Unbiased and efficient estimation of causal treatment effects in crossover trials

Jeppe Ekstrand Halkjær Madsen1,2 | Thomas Scheike1 | Christian Pipper3,4

Abstract
We introduce causal inference reasoning to crossover trials, with a focus on thorough QT (TQT) studies. For such trials, we propose different sets of assumptions and consider their impact on the modeling 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. Crossover trials and particularly TQT studies can be analyzed efficiently using simple working regression models without biasing the estimates for the causal parameters of interest.

KEYWORDS
bias, causal inference, crossover trials, efficiency, TQT studies

INTRODUCTION
A thorough QT (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 (Kenward &
FIGURE 1 ECG.

Roger, 2010; Lu, 2014; Orihashi & Kumagai, 2021; Orihashi et al., 2021; Schall & Ring, 2011). 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 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, that is, 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 QT corrected (QTc) measurement corresponding to a particular length of the RR interval (typically 1 s). An example of a commonly used correction is the Friderici correction, which is given by

$$\text{QTc} = \frac{QT}{\sqrt{RR}}.$$ 

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 & Jones, 2006).

TQT studies are predominantly conducted as crossover trials in an attempt to lower sample size and eliminate between subject variation by paired comparisons of different treatments. In a crossover 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 posttreatment QTc measurements obtained at predefined 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 carryover effects, that is, 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 crossover 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 washout period is required in order to avoid carryover effects (typically five 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 litera-
One of the central points made here are 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 toward 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 crossover 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., 2023). Moreover, appreciable gains in efficiency compared to marginal estimators for causal effects have been demonstrated (Bartlett, 2018; Hernández et al., 2006; Robinson and Jewell, 1991). We extend these results to crossover 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, which 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 carryover in a crossover 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 analyze 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 compose 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 $t$, $i = 1,...,n, p = 1,...,P, t = 1,...,T$. 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 $0,...,P - 1$, where $Z_{ip} = 0$ corresponds to subject $i$ receiving placebo in period $p$. Often, we will suppress the $i$ because we assume that the subjects are independent draws from the same distribution. Furthermore, let $Y_p = (Y_{p1},...,Y_{pT})^T$ denote the vector of postbaseline measurements in period $p$.

In line with the informed choice of washout period in TQT studies, we assume the washout period has been sufficient to ensure no carryover 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 postbaseline measurements, because we do not expect a causal effect of the baseline measurements. Instead, we expect any association between baseline measurements and postbaseline 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, that is, temporary traits. Despite the lack of arrows between baseline and postbaseline 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 postbaseline measurements only has an impact on Assumption 1 in the remaining part of the manuscript, and does not matter for the theoretical results such as Theorem 1.

The DAG in Figure 3 implies the following about the data distribution:

**Assumption 1 (No carryover).** Let $x = (x_1, ..., x_p)^T$, 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 satisfies the Markov factorization property with respect to the DAG in Figure 3 (Peters et al., 2017), that is, the joint density of our data can be written as

\[
f(z, w, x, y) = f_z(z)f_w(w, w_1, ..., w_p)f_x(x|w, w_1, ..., w_p)f_y(y|z, w, w_1, ..., w_p)
\]

\[= f_z(z)f_w(w, w_1, ..., w_p)\prod_{p=1}^p f_{x_p}(x_p|w, w_p)f_{y_p}(y_p|z_p, w, w_p).
\]

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 3, 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 modeling 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 modeling stage of the estimation.

Let $Y_{z_1, ..., z_p}$ denote the postbaseline QTc measurements we would have made in period $p$ if, possibly counter to fact, the subject had received treatments $z_1, ..., z_p$. 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 $Y_{z_1, ..., z_p}$. 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, that is, $Y_{z_1, ..., z_p} = Y_{z_1, ..., z_p}$ for all treatment regimes, $z_1, ..., z_p, ..., z_P$ and $q_1, ..., q_p, ..., q_P$, and periods, $p$. Thus, from here on, $Y_p$ 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

\[
E(Y^z_t - Y^0_t), \quad t = 1, \ldots, T,
\]
as the causal contrasts of interest. Note that the first period in a crossover trial corresponds to such a trial, and as a consequence, the targeted causal contrasts can be identified as:

\[
E(Y^z_{1t} - Y^0_{1t}), \quad t = 1, \ldots, T.
\]

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

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

Clearly, this is not an efficient use of all the data collected in the crossover 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 postbaseline measurements. To this end, it would be natural to assume the same treatment effect in all periods:

**Assumption 2** (Same treatment effect).

\[
E(Y^z_p - Y^0_p) = E(Y^z_q - Y^0_q)
\]

for all \( p \) and \( q \), that is, 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).

\[
(Y^0_p, \ldots, Y^{p-1}_p, X_p, Z_p) \overset{D}{=} (Y^0_q, \ldots, Y^{q-1}_q, X_q, Z_q),
\]

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 modeling and design practice in crossover trials, where potential period effects are modeled 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 postbaseline measurements, given covariates, is the same in all periods:

**Assumption 4** (Same relationship).

\[
(Y_p | X_p = x, Z_p = z) \overset{D}{=} (Y_q | X_q = x, Z_q = z)
\]

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 \( E(Y^z_p - Y^0_p) \).

Assumption 4 may initially appear paradoxical in light of the DAG depicted in Figure 3, along with Assumption 1. Specifically, it may seem surprising that it does not 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, that is, we specifically develop estimators for $E(Y_1^T - Y_0^T)$ 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 nonparametric estimator:

$$\hat{\mu}_1(z) = \frac{1}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} I(Z_{ip} = z)Y_{ipt} - I(Z_{ip} = 0)Y_{ipt},$$

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 postbaseline measurement time point $t$. Naturally, an effect will be estimated for each $t = 1, \ldots, T$. Note that $\hat{\mu}_1(z)$ 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 postbaseline 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):

$$\hat{\mu}_2(z) = \frac{1}{n \cdot P} \sum_{i=1}^{n} \sum_{p=1}^{P} [h_{pt}(X_{ip}, z, \hat{\beta}) - h_{pt}(X_{ip}, 0, \hat{\beta})].$$

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 have to be missing completely at random for the simple estimator $\hat{\mu}_1(z)$ 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 semiparametric estimator:

$$\hat{\mu}_3(z) = \hat{\mu}_1(z) - \frac{1}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} \left[(I(Z_{ip} = z) - \frac{1}{P})h_{pt}(X_{ip}, z, \hat{\beta}) - (I(Z_{ip} = 0) - \frac{1}{P})h_{pt}(X_{ip}, 0, \hat{\beta})\right].$$

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 nonparametric estimator, namely, the true causal effect. The following theorem shows that $\hat{\mu}_3(z) = \hat{\mu}_2(z)$ for certain types of working models.

**Theorem 1** (Unbiasedness of $\hat{\mu}_3(z)$). Assume the data structure of this paper, and assume that we use a working model with postbaseline measurements as outcome, and with main effects of treatment specific to each postbaseline time point. Let $Y_i$ be
the vector of all postbaseline 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 that the working model parameters, $\beta$, are estimated from the following weighted least squares estimating equation:

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

(1)

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

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

(2)

where $A$ and $B$ are $T \times T$ matrices. If $A - B$ is nonsingular, 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, because this estimator can be achieved as a very special case of Theorem 1. However, adjustment for baseline measurements is a good idea 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 that the working model is a Gaussian linear mixed model with main effects of treatment specific to each postbaseline 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 maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (REML), although variance parameters, and thereby the $V$ matrix from Theorem 1 differs (see page 14 in Jiang, 2007).

When the working model is a Gaussian linear mixed model, the requirement (2) corresponds to modeling 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 modeled 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 corresponds 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):

\[ h_{pt}(x, z, \beta) = \beta_{pt} + \beta_x x + \beta_{zt}. \] (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 postbaseline 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 \( \hat{\mu}_{zt}(z) \). Thus, the estimates of \( \beta_{zt} \) are unbiased for the treatment effects of interest under arbitrary model misspecification.

To enable inference fully, we further need to characterize the large sample behavior. 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 \( \hat{\mu}_{zt}(z) \) 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, \( \hat{\mu}_{zt}(z) \) is still unbiased, whereas \( \hat{\mu}_{zt}(z) \) may be biased. Therefore, it would be preferable to use \( \hat{\mu}_{zt}(z) \) in such cases. Accordingly, we also derive the influence function for \( \hat{\mu}_{zt}(z) \) 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, that is, the null-hypothesis is

\[ H_0 : \bigcap_{t=1}^{T} \mu_t(z) \geq \Delta, \]

where

\[ \mu_t(z) = E(Y_{zt}^z - Y_{zt}^0), \]

and \( \Delta \) is some reasonable amount of QT prolongation, such as 10 ms (Patterson & 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 behavior 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 (Hothorn et al., 2008; Pipper et al., 2012). We can derive the joint asymptotic distribution of our estimators from the influence functions as

\[ \sqrt{n}(\hat{\mu}_z(z) - \mu(z)) = \begin{bmatrix} \hat{\mu}_{z1} - \mu_1(z) \\ \vdots \\ \hat{\mu}_{zT} - \mu_T(z) \end{bmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{bmatrix} \varphi_1(d_i) \\ \vdots \\ \varphi_T(d_i) \end{bmatrix} + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(d_i) + o_p(1), \]

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

\[ \frac{E(\varphi^T \varphi)}{n}. \] (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 crossover 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 & McCaffrey, 2002; Colin Cameron & Miller, 2015; MacKinnon & White, 1985; Pustejovsky & 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, that is, to estimate the variance by

\[
\frac{1}{n-1} \sum_{i=1}^{n} \hat{\phi}_i^2
\]

instead of

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_i^2
\]

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 reanalyze a standard TQT trial also analyzed 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, that is, treatment E is known to mildly prolong the QT interval. The subjects’ QT intervals are measured in triplicates at baseline, 0.5, 1, 1.5, 2.5, and 4 h posttreatment. The triplicates are averaged at each time point.

In accordance with the developments presented in Section 3, we analyze 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 postbaseline 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:

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

(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:

\[
h_{pt}(x, z, \beta) = \beta_{pt} + \beta_{zt},
\]

that is, 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

\[
h_{pt}(x, z, \beta) = \beta_{t} + \beta_{zt},
\]

that is, 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 modeling flexibility facilitated by Theorem 1, we fit a model with interaction between baseline and
TABLE 1  Estimates and standard errors for data example.

| Mean structure | Covariance structure | Estimate | Standard Error | 95% CI |
|----------------|----------------------|----------|---------------|-------|
| \(\beta_{pt} + \beta_{xz}x + \beta_{zi}z\) | 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) |
| \(\beta_{pt} + \beta_{xz}x + \beta_{zi}z\) | 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) |
| \(\beta_{i} + \beta_{xz}x + \beta_{zi}z\) | 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) |
| \(\beta_{i} + \beta_{xz}x + \beta_{zi}z\) | Unspecified          | 8.43     | 1.40          | (5.60, 11.26) |
|                 | AR(1)                | 8.29     | 1.43          | (5.40, 11.19) |
|                 | Independence         | 8.42     | 1.42          | (5.55, 11.30) |
| \(\hat{\mu}_{1t}(z)\)          |                      | 8.18     | 2.05          | (4.03, 12.33) |

Treatment:

\[ h_{pt}(x, z, \beta) = \beta_{i} + \beta_{xz}x + \beta_{zi}z. \]

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:

\[
A = \sigma^2 \begin{pmatrix}
1 & \rho & \ldots & \rho^{T-1} \\
\rho & 1 & \ldots & \rho^{T-2} \\
\vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \ldots & 1
\end{pmatrix},
\]

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 nonparametric estimator \(\hat{\mu}_{1t}(z)\). The estimates of the effect of treatment \(E\) compared to placebo at postbaseline 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 nonparametric 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.
### TABLE 2 Bias, standard deviation of estimates, and coverage of confidence intervals in simulations.

| Mean structure | Covariance structure | Bias | SD  | Avg. SE | Coverage |
|----------------|----------------------|------|-----|---------|----------|
| $\beta_{pt} + \beta_{x\bar{x}} + \beta_{\bar{x}\bar{x}} + \beta_{zt}$ | Unspecified          | 0.04 | 1.49| 1.43    | 0.944    |
|                | AR(1)                | 0.04 | 1.49| 1.43    | 0.944    |
|                | Independence         | 0.04 | 1.49| 1.48    | 0.951    |
| $\beta_{pt} + \beta_{x\bar{x}} + \beta_{zt}$ | Unspecified          | 0.04 | 1.50| 1.45    | 0.947    |
|                | AR(1)                | 0.04 | 1.50| 1.45    | 0.947    |
|                | Independence         | 0.04 | 1.52| 1.52    | 0.952    |
| $\beta_{t} + \beta_{x\bar{x}} + \beta_{zt}$ | Unspecified          | 0.04 | 1.49| 1.46    | 0.948    |
|                | AR(1)                | 0.04 | 1.48| 1.46    | 0.947    |
|                | Independence         | 0.04 | 1.51| 1.51    | 0.952    |
| $\beta_{t} + \beta_{x\bar{x}} + \beta_{zt}$ | Unspecified          | 0.03 | 1.48| 1.38    | 0.934    |
|                | AR(1)                | 0.03 | 1.48| 1.38    | 0.934    |
|                | Independence         | 0.04 | 1.51| 1.49    | 0.948    |
| $\hat{\mu}_{1}(z)$ | 0.03                | 1.96 | 1.94| 0.950    |

### 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 postbaseline measurements are simulated from that model.

There are at least two advantages to this approach: first, the simulation must be considered realistic because 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 10,000 simulations in the statistical program R (R Core Team, 2021), and the code is available at [https://github.com/Jeepen/TQTpaper](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, because 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 nonnegligible 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 & Rubin, 2020; Tsiatis, 2006).

We have focused on how to estimate causal effects under Assumption 2, that is, assuming that 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

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

that is, 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:

$$\frac{1}{n} \sum_{i=1}^{n} h(X_{iPt}, z, \hat{\beta}) - h(X_{iP0}, 0, \hat{\beta}).$$

(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, whereas it is unclear what we are emulating by ignoring the period-specific treatment effects (Hernández 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 nonparametric 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, e.g., chapter 8 of Patterson & Jones, 2006). It is duly noted that differences in least squares means are equivalent to $\hat{\mu}_{2t}(z)$ when the treatment effects are modeled 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 STATEMENT**

The authors have declared no conflict of interest.

**DATA AVAILABILITY STATEMENT**

The data that support the findings of this study are available in the Supporting Information of this article and at https://github.com/Jeepen/TQTpaper.

**OPEN RESEARCH BADGES**

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.
REFERENCES

Bartlett, J. W. (2018). Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical Statistics, 17*, 648–666.

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology, 28*, 169–182.

Colin Cameron, A., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources, 50*, 317–372. http://jhr.uwpress.org/lookup/doi/10.3368/jhr.50.2.317

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. 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., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology, 19*, 766–779. https://journals.lww.com/00000648-200811000-00002

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., & 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., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*, 346–363. https://doi.org/10.1002/bimj.200810425

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., & Roger, J. H. (2010). The use of baseline covariates in crossover studies. *Biostatistics, 11*, 1–17.

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data*, 3rd ed. Wiley.

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. https://onlinelibrary.wiley.com/doi/10.1002/pst.1647

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*, 305–325. https://linkinghub.elsevier.com/retrieve/pii/0304407685901587

Meng, Z., Quan, H., Fan, L., Kringle, R., & 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. https://www.tandfonline.com/doi/full/10.1080/10543409093582067

Orihashi, Y., & Kumagai, Y. (2021). Concentration-QTc analysis with two or more correlated baselines. *Journal of Pharmacokinetics and Pharmacodynamics, 48*, 615–622. https://doi.org/10.1080/07904383.2021.197578-9

Orihashi, Y., Kumagai, Y., & Shiosakai, K. (2021). Novel concentration-qtc models for early clinical studies with parallel placebo controls: A simulation study. *Pharmaceutical Statistics, 20*, 375–389.

Patterson, S. D., & Jones, B. (2006). *Bioequivalence and statistics in clinical pharmacology*. Chapman & Hall/CRC.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pipper, C. B., Ritz, C., & 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. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Biopharmaceutical Statistics, 28*, 672–683. https://www.tandfonline.com/doi/full/10.1080/07904383.2016.1247004

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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. https://linkinghub.elsevier.com/retrieve/pii/0270025890900886

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique, 59*, 227–240. http://www.jstor.org/stable/1403444

Rosenblum, M., & 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., & Ring, A. (2011). Mixed models for data from thorough QT studies: Part 1. Assessment of marginal QT prolongation. *Pharmaceutical Statistics, 10*, 265–276. https://onlinelibrary.wiley.com/doi/10.1002/pst.463

Senn, S. (2002). *Cross-over trials in clinical research*, 2nd ed. John Wiley & Sons.
Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted Learning*. Springer. [http://link.springer.com/10.1007/978-1-4419-9782-1](http://link.springer.com/10.1007/978-1-4419-9782-1)

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., & Rosenblum, M. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association, 118*, 1152.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Madsen, J. E. H., Scheike, T., & Pipper, C. (2023). Unbiased and efficient estimation of causal treatment effects in crossover trials. *Biometrical Journal, 65*, 2200170.

[https://doi.org/10.1002/bimj.202200170](https://doi.org/10.1002/bimj.202200170)