Meta TCT: Towards Interpretable Treatment Effects in Clinical Trials for Progressive Diseases

Florian Stijven*
I-BioStat, KU Leuven, Belgium

Craig Mallinckrodt
Pentara Corporation, Utah, United States

Geert Molenberghs
KU Leuven, I-BioStat, B-3000 Leuven, Belgium
Universiteit Hasselt, I-BioStat, B-3500 Hasselt, Belgium

Ariel Alonso
KU Leuven, I-BioStat, B-3000 Leuven, Belgium

Sam Dickson
Pentara Corporation, Utah, United States

Suzanne Hendrix
Pentara Corporation, Utah, United States

May 31, 2024

Abstract

In progressive diseases, like Alzheimer's disease, treatments that slow progression should start early in the disease course to longer maintain higher levels of functioning. In corresponding clinical trials, the treatment effect is usually expressed in terms of mean differences on a clinical scale. Early in the disease course, however, treatment effects expressed on a clinical scale are often small but may nonetheless correspond to an important slowing of disease progression. This complicates the appreciation of the relevance of observed treatment effects. For example, it may be difficult to determine whether a 2-point improvement on a clinical scale is relevant for clinical practice. In this paper, we propose the meta Time-Component Tests (meta TCT). This new approach leads to estimators of treatment effects on the time scale, in terms of time saved or percentage slowing of progression, that are easy to interpret. This approach is based on estimates obtained from an arbitrary model for longitudinal data and is, therefore, very flexible. Asymptotic properties of the Meta TCT estimators are

*F. Stijven gratefully acknowledges funding from Agentschap Innoveren & Ondernemen (VLAIO) and Janssen through a Baekeland Mandate [grant number HBC.2022.0145].

derived and evaluated in an extensive simulation study. Meta TCT is then applied to a phase 2/3 clinical trial for Alzheimer's disease, which was first analyzed with a mixed model. In this trial, meta TCT leads to important additional insights into the treatment effect. We believe that meta TCT will facilitate the estimation of interpretable treatment effects in clinical trials for progressive diseases, and that this, in turn, will fine-tune the evaluation of the clinical relevance of new treatments.

Keywords: Time-Component Test, Longitudinal Data, Progressive Diseases, Alzheimer's Disease.

1 Introduction

It is often recommended to start disease-modifying treatments (DMT) early for progressive diseases to maintain lower symptom severity and higher functioning levels for a longer time[1]. However, disease progression is typically slow in the early disease stages, resulting in small effect sizes for DMTs. Nonetheless, small effect sizes early on can imply a considerable slowing of the disease progression. In contrast, symptomatic treatments may have a large effect early on, without slowing disease progression on the long term. This poses considerable challenges for the design, analysis, and interpretation of clinical trials for DMTs in progressive diseases.

A recent trial in Alzheimer's disease (AD) highlights these challenges[2]. In this study, patients with mild cognitive impairment (MCI) – a preclinical stage of AD – progressed more slowly than patients with mild AD (as expected in early AD). In the placebo group, MCI patients deteriorated on average 1.25 points on the Clinical Dementia Rating Sum of Boxes (CDR-SB) scale[3], while mild AD patients deteriorated on average 2.30 points. A similar pattern emerged for the treatment effect: a mean drug-placebo difference of -0.62 in mild AD patients but only -0.35 in the MCI group. Although the absolute differences suggest that the drug is more effective in later AD stages, the mean changes under active treatment were about 28% less than those under placebo in both groups. This finding implies that the drug's relative effectiveness may be comparable across different stages of AD, even though the absolute effects (which emphasize symptom reduction) are larger in patients with mild AD.

Due to the small absolute effect sizes and the challenges in interpreting mean differences on clinical scales, there is a growing interest in population-level summary statistics that quantify treatment effects on the time scale. These statistics express effects in terms of time saved or percentage slowing of progression[4,5,6]. Raket[4] introduced several non-linear models using individual-patient data to achieve this. Alternatively, the meta time-component tests (TCTs)[7,8] provide a framework that uses summary-level information to quantify treatment effects on the time scale. These approaches are appealing because the time scale is intuitive and universal, in contrast to clinical scales that may be disease specific and difficult to interpret. It is, therefore, no surprise that these methods are receiving attention in the literature, at conferences, and in interactions with sponsors and regulatory authorities.

While empirical demonstrations of meta TCTs have already been published[7,8], rigorous studies of their statisical properties – both theoretical and empirical – are currently limited.[9] The primary aim of this paper is to introduce a general framework for meta TCTs, leading to several estimators for effects on the time scale for which asymptotic properties are derived. A secondary aim is to investigate the finite-sample properties of these estimators. Finally, we also want to elucidate the assumptions underlying meta TCTs and identify situations where meta TCTs may yield unstable results – a contribution that will facilitate the adoption of this approach in practice. On a related note, these estimators have been implemented in an easy-to-use R package: *TCT*.

The remainder of this paper is organized as follows. In Section 2, we introduce the motivating data along with a description of the standard analysis method of such data. Next, we outline a framework for defining treatment effects on the time scale in Section 3. We develop the meta TCT approach and present asymptotic results in Section 4. The corresponding finite-sample properties are evaluated in a simulation study, presented in

Section 5. We apply the meta TCT estimators to the motivating data in Section 6. The paper ends with reflections on the limitations and strengths of the meta TCT approach and avenues for future research.

2 Motivating Data

2.1 Notation

In this paper, we consider data from a randomized clinical trial with two parallel treatment groups and an outcome variable that is measured at fixed measurement occasions. For each patient, there are $K+1$ measurements of the outcome variable at fixed time points t_j for $j = 0, 1, \ldots, K$ where $t_0 = 0$. These time points correspond to the time since randomization. The first measurement has thus been made just before or after randomization. Let Y_{it_j} denote the observed outcome for patient i at time since randomization t_j. The $K+1$ dimensional vector $\boldsymbol{Y}_i = (Y_{it_0}, \ldots, Y_{it_K})'$ contains all observations for patient i. Let Z_i be the treatment indicator for patient i where $Z_i = 0$ corresponds to the control treatment and $Z_i = 1$ to the experimental treatment. In the meta TCT framework, it is also possible to adjust for baseline covariates, denoted by $\boldsymbol{X}_i$. Throughout this paper, we assume that data from each patient, $(\boldsymbol{X}_i', Z_i, \boldsymbol{Y}_i')'$ are i.i.d. To simplify the presentation, we will ignore the baseline covariates until the data application.

2.2 Motivating Data

The motivating data come from a multi-center phase 2/3 clinical trial (NCT03823404) in AD where a potentially DMT was compared to placebo[10]. Our analysis of these data focused on a pre-specified subgroup of patients in which, based on the mechanism of action, it was thought the drug would be most effective. Within this subgroup, 148 patients were randomized to active treatment and 79 to placebo. The primary outcome measure is the ADAS-Cog scale: a 13-item cognitive subscale of the Alzheimer's disease assessment scale, lower scores indicating less impairment[11]. The ADAS-Cog score is measured at baseline and at 12, 24, 40, and 48 weeks after randomization. Several baseline covariates were also measured: Mini Mental State Examination (MMSE) score at baseline, ApoE4 status, and treatment with acetylcholinesterase inhibitors.

There is a substantial amount of missingness in the outcome variable that also depends on the treatment group, as shown in Table 1. There is no missingness in the baseline covariates. We further assume missing at random (MAR). A mixed model for repeated measures (MMRM) leads to valid inferences under MAR regardless of the proportion of missing values; however, the results may be sensitive to violations of MAR because there are many missing values. Analyses of these data that rely on MAR should therefore be interpreted with care, and ideally, sensitivity analyses should be conducted with respect to violations of MAR.

2.3 Mixed Model for Repeated Measures

The MMRM is a standard method to analyze longitudinal data in clinical trials[12] and is the most common primary analysis in AD[2,7,8]. It fits within the ignorable likelihood analysis

Table 1: Number and proportion of missing values for the ADAS-Cog scores by measurement occasion and treatment arm. The p-values correspond to Fisher's exact tests for equal proportions of missingness between the treatment groups at each measurement occasion.

Measurement Occassion	Active Treatement, $n = 148$	Placebo, $n = 79$	p-value
Week 12	7 (4.7%)	1 (1.3%)	0.30
Week 24	24 (16%)	4 (5.1%)	0.018
Week 40	39 (26%)	8 (10%)	0.004
Week 48	52 (35%)	10 (13%)	<0.001

paradigm when MAR holds. It is, therefore, used as the reference model throughout this paper. Nonetheless, as will become clear in Section 4, the meta TCT framework can be used in conjunction with many other models for longitudinal data such as, for example, generalized estimating equations[13].

The MMRM is a linear mixed model that makes few assumptions about the mean and covariance structure. The linear predictor contains the interaction effect between time, treated a as a categorical covariate, and treatment, except for $t = 0$ where we assume that the mean outcome is equal in both treatment groups. Formally, the mean outcome is modeled as

$$E(Y_{t_j} \mid Z = z) = \begin{cases} \alpha_j \text{ if } z = 0 \\ \beta_j \text{ if } z = 1 \end{cases} \quad \text{for } \; j \in \{1, \ldots, K\}. \tag{1}$$

and $E(Y_{t_0} \mid Z = 0) = E(Y_{t_0} \mid Z = 1) = \alpha_0$. This essentially means that we have a mean parameter for each time-treatment combination, except at baseline. We can also adjust for baseline covariates by including them in the linear predictor in (1). The time-treatment specific means then depend on baseline covariates, but we can still define and estimate population-level means through the estimated marginal means (EMMs, also known as least-squares means)[14,15]. The covariance matrix is further assumed to be unstructured but common to all patients.

All MMRMs considered in this paper are fitted with restricted maximum likelihood (REML) using the `mmrm()` function from the `mmrm` R package[16].

Conventionally, the treatment effect is expressed in terms of the mean differences at the post-randomization measurement occasions: $(\alpha_1 - \beta_1, \ldots, \alpha_K - \beta_K)'$. This treatment effect is further termed the vertical treatment effect, as opposed to horizontal treatment effects that are defined on the time scale and estimated by meta TCTs. To test for a treatment effect in the MMRM, we formulate the following null and alternative hypotheses:

$$H_0 : \alpha_j = \beta_j \; \forall \; j \in \{1, \ldots, K\}$$
$$H_1 : \alpha_j \neq \beta_j \text{ for at least one } j \in \{1, \ldots, K\}.$$

In an MMRM, this null hypothesis can be tested with several methods, among which we will use the F-test with the Kenward-Roger approximate degrees-of-freedom[17]. Although this is a commonly used model in clinical trials with a longitudinal endpoint, interpreting the results of this model is not straightforward when the endpoint is a clinical scale. Indeed, it can be difficult to appraise the clinical relevance of a certain mean difference on a clinical scale[18]. This problem is compounded by the use of different clinical scales for the same disease, e.g., ADAS-Cog and CDR-SB for AD.

3 Progression Models for Repeated Measures

The progression models for repeated measures (PMRMs) are a class of models, introduced by Raket[4], that allows one to express treatment effects on the time scale. Raket[4] developed this class based on the mean outcome as a function of time. In this section, we generalize this class to arbitrary functionals such as the quantiles. Next, we discuss possible parameterizations of treatment effects on the time scale.

3.1 General Formulation

The PMRMs center around the *trajectories* in both treatment groups, as defined next, and the relations between these two trajectories.

Definition 1 (Trajectory)**.** *Let (i) $F_{Y|t,Z=z}$ denote the distribution of Y at time t in treatment group $Z = z$ and (ii) $h(\cdot)$ denote a relevant functional, such as the mean. The trajectory for treatment group $Z = z$ is then defined as the real-valued function*

$$f_z(t) \stackrel{def}{=} h(F_{Y|t,Z=z})$$

with $t \in [0, t_K]$ as domain.

In most applications and in the remainder of this paper, $h(\cdot)$ will be the mean; however, this can in principle be any functional that is deemed relevant for the application at hand. As before, let $\boldsymbol{\alpha}$ be the vector of time-specific mean parameters in the control group (although these parameters could also represent functionals such as the median). The trajectory in the control group is further termed the *reference trajectory* and is parameterized by $\boldsymbol{\alpha}$:

$$f_0(t; \boldsymbol{\alpha}) \stackrel{def}{=} h(F_{Y|t,Z=0}).$$

By construction, we have that $f_0(t_j; \boldsymbol{\alpha}) = \alpha_j$; however, the reference trajectory is not observable for $t \notin \{t_0, \ldots, t_K\}$. The reference trajectory between the fixed measurement occasions has to be based on interpolation between $\{(t_0, \alpha_0)', \ldots, (t_K, \alpha_K)'\}$.

The defining characteristic of the PMRMs class is that the trajectory in the active treatment group, f_1, is related in a very specific way to the reference trajectory:

$$h\left(F_{Y|t,Z=1}\right) \stackrel{def}{=} f_1(t; \boldsymbol{\alpha}, \boldsymbol{\gamma}) = f_0\left\{g(t; \boldsymbol{\gamma}); \boldsymbol{\alpha}\right\}.$$

The time mapping function $g(t; \boldsymbol{\gamma})$ maps the time point t in the active treatment group to the corresponding time point in the placebo group. This function is parameterized by $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}$ summarizes the treatment effect on the time scale in a real-valued vector. Depending on the functional form of $g(\cdot; \boldsymbol{\gamma})$, the PMRM can describe different types of horizontal treatment effects, which are discussed next. We further assume that $g(0; \boldsymbol{\gamma}) = 0$; this is justified by randomization at $t_0 = 0$.

3.2 Time-based Changes in Disease Progression

Raket[4] described the *time-based changes in disease progression* model that is parameterized as follows:

$$f_1(t_j; \boldsymbol{\alpha}, \boldsymbol{\gamma}) = f_0\left(\gamma_j \cdot t_j; \boldsymbol{\alpha}\right), \tag{2}$$

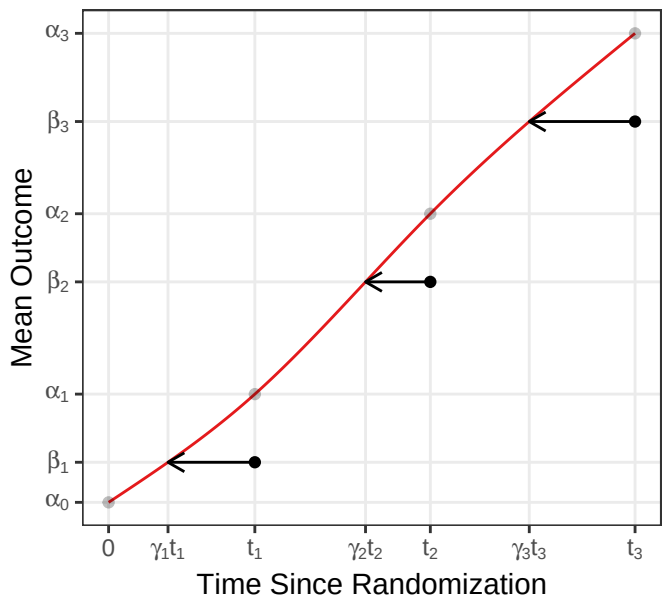

Figure 1: Illustration of the time mapping that underlies the time-specific acceleration factors. The red line is the reference trajectory. The gray dots on this line are the mean outcomes in the control group at the measurement occasions. The black dots are the corresponding mean outcomes in the active treatment group. The arrows visualize the time mapping from t_j to $\gamma_j t_j$. The length of these arrows corresponds to "the amount of time saved" at t_j.

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$. We further term γ_j the *time-specific acceleration factor* at t_j. In the above parameterization, the mean outcome in the active treatment group at t_j is $f_0(\gamma_j \cdot t_j; \boldsymbol{\alpha})$. In other words, the mean in the active treatment group at t_j is equal to the mean in the control group at $\gamma_j \cdot t_j$[4]. This time mapping is illustrated in Figure 1.

Effective DMTs slow progression and thus imply that $\gamma_j \in (0, 1)\ \forall\ j$. Treatments that reverse progression would have $\gamma_j < 0$, but we do not consider this possibility further.

In this parameterization, the time mapping function is only defined for $t \in \{t_0, t_1, \ldots, t_K\}$. Formally, we have that

$$g(t; \boldsymbol{\gamma}) = \begin{cases} 0 & \text{if } t = 0 \\ \gamma_j \cdot t & \text{if } t \in \{t_1, \ldots, t_K\} \\ \text{not defined} & \text{elsewhere.} \end{cases}$$

In principle, no special models are needed to estimate $\boldsymbol{\gamma}$. As before, let α_j and β_j be the mean outcome in, respectively, the placebo and active treatment group at t_j. We can then find γ_j by solving $\beta_j = f_0(\gamma_j \cdot t_j; \boldsymbol{\alpha})$ for γ_j. Consequently, γ_j can be expressed as a function of t_j, $\boldsymbol{\alpha}$, and β_j:

$$\gamma_j = \gamma(\boldsymbol{\alpha}, \beta_j; t_j).$$

With slight abuse of notation, we define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ as the vector-valued function

$$\boldsymbol{\gamma}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{t}) = \{\gamma(\boldsymbol{\alpha}, \beta_1; t_1), \ldots, \gamma(\boldsymbol{\alpha}, \beta_K; t_K)\}'. \tag{3}$$

This shows that, given $f_0(t;\boldsymbol{\alpha})$ and $\boldsymbol{\beta}$, the treatment effect on the time scale is also identified. Consequently, we could use an MMRM to estimate $(\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ as $(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')'$. The estimate for $\boldsymbol{\gamma}$ would then follow as $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}; \boldsymbol{t})$. Simple as this may seem, many subtleties are involved here; these are further discussed in Section 4.

The parameterization of horizontal treatment effects in (2) has a very intuitive interpretation that closely aligns with the interpretation of accelerated failure time models; hence, the adoption of the term *acceleration factor*. Indeed, γ_j quantifies how much the progression rate has increased or decreased by t_j in the active treatment group relative to the control group. For instance, $\gamma_j = 0.4$ means that the progression rate in the active treatment group, between randomization and t_j, was only 40% of the progression rate in the control group. In other words, after $t_j = 1$ year of treatment, we expect patients to only progress $0.40 \cdot 12 = 4.8$ months instead of 12 months if they receive the active treatment.

3.3 Proportional Slowing of Disease Progression

To reduce the number of parameters that describe the treatment effect, we could assume that the treatment causes a *proportional slowing* of disease progression. This is especially relevant for DMTs. Proportional slowing corresponds to additional restrictions on the $\boldsymbol{\gamma}$ vector in (2). If we assume that $\gamma_j = \gamma \; \forall \; j \in \{1, \ldots, K\}$, then we have that

$$f_1(t; \boldsymbol{\alpha}, \gamma) = f_0\left(\gamma \cdot t; \boldsymbol{\alpha}\right). \tag{4}$$

By the proportional slowing assumption, the time mapping function $g(t; \boldsymbol{\gamma})$ is now also well-defined in between measurement occasions.

When $\gamma \in (0, 1)$, the treatment can be said to *proportionally* slow the disease progression, although this provides no guarantee that the disease will continue to progress *more slowly* beyond the clinical trial's time window. In accelerated failure time model terminology, we now have a constant acceleration factor γ. We further refer to γ as the *common acceleration factor*.

4 Meta Time-Component Tests

In the previous section, we introduced the PMRMs as a general framework for formally defining treatment effects on the time scale. We now introduce the meta-TCT framework, which naturally leads to estimators for these effect measures. In this section, we examine the assumptions underlying meta TCT, describe two meta-TCT estimators, and finally, introduce a parametric bootstrap based on summary-level information.

4.1 Assumptions

The time-specific acceleration factor at t_j, defined in Section 3.2 as $\gamma_j = \gamma(\boldsymbol{\alpha}, \beta_j; t_j)$, is only well-defined if $\beta_j = f_0(\gamma_j \cdot t_j; \boldsymbol{\alpha})$ has a unique solution. We, therefore, make the *regular reference trajectory* assumption, a sufficient condition for the existence of a unique solution:

Assumption 1 (Regular reference trajectory)**.** *The reference trajectory, $f_0(t; \boldsymbol{\alpha})$, is regular if and only if it is*

1. *strictly increasing (or decreasing) with respect to* $t \in [t_0, t_K]$ *for* $\boldsymbol{\alpha} \in \Omega_{\boldsymbol{\alpha}}$

2. *twice continuously differentiable with respect to* $(t, \boldsymbol{\alpha}')'$ *for* $t \in [t_0, t_K]$ *and* $\boldsymbol{\alpha} \in \Omega_{\boldsymbol{\alpha}}$,

where $\Omega_{\boldsymbol{\alpha}}$ *is a compact set containing the true value* $\boldsymbol{\alpha_0}$.

Among other things, this assumption ensures that every point on the active treatment's trajectory can be mapped uniquely to the reference trajectory. If this time mapping were not unique, the acceleration factor γ_j would be ill-defined.

Because we have fixed measurement occasions, the estimation of the reference trajectory inevitably relies on interpolation. To avoid extrapolation, we make the range-restriction assumption.

Assumption 2 (Range restriction)**.** *The range of the trajectory in the active treatment group is a subset of the range of the reference trajectory, that is,*

$$f_1\left([t_0, t_K]; \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) \subseteq f_0\left([t_0, t_K]; \boldsymbol{\alpha}\right).$$

In finite samples, however, extrapolation may be needed because this assumption may fail for the *estimated* trajectories. Throughout this paper, the linear function through $(t_0, \hat{\alpha}_0)$ and $(t_K, \hat{\alpha}_K)$ is used for extrapolation.

Finally, we assume that $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$, where n is the total sample size, is asymptotically normal (Assumption 3). Although this holds for most estimators, the small-sample sampling distribution may still be far from normal—meta-TCT estimates and inferences should then be interpreted cautiously.

Assumption 3 (Asymptotically normal sampling distribution)**.** *The sampling distribution of* $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$ *has the* $n^{1/2}$ *convergence rate, and* $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$ *is consistent and asymptotically normal:*

$$n^{1/2}\left(\begin{pmatrix}\hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n\end{pmatrix} - \begin{pmatrix}\boldsymbol{\alpha_0} \\ \boldsymbol{\beta_0}\end{pmatrix}\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma\right),$$

where $(\boldsymbol{\alpha_0}', \boldsymbol{\beta_0}')'$ *are the true parameters and* $\xrightarrow{d}$ *denotes convergence in distribution.*

We further assume that a consistent estimator $\hat{\Sigma}_n$ for Σ is available.

In the remainder of this section, we introduce estimators and inferential procedures together with heuristic justifications that allow the reader to understand why they work. Rigorous justifications are provided in Web Appendices A and B.

4.2 Score-based Inference

The *score-based meta TCT* is based on score functions that quantify the discrepancy between the observed data and a particular value for $\boldsymbol{\gamma}$. The corresponding estimators minimize this discrepancy. Inference relies on the asymptotic normal distribution of these scores.

4.2.1 Score Functions

To construct a score function, we start from the following null hypothesis: $H_{\gamma_0} : \gamma_j = \gamma_0$. Under this null hypothesis, and assuming we know $\boldsymbol{\alpha}$, the mean in the experimental treatment group at t_j is

$$\beta_{0,j} = f_0(\gamma_0 \cdot t_j; \boldsymbol{\alpha_0}).$$

Consequently, a test for $H_0 : \beta_j = \beta_{0,j}$ is also a valid test for H_{γ_0}. Tests are often available for $H_0 : \beta_j = \beta_{0,j}$, e.g., F-tests in an MMRM, but these cannot be applied directly since $\boldsymbol{\alpha_0}$ is not known but estimated. Indeed, $\beta_{0,j}$ is itself estimated by $\hat{\beta}_{0,j} = f_0(\gamma_0 \cdot t_j; \hat{\boldsymbol{\alpha}}_n)$. This provides the basis for a score function and corresponding score test that draws upon well-established statistical tests.

We define the *score function* as follows:

$$\boldsymbol{s}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \stackrel{def}{=} \hat{\boldsymbol{\beta}}_n - \boldsymbol{f_0}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n), \tag{5}$$

where $\boldsymbol{\gamma t} = (\gamma_1 \cdot t_1, \ldots, \gamma_K \cdot t_K)'$ and $\boldsymbol{f_0}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n) = \{f_0(\gamma_1 \cdot t_1; \hat{\boldsymbol{\alpha}}_n), \ldots, f_0(\gamma_K \cdot t_K; \hat{\boldsymbol{\alpha}}_n)\}'$. Further, we will often assume proportional slowing, which corresponds to the following score function:

$$\boldsymbol{s}(\gamma \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) = \boldsymbol{s}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \text{ with } \boldsymbol{\gamma} = (\gamma, \ldots, \gamma)'. \tag{6}$$

Under H_{γ_0} and using the delta method, we have that

$$n^{1/2} \cdot \boldsymbol{s}(\boldsymbol{\gamma} \cdot \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \xrightarrow{d} N(\boldsymbol{0}, D_{\boldsymbol{s}}) \text{ if } \boldsymbol{\gamma} = \boldsymbol{\gamma_0}, \tag{7}$$

where

$$D_{\boldsymbol{s}} = J_{\boldsymbol{s}}(\boldsymbol{\alpha_0}, \boldsymbol{\beta_0}; \boldsymbol{\gamma_0}) \cdot \Sigma \cdot J_{\boldsymbol{s}}(\boldsymbol{\alpha_0}, \boldsymbol{\beta_0}; \boldsymbol{\gamma_0})'$$

and $J_{\boldsymbol{s}}(\boldsymbol{\alpha_0}, \boldsymbol{\beta_0}; \boldsymbol{\gamma_0})$ is the Jacobian of $\boldsymbol{s}(\boldsymbol{\gamma_0 t}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $(\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ evaluated in $(\boldsymbol{\alpha_0}', \boldsymbol{\beta_0}')'$. In practice, $D_{\boldsymbol{s}}$ is replaced with a consistent estimator $\hat{D}_{\boldsymbol{s},n}$ that is obtained by replacing the unknown parameters with their respective estimators: $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$ and $\hat{\Sigma}_n$.

4.2.2 Univariate Estimator

Every element of the score function in (5) corresponds to the time mapping at a particular measurement occasion. These elements are termed univariate score functions. Each univariate score function can be used to estimate the acceleration factor at a specific measurement occasion. Let $s(\gamma \cdot t_j; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) = \hat{\beta}_j - f_0(\gamma \cdot t_j; \hat{\boldsymbol{\alpha}}_n)$ be the univariate score function at t_j. The value of γ that solves $s(\gamma \cdot t_j; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) = 0$, further denoted as $\hat{\gamma}_{j,n}$, is an estimator of $\gamma_{j,0}$. A corresponding test statistic follows from the distributional result in (7):

$$z(\gamma) = \frac{n^{1/2} \cdot s(\gamma \cdot t_j; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)}{\sqrt{\hat{D}_{s,n}^{jj}}} \xrightarrow{d} N(0,1) \text{ if } \gamma = \gamma_{j,0}, \tag{8}$$

where $\hat{D}_{s,n}^{jj}$ is the j'th diagonal element of $\hat{D}_{s,n}$. For the construction of corresponding score-based confidence intervals, we use the duality of confidence intervals and hypothesis tests; hence, the $1-\alpha$ confidence interval is obtained as follows:

$$\left\{\gamma : z(\gamma)^2 \leq \chi^2_{1-\alpha,1}\right\} \tag{9}$$

where $\chi^2_{1-\alpha,1}$ is the $1-\alpha$ percentile of the chi-squared distribution with one degree of freedom. A numerical procedure for finding these confidence limits is described in Web Appendix A.4.1 and is implemented in the *TCT* R package.

4.2.3 Multivariate Estimator

Whereas the univariate score-based estimator is uniquely defined, the estimator for the common acceleration factor is not because the score function (6) can be reduced to distinct univariate equations, the solutions of which correspond to distinct score-based estimators. Next, we describe three such estimators.

Weights-based Estimator The weights-based estimator for the common acceleration factor, denoted by $\hat{\gamma}_{\boldsymbol{w},n}$, is the solution of

$$\boldsymbol{w}' \cdot \boldsymbol{s}(\gamma \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) = 0,$$

where $\boldsymbol{w}$ is a K-dimensional column vector of predefined weights. This equation is solved numerically in the *TCT* R package. Test statistics and confidence intervals can be derived similarly as in Section 4.2.2; they are provided in Web Appendix A.4.2. The standard error of $\hat{\gamma}_{\boldsymbol{w},n}$ can be estimated based on Lemma A.4, or the parametric bootstrap.

It is difficult to prespecifiy the weights as required for $\hat{\gamma}_{\boldsymbol{w},n}$. We, therefore, suggest an adaptive approach for determining these weights. Specifically, we choose the weights that minimize the estimated variance of the corresponding weights-based estimator, $\hat{\sigma}^2(\hat{\gamma}_{\boldsymbol{w},n})$:

$$\hat{\boldsymbol{w}}_{adapt,n} = \arg\min_{\boldsymbol{w}} \hat{\sigma}^2(\hat{\gamma}_{\boldsymbol{w},n}).$$

In Web Appendix A.5.1, an algorithm is proposed for finding $\hat{\boldsymbol{w}}_{adapt,n}$. This algorithm is also implemented in the *TCT* R package. For inference about $\hat{\gamma}_{\boldsymbol{w},n}$ when the weights have been adaptively determined, we can proceed as if the weights were fixed. This is justified in Web Appendix A.5.2.

Chi-squared-based Estimator Instead of a weighted sum of the univariate scores, we can also use well-known test statistics. A classic chi-squared statistic is obtained as follows:

$$t(\gamma)^2 = n \cdot \boldsymbol{s}(\gamma \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)' \cdot \hat{D}_{\boldsymbol{s}}^{-1} \cdot \boldsymbol{s}(\gamma \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \xrightarrow{d} \chi^2_K \text{ if } \gamma = \gamma_0.$$

The chi-squared-based estimator for the common acceleration factor is then defined as $\hat{\gamma}_{\chi^2,n} = \arg\min_\gamma t(\gamma)^2$. As before, a $1-\alpha$ confidence interval is obtained as follows:

$$\left\{\gamma : t(\gamma)^2 \leq \chi^2_{1-\alpha,K}\right\},$$

where $\chi^2_{1-\alpha,K}$ is the $1-\alpha$ percentile of the chi-squared distribution with K degrees-of-freedom.

Finally, we highlight a relationship between hypothesis tests based on $\boldsymbol{s}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ and hypothesis tests based on $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$. The hypothesis of no treatment effect, $H_0 : \beta_j = \alpha_j \; \forall \, j = 1, \ldots, K$, corresponds to the following linear hypothesis:

$$H_0 : L \cdot \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \beta_1 - \alpha_1 \\ \vdots \\ \beta_K - \alpha_K \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{10}$$

where L is an appropriate contrast matrix. A Wald test for (10), or corresponding subhypotheses, is based on the following distributional result:

$$n^{1/2} \cdot L \cdot \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, L\Sigma L') \text{ if } H_0 \text{ is true.}$$

Because $s(\gamma \cdot t_j; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \beta_j - \alpha_j$ for $\gamma = 1$, we have that

$$n^{1/2} \cdot L \cdot \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix} = n^{1/2} \cdot \boldsymbol{s}(\boldsymbol{\gamma t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \text{ for } \boldsymbol{\gamma} = (1, \ldots, 1)'.$$

Consequently, test statistics based on $L \cdot (\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$ are equivalent to corresponding test statistics based on $\boldsymbol{s}(\gamma \boldsymbol{t}; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ with $\gamma = 1$.

For instance, the Wald chi-squared test statistic for $H_0 : L \cdot (\boldsymbol{\alpha}', \boldsymbol{\beta}')' = 0$ equals $t(1)^2$ as defined above; consequently, inference based on $\hat{\gamma}_{\chi^2,n}$ and $t(\gamma)^2$ is always consistent with the Wald chi-squared test for a treatment effect. Using $\hat{\gamma}_{\chi^2,n}$ and $t(\gamma)^2$ will not increase power, but it does quantify the treatment effect in an easy-to-interpret measure. Similarly, the Wald z-statistic for $H_0 : \alpha_j = \beta_j$ equals $z(1)$ as defined in (8).

4.3 Non-linear Generalized Least Squares

For estimating the common acceleration factor, we propose a second approach based on non-linear generalized least squares, further termed *NL-GLS meta TCT.*

4.3.1 Estimation

In the NL-GLS meta TCT, we treat $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$ as the *observed data* and fit a non-linear regression function, parameterized by $(\gamma, \boldsymbol{\alpha}')'$, to these data:

$$E\left\{(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'\right\} = \{\boldsymbol{f_0}(\boldsymbol{t}; \boldsymbol{\alpha})', \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t}; \boldsymbol{\alpha})'\}' = \{\boldsymbol{\alpha}', \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t}; \boldsymbol{\alpha})'\}'. \tag{11}$$

The second equality is a consequence of f_0 being an interpolating function. This regression function is fitted using non-linear generalized least squares, that is, the estimator of $(\gamma_0, \boldsymbol{\alpha_0}')'$ minimizes the generalized least-squares criterion:

$$(\hat{\gamma}_{NL,n}, \hat{\boldsymbol{\alpha}}_{NL,n}')' = \arg\min_{(\gamma,\boldsymbol{\alpha}')'} \left\{ \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t}; \boldsymbol{\alpha}) \end{pmatrix} \right\}' \left(\frac{\Sigma}{n}\right)^{-1} \left\{ \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t}; \boldsymbol{\alpha}) \end{pmatrix} \right\}.$$

In practice, we replace Σ with a consistent estimator $\hat{\Sigma}_n$.

4.3.2 Inference

Inference based on $\hat{\gamma}_{NL,n}$ is not straightforward since the observed data, $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$, have a fixed "sample size." Asymptotic arguments still work because the observed data are asymptotically normal by Assumption 3. In Web Appendix B, we derive asymptotic results for $\hat{\gamma}_{NL,n}$. In the following paragraphs, we summarize these results.

The regression function in (11) can be locally approximated by a linear regression function using a Taylor series expansion about $(\gamma_0, \boldsymbol{\alpha_0}')'$. This leads to an approximate

general linear model. If the covariance matrix is known and the residuals are multivariate normal, exact inference in a general linear model is possible based on the generalized least-squares criterion. Let Q be the minimized generalized least-squares criterion under no restrictions and $Q_0(\gamma)$ the minimized criterion when fixing the common acceleration factor at γ. It then follows that $Q_0(\gamma) - Q \sim \chi^2_1$ for $\gamma = \gamma_0$. As before, a $1-\alpha$ confidence interval is obtained as follows:

$$\left\{\gamma : Q_0(\gamma) - Q \leq \chi^2_{1-\alpha,1}\right\}. \tag{12}$$

The standard error is computed similarly as $(nX'\Sigma^{-1}X)^{-1}$, where X is the design matrix of the approximate general linear model. Nevertheless, these results are only approximate because the non-linear regression function in (11) is only approximated by a linear function.

Because $(\hat{\gamma}_{NL,n}, \hat{\boldsymbol{\alpha}}'_{NL,n})'$ is a consistent estimator (shown in Web Appendix B), the relevant neighbourhood about $(\gamma_0, \boldsymbol{\alpha_0}')'$ becomes smaller with an increasing sample size. The error in the Taylor approximation converges to zero as the relevant neighborhood becomes more concentrated about $(\gamma, \boldsymbol{\alpha_0}')'$. Consequently, as $n \to \infty$, the non-linear regression model and its linear approximation will become equivalent; exact inferential results in the general linear model are, therefore, asymptotic inferential results in the non-linear model.

4.4 Parametric Bootstrap

Inference for the score-based and NL-GLS meta TCT estimators is based on asymptotic arguments, which provide no guarantee in finite samples. We, therefore, also propose a parametric bootstrap procedure that does not require individual-patient data and may perform better in small samples. We resample the estimated mean parameters from the estimated multivariate normal sampling distribution:

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}_b \\ \hat{\boldsymbol{\beta}}_b \end{pmatrix} \sim N\left(\begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix}, \frac{\hat{\Sigma}_n}{n}\right),$$

where $(\hat{\boldsymbol{\alpha}}'_b, \hat{\boldsymbol{\beta}}'_b)'$ is the b'th bootstrap replicate of the mean vector. The b'th bootstrap replicate of the common acceleration factor $\hat{\gamma}^b$ is obtained by applying the meta TCT estimator to $(\hat{\boldsymbol{\alpha}}'_b, \hat{\boldsymbol{\beta}}'_b)'$ and $\hat{\Sigma}_n$. Subsequent inference is based on the $1-\alpha$ percentile confidence interval $(\hat{\gamma}^b_{\alpha/2}, \hat{\gamma}^b_{1-\alpha/2})$ where $\hat{\gamma}^b_p$ is the p-th percentile of the bootstrap distribution. This parametric bootstrap can also be used for the time-specific acceleration factors.

5 Simulations

In this section, we summarize the results of our simulation study for assessing the finite sample properties of the meta TCT estimators and inferential procedures. This simulation study has three goals, ordered by importance:

1. Assess to what degree the theoretical asymptotic results translate to finite samples.
2. Identify finite sample settings where the meta TCT methods may not be trustworthy.
3. Assess the correctness of the implementation of the meta TCT methods in the *TCT* R package.

5.1 Data-generating Model

The data-generating model for the simulations is based on data from the Alzheimer's disease neuroimaging initiative (ADNI)[19], closely resembling the data-generating model used by Raket[4]. Specifically, the data-generating model for the control group is based on a selection of 556 patients from the ADNI for whom the ADAS-cog scores are available at baseline and 6, 12, 18, 24, and 36 months after baseline. To allow for equally-spaced measurement occasions in our data-generating model, we added a measurement at 30 months. Given these seven potential measurement occasions, we consider the following three measurement patterns:

- *24 Months.* One follow-up visit every 6 months until 24 months after randomization.
- *36 Months.* One follow-up visit every 6 months until 36 months after randomization.
- *36(-30) Months.* Same pattern as 36 Months but leaving out the measurement at 30 months.

We consider a normal and fast progression rate in the placebo group. The normal progression rate is based on the estimated means from the ADNI patients augmented with the fictitious mean ADAS-Cog score at 30 months, leading to the following mean vector: $(19.6, 20.5, 20.9, 22.7, 23.8, 25.8, 27.4)'$. The fast progression rate is not based on data; here, we used the following mean vector: $(18.0, 19.7, 20.9, 22.7, 24.7, 27.1, 29.2)'$. In all scenarios, the covariance matrix for the 36 Months pattern is

$$\begin{pmatrix} 45.1 & 40.0 & 45.1 & 54.9 & 53.6 & 53.6 & 60.8 \\ 40.0 & 57.8 & 54.4 & 66.3 & 64.1 & 64.1 & 74.7 \\ 45.1 & 54.4 & 72.0 & 80.0 & 77.6 & 77.6 & 93.1 \\ 54.9 & 66.3 & 80.0 & 109.8 & 99.3 & 99.3 & 121.7 \\ 53.6 & 64.1 & 77.6 & 99.3 & 111.4 & 99.1 & 127.8 \\ 53.6 & 64.1 & 77.6 & 99.3 & 99.1 & 111.4 & 127.8 \\ 60.8 & 74.7 & 93.1 & 121.7 & 127.8 & 127.8 & 191.4 \end{pmatrix}. \tag{13}$$

This is the estimated covariance matrix from the ADNI patients augmented with an extra row and column for the hypothetical measurement at 30 months. The covariance matrices for the other measurement patterns are the corresponding subsets of this matrix.

To simulate proportional slowing, we first consider the *reference trajectory.* In the control group, we can only observe the mean at seven distinct time points; we interpolate between these seven points with natural cubic spline interpolation. The interpolated reference trajectories are plotted as purple lines in Figure 2. To simulate data for the treated group ($Z = 1$), we consider trajectories of the following form:

$$E(Y_t|Z = 1) = f_0(\gamma \cdot t; \boldsymbol{\alpha}),$$

where $\gamma \in \{1, 0.90, 0.75, 0.50\}$, representing four different proportional slowing treatment effects.

Finally, we consider four different samples sizes: $n \in \{50, 200, 500, 1000\}$, where n is the total sample size and 1:1 randomization is assumed. The data are sampled from multivariate normal distributions with the means determined by the above trajectory function ($\gamma = 1$ in the placebo group) and with (13) or subsets thereof as covariance matrix.

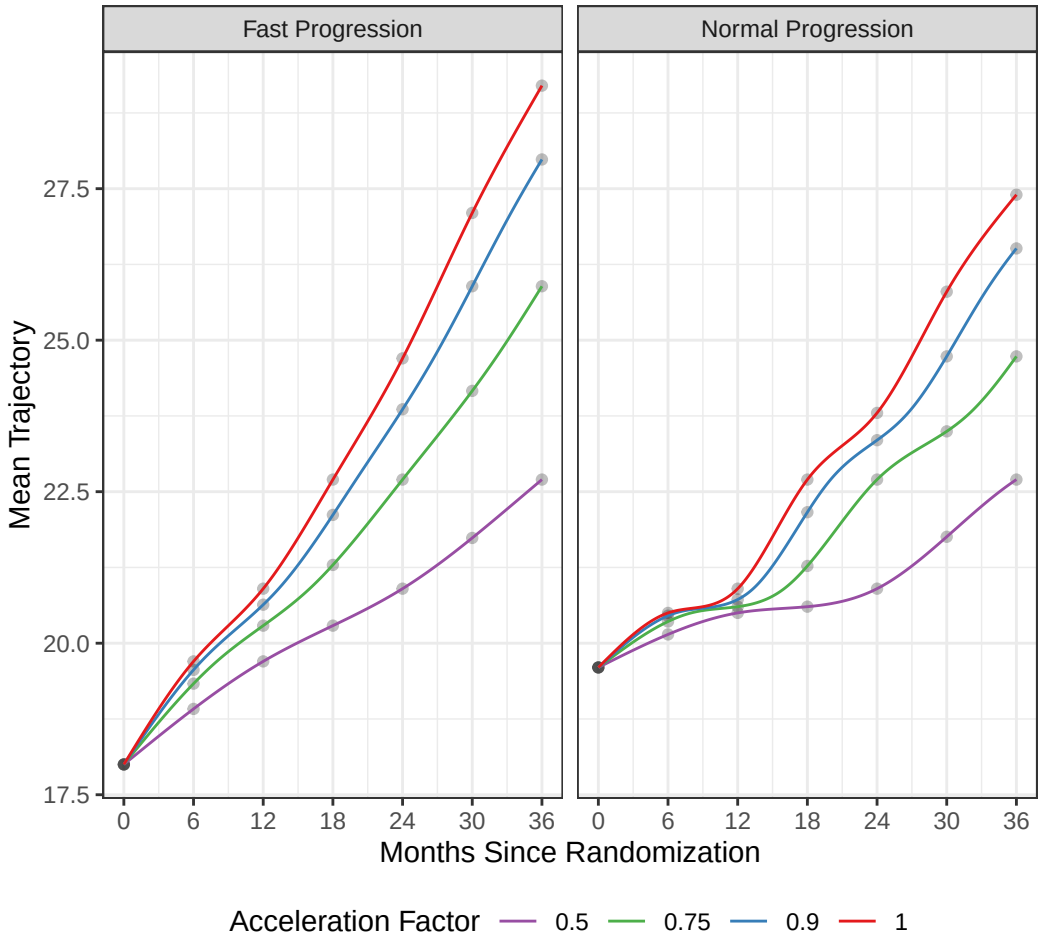

Figure 2: Trajectories underlying the data-generating models. An acceleration factor equal to one corresponds to no treatment effect. The dots correspond to measurement occasions, and the lines connecting the dots are based on natural cubic spline interpolation.

5.2 Results

In this section, the results of the simulation study are summarized only for the NL-GLS version of meta TCT, using either natural cubic splines or linear interpolation to estimate the reference trajectory. To obtain $(\hat{\boldsymbol{\alpha}}_n', \hat{\boldsymbol{\beta}}_n')'$, we used the MMRM as described in Section 2.3. The other versions of meta TCT are generally outperformed by NL-GLS meta TCT and are, therefore, not presented in the main text. Additional results for NL-GLS meta TCT and the adaptive weights-based estimator are presented in Web Appendix E. In each scenario, we perform 5000 Monte-Carlo replications. This leads to a standard error of $\sqrt{0.05 \cdot 0.95}/\sqrt{5000} = 0.003$ for estimating the empirical coverage of the 95% confidence intervals, assuming nominal coverage.

The presentation of the results is divided into two parts. First, we present the results regarding estimation of the common acceleration factor. This is the primary goal of the meta TCT methods: transforming the treatment effect on a difficult-to-interpret clinical scale to the time scale. Second, we present the results regarding inference based on the generalized least-squares criterion and the parametric bootstrap.

5.2.1 Estimation

In Figure 3, the means of the estimated acceleration factors are plotted as functions of the sample size. For cubic interpolation, the bias decreases to almost 0 as the total sample size increases to 1000. For linear interpolation, the bias does not decrease to zero everywhere, a small bias remains in some settings. This is expected because this estimator relies on a misspecified model: the data are generated not under linear interpolation but under natural cubic spline interpolation.

Figure 3 also shows that, when there is bias, the bias is smaller for the fast progression rate than for the normal progression rate. The bias also tends to be smaller when there are more measurement occasions.

In Figure 4, the empirical standard deviations are plotted alongside the median estimated standard errors. For small sample sizes, the standard error estimators underestimate the empirical standard error, but this underestimation largely disappears for larger sample sizes. Figure 4 also shows that a longer follow-up and faster progression lead to a smaller empirical standard deviation.

5.2.2 Inference

The empirical coverage rates of the 95% confidence intervals based on (12) are plotted in Figure 5. This reveals that there is undercoverage for the sample sizes between 50 and 500, but the empirical coverage rate approaches 95% for a sample size of 1000.

In Figure 6, the empirical coverage rates are presented for the 95% percentile confidence intervals based on the parametric bootstrap. To limit the computational burden, we only used $B = 500$ bootstrap replications throughout and only reanalyzed the normal progression scenarios (because the largest deviations from nominal were observed there). Figure 6 reveals an overcoverage that disappears with an increasing sample size; for $n = 1000$, coverage is close to nominal. These results indicate that the parametric bootstrap permits valid, but generally conservative, inferences in small samples.

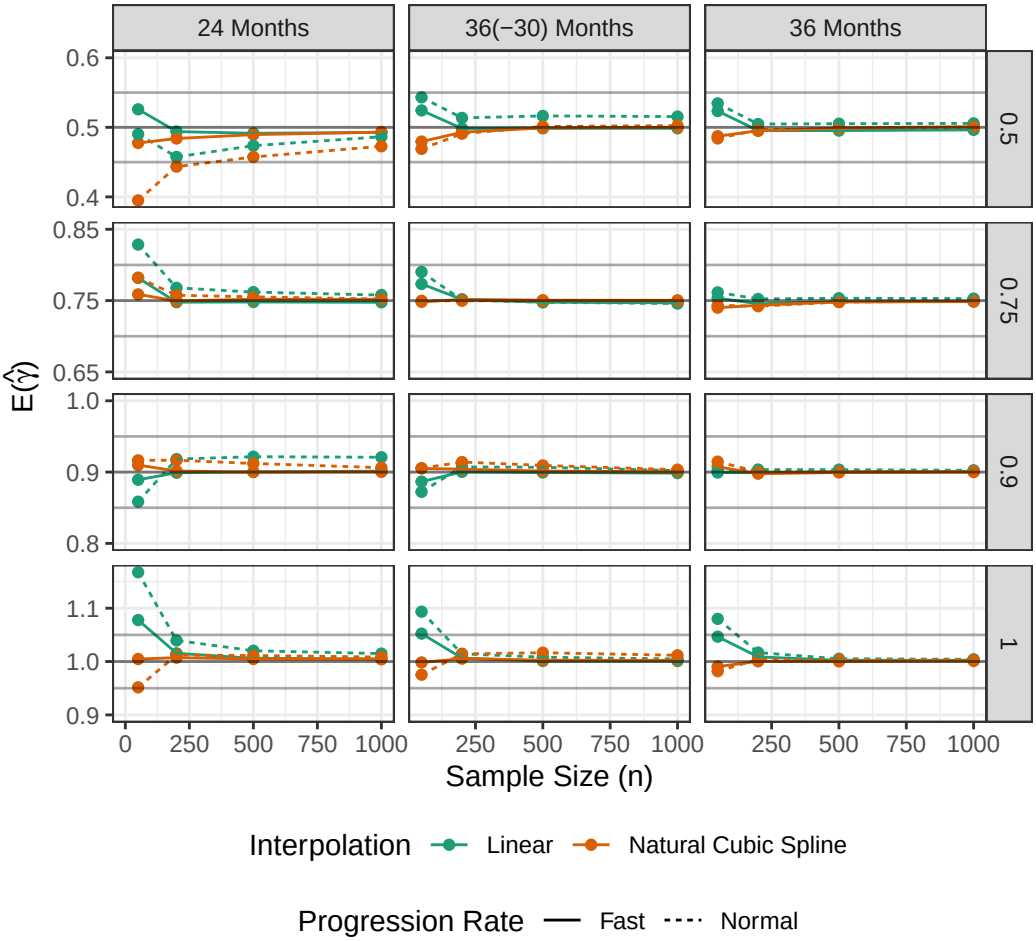

Figure 3: The means of the estimated acceleration factors as functions of the sample size. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factors, the columns correspond to the measurement patterns. In each subplot, a black horizontal line represents the true acceleration factor and gray horizontal lines represent the 0.05-margin around the true value. The maximum standard errors of the mean for a sample size of 50, 200, 500, and 1000 are 0.027, 0.003, 0.002, and 0.001, respectively.

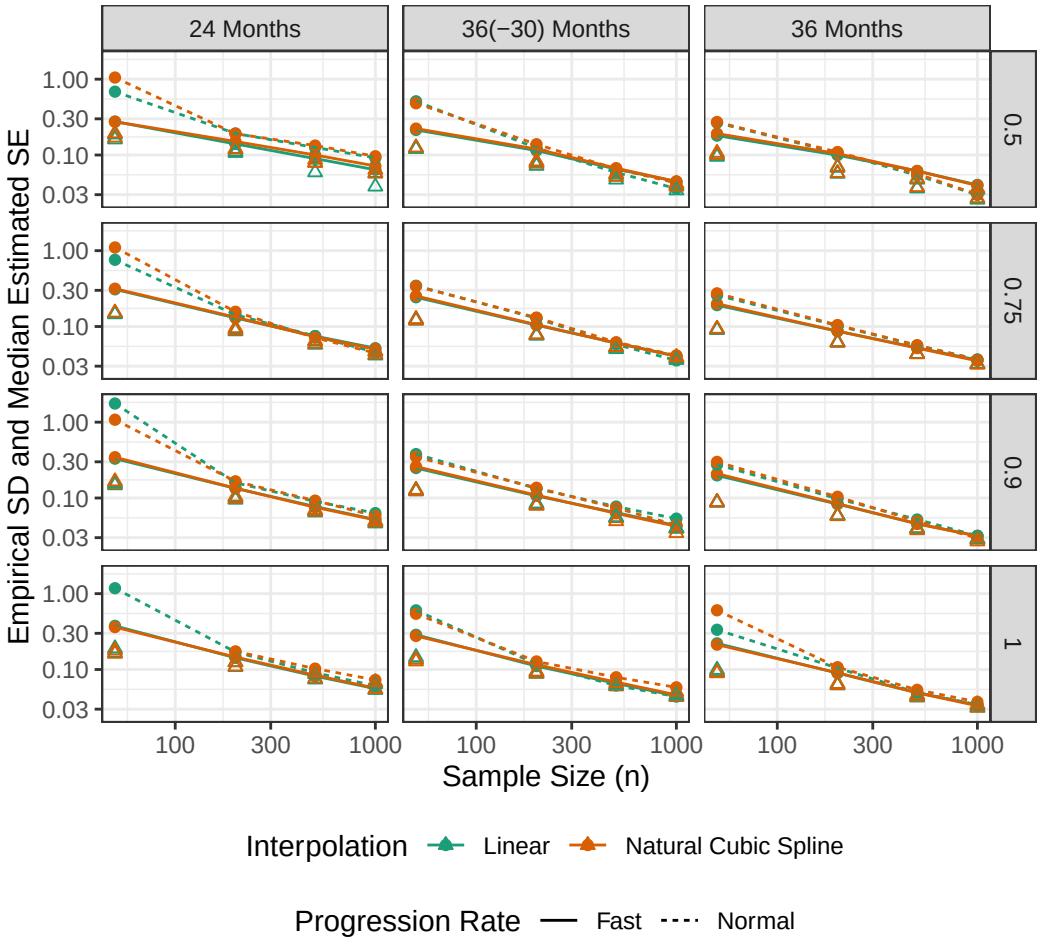

Figure 4: The empirical standard deviations and the median estimated standard errors of the estimator for the common acceleration factor as functions of the sample size. The dots and connecting lines represent the empirical standard deviations. The triangles represent the median estimated standard errors. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factors, the columns correspond to the measurement patterns. Note that both axes are log10-transformed.

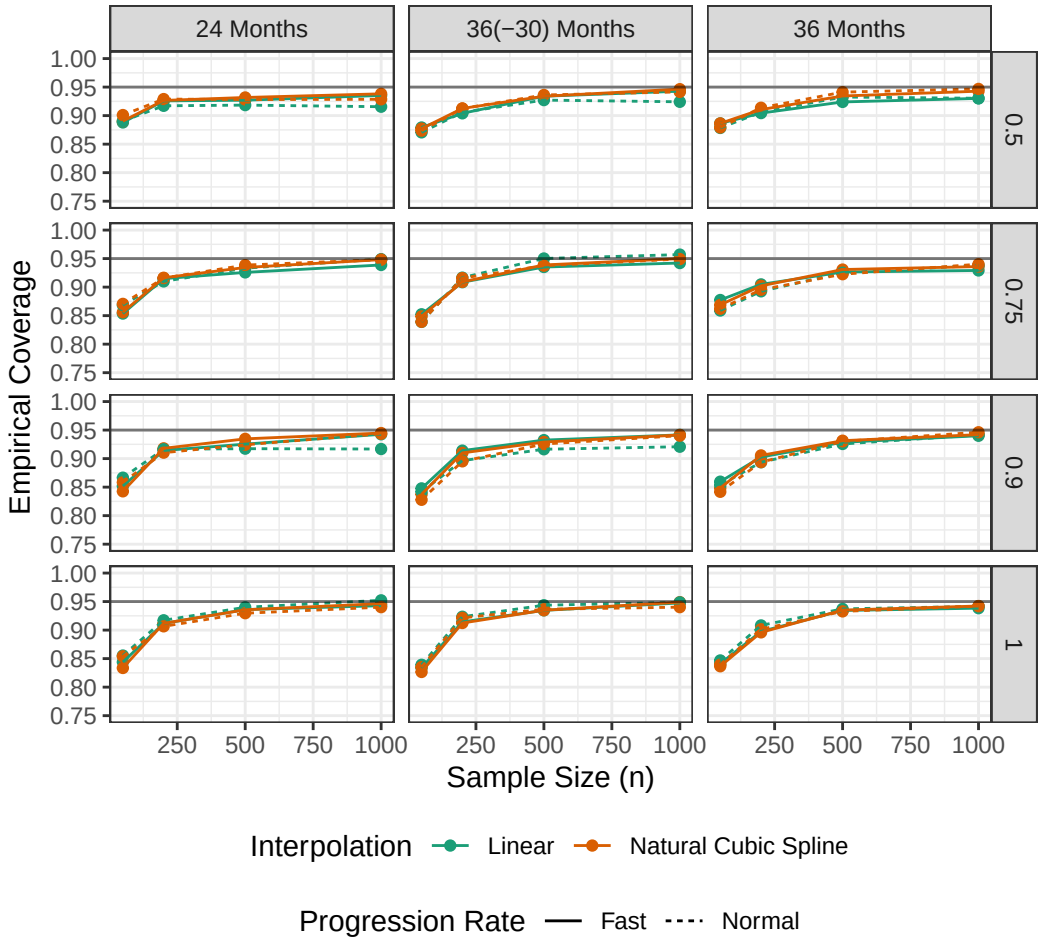

Figure 5: The empirical coverage rates of the confidence intervals based on (12) as functions of the sample size. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factors, the columns correspond to the measurement patterns. The black horizontal lines indicate 95% coverage.

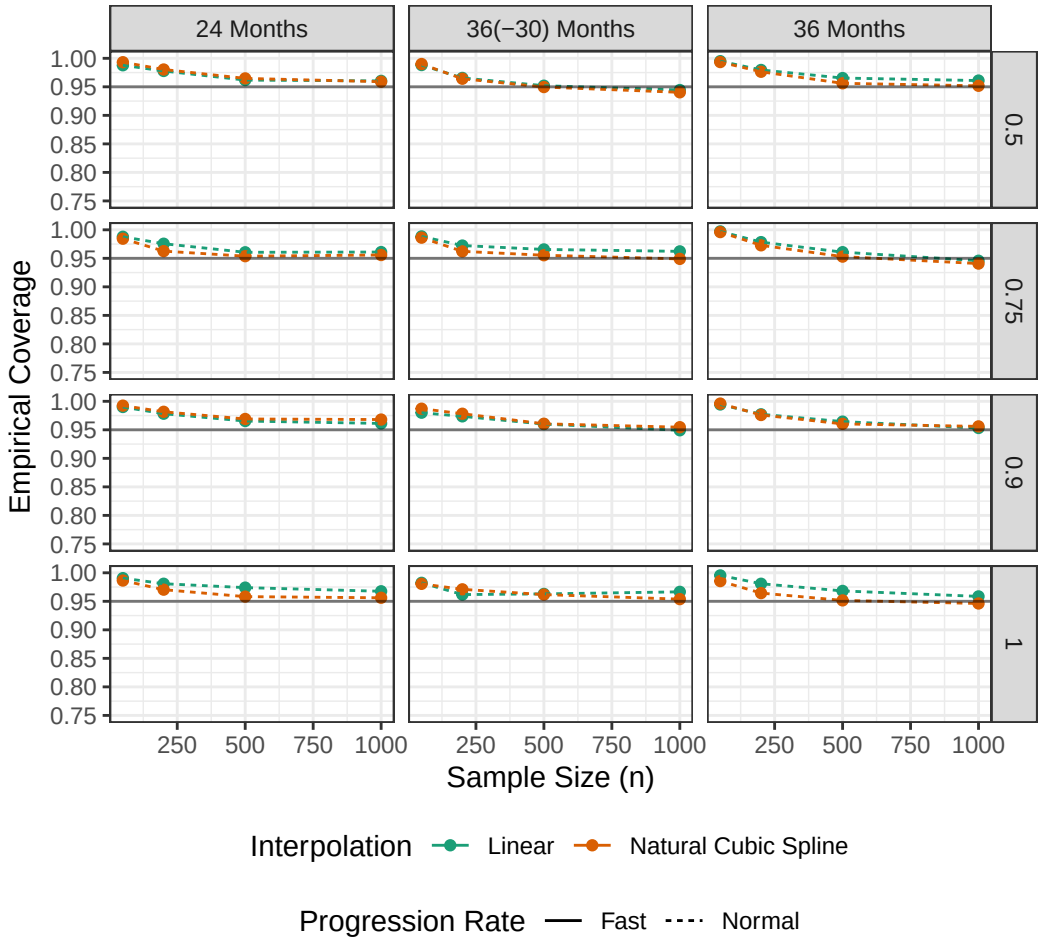

Figure 6: The empirical coverage rates of the percentile confidence intervals based on the parametric bootstrap as functions of the sample size. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factors, the columns correspond to the measurement patterns. The black horizontal lines indicate 95% coverage.

6 Data Application

6.1 MMRM

The data described in Section 2 are usually analyzed with an MMRM, typically including additional baseline covariates. We fitted (using REML) the MMRM described in Section 2.3 to these data where the vector of ADAS-Cog scores at baseline, 12, 24, 40, and 48 weeks after randomization is the response variable, and the covariance matrix is unstructured but common to all patients. We also added an interaction between time and ApoE4 status, and a main effect for the study site. The latter is necessary to maintain the independent observations assumption.

The null hypothesis of no treatment effect is tested using the F-test with the Kenward-Rogers degrees-of-freedom approximation, which results in $p = 0.015$. Thus, there is a significant treatment effect, but it is unclear how relevant this effect really is. A first step towards interpreting the magnitude of the treatment effect is looking at the estimated mean differences in ADAS-Cog scores. Because we have included baseline covariates in the MMRM, we have to carefully define the treatment group-specific mean ADAS-Cog scores at the various measurement occasions. In what follows, we consider the EMMs where the centers and the ApoE4 levels receive weights proportional to their observed frequencies in the data. Alternatively, we could have used equal weights for these factors. Which weights are most appropriate depends on whether these factors were controlled by the researchers and on the population one wishes to draw inferences from.

The EMMs with proportional weights and the corresponding contrasts by measurement occasion are estimated with `emmeans()` and `contrast()` from the *emmeans* R package[20] and are presented in Table 2. These contrasts indicate that the treatment effect increases with time, but it is difficult to appreciate the relevance of these effects: is a 2 or 3 points mean difference on the ADAS-Cog scale meaningful? This question cannot be answered without also taking the duration of treatment and the disease severity into account (as minimal clinically important differences also depend on these factors)[18].

Table 2: Estimated marginal means and treatment contrasts by measurement occasion. The numbers within brackets are 95% confidence intervals based on the t-distribution with Kenward-Roger approximate degrees-of-freedom. The last column contains the t-statistic for testing $H_0 : \alpha_j = \beta_j$.

Measurement Occasion	Active Treatment	Placebo	Contrast	t-statistic (df)
Baseline	-	23.2 $(22.3, 24.2)$	-	
Week 12	23.4 $(22.2, 24.6)$	24.1 $(22.7, 25.5)$	-0.68 $(-1.94, 0.59)$	-1.06 (217)
Week 24	24.5 $(23.1, 25.9)$	26.0 $(24.4, 27.6)$	-1.43 $(-2.91, 0.06)$	-1.89 (208)
Week 40	25.4 $(24.0, 26.9)$	28.2 $(26.5, 29.9)$	-2.76 $(-4.36, -1.16)$	-3.41 (201)
Week 48	26.4 $(24.7, 28.2)$	29.0 $(26.9, 31.0)$	-2.54 $(-4.60, -0.48)$	-2.43 (184)

6.2 Meta TCT

In this section, we use the EMMs from the previous section to estimate the treatment effect on the time scale using the meta-TCT methods implemented in the *TCT* R package.

First, we extract the EMMs at all time-treatment combinations. We also need the corresponding covariance matrix. The `emmeans()` function returns these quantities. Second, we use the `TCT_meta()` function to estimate the time-specific acceleration factors; these estimates are presented in Table 3. It is no coincidence that the t-statistics in Table 2 and the z-statistics in Table 3 are identical; we explained in Section 4.2.3 why that is. The estimated time-specific acceleration factors do not contradict proportional slowing; the estimates are all within one standard error of 0.50. We can, therefore, proceed to estimating the common acceleration factor.

Table 3: Estimated time-specific acceleration factors using the univariate score estimator described in Section 4.2.2. The numbers within brackets are 95% confidence intervals based on (9). The last column contains the z-statistic for testing $H_0 : \gamma_j = 1$.

Measurement Occasion	$\hat{\gamma}_j$	p-value	z-statistic
Week 12	0.32 $(-0.39, 1.37)$	0.291	-1.06
Week 24	0.65 $(0.19, 1.01)$	0.058	-1.89
Week 40	0.52 $(0.29, 0.76)$	0.001	-3.41
Week 48	0.56 $(0.35, 0.83)$	< 0.001	-2.43

We use the adaptive weights-based estimator and the NL-GLS estimator for the common acceleration factor. Because the sample size is relatively small, we further use the parametric bootstrap ($B = 1000$) for inference. The adaptive weights-based estimate (95% CI) is 0.54 $(0.31, 0.79)$, the NL-GLS estimate (95% CI) is 0.54 $(0.36, 0.91)$. Both estimates correspond to slowing disease progression with 50%—a very relevant effect. The p-values and confidence intervals vary a bit more between these estimators and also depend on whether the parametric bootstrap is used. The 95% confidence interval is (0.36, 0.77) based on (12) and (0.32, 0.77) based on the weighted-score test.

7 Discussion and Conclusions

In this paper, we formalized the meta TCT framework to estimate treatment effects on the time scale. We derived asymptotically valid estimators and inferential procedures whose finite-sample properties were studied in a simulation study based on the ADNI data. These methods were further applied to data from a phase 2/3 clinical trial in AD, revealing a significant and clinically relevant treatment effect. In contrast, the analysis with an MMRM revealed a statistically significant effect, but leaves the question of clinical relevance unaddressed.

Raket[4] proposed similar methods, using non-linear models for individual-patient data, to estimate interpretable treatment effects. However, there are two key differences between his methods and the meta-TCT approach. First, the methods of Raket[4] require individual-patient data, limiting the usefulness of this approach for meta-analysis. Second, these methods rely on non-linear regression models that have to be adapted to the data structure and the specific aims of the researcher. Meta TCT is more general: Although we focused on MMRMs as the basis for meta TCT, this approach can be combined—without any modifications—with other methods for longitudinal data such as longitudinal quantile regression[21] or marginal modeling using generalized estimating equations[13].

The simulation study revealed several insights that are relevant for the application of meta TCT in practice. First, inference is difficult for small to medium sample sizes: the estimators may be biased and confidence intervals do not contain the true value at the nominal rate. Raket[4] encountered similar issues with his non-linear regression models for individual-patient data. Our parametric bootstrap solves some of these issues. Second, faster progression and more measurements lead to less bias and more efficient estimators. Third, model misspecification cannot be avoided in practice because meta TCT relies on interpolation between measurement occasions. The only solution here is having the outcome measured at irregular time points, measured very frequently, or both; however, this may not be feasible in clinical trials.

The meta TCT-framework is a promising tool for meta-analysis for two reasons. First, meta TCT only requires summary-level information: the estimated means for each time-treatment combination and the corresponding covariance matrix. This covariance matrix is, however, usually not reported in publications; hence, cooperation from the owners of the data is required as they have to supply this matrix. Second, meta TCT estimates the common acceleration factor, a measure whose scale is shared across trials with different designs (e.g., different timing of measurements or different clinical scales). In addition, the common acceleration factor is more readily interpretable than mean differences on clinical scores, increasing the possible impact of the meta-analysis on policy.

The meta TCT framework can be extended in multiple directions. For instance, a statistical test for proportional slowing could be developed (e.g., as in Wang et al.[22]). Further, although we presented several estimators and inferential procedures, others could be developed as well. For example, inference for NL-GLS meta TCT relies on the minimized generalized least-squares criterion, but a corresponding score test based on the derivative of the least-squares criterion can also be used for inference. The methods can also be extended to multivariate longitudinal data, i.e., multiple clinical scales measured over time. If the acceleration factor is constant across time and clinical scales, power may increase considerably. In addition, similar treatment effects on the time scale across multiple domains is also evidence that the treatment is disease modifying.

To conclude, the meta TCT framework and the understanding of its finite sample properties will facilitate the estimation of interpretable treatment effects as a pre-specified primary or secondary objective of clinical trials in progressive diseases. In turn, this will refine the evaluation of the clinical relevance of new treatments.

Acknowledgments

We thank Lighthouse Pharma for allowing us to use their data in this paper.

The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

Financial disclosure

S. Dickson, C. Mallickrodt, and S. Hendrix are all full time employees of Pentara, and S. Hendrix is owner. No client contributed financially to this work.

Conflict of interest

There are no conflicts of interest to report.

Ethical Approval

Ethical Committee Research UZ Leuven declared that this study is exempt from ethical oversight according to local guidelines.

References

[1] Sperling RA, Karlawish J, Johnson KA. Preclinical Alzheimer disease—the challenges ahead. *Nature Reviews Neurology.* 2013;9(1):54–58.

[2] Van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. *New England Journal of Medicine.* 2023;388(1):9–21.

[3] Hughes CP, Berg L, Danziger W, Coben LA, Martin RL. A new clinical scale for the staging of dementia. *The British journal of psychiatry.* 1982;140(6):566–572.

[4] Raket LL. Progression models for repeated measures: Estimating novel treatment effects in progressive diseases. *Statistics in Medicine.* 2022;41(28):5537-5557.

[5] Landhuis E. Researchers call for a major rethink of how Alzheimer's treatments are evaluated. *Nature.* 2024;627(8003):S18–S20.

[6] Jönsson L, Ivkovic M, Atri A, et al. Progression analysis versus traditional methods to quantify slowing of disease progression in Alzheimer's disease. *Alzheimer's Research & Therapy.* 2024;16(1):48.

[7] Dickson S, Wessels AM, Dowsett S, et al. 'Time saved'as a demonstration of clinical meaningfulness and illustrated using the donanemab TRAILBLAZER-ALZ study findings. *The journal of prevention of Alzheimer's disease.* 2023;10(3):595–599.

[8] Dickson S, Solomon A, Kivipelto M, et al. Evaluation of Clinical Meaningfulness of Fortasyn Connect in Terms of "Time Saved". *The Journal of Prevention of Alzheimer's Disease.* 2024:1–6.

[9] Dickson S, Haaland B, Mallinckrodt C, et al. "Time Saved" Calculations to Improve Decision-Making in Progressive Disease Studies. *The Journal of Prevention of Alzheimer's Disease.* 2024:1–8.

[10] Michael J D. GAIN trial: evidence of target engagement and activity with atuzaginstat in Alzheimer's disease. In: AD/PD Conference 2022.

[11] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease.. *The American journal of psychiatry.* 1984;141(11):1356–1364.

[12] Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal.* 2008;42(4):303–319.

[13] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73(1):13–22.

[14] Searle SR, Speed FM, Milliken GA. Population marginal means in the linear model: an alternative to least squares means. *The American Statistician.* 1980;34(4):216–221.

[15] Cai W. Making comparisons fair: how LS-means unify the analysis of linear models. Tech. Rep. SAS060-2014, SAS Institute Inc.; 600 Research Drive, Cary, NC 27513: 2014.

[16] Sabanes Bove D, Li L, Dedic J, et al. *mmrm: Mixed Models for Repeated Measures.* ; : 2024. R package version 0.3.11.

[17] Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997(3):983–997.

[18] Muir RT, Hill MD, Black SE, Smith EE. Minimal clinically important difference in Alzheimer's disease: Rapid review. *Alzheimer's & Dementia.* 2024;20(5):3352–3363.

[19] Veitch DP, Weiner MW, Aisen PS, et al. Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia.* 2019;15(1):106–152.

[20] Lenth RV. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* ; : 2023. R package version 1.8.8.

[21] Koenker R, Chernozhukov V, He X, Peng L. *Handbook of quantile regression.* Boca Raton, FL: CRC Press. first edition ed., 2017.

[22] Wang G, Liu L, Li Y, et al. Proportional constrained longitudinal data analysis models for clinical trials in sporadic Alzheimer's disease. *Alzheimer's & Dementia: Translational Research & Clinical Interventions.* 2022;8(1):e12286.

[23] Harvey WR. Least-squares analysis of discrete data. *Journal of Animal Science.* 1982;54(5):1067–1071.

[24] The Lancet . Lecanemab for Alzheimer's disease: tempering hype and hope. *The Lancet.* 2022;400(10367):1899.

Supporting information

The *TCT* R package is available from `github.com/florianstijven/TCT`. The code used for the simulations and the data application are available from `github.com/florianstijven/meta-TCT-simulations`. The data described in Section 2 cannot be made available, but the GitHub repo contains a simulated mock dataset based on the case study (and code to analyze these mock data).

The Web Appendix referenced throughout the paper is available online.