# Results Simulation Study

Florian Stijven

2024-01-05

## Contents



# 1 Introduction

In this document, we present the results of the simulation study for assessing the finite sample properties of the meta TCT methods. This simulation study has three goals, ordered by importance:

1. The primary goal is to assess to what degree the theoretical asymptotic results translate to finite samples.
2. The secondary goal is to identify finite sample settings where the meta TCT methods may not be trustworthy.
3. The tertiary goal is to assess the correctness of the implementation of the meta TCT methods in the `TCT` R-package.

All simulated data sets were generated from multivariate normal distributions matching the data generating model (DGM) used by Raket (2022, sec. 5.1). This DGM represents a realistic 36-months clinical trial in prodromal Alzheimer's disease. The control group is based on the analysis of a selection[1] of 556 patients from the Alzheimer's disease neuroimaging initiative (ADNI) (Veitch et al. 2019).

We first present the DGM in more details together with a visualization of the relevant mean trajectories. Second, the analysis methods we consider in this simulation study are summarized. Finally, the simulation results are presented.

# 2 Data Generating Model

For the 556 patients selected from the ADNI, the ADAS-cog scores were[2] available at baseline visits and 6, 12, 18, 24, and 36 months after baseline. The estimated means at the corresponding time points are

$$(19.6, 20.5, 20.9, 22.7, 23.8, 27.4)'.$$

These data represent follow-up with non-equally spaced measurement occasions and a "normal" progression rate. To also allow for equally-spaced measurement occasions in our DGM - with possibly faster progression - we start from the following two mean profiles.

Table 1: Mean profiles. The means under normal progression are based on a selection of patients from the ADNI, except for the mean at 30 months. The means under fast progression are not based on external data.

| Progression | Baseline | 6 Months | 12 Months | 18 Months | 24 Months | 30 Months | 36 months |
|---|---|---|---|---|---|---|---|
| Normal | 19.6 | 20.5 | 20.9 | 22.7 | 23.8 | 25.8 | 27.4 |
| Fast | 18.0 | 19.7 | 20.9 | 22.7 | 24.7 | 27.1 | 29.2 |

Besides different progression rates, we consider the following three measurement patterns:

1. "24 Months". One follow-up visit every 6 months until 24 months after randomization.
2. "36 Months". One follow-up visit every 6 months until 36 months after randomization.
3. "36(-30) months". Same pattern as "36 Months", but leaving out the measurement at 30 months.

The covariance matrix for the "36 Months" pattern is

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

[1] The following inclusion criteria were used by Raket (2022): "at the baseline visit, patients must be diagnosed as having mild cognitive impairment, score less than or equal to 28 on the mini mental state examination (MMSE, range 30-0, higher scores indicate less impairment) and be amyloid positive according to a brain positron emission tomography scan or analysis of cerebrospinal fluid."

[2] 13-item cognitive subscale of the Alzheimer's disease assessment scale, lower scores indicate less impairment (Rosen, Mohs, and Davis 1984)

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

Patient-level data in the control group are generated from a multivariate normal distribution with the above mean vector and covariance matrix, as in Raket (2022, sec. 5.1). Whereas Raket (2022) considered multiple types of treatment effects, we only consider treatment effects that correspond to proportional slowing.

To simulate proportional slowing, we first consider the *reference trajectory*. Let $Y_t$ be the ADAS-cog score $t$ months after randomization and $Z = 0$ denote the control group. The *reference trajectory* is then the mean ADAS-cog score in the control group as a function of time since randomization,

$$E(Y_t|Z = 0) = f_0(t; \boldsymbol{\alpha})$$

where $\boldsymbol{\alpha}$ is a parameter vector indexing the reference trajectory. This trajectory should be a continuous function over the relevant range. In our DGM, this is $[0, 36]$ which is the total duration of the ADNI study. In our control group, we only know the mean at 7 distinct time points. Therefore, we interpolate between these 7 points with natural cubic spline interpolation. This *interpolated* reference trajectory is plotted in Figure 1.

For simulating 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 $\gamma$ is the *acceleration factor*. These data are simulated from a multivariate normal distribution with the means determined by the above trajectory function and with the same covariance matrix as in the control group.

We consider a set of DGMs where the following elements are varied to represent a range of realistic scenarios:

- **Progression Rate**. We consider a *normal* and *fast* progression rate. The corresponding mean vectors are presented in the above table. Alternatively, the fast progression scenario could also be interpreted as corresponding to trials where patients have been followed up longer.
- **Treatment Effect**. We consider 4 different (proportional slowing) treatment effects, that is, $\gamma \in \{1, 0.90, 0.75, 0.50\}$.
- **Sample Size**. We consider 4 different total sample sizes, $n \in \{50, 200, 500, 1000\}$. Note that $n$ is the *total* sample size, and we assume $1 : 1$ randomization in all settings.
- **Measurement Pattern**. We consider the three scenarios that were explained before: "24 Months", "36 Months", and "36(-30) Months".

# 3 Analysis Methods

In this section, two analysis methods are briefly outlined. We first present the mixed model for repeated measures (MMRM) that is fitted to the simulated data sets. Next, we discuss the non-linear generalized least squares (NL-GLS) version of meta TCT. While we consider other versions of meta TCT as well, we will focus on the NL-GLS version.

## 3.1 Mixed Model for Repeated Measures

For each simulated data set, a MMRM is fitted. This is a linear mixed model where time is treated as a categorical covariate. The **systematic part** consists of the interaction between treatment and time, except for $t = 0$ where we assume that the mean outcome is equal in both treatment groups. Let $j = 0, 1, ..., K$

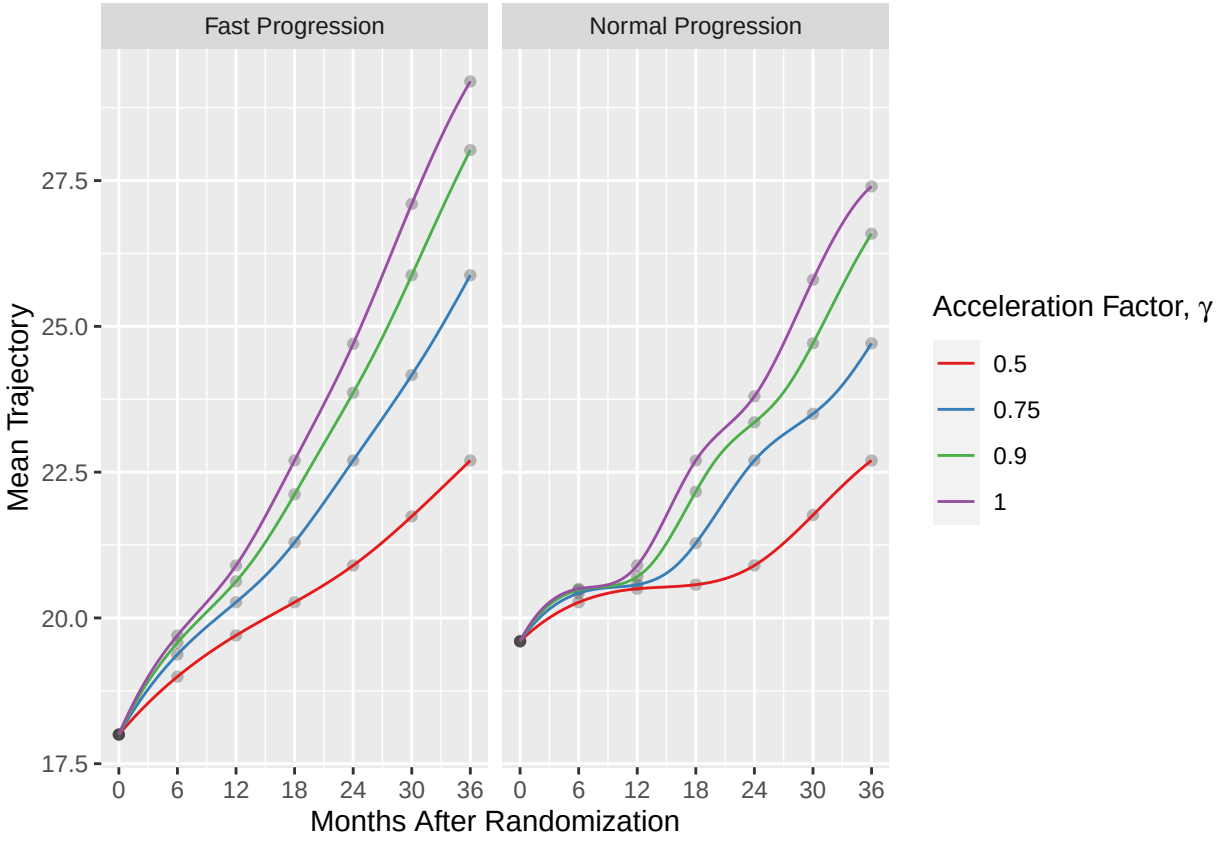


Figure 1: Plot of the trajectories used in the DGMs. An acceleration factor equal to 1 corresponds to no treatment effect. The dots correspond to measurement occasions while the lines between dots are based on interpolation.

denote the measurement occasions, and $t_j$ the corresponding months after baseline. The mean outcome is then modeled as

$$E(Y_{t_j}|Z=z) = \begin{cases} \alpha_j \text{ if } z=0 \\ \beta_j \text{ if } z=1 \end{cases} \quad \forall\, j \in \{1,...,K\}$$

and $E(Y_0|Z=0) = E(Y_0|Z=1) = \alpha_0$. This essentially means that we have a mean parameter for each measurement occasion-treatment combination, except for baseline. The **covariance matrix** is assumed to be unstructured, but common for both treatment groups. For all simulation scenarios, this is a correctly specified model.

These models are fitted with restricted maximum likelihood (REML) by the `mmrm()` function from the `mmrm` R-package (Sabanes Bove et al. 2022). The parameter estimates and corresponding variance-covariance matrix are obtained by the `coef()` and `vcov()` methods, respectively. These two components are the only inputs required for the meta TCT methods. The hypothesis test for

$$H_0 : \alpha_j = \beta_j \;\forall\, j \in \{1,...,K\}$$

in the MMRM is based on the F-test with the Kenward-Roger approximate degrees of freedom.

## 3.2 Meta TCT

The meta TCT methodology starts from the assumption that the estimator for the mean vector has a multivariate normal sampling distribution,

$$(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')' = (\hat{\alpha}_0, ..., \hat{\alpha}_K, \hat{\beta}_1, ..., \hat{\beta}_K)' \sim N((\boldsymbol{\alpha_0}', \boldsymbol{\beta_0}')', D).$$

In principle, we can replace the mean with any other functional of the distribution such as the median, as long as the sampling distribution of the estimator is (approximately) multivariate normal.

In the NL-GLS version of meta TCT, we treat this estimated vector, $(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')$, as the "observed data" and model these data with the following non-linear model,

$$\begin{aligned} E\left((\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')\right) &= (\boldsymbol{f_0}(\boldsymbol{t};\boldsymbol{\alpha})', \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t};\boldsymbol{\alpha})') \\ &= (\boldsymbol{\alpha}', \boldsymbol{f_0}(\gamma \cdot \boldsymbol{t};\boldsymbol{\alpha})') \end{aligned}$$

where $\boldsymbol{f_0}(\gamma \cdot \boldsymbol{t};\boldsymbol{\alpha}) = (f_0(\gamma \cdot t_1;\boldsymbol{\alpha}), ..., f_0(\gamma \cdot t_K;\boldsymbol{\alpha}))'$. The second equality is a consequence of $f_0$ being an interpolating function. While this function is a natural cubic splines interpolation in the DGM, we consider natural cubic spline and linear interpolation for estimation.

The non-linear regression function is fitted using non-linear generalized least squares. Hence, the estimator for $(\boldsymbol{\alpha_0}', \gamma_0)'$ minimizes the generalized least squares criterion,

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

In practice, we replace $D$ with its estimate, $\hat{D}$. As part of this procedure, $\boldsymbol{\alpha_0}$ is re-estimated by $\hat{\boldsymbol{\alpha}}_{NL}$. However, we do not use $\hat{\boldsymbol{\alpha}}_{NL}$ further on.

Inference is based on the generalized least-squares criterion. Let $l_1$ be the minimized generalized least-squares criterion under no restrictions, and $l_0$ the criterion under the restriction that $\gamma = \gamma^*$. It then follows that $l_1 - l_0 \overset{\sim}{\sim} \chi^2_1$ under the null that $\gamma_0 = \gamma^*$. Note that this result is asymptotic with respect to $n \to \infty$ while the length of $(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')$ remains constant.

The standard errors for the estimates from the above approach can be obtained analytically (by linear approximation) or through a parametric bootstrap. In the remainder of this report, the standard error is estimated analytically unless mentioned otherwise.

# 4 Results

In this section, the results of the simulation study are presented. For each setting, we consider 5000 replications. For estimating the empirical coverage of the 95% CIs and the empirical type 1 errors, this leads to a standard error of $\frac{\sqrt{0.05 \cdot 0.95}}{\sqrt{5000}} = 0.003$ assuming nominal coverage and a nominal type 1 error. This is sufficiently precise for the purposes of this simulation study.

In this section, we only present the results for the NL-GLS version of meta TCT. In the Appendix, the results for other estimators are presented in an interactive table.

The analyses and the presentation of the simulation results are divided into two parts. First, we present the results regarding estimation of the acceleration factor. This is the primary goal of the meta TCT methods: Transforming the treatment effect on a difficult to interpret scale to the time scale. Second, we present the results regarding inference, that is, hypothesis tests and confidence intervals. The operating characteristics of the meta TCT methods are also compared with those of the MMRMs.

## 4.1 Estimation

To evaluate estimation, we look at 3 key performance measures,

1. **Bias of the estimator.** We estimate $E(\hat{\gamma})$ across different settings and compare the estimated expectation with the true value, $\gamma_0$.
2. **Mean squared error (MSE).** The MSE is defined as $E\{(\hat{\gamma} - \gamma_0)^2\}$. This quantity is estimated as the mean of the squared differences between $\hat{\gamma}$ and $\gamma_0$. This quantity quantifies the average distance between the estimator and the estimand, which depends on the variance and bias.
3. **Empirical standard deviation.** The empirical standard deviation of the estimator simply is the standard deviation of the estimator. This measure is estimated as the sample standard deviation of the estimates in each setting. This value is compared with the median *estimated standard error*.

In Figure 2, the mean of the estimated acceleration factors is presented across a set of scenarios. For cubic interpolation, the bias decreases to almost 0 as the total sample size increases to 1000 in all but one scenario. This does not hold up only for the scenario with $\gamma = 0.90$, normal progression and 36(-30) months of follow up. A possible explanation is that - looking at Figure 1 - the estimated mean at 36 months in the experimental group is mapped into a time interval with no measurement occasions nearby; taking into account that there is no measurement at 30 months in this scenario. This time mapping is therefore rather sensitive to the interpolation method. This is corroborated by the fact that there is no bias in the corresponding scenarios with 24 or 36 months of follow up.

The estimator with linear interpolation exhibits in some settings a bias that does not decrease to zero. A small amount of bias is expected here because this estimator uses a slightly misspecified model. Indeed, the data are generated under cubic spline interpolation instead of linear interpolation.

In any case, the bias is generally smaller than 0.05 when we use the "correct" interpolation method: cubic spline interpolation. However, the correct interpolation method is in principle not identifiable. Using cubic spline interpolation, there is only one scenario with a bias larger than 0.05: $\gamma = 0.5$, normal progression, and 24 months of follow up.

Figure 2 also shows that, when there is bias, the bias is smaller for the fast progression rate as compared to the normal progression rate. The bias also tends to be smaller when there are more measurement occasions.

In Figure 3, the MSE of the estimator for the common acceleration factor is presented across the same set of scenarios as before. As expected, the MSE decreases as a function of the sample size. Moreover, a longer follow up leads to a smaller MSE. In small sample sizes, the MSE is considerably smaller for fast progression than for normal progression.

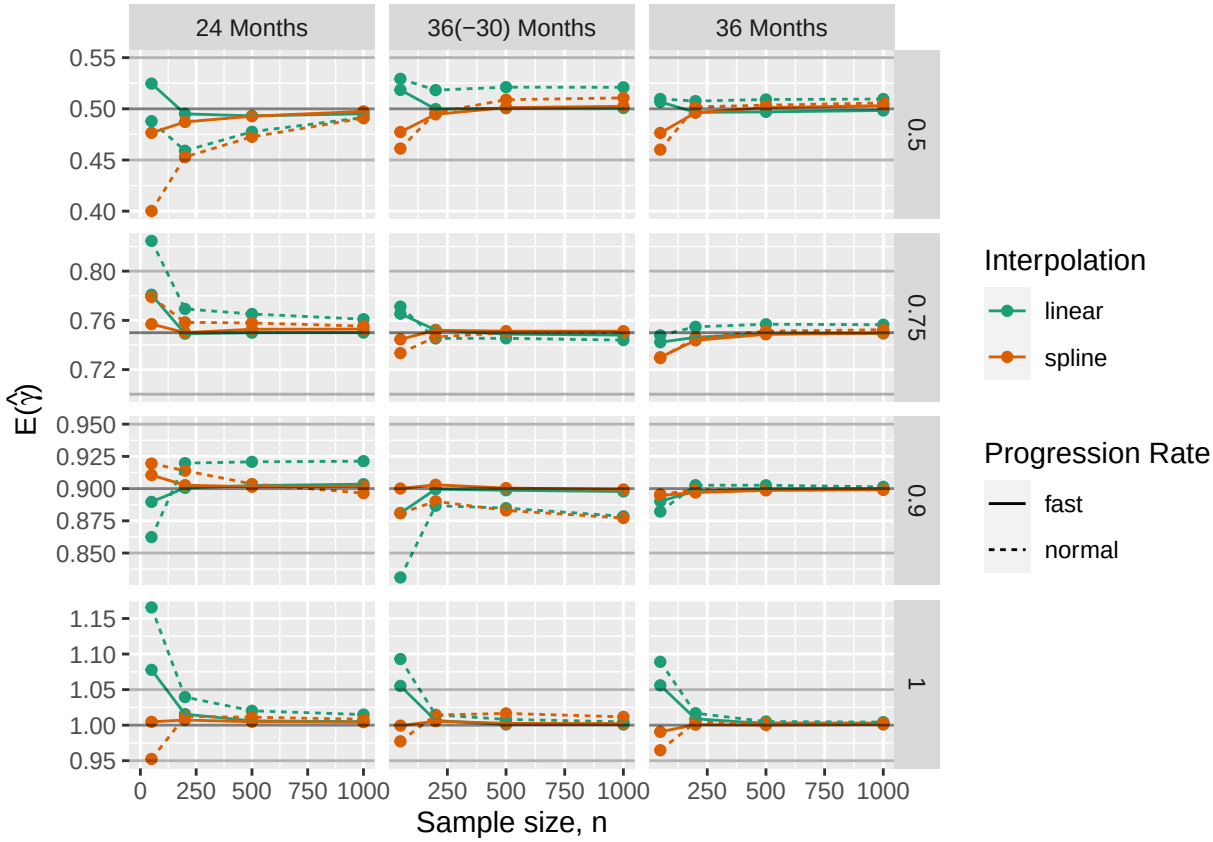


Figure 2: Graphs of the means of the estimated acceleration factors across a set of simulation settings. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns. In each subplot, a 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

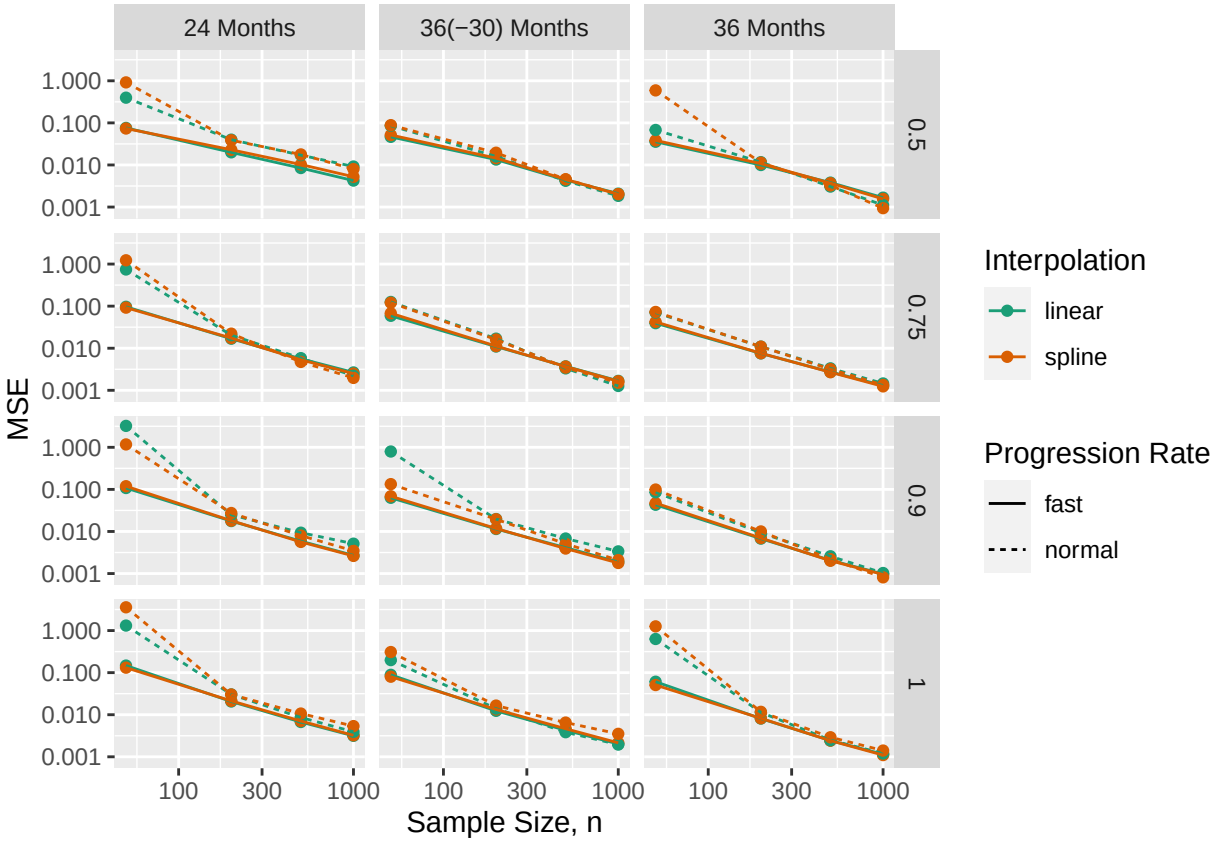


Figure 3: Graphs of the MSEs for the estimator of the common accleration factor across a set of simulation settings. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns. Note that both axes are log10-transformed.

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

Although inference is not based directly on the estimated standard error, the observed *underestimation* for smaller sample sizes would lead to issues if the estimated acceleration factor were used in meta-analyses. We have therefore also implemented a bootstrap-based estimator for the standard error, as presented in the Section 5.

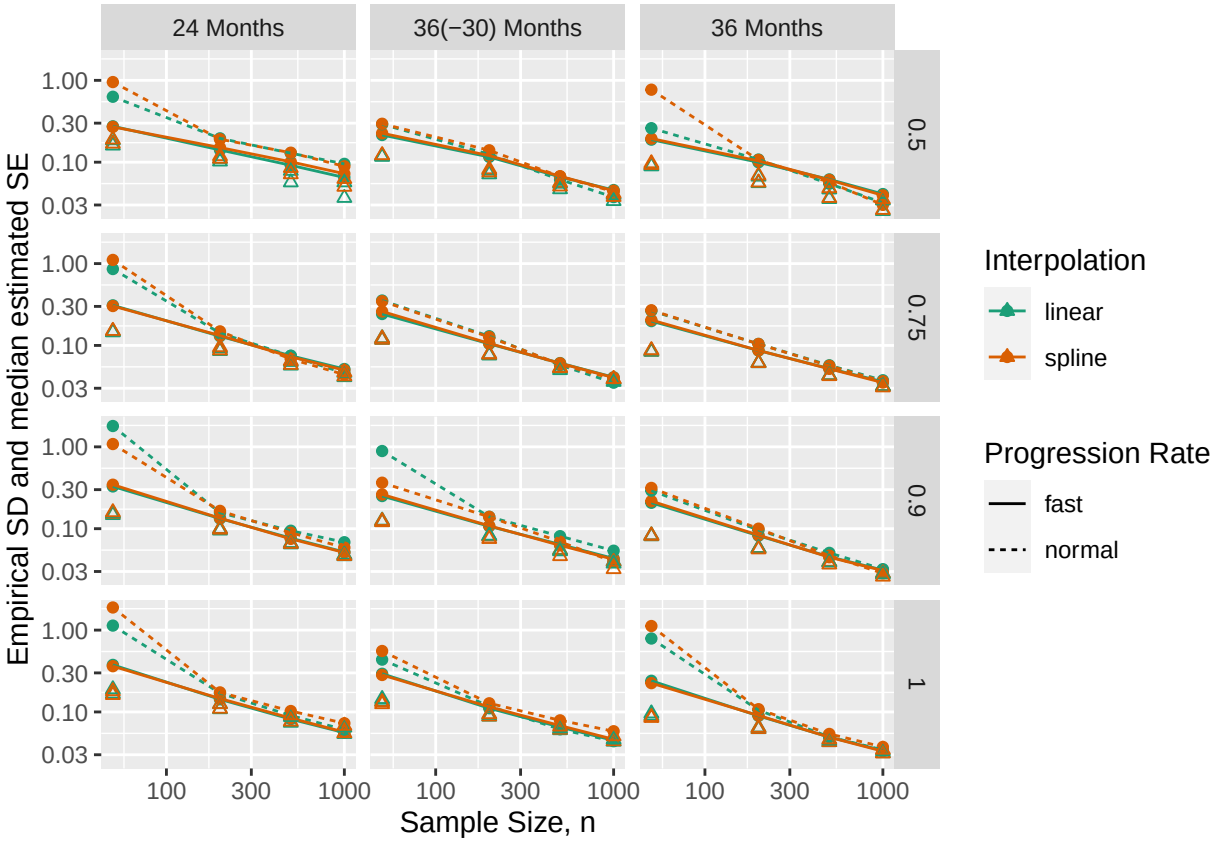


Figure 4: Graphs of the empirical standard deviations and the median estimated standard errors of the estimator for the common acceleration factor across a set of simulation settings. The dots, and connecting lines, represent the empirical standard deviations and the triangles represent the median estimated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

## 4.2 Inference

To evaluate inference, we look at 2 key performance measures,

1. **Type 1 error and power.** We compare the empirical type 1 error with the nominal rate, $\alpha = 0.05$. We do this for the meta TCT methods as well as for the MMRMs. The former is based on the latter, so we also expect that discrepancies between empirical and nominal type 1 error rates for the MMRM will be reflected in the meta TCT methods.

2. **Coverage.** We asses the empirical coverage rate of the estimated 95% CIs.

In Figure 5, we graph the empirical type 1 error rate and power. The corresponding empirical operating characteristics for the F-test in the MMRMs are superimposed in gray. This reveals that the type 1 error rate for meta TCT is inflated for the sample sizes between 50 and 500. Consequently, the power for the corresponding settings is not well-calibrated. For the settings with $n = 1000$, the empirical type 1 error is close to nominal, hence, the corresponding empirical powers are well-calibrated and can be interpreted as usual. In these scenarios, the power of the meta TCT test is larger than or equal to the power of the corresponding F-test in the MMRM. A longer follow up also leads to a larger power, which is consistent with the previous results.

The tests presented in Figure 5 are asymptotic. So, poor performance in small samples is not surprising. To provide robust inference for small samples, we also assess the performance of a parametric bootstrap in the next section.

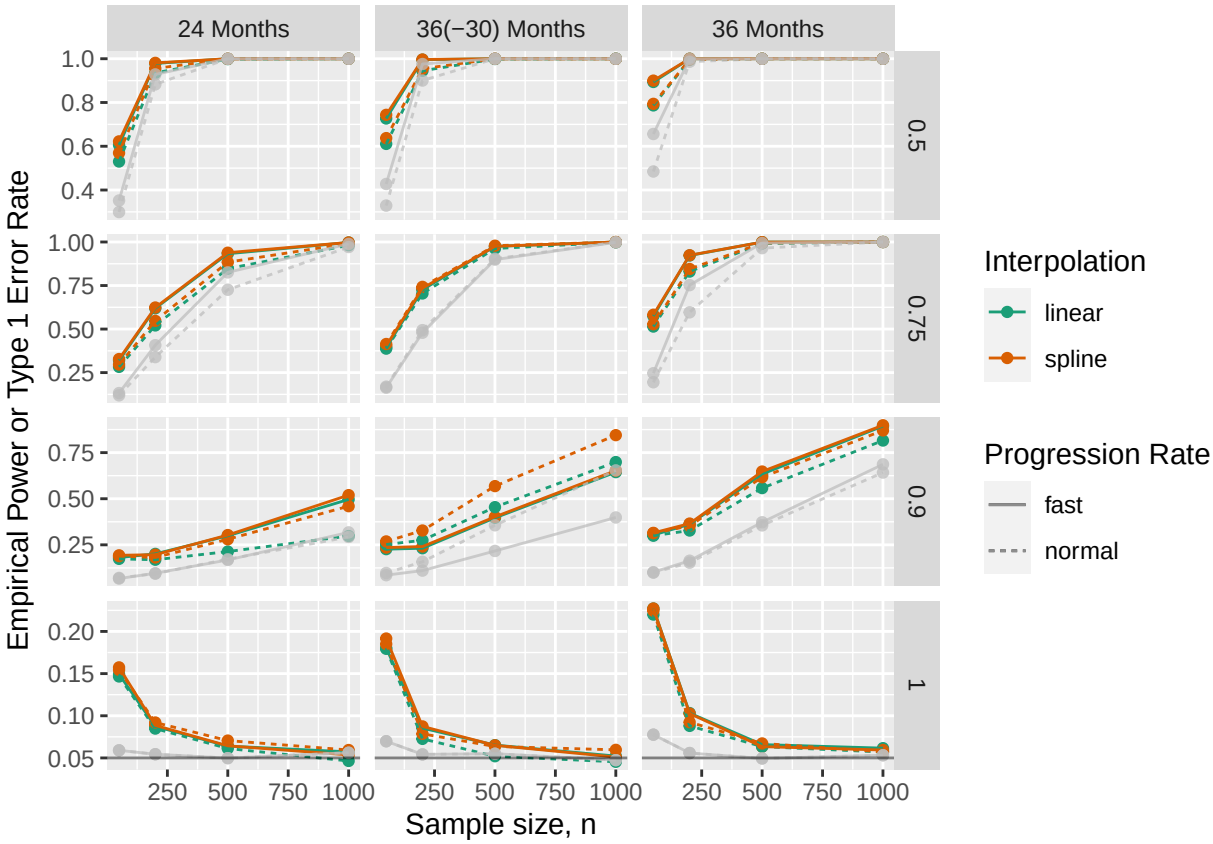


Figure 5: Graphs of the empirical type 1 error rate and power across a set of simulation settings. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns.

In Figure 6, the empirical coverage rates are presented for the same settings as before. This reveals that there is undercoverage for the sample sizes between 50 and 500. This undercoverage decreases with an increasing sample size, but does not disappear completely for some scenarios with a sample size of 1000.

There is one scenario where the coverage does not converge to 95% for an increasing sample size: the scenario with $\gamma = 0.90$, normal progression and 36(-30) months of follow up. This is also the scenario where the bias was not zero for a sample size of 1000. The explanation for this bias also explains the observed undercoverage.

As mentioned before, the coverage of alternative confidence intervals based on a parametric bootstrap is assessed in the next section.

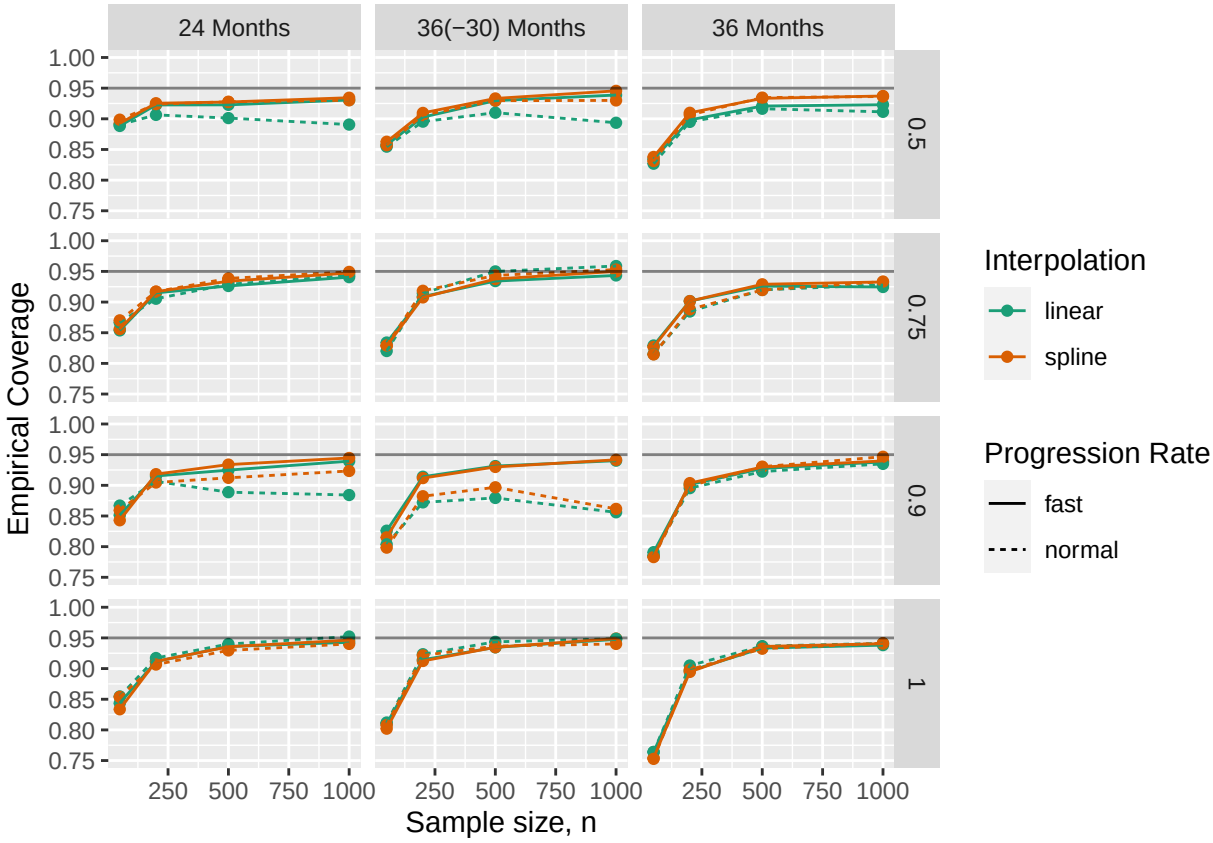


Figure 6: Graphs of the empirical coverage across a set of simulation settings. The presented results are based on the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns.

## 5 Parametric Bootstrap

In small samples, the operating characteristics of the NL-GLS meta TCT estimator deviate considerably from nominal. This is not surprising because the corresponding inferential procedures are only asymptotic. We provide an alternative inferential procedure based on a parametric bootstrap. The idea behind this parametric bootstrap is to resample the estimated parameters from the estimated multivariate normal sampling distribution, that is,

$$(\hat{\boldsymbol{\alpha}}^b, \hat{\boldsymbol{\beta}}^b)' \sim N\left((\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')', \hat{D}\right)$$

where

- $(\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')'$ is the estimated mean vector.
- $\hat{D}$ is the estimated variance-covariance matrix of the sampling distribution.
- $(\hat{\boldsymbol{\alpha}}^b, \hat{\boldsymbol{\beta}}^b)'$ 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 NL-GLS meta

TCT to $(\hat{\boldsymbol{\alpha}}^b, \hat{\boldsymbol{\beta}}^b)'$ and $\hat{D}$. Subsequent inference is based on the $1-\alpha$ percentile confidence interval, that is,

$$(\hat{\gamma}^b_{\alpha/2}, \hat{\gamma}^b_{1-\alpha/2})$$

where $\hat{\gamma}^b_p$ is the $p$-th percentile of the bootstrap distribution.

To limit the computational burden, we only use $B = 500$ bootstrap replications throughout. Also, we only consider the bootstrap for the normal progression scenarios because the largest deviations from nominal were observed there.

In Figure 7, the empirical standard deviations are plotted together with the median estimated standard errors that are based on the parametric bootstrap. The median estimated standard errors now closely match the empirical standard deviation. Only for a very small sample size, $n = 50$, is there some overestimation of the empirical standard deviation. So, the parametric bootstrap provides a standard error estimator that is valid even in small samples. However, some caution is warranted in very small sample sizes.

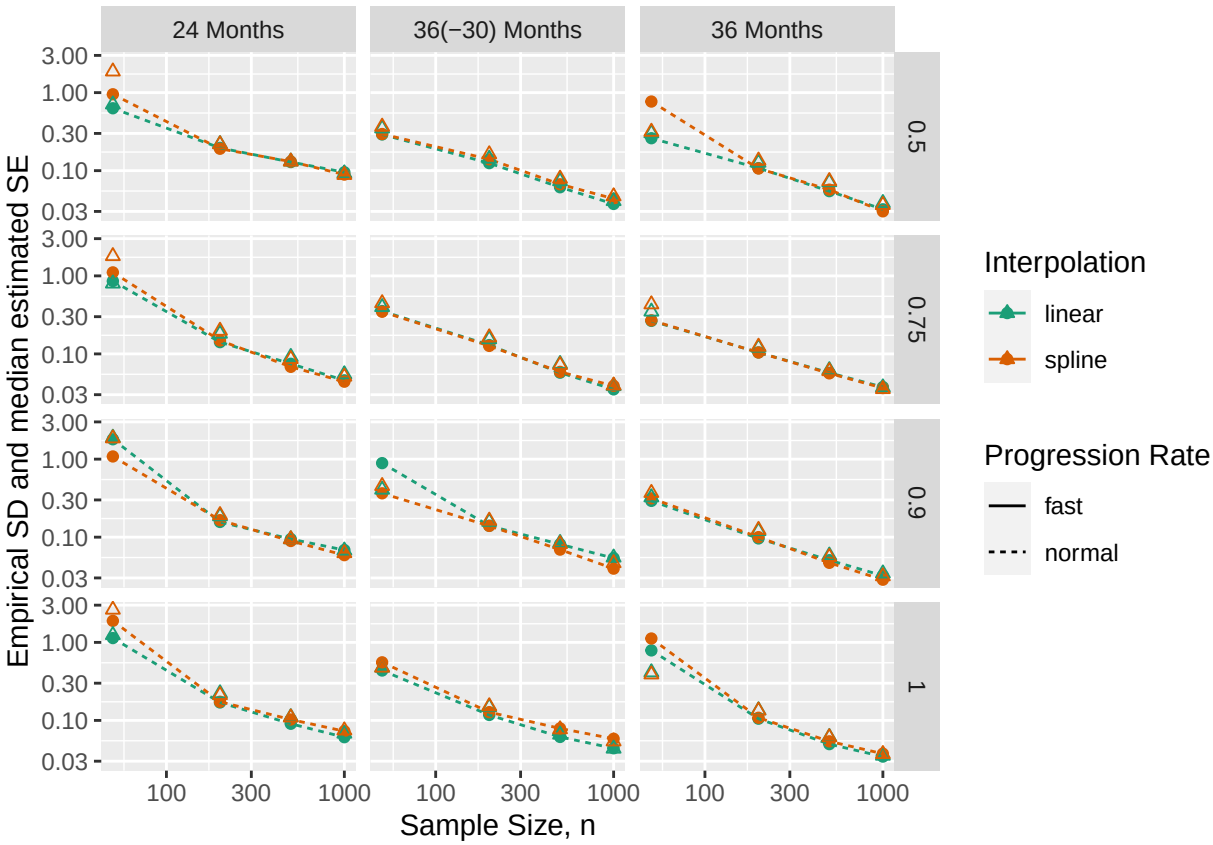


Figure 7: Graphs of the empirical standard deviations and the median estimated standard errors of the estimator for the common acceleration factor. The standard errors are estimated thrrough the parametric bootstrap as explained in the text. The dots, and connecting lines, represent the empirical standard deviations and 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 factor while the columns correspond to the measurement patterns. Note that both axes are log10-transformed.

In Figure 8, the empirical coverage rates are presented for the percentile confidence intervals. This reveals that there is overcoverage which disappears with an increasing sample size. For $n = 1000$, coverage is (close to) nominal, except in the scenario with $\gamma = 0.90$, normal progression and 36(-30) months of follow up. This is, again, the same scenario where we observed bias and non-nominal coverage for the standard CIs. So, the same explanation for the observed deviation from nominal holds here.

These results thus indicate that the parametric bootstrap permits valid, but generally conservative, inference in small samples.

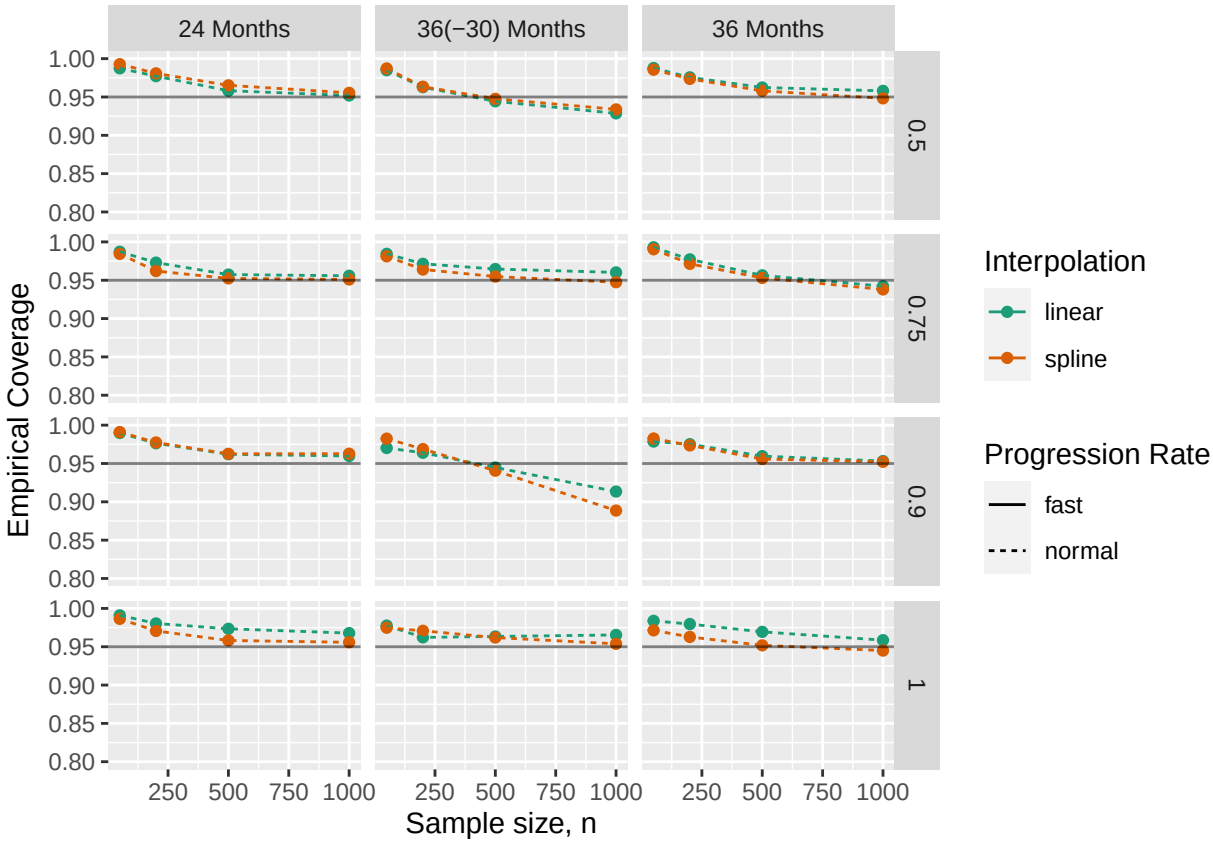


Figure 8: Graphs of the empirical coverage of the percentile bootstrap confidence intervals for the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns.

Finally, the empirical type 1 error rates and power for the parametric bootstrap are presented in Figure 9. The corresponding hypothesis tests are based on the percentile confidence intervals. First, the empirical type 1 error rate tends be be conservative for all sample sizes, but converges to nominal for an increasing sample size. Second, the empirical power for meta TCT is generally similar to the power of the F-tests in the MMRMs. However, there two settings where the meta TCT is more powerful:

1. $\gamma = 0.90$ and 36 months follow-up
2. $\gamma = 0.50$ and $n = 50$.

# 6 Summary

The results are summarized according to the goals of the simulation study. We only looked at the results with the NL-GLS version of meta TCT. A detailed analysis of the results with other versions of meta TCT is outside the scope of this report.

*Primary goal: Evaluate finite sample properties of the method, also in relation to asymptotic inferential procedures.*

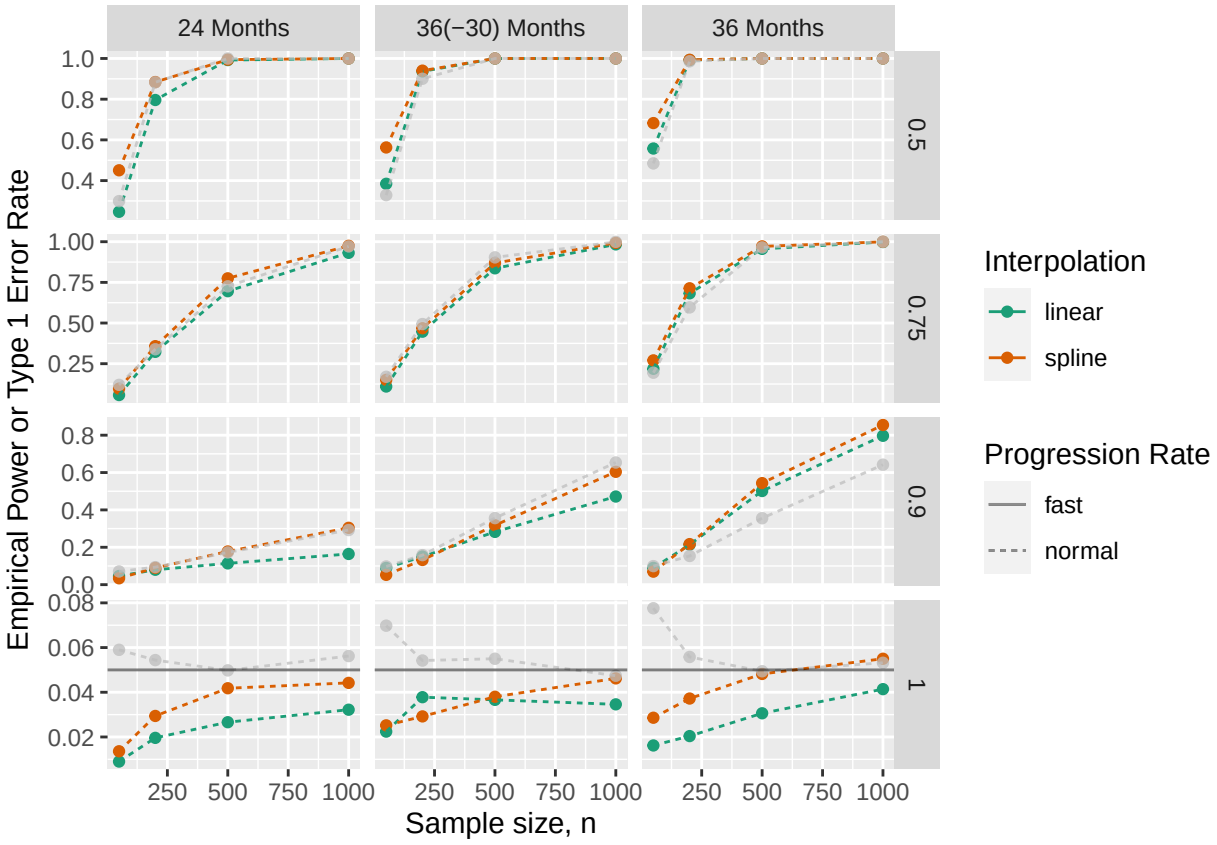


Figure 9: Graphs of the empirical type 1 error rate and power across a set of simulation settings. The hypothesis tests are based on the percentile confidence intervals for the NL-GLS version of meta TCT. The rows correspond to the true acceleration factor while the columns correspond to the measurement patterns.

- **Bias**
  - For small sample sizes, the estimator for the acceleration factor is in some settings slightly biased. However, this bias is generally within 0.05 units of zero and generally goes to zero as the sample size increases to 1000.
  - If there are time mappings to time points which do not have measurement occasions nearby, the estimator could remain biased, even for a sample size of 1000. This is explained by the fact that such time mappings heavily rely on the interpolation approach. The interpolated function may indeed be unstable for time points with no observed points nearby.
    - In these settings, similar issues are observed with the empirical coverage of the confidence intervals.
  - The "wrong" interpolation method can lead to a small bias, even in large samples. The results show that, when the data are generated with cubic spline interpolation, but the meta TCT estimator uses linear interpolation, there can be a small amount of bias that that does not disappear with an increasing sample size.
  - The bias also generally decreases with
    - an increasing sample size,
    - an increasing duration of follow up,
    - and a faster progression rate.
- **MSE**
  - The MSE decreases with
    - an increasing sample size,
    - an increasing duration of follow up,
    - and a faster progression rate (in small samples).
- **Estimated SE**
  - The analytic SE estimator underestimates the true SE for the small sample sizes. However, this underestimation largely disappears for a sample size of 1000.
  - The parametric bootstrap SE estimator performs better than the analytic estimator. Only with a sample size of 50 are there some settings where the true SE is overestimated.
    - The availability of an estimator for the SE that is generally valid and applicable is important if the results of the meta TCT analysis were to be used in meta-analyses.
- **Confidence Intervals**
  - There is undercoverage for the confidence intervals based on the least-squares criterion in the smaller sample sizes. This undercoverage largely disappears for a sample size of 1000.
  - There is overcoverage for the confidence intervals based on the parametric bootstrap in the smaller sample sizes. This overcoverage largely disappears for a sample size of 1000.
- **Type 1 Error and Power**
  - For the smaller sample sizes, the hypothesis test based on the least-squares criterion has an inflated type 1 error, whereas the tests based on the parametric bootstrap are conservative.
    - Tests based on the parametric bootstrap should thus be preferred for small sample sizes.
  - For a sample size of 1000, the least-squares and parametric bootstrap tests have an empirical type 1 error that is close to nominal.
    - For this sample size, the power of the least-squares test is larger than the power of the bootstrap test and the F-test in the MMRM.
  - There are two general settings where the bootstrap test outperforms the F-test in the MMRM.
    - Small sample size but large treatment effect.
    - Small treatment effect and a large number of measurement occasions.

*Secondary goal: Identify settings where the meta TCT methods may yield unstable results.*

- **Sample size**
  - The meta TCT method is generally less stable in the scenarios with a sample size of 50. However, for a sample size of 200 or larger, the method is quite stable.
    * For inference in small samples, the parametric bootstrap method should be preferred.
    * Stability also depends on many other factors besides the sample size. For instance,
      · Whether the assumption of a multivariate normal sampling distribution holds for the given sample size. Most methods for analyzing longitudinal data only lead to a multivariate normal sampling distribution in an asymptotic sense. So, one should always be skeptical of this assumption. The parametric bootstrap also relies on this normality assumption and will thus not "solve" violations of this assumption.
- **Measurement Pattern**
  - A longer follow up generally leads to more stable results.
  - There is, however, an important caveat. If there are time mappings to points in time with no measurement occasions nearby, the methods may become *less stable*. For instance, if there are regular measurements in the first two years of the study, one should be careful with including a measurement 5 years post-randomization. If this measurement maps to, e.g., 3.5 years, this mapping will be very sensitive to the interpolation method and could lead to unstable results.
- **Progression rate**
  - A faster progression rate leads to more stable results. This is expected because faster progression translates to a steeper reference trajectory. It is furthermore easier to map estimated means to a steeper reference trajectory.

*Tertiary Goal: Evaluate correctness of the implementation in the `TCT` R-package.*

- The functions implemented in `TCT` behaved as expected. For the NL-GLS version of meta TCT, we analyzed 480,000 simulated data sets without any failures. We can thus be confident that the methods implemented in the `TCT` R-package will not fail in similar data sets.

# 7 Appendix

## 7.1 All Simulation Settings

In the following table, we summarize the results for all simulation settings. In the html version of the file, it is possible the interactively search through and filter this table. In the pdf version of the file, we did not include the table.

# References


Raket, Lars Lau. 2022. "Progression Models for Repeated Measures: Estimating Novel Treatment Effects in Progressive Diseases." *Statistics in Medicine* 41 (28): 5537–57. https://doi.org/10.1002/sim.9581.

Rosen, Wilma G, Richard C Mohs, and Kenneth L Davis. 1984. "A New Rating Scale for Alzheimer's Disease." *The American Journal of Psychiatry* 141 (11): 1356–64.

Sabanes Bove, Daniel, Julia Dedic, Doug Kelkhoff, Kevin Kunzmann, Brian Matthew Lang, Liming Li, and Ya Wang. 2022. *Mmrm: Mixed Models for Repeated Measures*. https://openpharma.github.io/mmrm/.

Veitch, Dallas P, Michael W Weiner, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, et al. 2019. "Understanding Disease Progression and Improving Alzheimer's Disease Clinical Trials: Recent Highlights from the Alzheimer's Disease Neuroimaging Initiative." *Alzheimer's & Dementia* 15 (1): 106–52.