delta) = 0, T_j > t | U_j) &= \\ P(Z_j(t) = 1, Z_j(t - \Delta) = 0) \cdot P(T_j > t | Z_j(t) = 1, Z_j(t - \Delta) = 0, U_j) &\approx \\ P(Z_j(t) = 1, Z_j(t - \Delta) = 0). \end{aligned}$$

Then we also have

$$P(Z_j(t) = 1, Z_j(t - \Delta) = 0, T_j > t) \approx P(Z_j(t) = 1, Z_j(t - \Delta) = 0).$$

Likewise, we have

$$P(Z_j(t) = 0, Z_j(t - \Delta) = 1, T_j > t) \approx P(Z_j(t) = 0, Z_j(t - \Delta) = 1),$$

$$P(T_j > t, L_j < t) \approx P(L_j < t).$$

Last, but not least:

$$P(T_i = t, L_i < t | U_i) =$$

$$P(L_i < t | U_i) \cdot P(T_i = t | L_i < t, U_i) =$$

$$\begin{aligned}
& P(L_i < t) \cdot (P(T_i = t \mid L_i < t, U_i, Z_i(t) = 1) \cdot P(Z_i(t) = 1 \mid L_i < t, U_i) \\
& \quad + P(T_i = t \mid L_i < t, U_i, Z_i(t) = 0) \cdot P(Z_i(t) = 0 \mid L_i < t, U_i)) \approx \\
& P(L_i < t) \cdot (P(Z_i(t) = 1 \mid L_i < t) \cdot \lambda(t \mid 1) \cdot U_i + P(Z_i(t) = 0 \mid L_i < t) \cdot \lambda(t \mid 0) \cdot U_i)
\end{aligned}$$

Again, we integrate the frailties out to get  $P(T_i = t, L_i < t)$ . Note that we have the following inequality:

$$P(L_i < t) \cdot \lambda(t \mid 0) < P(T_i = t, L_i < t) < P(L_i < t) \cdot \lambda(t \mid 1),$$

which can be used to show

$$OR_{\text{ctrl}} < \frac{\int_0^\tau \lambda(t \mid 1) \cdot P(Z_j(t) = 1, Z_j(t - \Delta) = 0) dt}{\int_0^\tau \lambda(t \mid 0) \cdot P(Z_j(t) = 0, Z_j(t - \Delta) = 1) dt} \approx OR_{\text{case}}.$$

This implies  $\exp(\beta_2) = \frac{OR_{\text{case}}}{OR_{\text{ctrl}}} > 1$ , which means that an association between treatment and event, corresponding to  $\lambda(t \mid 1) > \lambda(t \mid 0)$  for all  $t$ , implies a trend-adjusted OR strictly greater than 1.

## Appendix C: Within cluster dependence

The proposed sampling method gives rise to a dependence between the case and its matched control. This is because both the case and its matched control initiate treatment before the event time of the case, which is not necessarily true for all other subjects. Thus, the independence assumption in the conditional logistic regression is false. This does not influence the estimation of the trend-adjusted OR, but we have to take the dependence into account when we estimate the standard error (SE) of  $\beta_2$  from (1).<sup>1</sup> Here we derive a sandwich estimator that does that. This estimator is easy to use with existing software.<sup>2</sup> Let  $\beta = (\beta_1, \beta_2)$  from model (1), let  $S_i^1(\beta)$  be the contribution to the score function from the  $i$ 'th case, and let  $S_i^2(\beta)$  be the contribution to the score function from the  $i$ 'th control. Then the likelihood equation can be written as:

$$0 = \sum_{i=1}^n S_i^1(\beta) + \sum_{i=1}^n S_i^2(\beta) =: S_n(\beta)$$

This is exactly the equation that is solved by the conditional logistic regression. However,  $S_i^1(\beta)$  and  $S_i^2(\beta)$  for  $i = 1, \dots, n$  are treated as independent by the standard conditional logistic regression when estimating the variance, which they are not due to the fact that the first exposure has to be before the event time of the case for both the case and its matched control, but not necessarily for other observations. Instead the true asymptotic variance is derived as follows: use a Taylor expansion and rearrange to get:

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = -\left(\frac{1}{n}D_{\beta}S_n(\beta_0)\right)^{-1} \cdot \frac{1}{\sqrt{n}} \cdot S_n(\beta_0) + o(1)$$

where  $D_{\beta}S_n(\beta_0)$  is the derivative of  $S_n(\beta_0)$  wrt.  $\beta$ , and  $\widehat{\beta}_n$  is the MLE based on the first  $n$  pairs of cases and controls. Make the following definitions:

$$A_n = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 D_{\beta} S_i^j(\beta_0)$$

$$B_n = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^2 S_i^j(\beta_0) \right)^{\otimes 2}$$

$$A = \lim_{n \rightarrow \infty} A_n$$

$$B = \lim_{n \rightarrow \infty} B_n$$

Then we get the asymptotic distribution of  $\widehat{\beta}_n$ :

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = -A_n^{-1} \frac{1}{\sqrt{n}} S_n(\beta_0) + o(1) \rightarrow \mathcal{N}(0, A^{-1}BA^{-1})$$

The asymptotic variance,  $A^{-1}BA^{-1}$ , is exactly the sandwich estimator.

The sandwich estimator and the default conditional logistic regression estimator of the SE of  $\widehat{\beta}_2$ , i.e. the log of the trend-adjusted OR estimator, are compared in Table 1 based on the simulations from the main text. The default estimator of the SE turns out to be very close to

the standard deviation (SD) of the estimates of  $\beta_2$ . Thus, the sandwich estimator does not improve the precision of the SE estimate considerably in this setup. However, we do not know if this is the case in all practical applications. Furthermore, the sandwich estimator consistently produces a lower SE than the default estimator. Consequently, we gain efficiency by taking the cluster dependence into account. The 95 % (Wald) confidence intervals cover the true trend-adjusted OR very close to 95 % of the time in these simulations when using the sandwich estimator. The confidence intervals cover the true trend-adjusted OR too often when using the default estimator of the SE. Thus, it is still advisable to use the sandwich estimator.

Table 1 about here

## Appendix D: Time-trend in exposure despite uniform treatment initiation

The attentive reader might wonder why we see such a strong time-trend among cases when treatment initiation is simulated from a uniform distribution and treatment and event are independent. The OR for cases in Table 4 is  $355/187 = 1.9$  in a setup where one might expect no time-trend and hence an OR of 1. The time-trend emerges because cases with an event time less than  $D_1 + D_2$ , i.e. 6.75 in this simulation, have a higher chance of being exposed at event than at reference. The extreme case is for those whose event time is before  $D_2$  where the reference time is negative. These cases will by design be exposed at event but not at reference since we only observe drug users, i.e. subjects that have their first treatment in the study period. In this extreme case it is impossible for them to end up in the first three scenarios in Figure 1. Those with an event time between  $D_2$  and  $D_1$  will always be exposed at event but only sometimes at reference so these cases will also, in theory, have an infinite time-trend. They can only end up in the last two scenarios in Figure 1. Those with an event

time between  $D_1$  and  $D_1 + D_2$  can end up in the last three scenarios. They will still have a time-trend of exposure since they are more likely to have treatment initiation between the event and the reference time, and consequently be exposed at event but not at reference, than they are of initiating treatment between time 0 and time  $T - D_1$  where they would be exposed at reference but not at event.

The uniform distribution ensures no time-trend for those with an event time greater than  $D_1 + D_2$ . The exposure patterns for those cases are displayed in Table 7 and, as expected, we get  $OR = 172/169 = 1.02$ , which is close to 1.

Table 2 about here

## References

1. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: Vol 1. University of California Press; 1967:221-233.
2. Holst KK, Scheike TH, Hjelmberg JB. The liability threshold model for censored twin data. *Comput Stat Data Anal.* 2016;93:324-335.

## Tables

**Table 1: Sample means of SEs and coverage probabilities for 95% (Wald) confidence intervals.**

True HR	Sample SD	Sandwich (coverage)	Default (coverage)
1.0	0.39	0.39 (95.3 %)	0.40 (96.0 %)
0.5	0.48	0.47 (95.1 %)	0.49 (96.1 %)
2.0	0.33	0.33 (95.3 %)	0.34 (95.8 %)
5.0	0.29	0.28 (95.5 %)	0.29 (95.7 %)

**Table 2: Exposure patterns among cases with an event time greater than 6.75.**

		Reference	
		Unexposed	Exposed
Event	Unexposed	759	169
	Exposed	172	1327

# Manuscript II

## Unbiased and Efficient Estimation of Causal Treatment Effects in Cross-over Trials

Jeppe Ekstrand Halkjær Madsen, Thomas Scheike & Christian Phipper

**Details:** Submitted to *Biometrical Journal* in 2022.

## Unbiased and Efficient Estimation of Causal Treatment Effects in Cross-over Trials

Jeppe Ekstrand Halkjær Madsen<sup>\*,1,2</sup>, Thomas Scheike<sup>1</sup>, and Christian Pippert<sup>2,3</sup>

<sup>1</sup> Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen K, Denmark.

<sup>2</sup> Biostatistics and Pharmacoepidemiology, Medical Sciences, Leo Pharma A/S, Ballerup, Denmark.

<sup>3</sup> Epidemiology, Biostatistics & Biodemography, Dept. of Public Health, University of Southern Denmark, Denmark.

Received zzz, revised zzz, accepted zzz

We introduce causal inference reasoning to cross-over trials, with a focus on Thorough QT (TQT) studies. For such trials, we propose different sets of assumptions and consider their impact on the modelling strategy and estimation procedure.

We show that unbiased estimates of a causal treatment effect are obtained by a g-computation approach in combination with weighted least squares predictions from a working regression model. Only a few natural requirements on the working regression and weighting matrix are needed for the result to hold. It follows that a large class of Gaussian linear mixed working models lead to unbiased estimates of a causal treatment effect, even if they do not capture the true data generating mechanism.

We compare a range of working regression models in a simulation study where data are simulated from a complex data generating mechanism with input parameters estimated on a real TQT data set. In this setting, we find that for all practical purposes working models adjusting for baseline QTc measurements have comparable performance. Specifically, this is observed for working models that are by default too simplistic to capture the true data generating mechanism.

Cross-over trials and particularly TQT studies can be analysed efficiently using simple working regression models without biasing the estimates for the causal parameters of interest.

*Key words:* Bias; Causal inference; Cross-over trials; Efficiency; TQT studies;

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1022/bimj.XXXXXXX>

### 1 Introduction

A TQT study is an essential component of drug development, ensuring drug safety for patients. Therefore, it is a regulatory requirement to conduct such trials (Food and Drug Administration, 2005). These trials have complicated designs in an attempt to minimize the sample size. This complexity has in turn led to a long-standing debate and several suggestions on how to best model the resulting data, and in particular how to use baseline measurements (Lu, 2014; Kenward and Roger, 2010; Orihashi et al., 2021; Schall and Ring, 2011; Orihashi and Kumagai, 2021). For a standard TQT study, healthy volunteers (International Council for Harmonisation, 2019) are enrolled with the purpose of obtaining electrocardiograms (ECGs) from each subject under different treatment conditions. An ECG measures the electrical output from the heart, resulting in replicates of graphical outputs as illustrated in Figure 1. The output is characterized by different waves, complexes, and intervals. In Figure 1 we see the P, Q, R, S, and T waves. The interval from the beginning of the Q wave to the end of the T wave is called the QT interval, and is measured in milliseconds (ms). The QT interval measures the time it takes the heart to repolarize and prepare for the next beat. The longer the QT interval, the longer the time between heart beats, and the less oxygen is

\*Corresponding author: e-mail: [jehm@sund.ku.dk](mailto:jehm@sund.ku.dk), Phone: +45-31499927

transported to cells in the body. Specifically, prolongation of the QT interval has been shown to be related to an increased risk of Torsades de Pointes, a malignant ventricular arrhythmia. Thus, it is undesirable for the QT interval to be prolonged due to drug exposure. The length of the QT interval is positively associated with the length of the RR interval, i.e. the interval from the R wave on an ECG until the next R wave on the ECG. Therefore, QT intervals are standardized in order to get a QTc (QT corrected) measurement corresponding to a particular length of the RR interval (typically 1 second). An example of a commonly used correction is the Fridericia correction, which is given by

$$QTc = \frac{QT}{\sqrt[3]{RR}}.$$

The purpose of the statistical analysis in a TQT study is to formally assess if clinically relevant prolongation is present based on the QTc measurements (Patterson and Jones, 2006).

TQT studies are predominantly conducted as cross-over trials in an attempt to lower sample size and eliminate between subject variation by paired comparisons of different treatments. In a cross-over trial, each subject is randomized to one of several treatment sequences that uniquely determines the treatment they receive at any given treatment period throughout the trial. Within each period, a baseline QTc measurement is obtained just prior to treatment, followed by a number of post treatment QTc measurements obtained at pre-defined time points following treatment (see Figure 2 for the two-period case). Each pair of consecutive periods will be separated by a washout period to minimize the risk of carry-over effects, i.e. any effects of treatment from the previous period on the QTc measurements in the current period. From a practical point of view, the main challenge with cross-over designs is that the washout period needs to be tailored to the half-life of drug concentration to reflect proper washout of the drug. Specifically, if the half life is long, an even longer wash-out period is required in order to avoid carry-over effects (typically 5 half lives). Ultimately, this may impose a very long study period for the subjects enrolled in the study. This is clearly not optimal and may also prove to be a challenge with regard to case retention. In such situations, a parallel arm design may be more feasible (Food and Drug Administration, 2005).

The causal inference literature, and recently also official regulatory recommendations, have increased the focus on clearly defining what we are actually trying to estimate. This has led to the endorsement of the so called estimand framework within regulatory guidance documents and the causal roadmap concept within the causal inference literature (International Council for Harmonisation, 2019; van der Laan and Rose, 2011). One of the central points made here is that we should enable our research question to be defined in terms of the trial data and not just in terms of a specific model. That said, we would still want to use models in order to gain efficiency or eliminate bias. This is in line with recent regulatory guidance documents that encourage the active use of baseline variables in randomized trials for gaining precision of estimated treatment effects (European Medicines Agency, 2015; Food and Drug Administration, 2021). The developments in this paper aim to support this push towards clearer and more focused statistical procedures. In particular, the causal inference approach we present facilitates a clear and transparent definition of our target of estimation in cross-over trials and specifically in TQT studies. Similar developments are already available in the literature for one-period trials. Here, estimators based on standard regression models have been shown to be unbiased for causal parameters Rosenblum and Steingrimsson (2016); Wang et al. (2021). Moreover, appreciable gains in efficiency compared to marginal estimators for causal effects have been demonstrated Robinson and Jewell (1991); Hernández et al. (2006); Bartlett (2018). We extend these results to cross-over trials, where we show that working mixed models with compound symmetry will allow practitioners to provide sound inference without having to resort to very complex models in an attempt to capture the true data generating mechanism with all the practical challenges, for instance convergence issues, that follow.

The outline of the paper is as follows. In the next section we introduce the basic notation used in the paper and in that context define the fundamental assumption of no carry-over in a cross-over trial. We briefly introduce the concept of counterfactual outcomes and define the causal quantities of interest alongside the data assumptions needed to identify these quantities directly from the data. Section 3 is dedicated

to deriving and assessing the theoretical performance of a number of causally motivated estimators. The section also contains the main result of this paper, which shows that certain types of working models yield unbiased estimates of the causal target parameters under arbitrary misspecification of the working model. In Section 4, we analyse data from a real TQT study using a range of working models that in theory lead to unbiased estimates of the causal target parameters according to the main result of this paper. Section 5 is dedicated to comparing the same working models in a simulation study cast around the data example in order to evaluate performance in a realistic scenario. We conclude the paper with a discussion in Section 6.

## 2 Notation and Assumptions

In the following, we introduce the notation and causal assumptions needed in order to identify the target of estimation. Note that TQT studies are concrete examples of crossover trials with baseline measurements and several outcomes per period. All the results in this paper apply to general crossover trials. Specifically,  $X_p$  in the following may comprise of covariates measured at baseline, or at least independent of treatment assignment, used for covariate adjustment for the outcome(s) in period  $p$ . This could also include period effects and baseline measurements from other periods. Let  $Y_{ipt}$  denote the QTc measurement for subject  $i$  in period  $p$  at time  $t$ ,  $i = 1, \dots, n, p = 1, \dots, P, t = 1, \dots, T$ . Denote the baseline measurement for subject  $i$  in period  $p$  by  $X_{ip}$ , and treatment by  $Z_{ip}$ . TQT studies tend to have as many treatments as periods. Thus, we will denote treatments by  $0, \dots, P-1$ , where  $Z_{ip} = 0$  corresponds to subject  $i$  receiving placebo in period  $p$ . Often we will suppress the  $i$  since we assume the subjects are independent draws from the same distribution. Furthermore, let  $Y_p = (Y_{p1}, \dots, Y_{pT})^T$  denote the vector of post-baseline measurements in period  $p$ .

In line with the informed choice of washout period in TQT studies, we assume the wash-out period has been sufficient to ensure no carry-over effects. Under this assumption, our data can be described by the Directed Acyclic Graph (DAG) in Figure 3 in the two-period case. Note that the DAG has no arrows from baseline measurements to post-baseline measurements, since we do not expect a causal effect of the baseline measurements. Instead, we expect any association between baseline measurements and post-baseline measurements to arise from the latent variables,  $W, W_1$ , and  $W_2$  from Figure 3. The latent variable  $W$  reflects the dependence owing to measurements being from the same subject, whereas the latent variables,  $W_1$  and  $W_2$  reflect the dependence between measurements from the same period, i.e. temporary traits. Despite the lack of arrows between baseline and post-baseline measurements, it still makes sense to adjust for baseline measurements, for example in a regression model, because we do not observe the latent variables, in which case the baseline measurements act as proxies for the latent variables. However, the lack of an arrow between baseline and post-baseline measurements only has an impact on Assumption 1 in the remaining part of the manuscript, and doesn't matter for the theoretical results such as Theorem 1. The DAG in Figure 3 implies the following about the data distribution:

**Assumption 1** (No carry-over) *Let  $x = (x_1, \dots, x_p)^T$ , and likewise for  $z$  and  $y$ , and let  $f_x(x)$  be the density for variable  $x$  and likewise for all other variables. The distribution of our data satisfy the Markov factorization property with respect to the DAG in Figure 3 (Peters et al., 2017), i.e., the joint density of our data can be written as*

$$\begin{aligned} f(z, w, x, y) &= f_z(z) f_w(w, w_1, \dots, w_P) f_x(x|w, w_1, \dots, w_P) f_y(y|z, w, w_1, \dots, w_P) \\ &= f_z(z) f_w(w, w_1, \dots, w_P) \prod_{p=1}^P f_{x_p}(x_p|w, w_p) f_{y_p}(y_p|z_p, w, w_p). \end{aligned}$$

In accordance with Figure 3, Assumption 1 states that the conditional distribution of  $y$  in period  $p$  only depends on treatment in period  $p$  and the latent variables  $w$  and  $w_p$ . Assumption 1 reflects the DAG in

Figure 1, and is a good starting point for exploring the theory behind crossover trials. Unfortunately, more assumptions are needed in order to be able to identify any quantity of interest from data. To do this, we take a causal approach to crossover studies, in the following.

By doing so, we provide a clearly stated research question that is completely disentangled from the modelling of the data. This exercise provides complete clarity on what assumptions about the data generating mechanism are necessary to answer the research question and sets them apart from purely technical assumptions made during the modelling stage of the estimation.

Let  $Y_p^{z_1, \dots, z_P}$  denote the post-baseline QTc measurements we would have made in period  $p$  if, possibly counter to fact, the subject had received treatments  $z_1, \dots, z_P$ . Up front, it seems reasonable to assume that the potential outcome in period  $p$  is independent of future treatments, so that the counterfactual outcomes could be written  $Y_p^{z_1, \dots, z_P}$ . However, according to Assumption 1, the QTc measurements in period  $p$  only depend on the treatment in period  $p$ , and therefore notation can be simplified further. I.e.,  $Y_p^{z_1, \dots, z_P, \dots, z_P} = Y_p^{z_1, \dots, z_P, \dots, z_P} = Y_p^{z_p}$  for all treatment regimes and periods. Thus, from here on,  $Y_p^{z_p}$  will denote the potential QTc measurements in period  $p$  if the subject, possibly counter to fact, had received treatment  $z_p$  in period  $p$ .

If we were particularly interested in treatment  $z$  and had ample resources in terms of money, time, and subjects, we would have made a two-arm trial and used

$$E(Y_t^z - Y_t^0), \quad t = 1, \dots, T,$$

as the causal contrasts of interest. Note that the first period in a cross-over trial corresponds to such a trial and as a consequence the targeted causal contrasts can be identified as:

$$E(Y_{1t}^z - Y_{1t}^0), \quad t = 1, \dots, T.$$

We can then easily estimate the contrasts based on data from the first period only, for example by

$$\frac{\sum_{i=1}^n I(Z_{i1} = z) Y_{i1t}}{\sum_{i=1}^n I(Z_{i1} = z)} - \frac{\sum_{i=1}^n I(Z_{i1} = 0) Y_{i1t}}{\sum_{i=1}^n I(Z_{i1} = 0)}, \quad t = 1, \dots, T.$$

Clearly this is not an efficient use of all the data collected in the cross-over trial. However, to enable full use of the data, stricter assumptions than Assumption 1 are needed. Specifically, we need to make assumptions about the distributions of the post-baseline measurements. To this end, it would be natural to assume the same treatment effect in all periods:

**Assumption 2** (Same treatment effect)

$$E(Y_p^z - Y_p^0) = E(Y_q^z - Y_q^0)$$

for all  $p$  and  $q$ . I.e. the treatment effect is the same in all periods.

Assumption 2 enables estimation of one overall treatment effect across all periods. A special case where Assumption 2 holds is when the distribution of period specific data does not depend on period:

**Assumption 3** (Same distribution)

$$(Y_p^0, \dots, Y_p^{P-1}, X_p, Z_p) \stackrel{\mathcal{D}}{=} (Y_q^0, \dots, Y_q^{P-1}, X_q, Z_q),$$

for all  $p$  and  $q$ .

Assumption 3 is rather restrictive compared to Assumption 2 in that it a priori excludes any systematic effect due to period. This contradicts current modelling and design practice in cross-over trials, where potential period effects are modelled and subjects are randomized according to a latin square in order to balance any period effect (Senn, 2002).

As an alternative, one can assume that the conditional distribution of the post-baseline measurements, given covariates, is the same in all periods:

**Assumption 4** (Same relationship)

$$(Y_p | X_p = x, Z_p = z) \stackrel{D}{=} (Y_q | X_q = x, Z_q = z)$$

for all  $p$  and  $q$ .

This facilitates a model fit based on data from all periods, which may in turn be used to infer the period specific causal contrast  $E(Y_p^z - Y_p^0)$ .

Assumption 4 may initially appear paradoxical in light of the DAG depicted in Figure 1, along with Assumption 1. Specifically, it may seem surprising that it doesn't involve the unmeasured confounders. Essentially, Assumption 4 posits a hypothesis about the structure of these unmeasured confounders and their impact on the outcome. This assumption can be satisfied if, for instance, the unmeasured confounders' distribution remains the same throughout all periods, and the covariates affect the outcome in the same way in all periods. Hence, Assumption 4 entails weaker causal assumptions, but necessitates stronger assumptions on the data distribution to guarantee the identifiability of a causal effect. In general we will make Assumption 2 in the remaining part of the manuscript, although we mention in the discussion what happens if it is not satisfied, and how Assumption 4 can be used to get an estimate of a causal effect.

### 3 Estimation

The causal framework from the last section enables us to clearly define our target of estimation. In this section, we provide inference procedures tailored to assess the causal target parameter in the context of a TQT study. For brevity, we only consider the case, where we have assumed the same average causal treatment effect in all periods. I.e., we specifically develop estimators for  $E(Y_1^z - Y_1^0)$  under Assumption 2. An outline of how to estimate period specific average causal effects without Assumption 2 is provided in Section 6.

Under Assumption 2 the fact that subjects receive both the placebo and active treatment initially motivates the following simple non-parametric estimator:

$$\hat{\mu}_{1t}(z) = \frac{1}{n} \sum_{i=1}^n \sum_{p=1}^P I(Z_{ip} = z) Y_{ipt} - I(Z_{ip} = 0) Y_{ipt},$$

where the indicator function,  $I(A)$ , maps elements of  $A$  to one and is zero otherwise. Note that this, and all other estimators throughout this paper, has  $t$  in the subscript to indicate that it is the estimator for the post-baseline measurement time point  $t$ . Naturally, an effect will be estimated for each  $t = 1, \dots, T$ . Note that  $\hat{\mu}_{1t}(z)$  simply takes the outcomes in the periods, where the subjects receive the treatment of interest, and subtracts the outcomes in the placebo periods, and averages across subjects. It is an unbiased estimator for the treatment effects of interest due to the randomization. However, it only uses post-baseline measurements, and may therefore lack precision. Alternatively, one can pursue fitting a working regression model to the data and thereby bring baseline measurements into play. We specifically assume that the possible outcome predictions in this working model are given by  $h_{pt}(x, z, \beta)$ , where  $\beta$  is a vector of regression parameters. Under Assumption 3 it is possible to ignore periods and use the simpler working regression model  $h_t(x, z, \beta)$ . Such a working model can be used to estimate the causal effects of interest simply by plugging into the g-computation formula (Robins, 1986):

$$\hat{\mu}_{2t}(z) = \frac{1}{n \cdot P} \sum_{i=1}^n \sum_{p=1}^P [h_{pt}(X_{ip}, z, \hat{\beta}) - h_{pt}(X_{ip}, 0, \hat{\beta})].$$

This estimator uses covariate information to gain efficiency. Moreover, in situations with missing endpoint data, the estimator is still unbiased under the missing at random assumption given that the regression

model is correctly specified. In comparison, the endpoint data has to be missing completely at random for the simple estimator  $\hat{\mu}_{1t}(z)$  to be unbiased.

Up front, the above developments depend on the fact that the working regression model is specified so that it captures the true mean value structure. Since this by no means is warranted, it is important to mitigate the impact in terms of bias if the working model is misspecified. Such a mitigation can be successfully achieved with the following semi-parametric estimator:

$$\hat{\mu}_{3t}(z) = \hat{\mu}_{1t}(z) - \frac{1}{n} \sum_{i=1}^n \sum_{p=1}^P \left[ \left( I(Z_{ip} = z) - \frac{1}{P} \right) h_{pt}(X_{ip}, z, \hat{\beta}) - \left( I(Z_{ip} = 0) - \frac{1}{P} \right) h_{pt}(X_{ip}, 0, \hat{\beta}) \right].$$

The estimator is derived in Web Appendix A. It uses covariates to gain precision, but is unbiased due to the independence between  $X_p$  and  $Z_p$ . This independence is ensured by the randomization and implies that the last term has a mean of zero. Thus, the estimator has the same mean as the non-parametric estimator, namely the true causal effect. The following theorem shows that  $\hat{\mu}_{3t}(z) = \hat{\mu}_{2t}(z)$  for certain types of working models.

**Theorem 1** (Unbiasedness of  $\hat{\mu}_{2t}(z)$ ) *Assume the data structure of this paper, and assume we use a working model with post baseline measurements as outcome, and with main effects of treatment specific to each post baseline time point. Let  $Y_i$  be the vector of all post baseline measurements for subject  $i$ , and let  $h(X_i, Z_i, \beta)$  be the vector of all predictions for subject  $i$  from the working model. Assume the working model parameters,  $\beta$ , are estimated from the following weighted least squares estimating equation:*

$$\sum_{i=1}^n D_i V^{-1} (Y_i - h_i(X_i, Z_i, \beta)) = 0, \quad (1)$$

where  $D_i$  is the design matrix for subject  $i$ , and  $V$  is a weight matrix on the form

$$\begin{pmatrix} A & B & \cdots & B \\ B & A & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & A \end{pmatrix}, \quad (2)$$

where  $A$  and  $B$  are  $T \times T$  matrices. If  $A - B$  is non-singular, then

$$\hat{\mu}_{2t}(z) = \hat{\mu}_{3t}(z).$$

**Proof.** The proof is provided in Web Appendix B. □

We know that  $\hat{\mu}_{3t}(z)$  is unbiased by construction. Hence, Theorem 1 implies unbiased estimation when using  $\hat{\mu}_{2t}(z)$  with a working model satisfying the conditions in Theorem 1, no matter how misspecified the working model happens to be. There have been a lot of discussions about the right choice of model for TQT studies, but Theorem 1 implies the existence of a whole range of models we can use without fear of biased estimation of treatment effects in crossover trials. We stress that  $h(X_i, Z_i, \beta)$  is a vector of all the predictions made for all the outcomes of subjects  $i$ . In particular  $X_i$  here, admittedly with some violation of notation, is just the covariates that are adjusted for. This could be the baseline measurements from the same periods, baseline measurements from all periods, as argued for by Kenward and Roger (2010), other covariates such as sex and age, or it could even be no covariates except for the treatment by time point effect required by the Theorem. This last point is partly reflected by the fact that  $\hat{\mu}_{1t}(z)$  is unbiased despite not adjusting for covariates, since this estimator can be achieved as a very special case of Theorem 1. However, adjustment for baseline measurements is a good idea in order to improve efficiency of the estimator despite bias not being a concern. However, it would be beneficial to know whether any

popular choices of models happen to satisfy the conditions in Theorem 1. The following Corollary shows that a big class of the most popular types of models for crossover trials satisfy the conditions of Theorem 1, and thereby ensure unbiased estimation of causal effects under arbitrary misspecification of the applied working model.

**Corollary 1** (Gaussian linear mixed models) *Assume the working model is a Gaussian linear mixed model with main effects of treatment specific to each post baseline time point, and a correlation structure satisfying (2). Then,  $\hat{\mu}_{2t}(z) = \hat{\mu}_{3t}(z)$  if model parameters are estimated by maximum likelihood or restricted maximum likelihood estimation.*

**Proof.** The estimating equation for Gaussian linear mixed models is given on page 10 in Jiang (2007) and can be rewritten to (1). This is the case both for MLE and REML, although variance parameters, and thereby the  $V$  matrix from Theorem 1 differs (see page 14 in Jiang (2007)).  $\square$

When the working model is a Gaussian linear mixed model, the requirement (2) corresponds to modelling the dependence within periods by some matrix  $A$ , and the dependence between periods by a matrix  $B$ . For example,  $B$  might be a matrix of constants corresponding to a random subject effect, and  $A$  can be modelled more flexibly, for example with an unstructured covariance structure, or an AR(1) covariance structure. When we only have one outcome per period, the assumption correspond to using compound symmetry for the correlation structure. In the special case of a Gaussian linear mixed model, consider  $A = \sigma^2 I$  and  $B$  a matrix of zeros. In this case, the working model is a standard linear regression model that ignores any dependence between observations. Corollary 1 then ensures that we are able to produce sound inference even with this simplistic model. We do, however, expect this working model to be less efficient than if we model the dependence structure in a Gaussian linear mixed model. A particular example of a much used working regression model is given by Patterson and Jones (2006):

$$h_{pt}(x, z, \beta) = \beta_{pt} + \beta_x x + \beta_{zt}. \quad (3)$$

Clearly the conditions of Theorem 1 are satisfied for the systematic part of this model, as it includes a main effect of treatment specific to each post baseline time point. Additionally, the covariance structure proposed in Patterson and Jones (2006) is AR(1) within periods, and assuming constant covariance between observations on the same subject in different periods. Accordingly, the proposed covariance structure complies with (2). The estimates of the time specific effects of treatment  $\beta_{zt}$  equal the estimates obtained if we were to plug model (3) into  $\hat{\mu}_{2t}(z)$ . Thus, the estimates of  $\beta_{zt}$  are unbiased for the treatment effects of interest under arbitrary model misspecification.

In order to enable inference fully, we further need to characterize the large sample behaviour. This is well established if the targeted treatment effects appear as parameters in the model, and assuming that the model is correctly specified. However, in more complex models, the target treatment effect is not readily identified as a parameter specified in the model. Moreover, model based standard errors may not be appropriate, unless the model is correctly specified. The influence function of  $\hat{\mu}_{2t}(z)$  is derived in Web Appendix C under the assumption that the  $\beta$  parameters are estimated using an M-estimator. For models not covered by Theorem 1,  $\hat{\mu}_{3t}(z)$  is still unbiased whereas  $\hat{\mu}_{2t}(z)$  may be biased. Therefore, it would be preferable to use  $\hat{\mu}_{3t}(z)$  in such cases. Accordingly, we also derive the influence function for  $\hat{\mu}_{3t}(z)$  in Web Appendix A.

One particular use of the above asymptotic results is when assessing QT prolongation in a TQT trial. In this context, the test for QTc prolongation for the drug of interest is carried out by use of an intersection-union test. I.e. the null-hypothesis is

$$H_0: \bigcup_{t=1}^T \mu_t(z) \geq \Delta,$$

where

$$\mu_t(z) = E(Y_{pt}^z - Y_{pt}^0),$$

and  $\Delta$  is some reasonable amount of QT prolongation, such as 10 ms (Patterson and Jones, 2006). Commonly speaking, the null-hypothesis dictates that there exists a time point where QT prolongation exceeds a prespecified clinically negligible threshold. Tests are carried out for each  $t$  based on the asymptotic behaviour of the standardized estimates of  $\mu_t(z)$ , and the null is rejected if all of these tests are rejected.

In addition, it is common practice to assess if prolongation can be detected for the positive control. The corresponding null-hypothesis is

$$H_0: \bigcap_{t=1}^T \mu_t(z) \leq \Delta.$$

The test of this hypothesis needs to be adjusted for multiple testing. This adjustment should be made in the most efficient way possible, which is possible because we can estimate the joint (asymptotic) distribution of our estimators (Pipper *et al.*, 2012; Hothorn *et al.*, 2008). We can derive the joint asymptotic distribution of our estimators from the influence functions as

$$\sqrt{n}(\hat{\mu}_2(z) - \mu(z)) = \begin{pmatrix} \hat{\mu}_{21} - \mu_1(z) \\ \vdots \\ \hat{\mu}_{2T} - \mu_T(z) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \varphi_1(d_i) \\ \vdots \\ \varphi_T(d_i) \end{pmatrix} + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(d_i) + o_p(1),$$

where  $\varphi_t(d_i), t = 1, \dots, T$  are the influence functions of the individual estimators. It follows that the asymptotic variance matrix of the joint distribution of our estimators equals

$$\frac{E(\varphi^T \varphi)}{n}. \quad (4)$$

The estimation of targeted treatment effects as outlined above is based on well known regression models and as we have seen the targeted treatment effects may sometimes even be identified directly as regression parameters in those models. Theorem 1 then guarantees that the identified regression parameters are estimated without bias, irrespective of whether the regression model is correctly specified or not. This guarantee, however, does not extend to the model based variance matrix of the estimates. For this to be appropriate, one also needs to assume that the regression model is correct. The asymptotic variance matrix (4) on the other hand is generally applicable. When the targeted treatment effects are identified as regression parameters, it may even be obtained by standard software (Pustejovsky, 2022).

The asymptotic theory developed is applicable as is, when the sample size is large. However, TQT trials and many cross-over trials have a rather small sample size, in which case the appropriateness of the asymptotic theory is questionable. For such cases, there is a substantial literature on how to improve asymptotic standard errors and confidence intervals (Bell and McCaffrey, 2002; Colin Cameron and Miller, 2015; MacKinnon and White, 1985; Pustejovsky and Tipton, 2018). One simple improvement of the asymptotic standard errors from Colin Cameron and Miller (2015) is to use the influence function in the same way as we would in the context of a linear normal model. I.e. to estimate the variance by  $\frac{1}{n-1} \sum_{i=1}^n \hat{\varphi}_i^2$  instead of  $\frac{1}{n} \sum_{i=1}^n \hat{\varphi}_i^2$ , and use the 97.5% quantile of the t-distribution with  $n - 1$  degrees of freedom instead of a standard normal distribution to construct confidence intervals. These modifications vanish as the sample size increases, and will consequently lead to correct asymptotic inference. These simple small sample modifications are used in the analyses and simulations throughout this paper.

## 4 Data Example

To illustrate the developments in this paper, we reanalyse a standard TQT trial also analysed in chapter 9 of Patterson and Jones (2006). The data set is freely available on the book website, and consists of 41

subjects, two of which we have excluded due to missingness. There are three single-dose treatments (C, D, E) and a placebo (F). Treatment E is included as a positive control, i.e. treatment E is known to mildly prolong the QT interval. The subjects' QT intervals are measured in triplicates at baseline, 0.5 hours, 1 hour, 1.5 hours, 2.5 hours, and 4 hours post treatment. The triplicates are averaged at each time point.

In accordance with the developments presented in Section 3, we analyse the data with a range of models of differing complexity that, in theory, facilitate unbiased estimates of the average causal effects under Assumption 2. We informally compare these models in terms of obtained estimates and standard errors.

Our benchmark model corresponds to the recommendation made in Lu (2014). This paper advocates a regression model including average baseline measurements as a covariate. It is shown in Lu (2014) that this approach is consistent with the joint baseline and post baseline measurement model advocated in Kenward and Roger (2010) and Meng et al. (2010). It is further argued in Lu (2014) that the resulting estimates of the treatment effects will be superior in terms of precision.

Specifically, the working regression model proposed in Lu (2014) is given by:

$$h_{pt}(x, z, \beta) = \beta_{pt} + \beta_{xt}x + \beta_{\bar{x}t}\bar{x} + \beta_{zt}, \quad (5)$$

where  $\bar{X}$  is the average baseline measurement. The effect of the baseline measurements and average baseline measurements are different at different time points.

Moreover, we fit a simpler model with mean structure:

$$h_{pt}(x, z, \beta) = \beta_{pt} + \beta_{xt}x + \beta_{zt},$$

i.e. a model without average baseline measurements, but still with interaction between the effect of baseline and time point.

Furthermore, we fit an even simpler model with mean structure

$$h_{pt}(x, z, \beta) = \beta_t + \beta_{xz}x + \beta_{zt},$$

i.e. without average baseline measurement, no interaction between time point and period, and the effect of the baseline measurement is the same at all time points. All the models above have the treatment effects as specific parameters. In order to illustrate the modelling flexibility facilitated by Theorem 1, we fit a model with interaction between baseline and treatment:

$$h_{pt}(x, z, \beta) = \beta_t + \beta_{xz}x + \beta_{zt}.$$

Note that with the complexity of this model, it is no longer possible to identify the average causal effect as a parameter in the regression model. Therefore, we can no longer rely on standard inference of regression models, but need to rely on the general inference procedures developed in Section 3.

On top of specifying a mean structure for the models, we also need to specify the working covariance structure. The working covariance structure has to be on the form (2), and in the following we will use a random subject effect corresponding to  $B$  being a matrix of constants unless otherwise specified. We consider three different specifications for the  $A$  matrix:

1. Unspecified: all the variances and covariances have to be estimated.
2. AR(1): variance matrix is on the form:

$$A = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{pmatrix},$$

where  $\rho$  and  $\sigma^2$  are parameters to be estimated.

3. Independence:  $A = \sigma^2 I$ , where  $I$  is the identity matrix. In this case,  $B$  will be a matrix of zeros corresponding to a standard linear regression model.

Last, we also fit the non-parametric estimator  $\hat{\mu}_{1t}(z)$ . The estimates of the effect of treatment E compared to placebo at post baseline time 4 from the models are displayed in Table 1. Note that the standard errors and confidence intervals are based on the small sample size adjustment discussed in the last section. The remaining estimates of treatment effects are presented in Web Appendix D. In practice, the main reason for choosing one model over another is in order to have as much efficiency as possible, and not because we expect to actually know the true data-generating mechanism. From Table 1 we note that estimates of the targeted treatment effect across all models seem comparable. The standard errors are substantially larger with the non-parametric approach, whereas for all other model based approaches, standard errors are comparable. We investigate these observations further in a simulation study mimicking the data example in the next section.

## 5 Simulation Study

Simulation studies have already demonstrated that it is theoretically possible to gain precision by including the average baseline measurement as a covariate (Lu, 2014; Meng *et al.*, 2010). However, it is unclear how much the addition of the average baseline covariates matters for the precision in realistic setups. Therefore, we have based our simulation on the data set from the previous section.

Specifically, we have simulated the data as follows: a joint normal distribution of the baseline measurements is fitted to the baseline measurements in the data set, and baseline measurements are simulated according to this fit. The model with mean structure (5) and unspecified covariance structure between observations from the same period is fitted to the data set, and the post baseline measurements are simulated from that model.

There are at least two advantages to this approach: first, the simulation must be considered realistic since it is based on parameters estimated on a data set from a real TQT trial. Second, we know that models ignoring the average baseline measurements are too simple to capture the true data-generating mechanism. Thereby, the simulations can show us how much precision we can expect to lose by not using average baseline measurements as covariates in real TQT studies.

We compare all the models from the last section. The models are misspecified both in terms of mean structure and in terms of correlation structure, but are all unbiased by Corollary 1. We ran 10000 simulations in the statistical program R (R Core Team, 2021), and the code is available at <https://github.com/Jeepen/TQTpaper>. The results of the simulations are displayed in Table 2. Standard errors and confidence interval coverages are again based on an adjustment for small sample size. As expected, all estimators have negligible bias. The standard error estimates seem approximately correct compared to the sample standard deviation of the estimates, and the coverages of the confidence intervals are consequently close to 95 %. The standard deviations in the table show that by far the majority of the gain in precision from using a model comes from the inclusion of the baseline measurements. Using the average baseline measurement as a covariate adds little precision, even over the linear regression model. However, the gain in precision from including the average baseline as a covariate and using a more flexible covariance structure than independence is for free, since Corollary 1 implies unbiased estimation in any case.

## 6 Discussion

We have introduced causal reasoning to the field of TQT studies. We have shown how typical choices of estimators can be given very clear causal interpretations in terms of the data, and not just in terms of specific models. Furthermore, we have shown that popular choices of working models in many circumstances yield

unbiased estimates of causal parameters under arbitrary model misspecification. We have illustrated these results in a data example and a simulation study.

However, the unbiasedness of the proposed estimators follows from the balancing induced by randomization. In practice, we may have missing data, which can invalidate the balancing initially ensured by the randomization. That said, the amount of missing data is typically negligible in TQT studies, owing to the fact that they are most often conducted on healthy subjects who are paid to participate (Food and Drug Administration, 2005). In cases with a non-negligible amount of missing data the working regression models can be fitted to the observed data with MLE assuming missingness at random (MAR), and distinct parameters for the missing data mechanism and for the outcome. As a consequence, it is straightforward to estimate the causal effects when using a linear mixed model, where the causal effects are identified as the main effects of treatment. However, this becomes challenging if we consider more complex models. In that case, a viable strategy could be to use imputation or weighting methods in order to go from model to estimates of the causal effects (Little and Rubin, 2020; Tsiatis, 2006).

We have focused on how to estimate causal effects under assumption (2), that is, assuming the effects are the same in all periods. To complement these developments, it is interesting to consider what we are estimating if that assumption does not hold. In general, the mean of  $\hat{\mu}_{1t}(z)$  equals

$$\frac{1}{P} \sum_{p=1}^P E(Y_{pt}^z - Y_{pt}^0),$$

i.e. the average of the causal effects for each period.  $\hat{\mu}_{3t}(z)$  has the same mean due to the randomization, and  $\hat{\mu}_{2t}(z)$  will equal  $\hat{\mu}_{3t}(z)$  when estimation is done according to Theorem 1. Thus, in general, the above strategy will lead to unbiased estimation of the average of the period specific causal effects. Alternatively, one may fit a working model to all data under Assumption 4, and subsequently apply g-computation for a single period:

$$\frac{1}{n} \sum_{i=1}^n h(X_{iPt}, z, \hat{\beta}) - h(X_{iPt}, 0, \hat{\beta}). \quad (6)$$

The estimation of  $h$  gains precision by using data from all periods, which in turn makes (6) more precise. The estimator (6) emulates a standard one-period trial, while it is unclear what we are emulating by ignoring the period specific treatment effects (Hernán et al., 2008). This is a topic for further research.

Two other theoretical issues also deserve more attention. First, we have not theoretically shown that  $\hat{\mu}_{2t}(z)$  or for that matter  $\hat{\mu}_{3t}(z)$  are in fact more efficient than the non-parametric estimator  $\hat{\mu}_{1t}(z)$ . We, however, suspect this to be the case in line with the results obtained for one-period trials in Bartlett (2018) and van der Laan and Rose (2011). Second, the impact of a violation of restrictions on the working model dictated by Theorem 1 deserves further investigation. Possibly, a more flexible weight matrix than what is warranted by Theorem 1 may lead to further efficiency gains.

Finally, we would like to point out the difference between the estimation procedure proposed in this paper and the traditional approach of reporting treatment effects based on differences in least squares means (see for example chapter 8 of Patterson and Jones (2006)). It is duly noted that differences in least squares means are equivalent to  $\hat{\mu}_{2t}(z)$  when the treatment effects are modelled as main effects in a linear mixed model. However, they are not equivalent to  $\hat{\mu}_{2t}(z)$  when the model is more complex. In those cases, least squares means lack a proper causal interpretation in a meaningful population and on those grounds  $\hat{\mu}_{2t}(z)$  or  $\hat{\mu}_{3t}(z)$  should be preferred for assessing causal treatment effects.

#### Conflict of Interest

*The authors have declared no conflict of interest.*

## References

- Bartlett, J. W. (2018). Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical statistics* **17**, 648–666.
- Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* **28**, 169–182.
- Colin Cameron, A. and Miller, D. L. (2015). A Practitioner’s Guide to Cluster-Robust Inference. *Journal of Human Resources* **50**, 317–372.
- European Medicines Agency (2015). European Medicines Agency Guideline on Adjustment for Baseline Covariates in Clinical Trials. *European Medicines Agency: CPMP/295050/2013*.
- Food and Drug Administration (2005). International Conference on Harmonisation; guidance on E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs; availability. Notice. *Federal Register* **70**, 61134–61135.
- Food and Drug Administration (2021). Adjusting for covariates in randomized clinical trials for drugs and biologics with continuous outcomes guidance for industry. URL <https://www.fda.gov/media/148910/download>.
- Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. *Epidemiology* **19**, 766–779.
- Hernández, A. V., Steyerberg, E. W., Butcher, I., Mushkudiani, N., Taylor, G. S., Murray, G. D., Marmarou, A., Choi, S. C., Lu, J., Habbema, J. D. F., and Maas, A. I. (2006). Adjustment for Strong Predictors of Outcome in Traumatic Brain Injury Trials: 25% Reduction in Sample Size Requirements in the IMPACT Study. *Journal of Neurotrauma* **23**, 1295–1303.
- Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal* **50**, 346–363.
- International Council for Harmonisation (2019). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9 (R1).
- Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer.
- Kenward, M. G. and Roger, J. H. (2010). The use of baseline covariates in crossover studies. *Biostatistics* **11**, 1–17.
- Little, R. J. A. and Rubin, D. B. (2020). *Statistical analysis with missing data*. Wiley, third edition.
- Lu, K. (2014). An efficient analysis of covariance model for crossover thorough QT studies with period-specific pre-dose baselines. *Pharmaceutical Statistics* **13**, 388–396.
- MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–325.
- Meng, Z., Quan, H., Fan, L., Kringle, R., and Sun, G. (2010). Use of the Average Baseline Versus the Time-Matched Baseline in Parallel Group Thorough QT/QTc Studies. *Journal of Biopharmaceutical Statistics* **20**, 665–682.
- Orihashi, Y. and Kumagai, Y. (2021). Concentration-QTc analysis with two or more correlated baselines. *Journal of Pharmacokinetics and Pharmacodynamics* **48**, 615–622.
- Orihashi, Y., Kumagai, Y., and Shiosakai, K. (2021). Novel concentration-qt models for early clinical studies with parallel placebo controls: A simulation study. *Pharmaceutical Statistics* **20**, 375–389.
- Patterson, S. D. and Jones, B. (2006). *Bioequivalence and statistics in clinical pharmacology*. Chapman & Hall/CRC.
- Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
- Pipper, C. B., Ritz, C., and Bisgaard, H. (2012). A versatile method for confirmatory evaluation of the effects of a covariate in multiple models: Evaluation of Effects of a Covariate in Multiple Models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**, 315–326.
- Pustejovsky, J. (2022). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.5.6.
- Pustejovsky, J. E. and Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics* **36**, 672–683.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.
- Robinson, L. D. and Jewell, N. P. (1991). Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique* **59**, 227–240.
- Rosenblum, M. and Steingrimsson, J. A. (2016). Matching the efficiency gains of the logistic regression estimator while avoiding its interpretability problems, in randomized trials. *Johns Hopkins University, Dept. of Biostatistics Working Papers Working Paper* 281.
- Schall, R. and Ring, A. (2011). Mixed models for data from thorough QT studies: part 1. assessment of marginal QT prolongation. *Pharmaceutical Statistics* **10**, 265–276.
- Senn, S. (2002). *Cross-over trials in clinical research*. John Wiley & Sons, 2nd edition.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.
- van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. Springer.
- Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2021). Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment. *Journal of the American Statistical Association* pages 1–12.

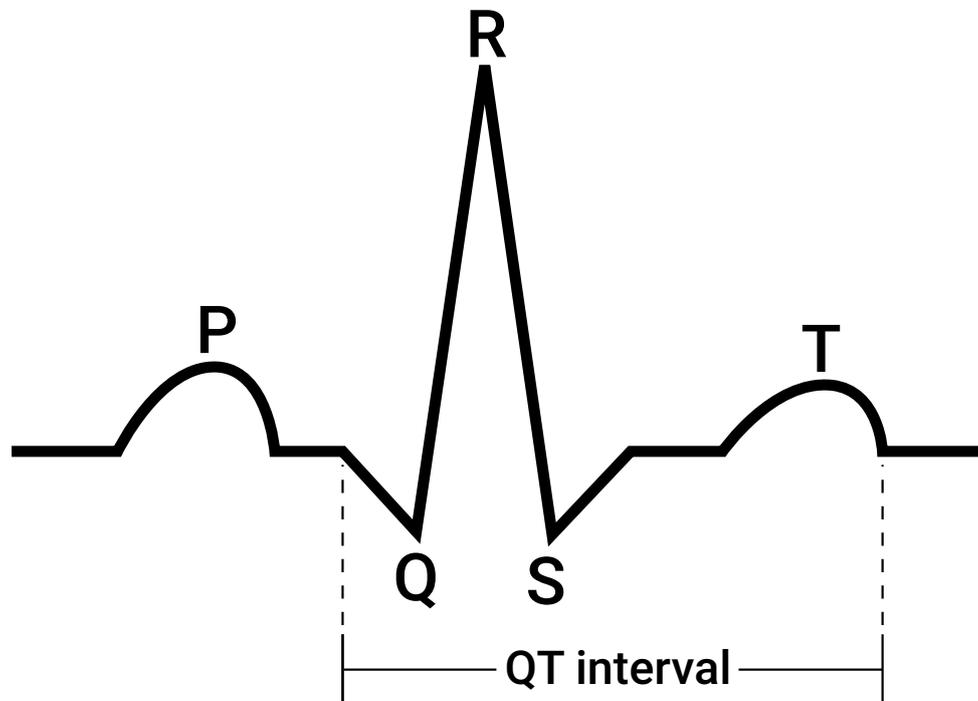


Figure 1 ECG.

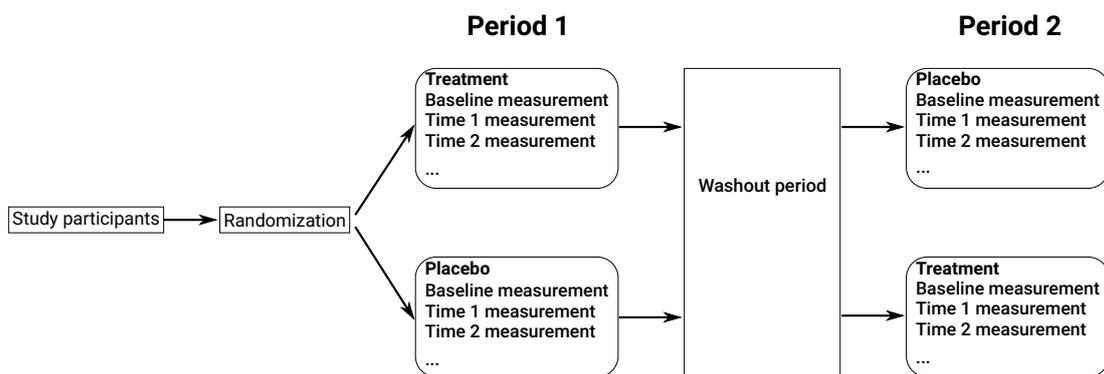
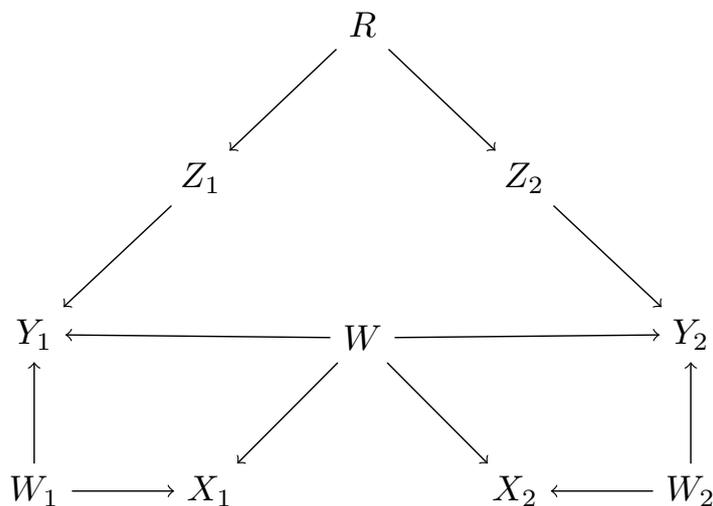


Figure 2 Cross-over trial design.



**Figure 3**  $W, W_1, W_2$  = subject specific latent variables,  $R$  = Randomization.

**Table 1** Estimates and standard errors for data example.

Mean structure	Covariance structure	Estimate	Standard Error	95% CI
$\beta_{pt} + \beta_{xt}x + \beta_{\bar{x}t}\bar{x} + \beta_{zt}$	Unspecified	8.32	1.53	(5.22, 11.42)
	AR(1)	8.11	1.49	(5.09, 11.14)
	Independence	8.19	1.54	(5.07, 11.32)
$\beta_{pt} + \beta_{xt}x + \beta_{zt}$	Unspecified	8.22	1.52	(5.14, 11.31)
	AR(1)	8.09	1.48	(5.11, 11.08)
	Independence	8.19	1.52	(5.12, 11.26)
$\beta_t + \beta_x x + \beta_{zt}$	Unspecified	8.49	1.43	(5.59, 11.40)
	AR(1)	8.30	1.44	(5.38, 11.22)
	Independence	8.43	1.44	(5.51, 11.34)
$\beta_t + \beta_{xz}x + \beta_{zt}$	Unspecified	8.43	1.40	(5.60, 11.26)
	AR(1)	8.29	1.43	(5.49, 11.10)
	Independence	8.42	1.42	(5.55, 11.30)
$\hat{\mu}_{1t}(z)$		8.18	2.05	(4.03, 12.33)

**Table 2** Bias, standard deviation of estimates, and coverage of confidence intervals in simulations.

Mean structure	Covariance structure	Bias	SD	Avg. SE	Coverage
$\beta_{pt} + \beta_{xt}x + \beta_{\bar{x}t}\bar{x} + \beta_{zt}$	Unspecified	0.02	1.48	1.43	0.947
	AR(1)	0.02	1.48	1.43	0.947
	Independence	0.02	1.48	1.48	0.954
$\beta_{pt} + \beta_{xt}x + \beta_{zt}$	Unspecified	0.01	1.49	1.45	0.945
	AR(1)	0.01	1.49	1.45	0.945
	Independence	0.01	1.52	1.52	0.952
$\beta_t + \beta_x x + \beta_{zt}$	Unspecified	0.02	1.48	1.46	0.951
	AR(1)	0.02	1.48	1.46	0.951
	Independence	0.02	1.51	1.51	0.954
$\beta_t + \beta_{xz}x + \beta_{zt}$	Unspecified	0.02	1.48	1.38	0.939
	AR(1)	0.02	1.47	1.38	0.939
	Independence	0.02	1.50	1.49	0.951
$\hat{\mu}_{1t}(z)$		0.02	1.94	1.94	0.952

## Supplementary Information for "Unbiased and Efficient Estimation of Causal Treatment Effects in Cross-over Trials"

Jeppe Ekstrand Halkjær Madsen<sup>\*,1,2</sup>, Thomas Scheike<sup>1</sup>, and Christian Pippert<sup>2,3</sup>

<sup>1</sup> Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen K, Denmark.

<sup>2</sup> Biostatistics and Pharmacoepidemiology, Medical Sciences, Leo Pharma A/S, Ballerup, Denmark.

<sup>3</sup> Epidemiology, Biostatistics & Biodemography, Dept. of Public Health, University of Southern Denmark, Denmark.

Received zzz, revised zzz, accepted zzz

**Key words:** Bias; Causal inference; Cross-over trials; Efficiency; TQT studies;  
Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1022/bimj.XXXXXXX>

### 1 Web Appendix A: Derivation of $\hat{\mu}_{3t}(z)$

The estimator is derived in the same way as the semi-parametric estimator for the pretest-posttest study in Tsiatis (2006), but with  $\hat{\mu}_{1t}(z)$  as the non-parametric starting point. Note that this won't result in the efficient estimator in our setup, since we also have the assumption of the same treatment effect in all periods, which is not used in the derivation.

The influence function of  $\hat{\mu}_{1t}(z)$  is given by

$$\varphi_1(D) = \sum_{p=1}^P [I(Z_p = z)Y_{pt} - I(Z_p = 0)Y_{pt}] - \mu_t(z),$$

where  $\mu_t(z) = E(Y_{pt}^z - Y_{pt}^0)$  is the true causal effect, which we remind the reader is independent of period under Assumption 2, and  $D$  consists of all the data from the subject. We derive the estimator  $\hat{\mu}_{3t}(z)$  by the same calculations as Tsiatis (2006), namely

$$\varphi_3(D) = \varphi_1(D) - (E(\varphi_1(D)|\bar{X}, \bar{Z}) - E(\varphi_1(D)|\bar{X})), \quad (1)$$

where  $\bar{X} = \{X_1, \dots, X_P\}$ , and likewise for  $\bar{Z}$ . We calculate the expectations:

$$\begin{aligned} E(\varphi_1(D)|\bar{X}, \bar{Z}) &= E\left(\sum_{p=1}^P I(Z_p = z)Y_{pt} - I(Z_p = 0)Y_{pt} - \mu_t(z)|\bar{X}, \bar{Z}\right) \\ &= \sum_{p=1}^P [I(Z_p = z)E(Y_{pt}|\bar{X}, \bar{Z}) - I(Z_p = 0)E(Y_{pt}|\bar{X}, \bar{Z})] - \mu_t(z) \\ &= \sum_{p=1}^P I(Z_p = z) [E(Y_{pt}|X_p, Z_p = z) - I(Z_p = 0)E(Y_{pt}|X_p, Z_p = 0)] - \mu_t(z), \end{aligned}$$

\*Corresponding author: e-mail: [jehm@sund.ku.dk](mailto:jehm@sund.ku.dk), Phone: +45-31499927

and

$$\begin{aligned} E(\varphi_1(D)|\bar{X}) &= \sum_{p=1}^P [\mathbb{P}(Z_p = z)E(Y_{pt}|X_p, Z_p = z) - \mathbb{P}(Z_p = 0)E(Y_{pt}|X_p, Z_p = 0)] - \mu_t(z) \\ &= \sum_{p=1}^P \left[ \frac{1}{P}E(Y_{pt}|X_p, Z_p = z) - \frac{1}{P}E(Y_{pt}|X_p, Z_p = 0) \right] - \mu_t(z), \end{aligned}$$

The expectations  $E(Y_{pt}|X_p, Z_p = z)$  and  $E(Y_{pt}|X_p, Z_p = 0)$  require a model,  $h$ .

When plugging the above into (1) we get an estimator of the influence function,  $\varphi_3(D)$ , and the estimator  $\hat{\mu}_{3t}(z)$  is obtained by simple isolation from the equation

$$\sqrt{n}(\hat{\mu}_{3t}(z) - \mu_t(z)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_3(D_i) + o_P(1),$$

which defines influence functions.

### 1.1 Influence Function for $\hat{\mu}_{3t}(z)$

It might seem like the influence function was derived above. However, one could imagine that the estimation of the models in the first step would change the influence function, so we get a term like  $G_\beta\psi(d)$  in the case of  $\hat{\mu}_{2t}(z)$ .

We can simply take the estimator  $\hat{\mu}_{3t}(z)$ , subtract  $\mu_t(z)$  and multiply by  $\sqrt{n}$  in order to get

$$\begin{aligned} \sqrt{n}(\hat{\mu}_{3t}(z) - \mu_t(z)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n -\mu_t(z) + \sum_{p=1}^P I(Z_p = z)Y_{pt} - I(Z_p = 0)Y_{pt} \\ &\quad - \left[ \left( I(Z_{ip} = z) - \frac{1}{P} \right) h_{pt}(X_{ip}, z, \hat{\beta}_n) - \left( I(Z_{ip} = 0) - \frac{1}{P} \right) h_{pt}(X_{ip}, 0, \hat{\beta}_n) \right]. \end{aligned} \quad (2)$$

This looks like an equation that gives us the influence function, but note that  $\hat{\beta}_n$  is estimated on all data, so the terms above are strictly speaking not independent. To proceed, we need to assume something about what happens to  $\hat{\beta}_n$  as  $n$  gets bigger. Assume that there exists  $\beta^*$  such that  $\sqrt{n}(\hat{\beta}_n - \beta^*)$  are bounded in probability and that  $h$  as a function of  $\beta$  is differentiable in a neighbourhood of  $\beta^*$ . Then it is possible to Taylor expand  $h$  in (2) to get

$$\begin{aligned} \sqrt{n}(\hat{\mu}_{3t}(z) - \mu_t(z)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n -\mu_t(z) + \sum_{p=1}^P I(Z_p = z)Y_{pt} - I(Z_p = 0)Y_{pt} \\ &\quad - \left[ \left( I(Z_{ip} = z) - \frac{1}{P} \right) \left( h_{pt}(X_{ip}, z, \beta^*) + 0.5 \frac{\partial h_{pt}(X_{ip}, z, \beta^*)}{\partial \beta} (\hat{\beta}_n - \beta^*) \right) \right. \\ &\quad \left. - \left( I(Z_{ip} = 0) - \frac{1}{P} \right) \left( h_{pt}(X_{ip}, 0, \beta^*) + 0.5 \frac{\partial h_{pt}(X_{ip}, 0, \beta^*)}{\partial \beta} (\hat{\beta}_n - \beta^*) \right) \right] + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n -\mu_t(z) + \sum_{p=1}^P I(Z_p = z)Y_{pt} - I(Z_p = 0)Y_{pt} \\ &\quad - \left[ \left( I(Z_{ip} = z) - \frac{1}{P} \right) h_{pt}(X_{ip}, z, \beta^*) - \left( I(Z_{ip} = 0) - \frac{1}{P} \right) h_{pt}(X_{ip}, 0, \beta^*) \right] + o_p(1), \end{aligned}$$

which gives us the influence function for  $\hat{\mu}_{3t}(z)$ . It looks a lot like what we would expect from (2) with the detail that we have  $h_{pt}(X_{ip}, z, \beta^*)$  and  $h_{pt}(X_{ip}, 0, \beta^*)$  instead of  $h_{pt}(X_{ip}, z, \hat{\beta}_n)$  and  $h_{pt}(X_{ip}, 0, \hat{\beta}_n)$ . The influence function for  $\hat{\mu}_{3t}(z)$  has the added advantage that it is easier to implement a variance estimator using that influence function than the influence function for  $\hat{\mu}_{2t}(z)$ . Therefore, it is preferable to use the influence function for  $\hat{\mu}_{3t}(z)$  to estimate the variance when the two estimators coincide.

## 2 Web Appendix B: Proof of Theorem 1

We base the proof on rewriting the estimator  $\hat{\mu}_{3t}(z)$  as

$$\begin{aligned} \hat{\mu}_{3t}(z) &= \frac{1}{n} \sum_{i=1}^n \sum_{p=1}^P I(Z_{ip} = z) Y_{ipt} - I(Z_{ip} = 0) Y_{ipt} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \sum_{p=1}^P \left[ \left( I(Z_{ip} = z) - \frac{1}{P} \right) h_{pt}(X_{ip}, z, \hat{\beta}) - \left( I(Z_{ip} = 0) - \frac{1}{P} \right) h_{pt}(X_{ip}, 0, \hat{\beta}) \right] \\ &= \frac{1}{n \cdot P} \sum_{i=1}^n \sum_{p=1}^P [h_{pt}(X_{ip}, z, \hat{\beta}) - h_{pt}(X_{ip}, 0, \hat{\beta})] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \sum_{p=1}^P I(Z_{ip} = z) (Y_{ipt} - h_{pt}(X_{ip}, z, \hat{\beta})) - I(Z_{ip} = 0) (Y_{ipt} - h_{pt}(X_{ip}, 0, \hat{\beta})). \end{aligned}$$

The last rewriting shows that  $\hat{\mu}_{3t}(z)$  is equal to  $\hat{\mu}_{2t}(z)$  plus an adjustment term, i.e.,  $\hat{\mu}_{2t}(z) = \hat{\mu}_{3t}(z)$  if the last term is zero.

We assume the  $\beta$ -parameters are estimated from the following weighted least squares estimating equation.

$$\sum_{i=1}^n D_i V^{-1} (Y_i - h(X_i, Z_i, \beta)) = 0, \tag{3}$$

where  $V$  is on the form:

$$\begin{pmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \dots & A \end{pmatrix}, \tag{4}$$

where  $A$  and  $B$  are  $T \times T$  matrices. The inverse of these matrices is also on the form

$$\begin{pmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \dots & A \end{pmatrix}^{-1} = \begin{pmatrix} C & D & \dots & D \\ D & C & \dots & D \\ \vdots & \vdots & \ddots & \vdots \\ D & D & \dots & C \end{pmatrix}, \tag{5}$$

where  $C$  and  $D$  are  $T \times T$  matrices. This can be realized by using the Woodbury identity (Theorem 18.2.8 in Harville (2008)). The Woodbury identity states that assuming  $X$  is nonsingular, then  $X + Y$  is nonsingular if and only if  $I + X^{-1}Y$  is nonsingular, and in that case:

$$(X + Y)^{-1} = X^{-1} - X^{-1}Y(I + X^{-1}Y)^{-1}X^{-1}.$$

For our purposes we will choose:

$$X = \begin{pmatrix} A - B & 0 & \cdots & 0 \\ 0 & A - B & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A - B \end{pmatrix},$$

and

$$Y = \begin{pmatrix} B & B & \cdots & B \\ B & B & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & B \end{pmatrix},$$

such that  $V = X + Y$ . To realize that (5) follows from the Woodbury identity, one needs to realize that  $X^{-1}$  is block diagonal with  $(A - B)^{-1}$  in the diagonal. Moreover,

$$X^{-1}Y = \begin{pmatrix} (A - B)^{-1}B & (A - B)^{-1}B & \cdots & (A - B)^{-1}B \\ (A - B)^{-1}B & (A - B)^{-1}B & \cdots & (A - B)^{-1}B \\ \vdots & \vdots & \ddots & \vdots \\ (A - B)^{-1}B & (A - B)^{-1}B & \cdots & (A - B)^{-1}B \end{pmatrix}.$$

The inverse of  $(I + X^{-1}Y)$  is then  $EF$ , where

$$E = \begin{pmatrix} (P - 1)(A - B)^{-1}B + I & -(A - B)^{-1}B & \cdots & -(A - B)^{-1}B \\ -(A - B)^{-1}B & (P - 1)(A - B)^{-1}B + I & \cdots & -(A - B)^{-1}B \\ \vdots & \vdots & \ddots & \vdots \\ -(A - B)^{-1}B & -(A - B)^{-1}B & \cdots & (P - 1)(A - B)^{-1}B + I \end{pmatrix},$$

$$F = \begin{pmatrix} (P(A - B)^{-1}B + I)^{-1} & 0 & \cdots & 0 \\ 0 & (P(A - B)^{-1}B + I)^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (P(A - B)^{-1}B + I)^{-1} \end{pmatrix}.$$

This can be checked by multiplying the matrices and see that they give you the identity (it might be easier if you substitute  $(A - B)^{-1}B$  with  $K$ ). Then  $X^{-1}Y(I + X^{-1}Y)^{-1}X^{-1}$  is a block matrix where all blocks are equal. Then the Woodbury identity gives us a block diagonal matrix minus a block matrix with all blocks equal, which is on the form (5).

For convenience, we will introduce

$$\varepsilon_{ipt} = Y_{ipt} - h_{pt}(X_{ip}, Z_{ip}, \beta).$$

There is an equation in (3) for each column in our design matrix, i.e. one equation for each  $\beta$ -parameter. However, as it turns out, we only need some equations to get  $\hat{\mu}_{2t}(z) = \hat{\mu}_{3t}(z)$ . We re-parameterise our model so that instead of having a  $\beta$ -parameter per combination of treatment and time point, we have main effects from time points, and main effects of non-placebo treatment per time point, i.e. placebo treatment becomes a reference level. First, we need the equations coming from having an effect of time point in our model. If we specifically look at the estimating equation coming from the effect of time point

$t_0 \in \{1, \dots, T\}$ , then it looks like this:

$$\begin{aligned} 0 &= \sum_{i=1}^n \sum_{p=1}^P \sum_{t=1}^T \varepsilon_{ipt} \cdot (c_{tt_0} + (P-1)d_{tt_0}) \\ &= \sum_{t=1}^T (c_{tt_0} + (P-1)d_{tt_0}) \sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt}, \quad t_0 = 1, \dots, T. \end{aligned} \tag{6}$$

The inverse variance matrix leads to all residuals from all time points, periods, and subjects, being included in the estimating equation (6) for time point  $t_0$ . However, (6) is T linear equations, one for each  $t_0$ , with T unknowns, all equalling zero. Hence, we can conclude

$$\sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} = 0, \quad t = 1, \dots, T. \tag{7}$$

The second set of estimating equations we need comes from having main effects of treatment per time point in our model. The estimating equation corresponding to the effect of treatment  $z$  at time point  $t_0$  is:

$$\begin{aligned} 0 &= \sum_{i=1}^n \sum_{p=1}^P \sum_{t=1}^T \varepsilon_{ipt} \cdot (c_{tt_0} \cdot I(Z_{ip} = z) + d_{tt_0} \sum_{q \neq p} I(Z_{iq} = z)) \\ &\stackrel{(*)}{=} \sum_{i=1}^n \sum_{p=1}^P \sum_{t=1}^T \varepsilon_{ipt} \cdot I(Z_{ip} = z) \cdot (c_{tt_0} - d_{tt_0}) + \varepsilon_{ipt} \cdot d_{tt_0} \\ &= \sum_{t=1}^T (c_{tt_0} - d_{tt_0}) \sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} \cdot I(Z_{ip} = z) + \sum_{t=1}^T d_{tt_0} \sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} \\ &\stackrel{(**)}{=} \sum_{t=1}^T (c_{tt_0} - d_{tt_0}) \sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} \cdot I(Z_{ip} = z), \quad t_0 = 1, \dots, T. \end{aligned}$$

(\*) comes from the fact that all subjects receive each treatment exactly once, so that  $\sum_{q \neq p} I(Z_{iq} = z) = 1 - I(Z_{ip} = z)$ . (\*\*) comes from (7). The rest is just interchanging the order of summation. Again, we have T equations with T unknowns and can conclude

$$\sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} \cdot I(Z_{ip} = z) = 0, \quad t = 1, \dots, T, z = 1, \dots, P-1, \tag{8}$$

as wanted. Note that the above argument only works for all other treatments than the placebo, since placebo is the reference treatment. However, we can rewrite (7) to

$$\begin{aligned} 0 &= \sum_{i=1}^n \sum_{p=1}^P \varepsilon_{ipt} \\ &= \sum_{i=1}^n \sum_{p=1}^P \sum_{z=0}^{P-1} I(Z_{ip} = z) \varepsilon_{ipt} \\ &= \sum_{i=1}^n \sum_{p=1}^P I(Z_{ip} = 0) \varepsilon_{ipt}, \end{aligned}$$

where the last equality comes from (8). Thus, the extra term in  $\hat{\mu}_{3t}(z)$  is zero, and  $\hat{\mu}_{3t}(z) = \hat{\mu}_{2t}(z)$ . Adding other covariates or terms to the regression will not change this fact as long as we have main effects of treatment specific to each post baseline time point.

### 3 Web Appendix C: Influence Function for $\hat{\mu}_{2t}(z)$

Assume the estimation in the first stage solves the following equation:

$$\sum_{i=1}^n m(D_i, \beta) = 0, \quad (9)$$

where  $D_i$  is all the information we have about subject  $i$ . This would for example be the case if we used MLE in which case  $m$  would be the score function. Denote the solution to (9) by  $\hat{\beta}_n$ , and the limit in probability of  $\hat{\beta}_n$  by  $\beta_0$ . Then the influence function of  $\hat{\mu}_{2t}(z)$  is found using Theorem 6.1 from Newey and McFadden (1994):

$$\varphi_t(D) = \frac{1}{P} \sum_{p=1}^P [h(X_p, z, \beta_0) - h(X_p, 0, \beta_0) - \mu_t(z)] + G_\beta \psi(D), \quad (10)$$

where

$$G_\beta = E \left( \nabla_\beta \frac{1}{P} \sum_{p=1}^P [h(X_p, z, \beta_0) - h(X_{ip}, 0, \beta_0)] \right),$$

$$\psi(D) = -(E(\nabla_\beta m(D, \beta_0)))^{-1} m(D, \beta_0).$$

The first term in (10) represents the uncertainty coming from the covariate distribution, and is zero if the model simply is a main term linear mixed model, where  $\hat{\mu}_{2t}(z)$  is a parameter in the model, and thereby independent of covariates. The second term represents the uncertainty arising from the estimation of the  $\beta$ -parameters in the first stage. The variance of the estimator is then

$$\frac{E(\varphi_t(D)^2)}{n}.$$

### 4 Web Appendix D: Extra Tables

In the following tables, columns correspond to the models also used for the data example in the same order. Names are removed in the interest of space.

Tables 1 and 2 show the estimated effects and standard errors for the data example. These are hard to judge up front, but should all be unbiased, although the standard error estimates probably are slightly biased due to the small sample size. Table 3 shows that all the models are unbiased for all the causal effects considered, as they indeed should be according to Corollary 1. Tables 4 and 5 show that the standard error estimates from the influence function in general do well, although a bit biased towards zero due to the small sample size. Table 6 shows that the different models have good coverage probabilities for the true parameter, as they indeed should according to Corollary 1.

#### Conflict of Interest

*The authors have declared no conflict of interest.*

### References

- Harville, D. A. (2008). *Matrix algebra from a statistician's perspective*. Springer.
- Newey, W. K. and McFadden, D. (1994). Chapter 36 Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

**Table 1** Estimates of the causal effect of the different treatments at the different time points for the models in the data example.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13
C	0.5	3.76	3.71	3.75	4.04	3.97	4.19	4.01	3.96	4.16	3.94	3.89	4.09	2.66
	1.0	7.95	7.69	7.83	7.93	7.60	7.93	7.96	7.58	7.94	7.96	7.58	7.94	6.44
	1.5	5.62	5.59	5.59	5.68	5.71	5.90	5.67	5.69	5.88	5.67	5.69	5.88	4.38
	2.5	3.99	3.47	3.48	4.08	3.58	3.77	4.17	3.65	3.83	4.16	3.65	3.83	2.33
D	4.0	4.32	4.70	4.72	4.53	4.86	5.05	4.57	4.86	5.03	4.55	4.86	5.03	3.53
	0.5	5.11	6.04	6.03	5.05	6.05	6.05	5.31	6.22	6.23	4.93	5.89	5.83	5.95
	1.0	9.80	9.96	10.09	9.90	9.94	10.09	9.95	9.95	10.11	9.96	9.94	10.11	9.83
	1.5	7.39	7.20	7.20	7.28	7.20	7.21	7.50	7.36	7.38	7.50	7.35	7.37	7.09
E	2.5	6.05	5.74	5.76	6.06	5.74	5.77	6.67	6.38	6.45	6.63	6.37	6.44	6.16
	4.0	5.79	5.53	5.66	5.69	5.51	5.68	5.98	5.76	5.96	5.89	5.75	5.95	5.68
	0.5	0.88	1.41	1.41	0.78	1.41	1.41	1.01	1.61	1.62	0.75	1.37	1.42	1.38
	1.0	7.16	7.19	7.34	7.24	7.15	7.34	7.19	7.09	7.31	7.19	7.08	7.30	7.06
	1.5	6.03	6.03	6.08	5.92	6.02	6.08	6.13	6.15	6.24	6.13	6.15	6.24	5.99
	2.5	6.87	6.50	6.56	6.87	6.49	6.56	7.43	7.09	7.20	7.41	7.09	7.19	6.95
	4.0	8.32	8.11	8.19	8.22	8.09	8.19	8.49	8.30	8.43	8.43	8.29	8.42	8.18

**Table 2** Estimated standard errors of the effect estimates of the different treatments at the different time points in the data example for the different models.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13
C	0.5	1.21	1.26	1.30	1.22	1.28	1.35	1.23	1.28	1.31	1.24	1.28	1.29	1.81
	1.0	1.08	1.21	1.25	1.08	1.22	1.24	1.11	1.26	1.25	1.22	1.26	1.24	2.02
	1.5	1.13	1.08	1.10	1.12	1.07	1.08	1.15	1.10	1.09	1.01	1.07	1.07	1.76
	2.5	1.08	1.12	1.16	1.07	1.09	1.13	1.10	1.13	1.17	1.13	1.14	1.15	1.83
D	4.0	1.35	1.36	1.40	1.34	1.35	1.39	1.39	1.37	1.40	1.34	1.39	1.38	2.03
	0.5	1.01	1.00	1.05	1.09	1.12	1.27	1.07	1.15	1.26	1.15	1.15	1.25	1.48
	1.0	1.21	1.23	1.25	1.23	1.23	1.26	1.14	1.12	1.14	1.06	1.10	1.13	1.90
	1.5	1.13	1.14	1.18	1.14	1.15	1.24	1.10	1.11	1.17	1.10	1.12	1.16	1.88
E	2.5	1.34	1.37	1.42	1.36	1.40	1.50	1.33	1.37	1.47	1.37	1.40	1.45	1.89
	4.0	1.31	1.40	1.45	1.30	1.40	1.46	1.27	1.37	1.40	1.34	1.37	1.38	2.10
	0.5	1.41	1.38	1.42	1.39	1.38	1.42	1.30	1.29	1.31	1.27	1.28	1.29	1.79
	1.0	1.20	1.22	1.24	1.21	1.22	1.24	1.14	1.14	1.14	1.07	1.11	1.12	1.68
	1.5	1.31	1.35	1.40	1.30	1.33	1.39	1.28	1.31	1.33	1.30	1.33	1.31	1.97
	2.5	1.21	1.26	1.31	1.18	1.24	1.28	1.32	1.40	1.41	1.38	1.40	1.39	2.04
	4.0	1.53	1.49	1.54	1.52	1.48	1.52	1.43	1.44	1.44	1.40	1.43	1.42	2.05

**Table 3** Bias of the estimates in the simulations.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13	
C	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	-0.01	-0.02	0.01	
	1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	
	1.5	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.00
	2.5	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
D	4.0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
	0.5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	-0.01	-0.01	-0.02	0.02	
	1.0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
	1.5	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
E	2.5	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	
	4.0	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	-0.00	
	0.5	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.03	
	1.0	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
	1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.01	
	2.5	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	
	4.0	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	

**Table 4** Standard deviation of estimates in simulations.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13
C	0.5	1.45	1.45	1.45	1.47	1.47	1.51	1.45	1.45	1.48	1.44	1.44	1.47	1.92
	1.0	1.22	1.22	1.22	1.22	1.22	1.22	1.21	1.21	1.22	1.21	1.21	1.22	1.97
	1.5	1.25	1.25	1.25	1.26	1.26	1.29	1.25	1.24	1.27	1.25	1.24	1.27	1.89
	2.5	1.42	1.42	1.42	1.42	1.42	1.44	1.43	1.43	1.46	1.43	1.43	1.46	1.95
	4.0	1.47	1.47	1.47	1.48	1.48	1.51	1.47	1.47	1.50	1.47	1.47	1.49	1.95
D	0.5	1.45	1.45	1.45	1.48	1.48	1.51	1.45	1.45	1.48	1.45	1.45	1.48	1.92
	1.0	1.23	1.23	1.23	1.22	1.22	1.23	1.21	1.21	1.22	1.21	1.21	1.22	1.96
	1.5	1.25	1.25	1.25	1.26	1.26	1.29	1.25	1.25	1.27	1.25	1.25	1.27	1.90
	2.5	1.41	1.41	1.41	1.41	1.41	1.44	1.44	1.44	1.47	1.44	1.44	1.47	1.94
	4.0	1.47	1.47	1.47	1.48	1.48	1.51	1.47	1.47	1.50	1.47	1.47	1.50	1.94
E	0.5	1.44	1.44	1.44	1.47	1.47	1.51	1.45	1.45	1.48	1.44	1.44	1.47	1.92
	1.0	1.24	1.24	1.24	1.23	1.24	1.24	1.23	1.23	1.23	1.23	1.23	1.23	1.96
	1.5	1.25	1.25	1.25	1.27	1.26	1.29	1.25	1.25	1.27	1.25	1.25	1.27	1.89
	2.5	1.41	1.41	1.41	1.42	1.42	1.44	1.44	1.44	1.46	1.43	1.43	1.46	1.93
	4.0	1.48	1.48	1.48	1.49	1.49	1.52	1.48	1.48	1.51	1.48	1.47	1.50	1.94

**Table 5** Average standard error estimates in simulations.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13
C	0.5	1.42	1.42	1.47	1.45	1.44	1.52	1.44	1.44	1.49	1.37	1.37	1.47	1.95
	1.0	1.19	1.19	1.23	1.20	1.20	1.23	1.20	1.21	1.23	1.11	1.16	1.21	1.96
	1.5	1.21	1.21	1.25	1.23	1.22	1.29	1.23	1.23	1.27	1.14	1.18	1.25	1.90
	2.5	1.37	1.37	1.41	1.38	1.37	1.44	1.41	1.41	1.46	1.33	1.35	1.44	1.94
	4.0	1.43	1.43	1.48	1.45	1.45	1.52	1.46	1.46	1.51	1.38	1.38	1.49	1.94
D	0.5	1.42	1.42	1.47	1.45	1.45	1.52	1.44	1.44	1.49	1.37	1.37	1.47	1.95
	1.0	1.19	1.19	1.23	1.19	1.19	1.23	1.20	1.20	1.23	1.10	1.15	1.21	1.96
	1.5	1.22	1.22	1.25	1.23	1.23	1.29	1.24	1.24	1.28	1.15	1.19	1.26	1.90
	2.5	1.37	1.37	1.41	1.38	1.38	1.44	1.43	1.43	1.48	1.34	1.36	1.45	1.95
	4.0	1.43	1.43	1.48	1.45	1.45	1.52	1.46	1.46	1.51	1.38	1.38	1.49	1.94
E	0.5	1.42	1.42	1.47	1.45	1.45	1.53	1.45	1.45	1.50	1.37	1.37	1.47	1.95
	1.0	1.19	1.19	1.23	1.19	1.19	1.23	1.20	1.20	1.23	1.10	1.15	1.21	1.96
	1.5	1.21	1.21	1.25	1.23	1.23	1.29	1.23	1.23	1.27	1.14	1.18	1.25	1.89
	2.5	1.36	1.36	1.41	1.37	1.37	1.44	1.41	1.41	1.46	1.33	1.35	1.44	1.93
	4.0	1.43	1.43	1.48	1.45	1.45	1.52	1.46	1.46	1.51	1.38	1.38	1.49	1.94

**Table 6** Coverage frequencies of confidence intervals in the simulations.

Treatment	Time	1	2	3	4	5	6	7	8	9	10	11	12	13
C	0.5	0.946	0.946	0.954	0.947	0.946	0.953	0.951	0.951	0.953	0.938	0.938	0.950	0.952
	1.0	0.948	0.948	0.954	0.946	0.947	0.953	0.951	0.951	0.953	0.937	0.938	0.951	0.955
	1.5	0.946	0.946	0.953	0.946	0.947	0.953	0.951	0.951	0.953	0.940	0.939	0.951	0.957
	2.5	0.948	0.948	0.954	0.949	0.948	0.955	0.951	0.951	0.953	0.927	0.940	0.950	0.949
	4.0	0.946	0.946	0.952	0.947	0.947	0.952	0.952	0.952	0.952	0.924	0.939	0.950	0.952
D	0.5	0.940	0.940	0.948	0.941	0.942	0.948	0.947	0.947	0.950	0.923	0.936	0.948	0.950
	1.0	0.945	0.945	0.952	0.945	0.945	0.953	0.951	0.951	0.954	0.930	0.940	0.950	0.951
	1.5	0.944	0.944	0.951	0.946	0.945	0.952	0.949	0.949	0.954	0.930	0.940	0.950	0.950
	2.5	0.947	0.947	0.952	0.946	0.946	0.953	0.949	0.949	0.952	0.928	0.938	0.948	0.951
	4.0	0.945	0.945	0.952	0.945	0.945	0.952	0.948	0.948	0.948	0.932	0.935	0.947	0.949
E	0.5	0.944	0.944	0.950	0.944	0.943	0.950	0.948	0.948	0.950	0.932	0.937	0.947	0.954
	1.0	0.941	0.941	0.948	0.943	0.942	0.949	0.946	0.947	0.951	0.931	0.935	0.948	0.952
	1.5	0.946	0.946	0.954	0.946	0.947	0.955	0.951	0.951	0.950	0.937	0.938	0.953	0.949
	2.5	0.944	0.944	0.951	0.947	0.947	0.955	0.953	0.953	0.953	0.935	0.937	0.952	0.954
	4.0	0.947	0.947	0.954	0.945	0.945	0.952	0.951	0.951	0.951	0.939	0.939	0.951	0.952



# Manuscript III

## Estimating causal effects while adjusting for unmeasured time-stable confounding

Jeppe Ekstrand Halkjær Madsen, Thomas Delvin, Thomas Scheike & Christian Pipper

**Details:** Submitted to *Statistical Methods in Medical Research* in 2023.

# Estimating causal effects while adjusting for unmeasured time-stable confounding

Statistical Methods in Medical Research

XX(X):2-17

©The Author(s) 2016

Reprints and permission:

[sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)

DOI: 10.1177/ToBeAssigned

[www.sagepub.com/](http://www.sagepub.com/)

SAGE

Jeppe Ekstrand Halkjær Madsen<sup>1,2</sup>, Thomas Delvin<sup>3</sup>, Thomas Scheike<sup>1</sup>, and Christian Pipper<sup>4,5</sup>

## Abstract

We propose a novel method to adjust for unmeasured time-stable confounding, when the time between consecutive treatment administrations is fixed. In this setup, we may eliminate all unmeasured time-stable confounding by adjusting for the potential time on treatment or equivalently the potentially unrealized number of treatment administrations. A challenge with this approach is that right censoring of the potential time on treatment occurs when treatment is terminated at the time of the event of interest, for example if the event of interest is death. We show how this challenge may be solved by means of the EM algorithm. The usefulness of the methodology is illustrated in a simulation study. We also apply the methodology to investigate the effect of depression/anxiety drugs on subsequent poisoning by other medications in the Danish population by means of national registries. Here, we find a protective effect of treatment with selective serotonin reuptake inhibitors on the risk of poisoning by various medications (one year risk difference of approximately  $-3\%$ ). A standard Cox model analysis shows a harming effect (one year risk difference of approximately  $2\%$ ). Unmeasured time-stable confounding can be entirely adjusted for when time between consecutive treatment administrations is fixed.

---

## Keywords

causal inference, EM algorithm, unmeasured confounding, confounding by indication, DAG

## Introduction

Confounding is a problem in observational studies.<sup>1</sup> Confounding means that there are common causes of exposure and outcome, thereby making treatment groups different in terms of prognostic factors.<sup>2</sup> In an ideal world, the subjects would be *exchangeable*. Exchangeability means that the outcomes among the untreated, had they instead been treated, would be similar to the outcomes of those who actually received treatment and vice versa. Let  $Y^z$  denote the outcome a subject would have got under treatment  $z$ , and let  $Z$  denote actual treatment. Mathematically, the exchangeability assumption states that

$$Y^z \perp\!\!\!\perp Z, \quad \forall z.$$

Confounding prevents this property from being true and thereby challenges any comparison between treatment groups. There are several standard ways to handle confounding. The basic idea behind these methods is to replace the exchangeability assumption with a conditional exchangeability assumption, that is, to assume subjects exchangeable within strata such as sex, age, and other measured potential confounders:

$$Y^z \perp\!\!\!\perp Z \mid X, \quad \forall z,$$

where  $X$  is a set of potential confounders that are sufficient for ensuring conditional exchangeability. Under this assumption, it is possible to proceed with the actual data analysis in several ways, the most well-known being direct adjustment in a regression model. Another popular solution is to use an estimated probability of treatment, also known as a propensity score. Propensity scores are useful because conditional exchangeability given  $X$  implies conditional exchangeability given the propensity score,

---

<sup>5</sup>Epidemiology, Biostatistics & Biodemography, Department of Public Health, University of Southern Denmark, Denmark.

### Corresponding author:

Jeppe Ekstrand Halkjær Madsen, Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen K, Denmark.

Email: jehm@sund.ku.dk

$P(Z | X)$ . Moreover, the propensity score is the simplest transformation of  $X$  that implies conditional exchangeability in the sense that for any other transformation of  $X$ ,  $b(X)$ , that implies conditional exchangeability we have

$$P(Z | X) = f(b(X))$$

for some function  $f$ .<sup>3</sup> Propensity scores are used in several ways, such as matched sampling, direct adjustment for the propensity score, or inverse probability of treatment weighting.<sup>3</sup> However, all of these methods rely on the existence of an observed set of confounders that are sufficient for ensuring conditional exchangeability. Under further assumptions, it is sometimes possible to make sensitivity analyses for the effect of unmeasured confounding.<sup>4,5</sup> Although this to some extent alleviates the problem with unmeasured confounding, it relies on extra assumptions that can easily be unrealistic. Furthermore, it might tell us that we are very sensitive to unmeasured confounding, in which case it is hard to understand what the true effect of treatment is. Alternatively, a self-controlled design can be applied. The basic idea in these designs is to compare subjects to themselves. Subjects are hypothesized to be very similar to themselves, and it is therefore argued that exchangeability is often a more reasonable assumption in these designs. The comparison is possible because subjects are observed over time, both with and without treatment. Thereby, it is possible to estimate whether the risk of event is higher at times of treatment than at times without treatment. As elegantly formulated, it changes the research question from "why me" to "why now".<sup>6</sup> Self-controlled designs have the advantage that they automatically adjust for time-stable confounding, even if the confounders are unobserved.<sup>7</sup> However, they do not adjust for time-dependent confounding, since that would make the compared time points non-exchangeable.<sup>8</sup> Another challenge with these designs is that it is not always clear what is estimated. Finally, each design has its own specific assumptions and challenges, such as sensitivity to time-trends in exposure in the case-crossover design,<sup>9</sup> or inability to handle terminal events in the case of the self-controlled case series analysis.<sup>10</sup> Yet another approach is based on so-called instrumental variables. Here, the idea is to find some unconfounded cause of treatment that has no direct effect on the outcome of interest (see Figure 1). Such a variable is called an instrument or instrumental variable.

Note that confounding *is* allowed for the instrument, as long as it is measured, which is reflected by the variable  $X$  in Figure 1. Then the association between the instrument and outcome is used as a proxy for the effect of treatment on outcome. This association

is unconfounded, but will be biased towards the null since the instrument doesn't predict the actual treatment exactly. The size of the bias depends on how strong an effect the instrument has on treatment, also known as the strength of the instrument. Fortunately, it is possible to estimate the bias and get an unbiased and unconfounded estimate of the treatment effect among compliers under further assumptions through the Wald estimator<sup>11</sup>

$$\frac{E(Y | I = 1) - E(Y | I = 0)}{P(Z = 1 | I = 1) - P(Z = 1 | I = 0)}.$$

A nice example is from randomized trials, where actual treatment received may be confounded due to some subjects not complying to the randomized treatment. Here the instrument would be the randomized treatment and an instrumental variable analysis where the randomized treatment is used in the statistical analysis instead of the actually received treatment corresponds to what is known as an intention-to-treat analysis. The main disadvantage with instrumental variables is that it can be very difficult, if not impossible, to find a good instrument.<sup>11,12</sup>

Unmeasured confounding refers to the situation where we don't have a set of observed confounders that are sufficient for ensuring exchangeability. In practice, we never know for sure whether we have such a set of confounders, although sometimes it can be argued based on subject matter knowledge. To make matters worse, unmeasured confounding is an issue without one generally accepted solution.<sup>13</sup> As an example, consider confounding by indication. In this instance, the reason for treatment, also known as the indication for treatment, is a confounder. For example, depression might be a confounder when studying the relationship between antidepressant drugs and suicide. A special case is when the severity of the indication is a confounder, and not just the indication itself.<sup>14</sup> These types of confounding are hard to adjust for since we usually don't have good, if any, measurements of disease severity in most available data sources.

We consider a method where we adjust for time-stable unmeasured confounding in a flexible way that doesn't necessitate any specific model, such as the Cox model, but is also applicable for accelerated failure type models, for example. We consider a setup where treatment is given under strict supervision, such that the time between consecutive treatment administrations is constant. We do this by restricting attention to a new-user cohort.<sup>15</sup> Under these circumstances, time-stable confounders can only affect treatment by affecting treatment duration through the number of treatment administrations, since this becomes the only determinant of exposure status at different time points in this setup. Thereby, adjusting for the number of treatment administrations effectively adjusts for all

time-stable confounding. We proceed to show that it is possible to fully adjust for the confounding, without further assumptions on the number of treatment administrations, even if the number of treatment administrations is censored at the event time, such as if the event is terminal. That is, by adjusting for an unrealized number of treatment administrations. Furthermore, we show how this enables estimation of a causal effect for time-to-event data despite the unmeasured time-stable confounding. We illustrate the usefulness of the method in a simulation study. Finally, we use the methodology to analyse the effect of treatment with selective serotonin reuptake inhibitors on the risk of poisoning from various medications.

## Notation and causal assumptions

In this section, we point out why adjusting for the number of treatment administrations effectively adjusts for all unmeasured time-stable confounding, and how this enables unconfounded estimation of causal effects.

The subjects are sampled at the time of first treatment, such that we have a new-user cohort.<sup>15</sup> The time of first treatment then becomes the start of follow-up. Treatment is assumed to be given under strict supervision, such that the times between consecutive treatment administrations are constant. In this case the total treatment duration,  $D$ , is given by

$$D = \Delta \cdot W,$$

where  $\Delta$  is the (constant) time between consecutive treatment administrations and  $W$  is the (random) number of treatment administrations. This corresponds to a scenario where the doctor based on unmeasured confounders such as disease severity determines how many treatment administrations the patient will need. Subjects are considered exposed as long as they are receiving treatment. Hence, treatment status at time  $t$ ,  $Z(t)$ , is given by

$$Z(t) = I(t \leq D),$$

where  $I(t \leq D)$  is one until the end of treatment. By definition, a confounder affects both treatment and outcome. Since all subjects are treated at the start of follow-up, and since the time between consecutive treatment administrations is constant, the only way a confounder can affect treatment is through the number of treatment administrations,  $W$ . Denote measured confounders by  $X$  and unmeasured confounders by  $U$ . Then we can describe our data with the Directed Acyclic Graph (DAG) in Figure 2.

We can define counterfactual variables,  $T^{\{z(t)\}}$  which are the counterfactual event times given treatment history  $\{z(t)\}$ . The DAG implies the following conditional exchangeability relation

$$T^{\{z(t)\}} \perp\!\!\!\perp \{Z(t)\} \mid W.$$

The conditional exchangeability follows from the fact that nothing can be a time-stable confounder without its effect on treatment going through the number of treatment administrations in our setup. Therefore, adjusting for the number of treatment administrations also adjusts for all time-stable confounders. In essence, we don't care about the number of treatment administrations per se, but they represent all time-stable confounders at baseline, including unmeasured confounders. Thus, it is possible to fit a model to data without confounding being a problem by adjusting for the number of treatment administrations. In that case we might still want to include the observed confounders,  $X$ , in the model to gain efficiency.

Often, a hazard ratio (HR) is reported as the estimate of the effect of being exposed vs. unexposed.<sup>16</sup> It is possible to get an unconfounded estimate of the HR by including the number of treatment administrations as a covariate due to the conditional exchangeability, at least assuming the Cox model is correctly specified. However, now that we have a causal framework and even conditional exchangeability, we might want to estimate a causal effect instead to avoid the problems with the HR.<sup>17,18</sup>

The counterfactual outcomes we wish to compare depend on the research question. One might want to quantify the effect of treatment by comparing subjects when treated until the end of some follow-up  $\tau$  to the same subjects if they never received treatment. Unfortunately, this is a positivity violation, since we never observe subjects without treatment from time zero to time  $\Delta$  in a new-user cohort. Thus, we have to consider this when defining our counterfactual outcomes. A simple solution would be to simply compare continuing treatment until the end of follow-up and terminate treatment after the first treatment period. For simplicity, we might term these potential outcomes  $T^1$  and  $T^0$  respectively.

In any case, the causal treatment effect could simply be the average treatment effect (ATE)

$$P(T^1 \leq \tau) - P(T^0 \leq \tau), \quad (1)$$

i.e.,  $\tau$  years risk of event, where  $\tau$  is some time-frame of interest for the event. This quantity equals

$$E(P(T^1 \leq \tau \mid W)) - E(P(T^0 \leq \tau \mid W)), \quad (2)$$

which we can estimate due to the conditional exchangeability, for example by using the g-formula.<sup>19</sup> The formula in (2) may seem paradoxical. The interventional treatment in  $T^0$  for example, is impossible to get while also receiving for example four treatment administrations, i.e. having  $W = 4$ . Therefore, we stress that  $P(T^0 \leq \tau | W = 4)$  is the probability in the population of subjects who actually got four treatment administrations. Thus, positivity doesn't require that they get one treatment administration and four treatment administrations at the same time, which would be impossible. The positivity assumption in this case states that the subjects who got four treatment administrations could have gotten for example one treatment administration with a strictly positive probability. A positivity violation would be if the unmeasured confounders we are trying to adjust for deterministically predict the number of treatment administrations. In order to gain efficiency, we might be interested in using that the ATE also equals

$$E(P(T^1 \leq \tau | W, X)) - E(P(T^0 \leq \tau | W, X)), \quad (3)$$

i.e., we might want to adjust for other covariates,  $X$ . In any case, in order to use the above methodology, we need to fit a model to the data. At first, this may seem impossible since the number of treatment administrations a subject should have received,  $W$ , is unobserved, for example due to a terminal event. Luckily, there is an abundance of literature on how to handle missing covariates in, for example, the Cox model.<sup>20,21</sup> We propose to use the EM algorithm,<sup>22</sup> which makes it possible to handle the missingness without any further assumptions on the marginal distribution of  $W$ . The main challenge in that context is if we want a causal effect from (3) since that requires estimation of the joint distribution of  $X$  and  $W$  which is hard non-parametrically when we have censoring in  $W$ . In the next section, we will show how to apply the EM algorithm in our setup in order to model the data in a way that allows estimation of a causal effect.

## EM algorithm

In this section, we describe the EM algorithm in our setup. The main advantage with this approach is that we don't have to assume anything about the marginal distribution of the number of treatment administrations. Let  $T^*$  be the event time and  $C$  be the censoring time. We observe  $T = \min\{T^*, C\}$  and the status indicator  $\Delta = I(T^* \leq C)$ , which tells us whether the event time was observed or censored. Let  $W^*$  be the true number of treatment administrations and  $W$  be the observed number of treatment administrations. Let  $\Psi = I(W = W^*)$  be an indicator telling us whether  $W^*$  is observed or censored.

Denote covariates by  $X$ . In the following, we will use lowercase letters to denote realizations of the variables above. We assume we observe  $n$  iid. copies of the data described above. Assume we can write the probability of data from subject  $i$  as

$$f(t_i | w_i^*, x_i, \theta),$$

where  $\theta$  is a vector of parameters. Normally,  $\theta$  can be estimated by MLE, i.e., by maximizing the following log-likelihood for right-censored data:

$$\ell(\theta) = \sum_{i=1}^n \log (f(t_i | w_i^*, x_i, \theta)^{\delta_i} \cdot S(t_i | w_i^*, x_i, \theta)^{1-\delta_i}),$$

where  $S(t_i | w_i^*, x_i, \theta) = P(T^* > t | w_i^*, x_i, \theta)$ . In our case, this is not possible up front due to the fact that  $w_i^*$  is censored for some subjects.

To ensure identification, we assume that the value of  $w_i^*$  only matters until a certain value  $M$ , i.e.,

$$f(t_i | w_i^* = w, x_i, \theta) = f(t_i | w_i^* = M, x_i, \theta), \forall w \geq M.$$

In that case, we can redefine  $\Psi$  to be  $I(W^* = W \vee W \geq M)$ . That is, we observe enough about  $W$  for the estimation if either the end of treatment is observed or if we observe  $W^* \geq M$  (see Figure 3).

Denote the probability of having  $W = w$  given  $X = x$  by  $p_{wx}$  for  $w = 1, \dots, M - 1$  and the probability of  $W \geq M$  given  $X = x$  by  $p_{Mx}$ . The EM algorithm needs the following weights for the subjects where we don't know the true value of  $W^*$ :

$$\begin{aligned} q_{ij} &= P(w_i^* = j | w_i = k, \psi_i = 0, x_i, t_i, \delta_i) \\ &= \frac{f(t_i | w_i^* = j, x_i, \theta)^{\delta_i} S(t_i | w_i^* = j, x_i, \theta)^{1-\delta_i} \cdot p_{jx_i}}{\sum_{l=k}^M f(t_i | w_i^* = l, x_i, \theta)^{\delta_i} S(t_i | w_i^* = l, x_i, \theta)^{1-\delta_i} \cdot p_{lx_i}} \cdot I(j \geq k), j = 1, \dots, M - 1, \\ q_{iM} &= P(w_i^* \geq M | w_i = k, \psi_i = 0, x_i, t_i, \delta_i) \\ &= \frac{f(t_i | w_i^* = M, x_i, \theta)^{\delta_i} S(t_i | w_i^* = M, x_i, \theta)^{1-\delta_i} \cdot p_{Mx_i}}{\sum_{l=k}^M f(t_i | w_i^* = l, x_i, \theta)^{\delta_i} S(t_i | w_i^* = l, x_i, \theta)^{1-\delta_i} \cdot p_{lx_i}}. \end{aligned} \quad (4)$$

This is only necessary for the subjects where we don't know the true value of  $W^*$ . For observations where  $W^*$  is observed, the probabilities will be either 1 or 0. The weights  $p_{wx}$  can be estimated non-parametrically if  $x$  is discrete or if the model doesn't use  $x$  at

all, which would be completely legitimate since we have exchangeability conditional on  $W$ . Then the EM algorithm is as follows:

1. Initialize  $\theta^{(0)}$ , and  $(p_{jx}^{(0)})_{j=1,\dots,M}$ , and  $r = 0$  is the iteration number.
2. E-step: Calculate

$$Q = E(\ell(\theta))$$

$$= \sum_{i=1}^n \sum_{j=1}^M [\log(f(t_i | w_i = j, x_i, \theta)^{\delta_i} S(t_i | w_i = j, x_i, \theta)^{1-\delta_i}) + \log(f_w(j | (p_{kx_i})))] \cdot q_{ij}^{(r)}$$

where  $f_w$  is the density for  $W$ . Estimates of  $q_{ij}^{(r)}$  are obtained from (4) using  $\theta^{(r)}$  and  $(p_{jx}^{(r)})_{j=1,\dots,M}$ .

3. M-step: Maximize  $Q$ . This estimates  $\theta^{(r+1)}$  and  $(p_{jx}^{(r+1)})$ . Set  $r = r + 1$ .
4. Repeat steps 2. and 3. until convergence.

As previously noted, the  $Q$  function is a weighted version of a complete data likelihood due to the fact that the missingness is in a discrete covariate.<sup>23</sup> Thus, the main coding task is the calculation of the weights needed in the E-step. The rest can be done with standard software implementations. Variance estimates can be obtained from the information matrix.<sup>24</sup> However, this is time-consuming since we have a baseline parameter for each unique time point with an observed failure time. Thereby, the information matrix becomes very big. Alternatively, and as we will do in the simulation study, variance estimates can be obtained through EM-aided differentiation.<sup>25</sup>

All of the above theory works with other covariates,  $X$ . However, the estimator (3) is hard to plug into the g-formula due to the need for the joint distribution of  $X$  and  $W$ . For discrete  $X$  the conditional distribution  $W | X$  is estimated non-parametrically by the EM algorithm, which enables estimation of the joint distribution of  $X$  and  $W$  as  $P(W | X)P(X)$ . Likewise, if we don't include  $X$  in the model, the g-formula is simply

$$\sum_{w=1}^M (P(T^1 \leq \tau | W = w) - P(T^0 \leq \tau | W = w)) \cdot P(W = w).$$

The (asymptotic) variance of the causal effect can be estimated from the influence function.<sup>26</sup> However, this is challenging in the Cox model in this context when combined with the EM algorithm, since the dimension of the variance matrix can be quite high due

to all the baseline hazard parameters. This numerical problem is hard to solve, but valid variance estimates can be obtained by the non-parametric bootstrap.

## Simulation

We illustrate the usefulness of the method in a simulation study by simulating from the following Cox model:

$$\lambda(t) = \lambda_0(t) \cdot e^{\beta \cdot (1-Z(t)) + \gamma \cdot \text{sex}} \cdot 2^{I(W=2)} \cdot 3^{I(W=3)} \cdot 4^{I(W \geq 4)}. \quad (5)$$

Note that we estimate the effect of going from treatment to no treatment to ensure identifiability of the baseline hazard in the first period where all subjects are treated. As can be seen from (5), the effect of the number of treatment administrations is capped at  $M = 4$ , the time between consecutive treatment administrations is  $\Delta = 100$ , the baseline hazard is constantly equal to 0.0001, censoring times are constant and equal to 1000,  $\gamma = 0$ , and  $\beta = -\log(1.5) \approx -0.41$  in the simulations. Sex is estimated from a Bernoulli variable with probability one half, and  $W$  is simulated from Poisson distributions with rate equal to 3.5 for men and 1.5 for women. As discussed previously, we recommend estimating something with a causal interpretation, but for the sake of illustration, we will focus on the estimation of  $\beta$  in the following.

As discussed previously, it might be desired to fit a model without other covariates than treatment and  $W$  in order to make it easier to estimate a causal effect. Therefore, we also fit a Cox model without sex in the simulations. In this case the conditional distribution of the event time given  $W$  is correctly specified because we have assumed  $\gamma = 0$ . Note, however, that the estimation procedure changes since the weights no longer are conditional on sex, but instead depends on the marginal distribution of  $W$ .

We compare the EM algorithm models to three naive analyses one might conduct upon receiving data. The first one fits the model with treatment and sex as covariates without adjusting for the number of treatment administrations,  $W$ , at all, which is what we would often expect researchers to do in practice. The second analysis adjusts for the observed  $W$  and ignores that it is sometimes censored. The third model uses the number of treatment administrations as a time-dependent covariate with value equal to the number of treatment administrations until time  $t$ . This would be a pure modelling attempt at approximating the true model without using the EM algorithm. Up front, the model makes little sense, since it essentially states that the hazard rate jumps every time a new treatment administration takes place. We have run 10000 simulations, each with

2000 subjects, in the statistical program R.<sup>27</sup> The results are summarized in Table 1. Clearly, the unadjusted analysis is biased, as we would expect given the confounding. The adjusted model is even more biased, illustrating the importance of using the EM algorithm to alleviate the bias. The time-dependent adjustment is still biased, but does a decent job at approximating the true model. The EM algorithm estimates entirely without bias, and the standard errors lead to correct coverage for both EM algorithm models, although the one adjusting for sex has slightly higher power.

In conclusion, the proposed EM approach performs well in the considered simulation scenario, and, in particular, avoids the systematic bias that is evident for the other methods considered.

## Data example

Here we will illustrate the method of this paper on a data example and compare to what would be obtained if instead a Cox model was applied. The data example is based on a population identified in the Danish healthcare registries covering the whole population.<sup>28</sup> National data on drug use in Denmark were extracted from the Danish National Prescription Database.<sup>29</sup> The Registry contains complete information, from 1 January 1995 and onwards, on all prescriptions filled by Danish residents at outpatient pharmacies, providing information on drug type, quantity, strength, date of purchase, person age, and sex. Registered drugs are categorized according to the Anatomic Therapeutic Chemical (ATC) index, a hierarchical classification system developed by the World Health Organization (WHO) for purposes of drug use statistics.<sup>30</sup> The quantity dispensed for each prescription is expressed by the defined daily dose (DDD) measure, also developed by the WHO.<sup>30</sup> The registry is reported to have a high completeness and validity.<sup>29</sup> The National Patient Register (for hospital contacts) was used for identification of diagnosis of disease and procedure codes or ATC codes. The National Patient Register contains information on persons who have been admitted to somatic hospital departments since 1977, and from 1995 also outpatient and emergency department patients. Individual information include admission and discharge information, the time of any incidents over the course of an illness, diagnosis, examinations and treatment information etc. The Danish Civil Registration System (CPR) is a national register containing basic personal information on all who have a civil registration number, which is used for linking the above data sources. The Danish Civil Registration System holds a unique identifier for all Danish residents since 1968

encoding sex and date of birth.<sup>31</sup> Data sources were linked by the civil registry number, a unique identifier assigned to all Danish residents since 1968.<sup>31</sup> The study population was defined as all residents in Denmark with a prescription of a selective serotonin reuptake inhibitor (SSRI) since 1995, identified by the Anatomical Therapeutic Chemical (ATC) code of N06AB\*, identified within the Danish National Prescription Database. Exposure was defined as incident use of an SSRI identified in the Danish National Prescription Database by ATC codes N06AB\* since 1995. The outcome event was defined as the first-ever diagnosis of a range of specific conditions identified by ICD-10 codes T39\* to T49\* in the National Patient Register. These ICD-10 codes include poisoning by various medications.<sup>32</sup> Both the ATC (used to categorize drugs) and the ICD-10 (used to categorize diagnoses) are hierarchical classification systems. By considering an increasing number of digits of the code, increasing precision in terms of detail is achieved. For example, N06A indicates the ATC code for all antidepressants, N06AB for all SSRIs and N06AB06 for the specific drug Sertraline.

SSRI medication is used for treating depression and anxiety. This is a situation where we would expect a lot of confounding by indication that is hard to adjust for. Furthermore, these treatments are prescribed with regular intervals of approximately 50 days. Thus, the methodology of this paper is applicable here.

Our sample size is 36122. Before making the actual data analysis, we need to decide what value we want to cap the effect of  $W$  at. We do this by inspecting the distribution of  $W$  in the data in Table 2 along with the censoring pattern of  $W$ . We see a decreasing distribution with most subjects having few treatment administrations. We choose to cap the effect of  $W$  at  $M = 4$ , which should give us a censoring indicator for  $W$  of one for most observations but should still allow us to adjust for some confounding.

We compare a Cox model of the form

$$\lambda(t) = \lambda_0(t) \cdot \exp(\beta_1 \cdot Z(t) + \beta_2 \cdot sex),$$

to an EM algorithm that only includes the number of treatment administrations,  $W$ , as a categorical variable, and treatment, which is written as  $1 - Z(t)$  for identification of the baseline in the first period. Note, that sex could have been included in the EM algorithm as well, but the estimation and the g-formula would have been more complicated. Standard Errors (SE) and p-values are obtained from the non-parametric bootstrap. We compare the following two treatment regimes

- Treated the entire year.

- Treated for fifty days, and then not the rest of the year.

For brevity we will term these "always treated" and "never treated" in the following. We compare these treatment regimes in terms of the one-year risk difference of experiencing an event.

The results are summarized in Table 3. As displayed in Table 3, the Cox model shows an increased risk among the always treated, which is highly significant, as we would expect given the confounding. The EM algorithm shows the exact opposite, namely a protective effect of treatment, which reflects our intuition about the true causal nature of data better. This is also highly significant. Therefore, one should also consider the clinical relevance of going from a 20% risk to 17%. However, a protective (side-)effect is positive in any case.

## Discussion

Unmeasured time-stable confounding can be fully adjusted for in a new-user cohort when the time between treatment administrations is fixed by including the number of treatment administrations as a covariate. Adjusting for the number of treatment administrations ensures conditional exchangeability because the only way a time-stable confounder can affect treatment status is by changing the number of treatment administrations in this particular setup. If the number of treatment administrations is censored at the event time, for example if the event is terminal, the EM algorithm can be employed without further assumptions on the number of treatment administrations. This is not restricted to any specific type of model, such as the Cox model, but works for any model where parameters are estimated with MLE. The choice of causal target parameter has to reflect that we are dealing with a new-user cohort, and the g-formula has to handle the censoring in the number of treatment administrations, for example by excluding other covariates.

One might argue that the regression model in (5) is overly restrictive. Why not, for example, stratify on  $W$ ? It turns out stratifying on  $W$  would make it impossible to identify the effect of interest. This is due to the fact that we can't identify the baseline hazard for  $W > 1$  in the first period, since we never observe  $W$  and have an event in the first period at the same time. Thereby, all the probability mass for the weights among subjects with an event in the first period would be distributed to  $W = 1$ . Similar problems arise in the other periods, particularly if the event time has a continuous distribution, in which case the baseline hazards will almost surely put probability mass at different time points. Thus, we are more or less guaranteed to estimate the distribution of  $W$  incorrectly.

A limitation of the methodology described in this paper is when the confounding is time-dependent. For example, if disease severity changes over time, and this time-dependent severity determines treatment status, then the method will lead to a biased estimate of the treatment effect. This limitation is shared by all self-controlled designs.<sup>7</sup> It can be argued that the methodology in this paper would be conditioning on the future in that case, since the number of treatment administrations would be determined by future confounders.

An important extension of the methodology would be to enable the use of models estimated differently than MLE, such as inverse probability of censoring weighted methods.<sup>33</sup> Thereby, it would be possible to use the methodology with any classification model, such as logistic regression or random forests.

Nevertheless, it is possible to adjust for unmeasured time-stable confounding in a new-user cohort when treatment duration is fixed with standard models.

## Software

R code for simulations, and the simulation results are available at <https://github.com/Jeepen/em>.

## Acknowledgements

The authors want to thank Professor Jesper Hallas, Dr Lars Christian Lund, and Data Manager Martin Thomsen Ernst from University of Southern Denmark without whom the data example in this paper wouldn't exist.

## Declaration of conflicting interests

None declared.

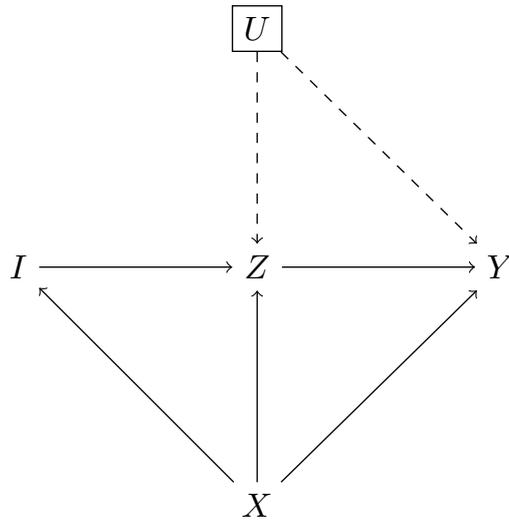
## References

1. Rothman KJ, Greenland S and Lash TL. *Modern epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008. ISBN 978-0-7817-5564-1.
2. Hernán MA RJ. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
3. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1): 41–55. DOI:10.1093/biomet/70.1.41. URL <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41>.

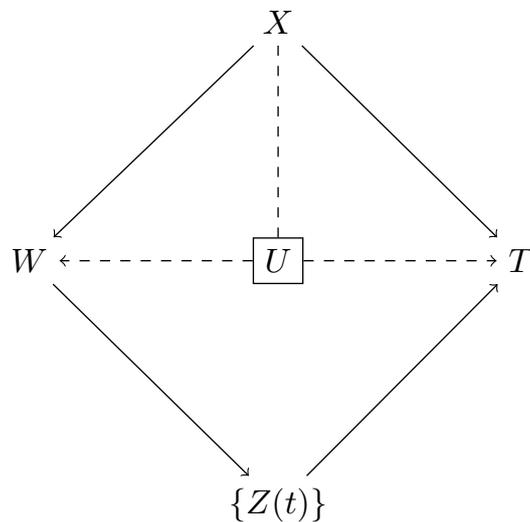
4. Lin DY, Psaty BM and Kronmal RA. Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies. *Biometrics* 1998; 54(3): 948. DOI: 10.2307/2533848. URL <https://www.jstor.org/stable/2533848?origin=crossref>.
5. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety* 2006; 15(5): 291–303. DOI:10.1002/pds.1200. URL <https://onlinelibrary.wiley.com/doi/10.1002/pds.1200>.
6. Maclure M. ‘Why me?’ versus ‘why now?’—differences between operational hypotheses in case-control versus case-crossover studies. *Pharmacoepidemiology and Drug Safety* 2007; 16(8): 850–853. DOI:10.1002/pds.1438. URL <https://onlinelibrary.wiley.com/doi/10.1002/pds.1438>.
7. Hallas J and Pottegård A. Use of self-controlled designs in pharmacoepidemiology. *Journal of Internal Medicine* 2014; 275(6): 581–589. DOI:10.1111/joim.12186. URL <https://onlinelibrary.wiley.com/doi/10.1111/joim.12186>.
8. Mittleman MA and Mostofsky E. Exchangeability in the case-crossover design. *International Journal of Epidemiology* 2014; 43(5): 1645–1655. DOI:10.1093/ije/dyu081. URL <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyu081>.
9. Maclure M. The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology* 1991; 133(2): 144–153. DOI:10.1093/oxfordjournals.aje.a115853. URL <https://academic.oup.com/aje/article/118507/The>.
10. Farrington CP. Relative Incidence Estimation from Case Series for Vaccine Safety Evaluation. *Biometrics* 1995; 51(1): 228–235. DOI:10.2307/2533328. URL <http://www.jstor.org/stable/2533328>.
11. Baiocchi M, Cheng J and Small DS. Instrumental variable methods for causal inference: Instrumental variable methods for causal inference. *Statistics in Medicine* 2014; 33(13): 2297–2340. DOI:10.1002/sim.6128. URL <https://onlinelibrary.wiley.com/doi/10.1002/sim.6128>.
12. Sjolander A and Martinussen T. Instrumental Variable Estimation with the R Package ivtools. *Epidemiologic Methods* 2019; 8(1): 20180024. DOI:10.1515/em-2018-0024. URL <https://www.degruyter.com/document/doi/10.1515/em-2018-0024/html>.
13. Bosco JL, Silliman RA, Thwin SS et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of Clinical Epidemiology* 2010; 63(1): 64–74. DOI:10.1016/j.jclinepi.2009.03.001. URL <https://linkinghub>.

- [elsevier.com/retrieve/pii/S0895435609000602](https://www.elsevier.com/retrieve/pii/S0895435609000602).
14. Salas M, Hotman A and Stricker BH. Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. *American Journal of Epidemiology* 1999; 149(11): 981–983. DOI:10.1093/oxfordjournals.aje.a009758. URL <https://academic.oup.com/aje/article-lookup/doi/10.1093/oxfordjournals.aje.a009758>.
  15. Ray WA. Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. *American Journal of Epidemiology* 2003; 158(9): 915–920. DOI:10.1093/aje/kwg231. URL <https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwg231>.
  16. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1972; 34(2): 187–202. DOI:10.1111/j.2517-6161.1972.tb00899.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1972.tb00899.x>.
  17. Hernán MA. The Hazards of Hazard Ratios. *Epidemiology* 2010; 21(1): 13–15. DOI:10.1097/EDE.0b013e3181c1ea43. URL <https://journals.lww.com/00001648-201001000-00004>.
  18. Martinussen T, Vansteelandt S and Andersen PK. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis* 2020; 26(4): 833–855. DOI:10.1007/s10985-020-09501-5. URL <https://link.springer.com/10.1007/s10985-020-09501-5>.
  19. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; 7(9-12): 1393–1512. DOI:10.1016/0270-0255(86)90088-6. URL <https://linkinghub.elsevier.com/retrieve/pii/0270025586900886>.
  20. Yi Y, Ye T, Yu M et al. Cox regression with survival-time-dependent missing covariate values. *Biometrics* 2020; 76(2): 460–471. DOI:10.1111/biom.13155. URL <https://onlinelibrary.wiley.com/doi/10.1111/biom.13155>.
  21. Rathouz PJ. Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics* 2007; 8(2): 345–356. DOI:10.1093/biostatistics/kxl014. URL <https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxl014>.
  22. Dempster AP, Laird NM and Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 1977; 39(1): 1–22. DOI:10.1111/j.2517-6161.1977.tb01600.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x>.

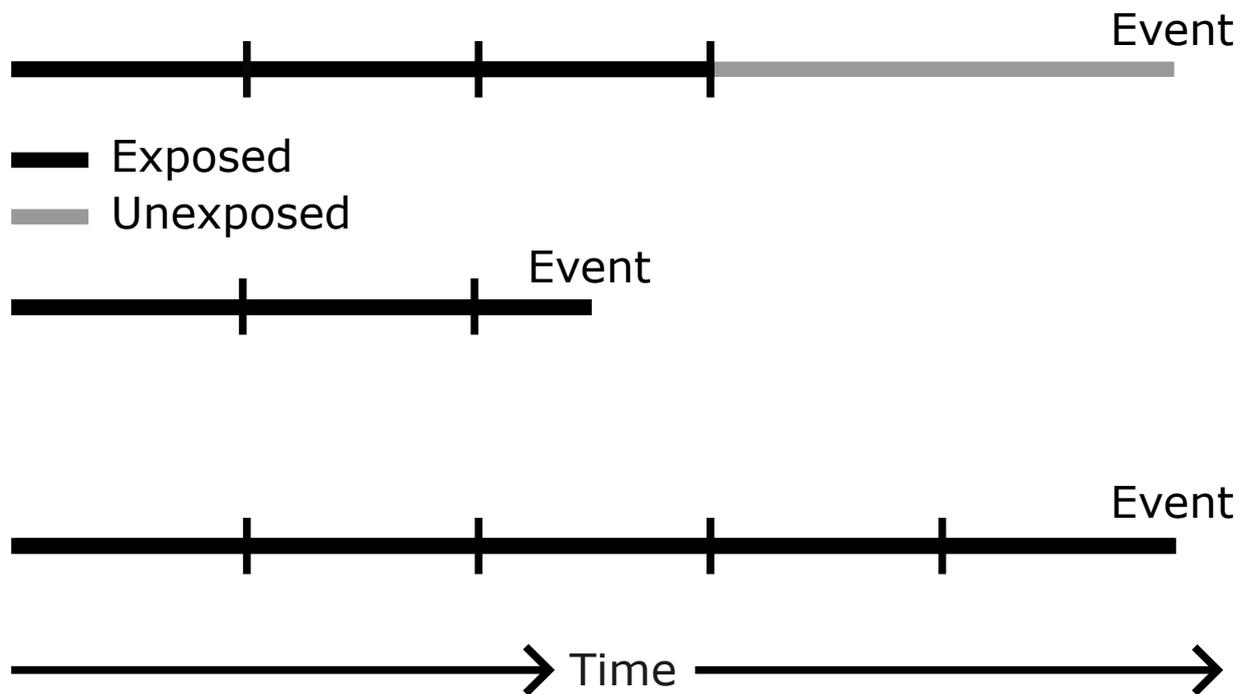
23. Ibrahim JG. Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association* 1990; 85(411): 765–769. DOI:10.1080/01621459.1990.10474938. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474938>.
24. Louis TA. Finding the Observed Information Matrix When Using the *EM* Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 1982; 44(2): 226–233. DOI: 10.1111/j.2517-6161.1982.tb01203.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1982.tb01203.x>.
25. Chen HY and Little RJA. Proportional Hazards Regression with Missing Covariates. *Journal of the American Statistical Association* 1999; 94(447): 896–908. DOI:10.1080/01621459.1999.10474195. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474195>.
26. Tsiatis AA. *Semiparametric theory and missing data*. Springer series in statistics, New York: Springer, 2006. ISBN 978-0-387-32448-7.
27. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL <https://www.R-project.org/>.
28. Schmidt M, Schmidt SAJ, Adelborg K et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clinical Epidemiology* 2019; Volume 11: 563–591. DOI:10.2147/CLEP.S179083. URL <https://www.dovepress.com/the-danish-health-care-system-and-epidemiological-research-from-health>
29. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H et al. Data Resource Profile: The Danish National Prescription Registry. *International Journal of Epidemiology* 2016; : dyw213 DOI: 10.1093/ije/dyw213. URL <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw213>.
30. WHO Collaborating Centre for Drug Statistics Methodology. Guidelines for atc classification and ddd assignment. [https://www.whocc.no/filearchive/publications/2020\\_guidelines\\_web.pdf](https://www.whocc.no/filearchive/publications/2020_guidelines_web.pdf), 2020. Accessed February 20, 2020.
31. Pedersen CB. The Danish Civil Registration System. *Scandinavian Journal of Public Health* 2011; 39(7\_suppl): 22–25. DOI:10.1177/1403494810387965. URL <http://journals.sagepub.com/doi/10.1177/1403494810387965>.
32. World Health Organization. Icd-10 version:2019. <https://icd.who.int/browse10/2019/en>. Accessed November 12, 2022.
33. Koul H, Susarla V and Van Ryzin J. Regression Analysis with Randomly Right-Censored Data. *The Annals of Statistics* 1981; 9(6): 1276–1288. URL <http://www.jstor.org/stable/2240417>. Publisher: Institute of Mathematical Statistics.



**Figure 1.** DAG for the instrumental variable setup.  $I$  is the instrument,  $Z$  is the exposure,  $U$  is a set of unmeasured confounders,  $X$  is a set of measured confounders, and  $Y$  is the outcome.



**Figure 2.** DAG of the setup.  $U$  is unobserved, but any confounding effect on treatment,  $Z(t)$ , and outcome,  $T$ , goes through the number of treatment administrations,  $W$ . Other confounders,  $X$ , can be adjusted for explicitly, although this is not necessary to get an unconfounded estimate of the effect of exposure.



**Figure 3.** In scenario number one, we know the number of treatment administrations because treatment has been terminated. In scenario two, we don't know the number of treatment administrations because the subject is still under treatment at the time of the event. In scenario three, we don't know the number of treatment administrations, but if the effect is capped at  $M = 4$ , then we do know that the number of treatment administrations is greater than or equal to four, which enables identification of the model.

**Table 1.** Results from simulation. Relative bias is bias / SD. SD is empirical standard deviation of estimates and Avg. SE is the average estimated SE from the simulations.

Model	Bias	Relative bias	SD	Avg. SE	CI Coverage
No adjustment	-0.43	-3.34	0.129	0.132	0.088
Adjusted analysis	-1.23	-6.77	0.182	0.156	0.000
Time-dependent adjustment	-0.04	-0.28	0.151	0.152	0.941
EM algorithm with sex	0.00	-0.01	0.160	0.157	0.945
EM algorithm without sex	0.00	-0.01	0.163	0.159	0.942

**Table 2.** The table shows the observed values of  $W$  along with the censoring pattern. Note that values of  $W > 10$  are observed but have been excluded from the table in the interest of space.

W	Censored	Observed
1	2436	3047
2	1320	2017
3	821	1640
4	631	1398
5	408	1304
6	293	1174
7	201	1052
8	134	924
9	99	847
10	65	824

**Table 3.** Estimated one-year risk difference of poisoning for new users of antidepressant medicine.

	Risk always treated	Risk never treated	Risk difference (95% CI)	p-value
Cox	19.3%	17.3%	1.9% (1.2%, 2.6%)	< 0.001
EM	16.7%	19.7%	-3.0% (-3.7%, -2.2%)	< 0.001