Statistical Analysis Plan: Quality of Life in the LIQPLAT Trial

Johannes Schwenke

2025-08-06

Table of contents

1 Abbreviations

Abbreviation	Full Term
ATE	Average Treatment Effect
BSC	Best Supportive Care
CDWH	Clinical Data Warehouse
ctDNA	Circulating Tumor DNA
DAG	Directed Acyclic Graph
ECOG	Eastern Cooperative Oncology Group
EORTC	European Organisation for Research and Treatment of Cancer
HR	Hazard Ratio
ITT	Intention-to-Treat
MAR	Missing At Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
NA	Not Applicable
OR	Odds Ratio
PMM	Predictive Mean Matching
PO	Proportional Odds
psATE	Principal Stratum Average Treatment Effect
QLQ-C15	Quality of Life Questionnaire - Core 15
QLQ-C30	Quality of Life Questionnaire - Core 30
QoL	Quality of Life
rcs	Restricted Cubic Splines
SAP	Statistical Analysis Plan
SAT	Single-Arm Trial
SE	Standard Error
SOP	State Occupancy Probability

2 Introduction

A primary goal of cancer therapy, especially in the advanced setting, is the maintenance or improvement of patient-reported Quality of Life (QoL). From a patient perspective a primary concern is 'What quality of life can I expect during the time that I am alive?'. Unfortunately, frequently used QoL endpoints in oncology trials, such as mean change from baseline, proportion of 'responders' at a single time point, or time to QoL-deterioration do not answer this question and have multiple methodological shortcomings [1], [2]. We thus propose modeling quality of life using a first-order Markov process [3]. While this solves many of the shortcomings of frequently used methods, several challenges remain.

3 Objectives, General Considerations & Assumptions

3.1 Primary Objective

To compare the effect of ctDNA-guided care versus standard of care (represented by a randomized external control group from the same research registry) on patients' overall QoL trajectory over a 26-week (6-month) follow-up period.

3.2 General Considerations

3.2.1 Ordinal Nature of QoL Data

In LIQPLAT, QoL is measured using the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 and QLQ-C15 questionnaires [4]. The QoL scale (question 30) is an ordinal outcome with seven levels, ranging from 1 (very poor) to 7 (excellent), as shown below:

How would you rate your overall quality of life during the past week?

1	2	3	4	5	6	7
Very poor						Excellent

The EORTC's scoring manual advocates for treating the responses as numeric [5], which we will not do. Treating such ordinal scores as if they were numeric (i.e., assuming equal intervals between categories and normally distributed errors) is a common but flawed practice. This approach leads to systematic errors, including inflated Type I error rates, loss of statistical power, and inversions of treatment effects where an analysis might indicate harm when there is benefit, or vice-versa [6].

For modeling purposes, we invert the scale. We will also assume that the categories that are not labeled according to EORTC can be labeled as follows.

1	2	3	4	5	6	7
Excellent	Very Good	Good	Average	Below Average	Poor	Very poor

We acknowledge that not everyone would assign precisely these labels, but believe them to be agreeable to the majority of patients. The labels will serve for easier communication of results.

3.2.2 Longitudinal and Irregularly Spaced Data

QoL changes over time, therefore its trajectory contains more information than a single measurement. In the LIQPLAT trial, QoL questionnaires are completed at clinical appointments, which occur frequently but at irregular intervals for each patient. We anticipate an average interval of about six weeks. This data structure has two key implications:

1. **Time-Dependent Correlation:** A patient's QoL at one time point is correlated with their QoL at preceding time points. This within-subject correlation should be accounted for [7]. In other words, multiple measurements within a subject are *not* exchangeable, as would be assumed by a simple random intercept and the associated compound symmetric correlation structure [7].
2. **Irregular Gaps:** The correlation between two measurements likely weakens as the time gap between them increases. The analytical model should explicitly incorporate the time gap between consecutive measurements to better approximate the decaying correlation structure. If we assume that the irregularity of follow-up is random conditional on a set of baseline covariates and the previous QoL state, the irregular spacing could be used to reconstruct QoL trajectories over the entire follow-up period.

3.2.3 The Competing Risk of Death

In a population with advanced cancer, death is a frequent outcome that prevents any further QoL measurement. This represents a competing risk. As detailed in Section 5.3 and Section 10.5 this competing risk remains a challenge with our modeling approach.

3.3 General Approach

We propose to make use of the irregular, pseudo-random follow-up of usual care and its QoL measurements to reconstruct the quality of life trajectories for the entire duration of follow-up (see Figure 1). These can then be used to derive easily interpretable estimands, such as the difference in time spent in a certain QoL state.

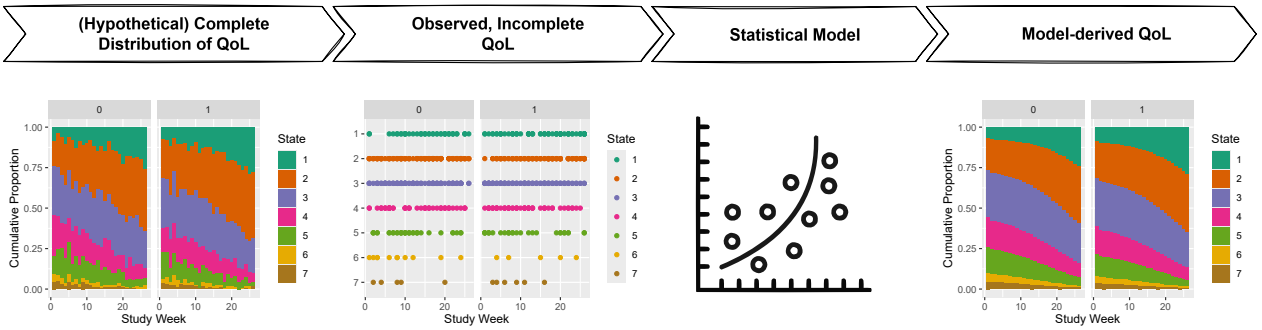

Figure 1: Conceptual overview of the statistical approach for reconstructing Quality of Life (QoL) trajectories. The analysis bridges the gap between the hypothetical, complete QoL data (far left), from which empirical state occupancy probabilities (SOPs) could be directly calculated, and the actual observed data, which is sparse and irregularly spaced (second from left). A statistical model (third from left) is fitted to these sparse observations to generate complete, model-derived SOPs for the entire follow-up period (far right), effectively reconstructing the full QoL trajectory for the population and providing associated uncertainty estimates.

3.4 Assumptions

We make assumptions, which must be clearly acknowledged.

! Key assumptions

1. QoL responses can be accurately modeled by using a first-order Markov process. (+)
2. If the effect of time is non-proportional, this can be well approximated using a linear constraint. (+)
3. A quality of life response from a single day is valid for an entire week. (+/-)
4. The treatment effect is proportional, i.e., the odds of changing to a higher category are the same for all thresholds. (+/-)
5. Participants who died during the 6 month follow-up would have died irrespective of their treatment assignment. (+/-)
6. How a previous state affects the current state does not change over time since randomization. (-)
7. The treatment effect, if present, does not change over time. (-)
8. Days where no QoL was assessed are missing at random (MAR). (-)

(+) Indicates that we are confident the assumption is valid, (+/-) indicates that we are uncertain regarding its validity, (-) indicates that we believe the assumption to be strenuous.

We explain the rationale and implications of the assumptions in Section 5.3, Section 5.5, and Section 7.1.2.3. Given some of the assumptions, especially assumption eight, we have **low confidence** in the results. Nevertheless, we believe that other analysis options we have explored are less optimal and would lead us to have very low confidence.

4 Analysis Populations

4.1 Intention-to-Treat (ITT) Population

The ITT population includes all patients from the prospective research registry who were eligible for LIQPLAT. It comprises two groups: 1. **Intervention Group:** All patients randomly selected for invitation to participate in the LIQPLAT single-arm trial (SAT). 2. **External Control Group:** All eligible patients who were considered for invitation but not randomly selected during the same recruitment period.

4.2 Principal Stratum Population (Primary Analysis Population)

The primary analysis will be conducted on the **Principal Stratum Population**. This population is defined as the subset of ITT patients who would have survived the entire 26-week follow-up period, regardless of their assignment to be invited to the SAT or not. In practice, this population will be approximated by including all patients from the ITT population who were alive at the end of the 26-week follow-up. The rationale for restricting the analysis to the Principal Stratum of 'always-survivors' is detailed in Section 5.3

5 Endpoints and Estimands

5.1 Primary Endpoint

Estimand	Analysis
Target Population	*Analysis Set*
Adult (≥ 18 years) with advanced solid cancer, who would survive >6 months irrespective of invitation to ctDNA-guided care	The subset of all participants who were selected for invitation to the SAT and the corresponding patients from the external control group, who survived at least 6 months. (Principal Stratum Population)
Variable	*Outcome measure*

Estimand	Analysis
Time (in weeks) spent with a good or better QoL during the first 6 months after randomization.	Ordinal QoL states measured using the 7-level EORTC QoL scale during routine follow-up at irregular intervals over 26 weeks, recorded in REDCap. 'Good' QoL is defined as the best three ordinal states (1 to 3).
Population-level summary measure	*Analysis approach*
Difference in mean number of weeks spent with good or better QoL	Estimated using a first-order Markov longitudinal ordinal transition model. This model conditions on the previous week's state and analyses the transitions as conditionally independent. Time is modeled non-linearly using restricted cubic splines. The proportionality assumption with respect to time is relaxed by using partial proportional odds with a linear constraint. The model is used to derive daily state occupancy probabilities (SOPs) for the three best QoL States (states 1 to 3), which are summed under both treatment groups and subtracted to derive the difference.

5.2 Secondary Endpoints

- **Estimand 1 (Principal Stratum):** The **difference in the mean number of weeks spent in a "Poor" QoL state** (defined as the bottom three categories, states 5, 6, or 7) over 26 weeks.
- **Estimand 2 (Principal Stratum):** The **transitional odds ratio (OR)** for being in a better QoL state versus a worse one at any given time, comparing treatment arms.

5.3 Rationale for Principal Stratum

Our analysis software of choice, the `rmsb` package [8], has built-in functions to handle interval censoring of the ordinal dependent variable. For LIQPLAT trial, the date of death is provided for all patients via the data warehouse (CDWH) and is virtually complete. QoL is assessed at irregular intervals. This means that the dependent variable ranging from 1-8, where 1-7 are QoL states and 8 is death, is heavily censored for states 1 to 7. In other words, we know that if a patient has not died, they have a state between 1 and 7. For a patient who has a QoL of 4 in week 0 and a QoL of 6 in week 6 and died in week 9, the states would look like this : 4 [1-7] [1-7] [1-7] [1-7] [1-7] 6 [1-7] [1-7] 8.

Unfortunately, we found that in such a situation, when the absorbing state is always known, and non-absorbing states are sparse (85% missing), our modeling approach produces falsely high estimates of the cumulative incidence of death. This bias in cumulative incidence then propagates to the estimates for all other state occupancy probabilities (SOPs), affecting the final estimand. Further research is needed to resolve this issue. For more details see Section 10.5.

Another option is to assume that ctDNA-guided care only affects QoL, but does not affect survival. This could be through the prolongation of imaging exam intervals, leading to fewer burdensome healthcare visits or earlier termination of ineffective therapies. We acknowledge that it seems unlikely that an improvement of QoL mediated by a reduction of treatment associated side effects and fewer healthcare encounters has no effect on mortality. However, the effect on mortality would plausibly be too low to meaningfully impact the 6-month survival status.

Concretely, we assume that patients who experience the intercurrent event death I during the six month follow-up, would do so no matter the treatment assignment a. $I^{a=1} = I^{a=0} = 1$ for people who always die and $I^{a=1} = I^{a=0} = 0$ for people who always survive. If one accepts this assumption, a related causal question we could ask is: 'What is the effect of ctDNA-guided care on the quality of life for those patients who would survive the follow-up period irrespective of treatment assignment?' This question is precisely addressed by a Principal Stratum analysis. By conditioning the analysis on the post-randomization outcome of survival, this approach provides an unbiased estimate of the QoL effect ***if*** the above assumptions are correct.

5.4 Rationale for Dichotomized Summary Estimands

While the primary analysis will leverage the full seven-level ordinal scale of the QoL measurement, the resulting treatment effect will be summarized as the difference in the mean number of weeks spent in 'Good' versus 'Poor' states. This approach is chosen primarily for clear and effective communication. An estimand like 'an average increase of 1.5 weeks in a state of good or better QoL' is more intuitive and directly interpretable for clinicians, patients, and policymakers than a transitional odds ratio. Crucially, this communicable summary is achieved without resorting to dichotomizing the outcome variable before analysis, as detailed in Section 7.1.

5.5 Handling of Intercurrent Events

Intercurrent events are events that occur after treatment initiation and affect either the interpretation or the existence of the measurements of interest [9]. Our strategy for handling them is defined below.

> ⚠ Warning
>
> A more significant challenge is that follow-up is in routine care. If ctDNA affects follow-up, e.g., fewer visits, or no more visits at all because of early switch to best supportive care, we record fewer or no more quality of life information about these patients. I think this could strongly bias the results, but I'm at a loss on how to handle this.

Intercurrent Event	Strategy	Rationale
Death	*Principal Stratum*	For the primary analysis, death defines the boundary of the Principal Stratum.
Discontinuation of ctDNA monitoring	*Treatment Policy*	Participant data will be analyzed as part of their assigned arm, regardless of adherence to the monitoring schedule. This preserves the ITT principle and estimates the effect of the *strategy* of ctDNA-guided care.
Switch to best supportive care (BSC)	*???*	This intercurrent presents a serious challenge. ctDNA-guided care could lead to a situation where a patient is switched earlier to BSC. Because QoL assessment happens in routine care, we would not know the QoL of these patients. Their QoL is likely to be different than if the patient had not been switched to BSC.
Treatment at another hospital	*???*	
Loss to follow-up	*???*	

5.6 Derivation of the Primary Estimand

To formally define the estimand for the Principal Stratum analysis, we first introduce notation for potential outcomes related to survival. Let S_i^a be an indicator for the survival outcome of patient i at 26 weeks under treatment a, where $a = 1$ for random selection for invitation to the SAT and $a = 0$ for the external control group. $S_i^a = 1$ if patient i would survive the 26-week period under treatment a, $S_i^a = 0$ if patient i would die during the 26-week period under treatment a.

The Principal Stratum of "always-survivors" is the sub-population of patients who would survive the 26-week follow-up regardless of treatment assignment, i.e., only if $S_i^1 = S_i^0 = 1$.

The primary estimand is the Average Treatment Effect (ATE) on the total number of weeks a patient spends in a "Good" QoL state j (categories 1, 2, or 3) over a 26-week period *within this stratum of always survivors.*

We define the potential outcome under treatment a for the total number of weeks spent in a good state, W_i^a, as:

$$W_i^a = \sum_{t=1}^{26} \sum_{j=1}^{3} \mathbb{I}(y_{it}^a = j)$$

where y_{it}^a is the potential QoL state for patient i at week t under treatment a and $\mathbb{I}(\cdot)$ is the indicator function that counts 1 for each week the condition is met and 0 otherwise.

The primary estimand, the average treatment effect in the Principal Stratum, denoted as τ_W^{psATE}, is the difference in the expected value of these potential outcomes, conditional on membership of the always-survivor stratum:

$$\tau_W^{psATE} = \mathbb{E}[W_i^1 \mid S_i^1 = 1, S_i^0 = 1] - \mathbb{E}[W_i^0 \mid S_i^1 = 1, S_i^0 = 1]$$

> **!** Identification and Key Assumption
>
> The estimand can only be correctly identified under the key assumption that the intervention has no effect on survival.
>
> For all patients i, $S_i^1 = S_i^0$. Under this assumption, the group of patients who are observed to survive 26 weeks is the Principal Stratum of always-survivors.

> **i** Note
>
> I'm actually not quite sure this is true. I think the effect on mortality would have to be so that patients who died within 26 weeks under treatment a = 0 would also do so under a = 1, not no effect. But it's probably very unlikely that the treatment has no effect on someone who dies during week 26, but some effect on someone who died in week 4, but just not enough to push them past 26 weeks.

6 General Statistical Considerations

6.1 Framework

All analyses will be conducted within a Bayesian framework. Our inferential goal is to estimate the full posterior distribution of the treatment effect (the difference in mean weeks with 'Good' or better QoL). This allows for direct probabilistic statements about the plausible magnitude of the effect. We will report key summaries of the posterior (e.g., mean, 95% credible interval) and the probability of any benefit, but we will not pre-specify a threshold for a dichotomous claim of success or failure.

6.2 Missing Data Handling

Missing data are expected for both baseline covariates and the longitudinal Quality of Life (QoL) measurements. All missing values will be handled using Multiple Imputation by Chained Equations (MICE) with predictive mean matching, under the assumption that the data are Missing At Random (MAR).

We will generate $m = 50$ complete datasets [10]. The primary analysis model will be fitted to each of these datasets, and the resulting posterior distributions for all parameters and derived estimands will be stacked.

The imputation model is specified as a multilevel model to account for the clustered nature of the data (i.e., repeated QoL measurements nested within patients). The model will use the following information as predictors: treatment assignment, patient age, gender, baseline ECOG performance status, cancer diagnosis, initial treatment plan, the time of the QoL assessment (in days), survival-status at month 6, and the Nelson-Aalen estimate of the cumulative hazard.

Specifically, we will handle missingness in:

- **Baseline Covariates:** Patient-level characteristics such as ECOG performance status, cancer diagnosis, and initial treatment plan will be imputed at the patient (cluster) level.
- **Longitudinal QoL:** Intermittently missing responses to the QoL question (QLQ-C30 question 30) will be imputed at the observation level. This includes timepoints where a questionnaire was provided but the question was unanswered, as well as for the day of random selection (day 0). Due to LIQPLAT's design, a large proportion of the participants will be randomly selected before their visit, i.e., before their first questionnaire. However, for marginalization a QoL state at day 0 is required. Multiple imputation is likely a more prudent choice than carrying the first observed value backward.

The technical specification of the imputation model is detailed in Section 10.1.

6.3 Software

Analyses will be performed using R (version 4.4.3 or later) [11]. Bayesian models will be fitted using Stan via the `rmsb` [8] and `brms` packages [12].

7 Analysis Methods

7.1 Primary Analysis Model: Bayesian First-Order Markov Ordinal Transition Model

7.1.1 Conceptual Framework

The primary analysis uses a Bayesian first-order Markov state transition model for ordinal data [3]. This class of model estimates the one-week transition probabilities between Quality of Life (QoL) states. By conditioning on the patient's most recent observed state, the model explicitly accounts for the longitudinal nature of the data and correctly handles the irregular measurement schedule by incorporating the time gap between observations as a covariate. The model can be used to predict the probability of a patient's QoL state at a given week, conditional on their *immediately preceding observed state* and the *time gap* since that observation.

7.1.2 Statistical Model

7.1.2.1 Mathematical Notation

Let y_{it} be the ordinal QoL state (from 1 to 7, where 1 is best) for patient i at week t. Let $y_{it'}$ be their last observed state at a prior week t', and let the time gap be $\Delta t = t - t'$. The model is specified as a cumulative logit model for the transition probabilities:

$$
\begin{aligned}
\text{logit }(P(y_{it} \geq j \mid y_{it'})) &= \alpha_j - \left(\eta_{it} + \gamma_{it,j}\right) \\
\eta_{it} &= \beta_{\text{tx}} \cdot \text{Treatment}_i + f(t) + \sum_{k=2}^{7} \beta_{\text{yprev }=k} \cdot \mathbb{I}(y_{it'} = k) \\
&\quad + \beta_{\text{gap}} \cdot \Delta t + \sum_{k=2}^{7} \beta_{\text{yprev }=k\times \text{ gap}} \cdot \mathbb{I}(y_{it'} = k) \cdot \Delta t + \mathbf{X}_i \beta_{\text{covars}} \\
\gamma_{it,j} &= (\tau \cdot t) \cdot j \\
\alpha &\sim \text{Dirichlet }(0.308) \\
\beta_k, &\sim \text{Normal }(0, 100) \\
\tau &\sim \text{Normal }(0, 100)
\end{aligned}
$$

Where:

- the log-odds of a patient's QoL being in state j or worse at week t (conditional on their prior state $y_{it'}$) is a function of the intercepts (α_j) and two linear predictor components.
- η_{it} is the main linear predictor for effects assumed to satisfy the proportional odds assumption (i.e., their effect is constant across the $j-1$ cumulative logits). It includes:
  - β_{tx}: The main effect of the treatment arm.
  - $f(t)$: A flexible function of study week t to model potentially non-linear time trends. This will be specified as a restricted cubic spline with four knots.
  - $\sum_{k=2}^{7} \beta_{\text{yprev }=k} \cdot I(y_{it'} = k)$: The effect of the previous QoL state, modeled as a categorical variable with 6 parameters relative to state 1 as the reference category. $I(\cdot)$ is an indicator function.
  - β_{gap}: The linear effect of the time gap since the last measurement.
  - $\sum_{k=2}^{7} \beta_{\text{yprev }=k\times \text{ gap}} \cdot I(y_{it'} = k) \cdot \Delta t$: An interaction term allowing the effect of the patient's previous QoL state to differ depending on the gap.
  - $\mathbf{X}_i \beta_{\text{covars}}$: A vector representing the effects of other baseline covariates, specifically functional status at baseline and diagnosis category.
- $\gamma_{it,j}$ models a deviation from the proportional odds assumption. We allow the effect of time to be non-proportional, but constrain the effect to be linear in the outcome category j. This means the odds ratio for t can change linearly across the different cutpoints of the QoL scale. We chose a linear constraint as a parsimonious way to relax the proportional-odds assumption, allowing for a simple trend in the effect of time without overfitting.
- **Prior Specifications**:
  - **Intercepts (α)**: The priors for the $j = 2, ..., 7$ intercepts are induced by a Dirichlet distribution on the baseline cell probabilities when all covariates are set to their mean or reference values. The concentration parameter as calculated by default by the `rmsb` package is $1/(0.8 + 0.35 \cdot \max(k, 3))$, which evaluates to 0.308 for a 7-level outcome. This enforces the necessary ordering of the cutpoints ($\alpha_2 < \alpha_3 < ... < \alpha_7$) for a typical subject, from which the model then estimates covariate effects.
  - **Coefficients** (β_k, τ): All regression coefficients, including the non-proportional odds parameter τ, are assigned the default almost non-informative Normal $(0, 100)$ priors. This serves to regularize the model by penalizing extreme parameter values while allowing the data to dominate the posterior inference.

This primary estimand is a derived quantity calculated from the model's posterior distribution. The expected values, $\mathbb{E}[W_i^1]$ and $\mathbb{E}[W_i^0]$, are computed by summing the SOPs for states 1 to 3 across the 26 weeks. These probabilities are marginalized over the observed distribution of baseline covariates and baseline QoL states to obtain the population-averaged result.

7.1.2.2 Model Specification (R Code)

The model will be implemented using the `blrm` function from the `rmsb` package:

```
model_primary <- blrm(
  formula = y ~ tx + rcs(week, 4) + yprev * gap + ecog_fstcnt + diagnosis,
  data = data_for_model_survivors_only,
  ppo = ~week,
  cppo = function(y) y,
  iter = 4000, chains = 4, seed = 1234
)
```

7.1.2.3 Discussion of modeling choices

- **Choice of one week as the discrete time unit:** The analysis discretizes time into weekly intervals rather than daily intervals for several pragmatic and methodological reasons. First, in the EORTC questionnaire patients are specifically asked about their quality of life in the last week. Second, given the high proportion of missing data inherent to routine care follow-up, daily estimates could imply a level of precision that is not supported by the data. Third, in LIQPLAT, it is highly unlikely that patients complete more than one questionnaire within a single week, making the week a natural and minimal unit of observation. Finally, this choice significantly reduces the computational burden of the simulation study and the primary analysis (e.g., 26 weekly transitions vs. 182 daily transitions per patient), enhancing the feasibility of the project without a substantial loss of relevant information. The impact of this aggregation on the final estimand is expected to be minimal.
- **Constant treatment effect:** A treatment effect, if present, is unlikely to be constant over time. CtDNA sampling can take weeks and is less likely to change first-line than second-line therapy of participants. An effect in the first weeks after baseline is thus very unlikely. However, given our sample size, we are likely not able to identify time-varying treatment effects reliably and choose a constant treatment effect for slightly smaller variance in our estimator.
- **Constant effect of the previous state:** it is unlikely that the effect of the previous state is constant over time. Very poor QoL at the beginning of the study might predict regression to the mean, whereas a very low-quality of life during later stages of the disease might not. However, an interaction of the six-level categorical variable (`yprev`) with flexibly modeled time would use too many degrees of freedom, given our sample size.
- **Linear effect of gap time:** The effect of the previous state (`yprev`) is unlikely to decay linearly over time. However, most questionnaires will be repeated after at least some weeks. Even if the decay shortly after a previous answer is non-linear, we anticipate the decay in the expected range of gaps to be approximately linear.
- **Inclusion of baseline-covariates:** Both the functional status at baseline (`ecog_fstcnt`) and the cancer diagnosis category (`diagnosis`) are likely to be predictive of quality of life based on our clinical experience and prior research, and are thus included to explain more variation in the outcome. If we observe fewer than five patients with a functional status (ECOG) of 4, we will merge categories 3 and 4 into '3plus', to avoid divergent transitions of the estimate of the effect of ecog 4 because of too few observations. The diagnosis categories will be refactored into a five level categorical variable, with the four most common diagnoses as distinct categories and the rest of diagnoses lumped together as 'other'.

7.1.2.4 Presentation of results

We will report the entire posterior distribution of the treatment effect as a half-eye plot which includes 66% and 95% credible intervals. See Figure 2.

Figure 2: Posterior distribution of treatment difference in mean time spent with good or better quality of life (defined as states 1 to 3), with 66% and 95% credible intervals.

We will also show the estimated SOPs for each state over time with 95% posterior credible bands. See Figure 3.

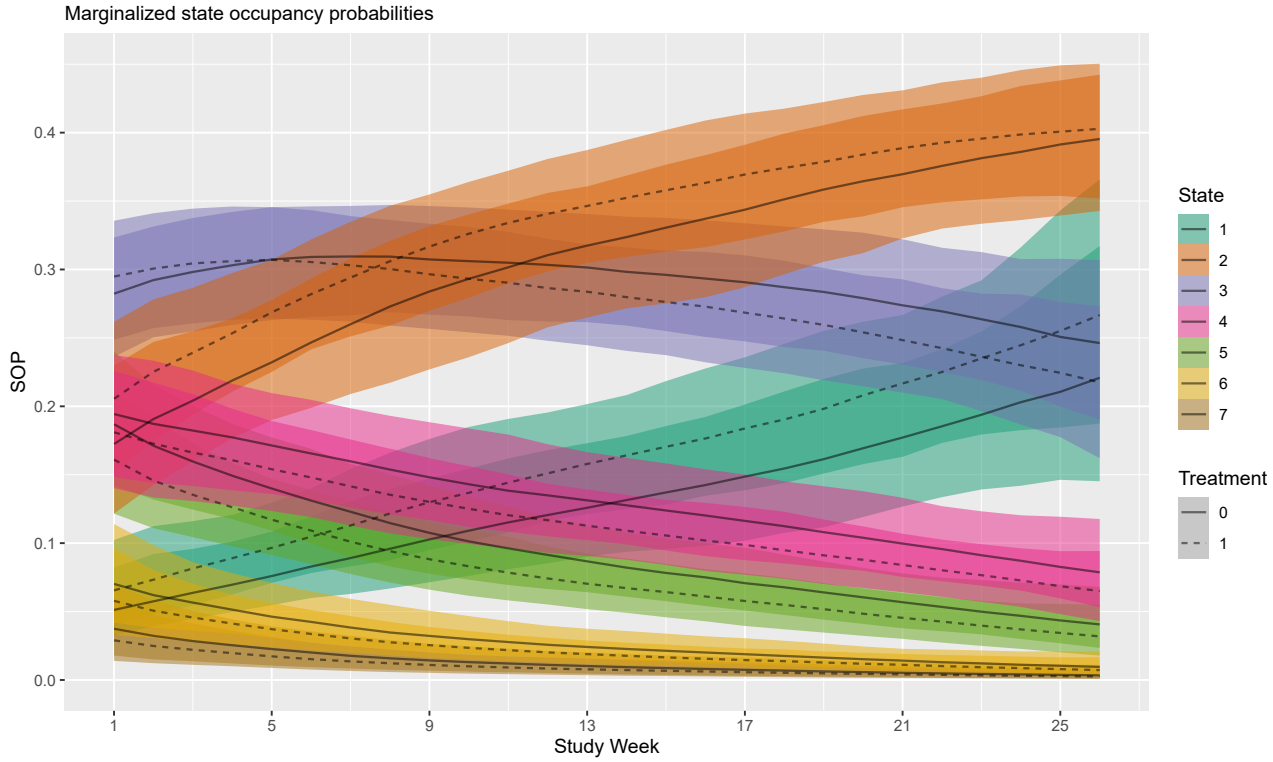

Figure 3: Model based marginal state occupancy probabilities (SOPs) for each ordinal state over time. The group of patients invited to the SAT is labeled with a dashed line, the external control group is labeled with a solid line. 95% posterior credible bands are shown.

8 Comparison of Common Endpoints

Common approaches to analysing quality of life in oncology are change from baseline, time-to-deterioration, or an ordinal analysis, e.g., a proportional odds model at a single timepoint [2]. Below we outline some theoretical considerations as to why these endpoints are less suitable than our primary analysis and show frequentist operating characterstics based on a simulation study. We also compare the Markov model with a model with a cumulative logit multilevel level with a random intercept only, at the patient level.

8.1 Assumptions Made for each Approach

Assumption	Markov Model	Cumulative Logit Multi-level Model	Time to Deterioration	Cumulative Logit Model at M6	Change from Baseline
Intervals are equal			X		X
Starting point does not matter			X		X
Quality of life after a substantial decrease doesn't matter			X		
Patients who can't experience substantial decrease don't matter			X		
Errors are normally distributed					X
QoL between baseline and month 6 can be ignored				X	X
Proportional odds for treatment effect	X	X		X	
Proportional hazards for treatment effect			X		
Data are MAR	X	X	??	??	??
Assumes a correlation structure	X	X	NA	NA	NA
Patients have no distinct trajectories	X		NA	NA	NA
Patient-specific effects are Gaussian		X	NA	NA	NA

Table 1: Table comparing the assumptions made by each method.

We deem change from baseline, time-to-deterioration, and to a lesser extent the ordinal QoL value at 26 weeks analyses as particularly problematic. Change from baseline and the ordinal analysis at 26 weeks ignore the patient's QoL trajectory and can be misleading. For example, an intervention might improve QoL between months 0 and 6 to a clinically relevant extent, even if that effect disappears by the 6-month endpoint. Furthermore, the irregular timing of clinic visits in LIQPLAT necessitates defining an arbitrary observation window (e.g., weeks 20-26) for these analyses, which incorrectly assumes that a patient's QoL is static within this window. A time-to-deterioration analysis systematically excludes patients with low baseline QoL who cannot experience the predefined deterioration event and discards all QoL information following the event. Moreover, both time-to-deterioration and change-from-baseline analyses treat the ordinal QoL scale as a numeric interval scale, assuming, for instance, that a two-point drop from category 7 to 5 is equivalent to a drop from category 4 to 2. This is a strong, unverifiable, and unrealistic assumption.

Collectively, these deficiencies make simpler endpoint or event-based analyses less appropriate for the LIQPLAT trial. An analysis that properly accounts for the full QoL trajectory, within-subject correlations, and irregular follow-up provides a more robust and defensible evaluation of the treatment effect.

8.2 Simulation

To evaluate the operating characteristics of the above approaches to analyze QoL data, if the underlying correlation structure is actually Markovian, we conducted a simulation study. The data for the simulation were generated by fitting a first-order Markov model to a historical dataset of cancer patients at the University Hospital Basel. We acknowledge that this data-generating mechanism inherently favors the Markov model for analysis. However, the purpose of the study was twofold: 1) to confirm that the Markov model could accurately recover the true, known treatment effect under ideal (but realistic) conditions of irregular follow-up and missingness, and 2) to quantify the specific failure modes and biases of commonly used but misspecified models when applied to such data. For more details on the design and results of the simulation study see Section 10.4.

The simulation tested the following models on the Principal Stratum of survivors:

1. **First-Order Markov Model**
2. **Ordinal Multilevel Model with a Random Intercept**
3. **Time-to-Deterioration Model**
4. **Cumulative Logit Model at Month 6**
5. **Change from Baseline Model**

8.2.1 Frequentist Operating Characteristics

Table 2 shows frequentist operating characteristics of the different analyses from our simulation study. The data-generating mechanism assumed an initial sample size of $N = 270$, before conditioning on the Principal Stratum, with an average cumulative incidence of death by 6 months of 17.8% for both groups, i.e., on average 49 patients were excluded. 85% of post-baseline observations are missing completely at random (MCAR), which equals about 4 post-baseline measurements per patient. Power and Type I error were assessed under a constant true odds ratio (OR) of 0.8 (H_A) and 1.0 (H_0).

For LIQPLAT, we will not use the posterior to make a dichotomous claim of benefit. However, for the purpose of evaluating the operating characteristics of our analysis plan, we assess power and Type I error under the hypothetical decision rule of requiring a posterior probability of benefit greater than 0.95. This allows for a standardized comparison of different analytical methods. Details on model specifications, convergence diagnostics, probability of benefit of each iteration can be found in Section 10.4.

Model	Power	Type I Error
Markov Model	0.450 (0.0157)	0.054 (0.0072)
Random Intercept	0.445 (0.0352)	0.044 (0.0101)
Time-to-Deterioration	0.112 (0.0100)	0.047 (0.0067)
6-Month Ordinal	0.197 (0.0126)	0.050 (0.0069)
Change from Baseline	0.176 (0.0120)	0.050 (0.0069)

Table 2: Bayesian Power and Type I error when $P(\text{benefit}) > 0.95$ is chosen as a cut-off for benefit for the five endpoints (Monte Carlo SE in parenthesis).

9 Limitations

9.1 Outcome assessment in routine care

⚠ Warning

@ Giusi: Is there anything we can do to debias the estimates a bit, given certain assumptions? Pretty confused by the causal inference literature on missing data.

A primary limitation arises because the treatment strategy is expected to influence post-randomization clinical decisions, which in turn affect whether the QoL outcome is observed (see Figure 4). These decisions affect both future QoL and the probability that QoL is measured, since data is collected during routine care appointments. For instance, a decision to switch a patient to best supportive care (BSC) may lead to fewer clinic visits, resulting in the cessation of QoL data collection.

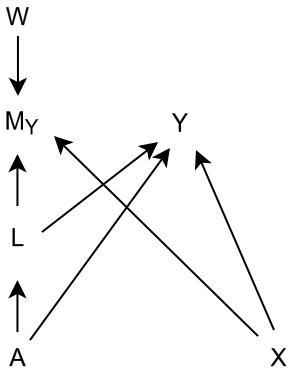

Figure 4: A missingness DAG (m-DAG) depicting assumptions regarding missingness in the quality of life outcome (Y). M_Y is the missingness indicator for Y. W is the vector of unmeasured causes of M_Y. A represents the selection for invitation to LIQPLAT (treatment indicator), X represents measured baseline covariates, L represents effects of A which affect both follow-up and thus missingness (M_Y) and QoL (Y), such as early switch to best supportive care. $A \rightarrow Y$ represent effects of A on Y which do not affect missingness, though we expect these to be rare. Realistically, ctDNA sampling would mostly Y through changes in the treatment regimen, which also affects missingness.

As a simple illustrative example, assume ctDNA management only improves QoL through an early switch to BSC of an ineffective treatment. Let's also assume that all the patients who are switched to BSC have no more follow-up visits, and their improved QoL is not recorded. Logically, this results in an inability to show the true treatment effect. A short R Code example is given below.

```r
n <- 1e6 # sample size
a <- sample(c(0, 1), size = n, replace = TRUE) # binary treatment assignment
# 20% of treated are switched early to BSC
bsc <- if_else(a == 1, rbinom(n, 1, prob = 0.2), 0)
# assume we can directly observe qol, which follows a standard normal. If switch to
BSC, massive improvement.
y <- rnorm(n) + bsc

# But bsc also causes missingness to 100%
missing <- if_else(bsc == 1, rbinom(n, 1, prob = 1), 0)
# Y with missing values
y_miss <- if_else(missing == 1, NA, y)

model_true <- lm(y ~ a)
model <- lm(y_miss ~ a)
```

	contrast	estimate
Estimate without missing data	1 - 0	0.200
Estimate with missing data	1 - 0	−0.001

Table 3

We will check whether the number of questionnaires handed out to patients, adjusted for their time alive, is similar between both groups.

It is important to emphasize that this missing data issue is a limitation of the trial's design rather than the specific analytical method chosen. By collecting QoL data only during routine appointments, the design inextricably links the observation process to the clinical pathway. This issue would affect any potential analysis.

- Single timepoint analyses (e.g., ordinal model at month 6, change from baseline) would be biased because patients who switched to best supportive care would likely be missing from the 6-month assessment.
- A time-to-deterioration analysis would be informatively censored, as deterioration events after a switch to best supportive care would go unrecorded.
- Alternative longitudinal models : ?????

Consequence: While our chosen Markov model cannot solve a problem rooted in the study design, its vulnerability is shared across all feasible alternatives. Under the null hypothesis, this issue is unlikely to inflate the Type I error rate. However, if a true treatment effect exists, our estimate of its magnitude will likely be biased because we are systematically losing outcome data from a non-random subset of patients.

9.2 Principal Stratum Assumption

Our primary analysis relies on the strong, untestable assumption that the intervention has no effect on mortality. The difficulty in handling the competing risk death is not unique to the Markov model but rather complicates all potential analytical approaches.

- A change-from-baseline analysis faces the challenge of how to code death. Assigning death an arbitrary numeric value (e.g., a value of '8') makes highly strenuous assumptions about the interval between 'Very Poor' QoL and death.

- A time-to-deterioration analysis could formally address death using a competing risks model (e.g., Fine-Gray). However, the resulting estimand — a subdistribution hazard ratio — is notoriously difficult for clinicians and patients to interpret, which would obstruct the translation of our findings.
- An ordinal model at month 6 that includes death as the worst outcome is an alternative, but would require multiple imputation for patients who did not fill out a questionnaire between week 14 and 26 to obtain accurate cell probabilities.

10 Appendices

10.1 Appendix A - Multiple Imputation

10.1.1 Overview

All imputations will be performed in R (version 4.4.3 or later) using the `mice` and `miceadds` package. The imputation procedure is designed to handle the multilevel structure of the data and uses predictive mean matching (PMM).

10.1.2 Imputation Model Specification

The imputation model will include all variables from the primary analysis model, as well as auxiliary variables thought to be predictive of missingness or the variables with missing data.

- **Variables in the Model:** The predictor matrix for the MICE algorithm will include treatment assignment (`tx`), baseline patient characteristics (`pat_age`, `gender`, `ecog_fstcnt`, `diagnosis`, `plan_fstcnt_coded`), time-varying information (`quest_day`), and survival status at month 6 (`status`.) We include the Nelson-Aalen estimate of the cumulative hazard (`na_est`) as a predictor instead of the raw survival time [13].
- **Clustering:** The patient identifier (`pat_id`) will be specified as the clustering variable (`-2` in the `mice` predictor matrix).

10.1.3 Imputation Method by Variable Type

Different imputation methods are used for baseline covariates and the longitudinal QoL outcome to reflect the data structure:

- **Baseline Covariates (`ecog_fstcnt`, `diagnosis`, `plan_fstcnt_coded`):** These are imputed using patient-level PMM (`2lonly.pmm`). This method imputes a single, consistent value for each patient across all of their longitudinal records within a single imputed dataset.
- **Longitudinal QoL (`q30`):** The ordinal QoL state is imputed using observation-level PMM within the multilevel framework (`2l.pmm`). This method is flexible and robust, borrowing information from other patients with similar observed characteristics to find a suitable donor value.
- **QoL at day of random selection:** To calculate SOPs for the entire duration of follow-up a QoL state at day 0 (day of random selection) is required. Given the design of LIQPLAT, a large proportion ($> 50\%$) of patients have their first visit and thus their first questionnaire in the week *after* random selection. Imputation for day 0 is thus required. Multiple imputation using observation-level PMM is likely a more prudent choice than carrying the first observed value backward.

10.1.4 MCMC Parameters

The MICE algorithm will be run to generate $m = 50$ imputed datasets with 50 iterations for each dataset (`maxit=50`). Convergence will be assessed visually using trace plots and density plots of the imputed values.

10.1.5 Pooling of Results

The primary Bayesian Markov model will be run on each of the 50 completed datasets. The posterior draws for every parameter and derived quantity of interest (e.g., the difference in mean time spent in a "Good" QoL state) will be extracted from each of the 50 analyses and combined (stacked) into a single posterior distribution for the final inference.

10.1.6 Code

Code that illustrates the imputation strategy.

```
# nelson aalen estimator
data$na_est <- nelsonaalen(data, timevar = time, statusvar = status)
```

```r
# add day 0 for everyone, even if first contact on later day
id_cols <- c(
  "id", "age", "gender", "ecog_fstcnt", "diagnosis",
  "plan_fstcnt_coded", "time", "status", "na_est", "tx"
)

data <- data |>
    group_by(pick(all_of(id_cols))) |>
    complete(quest_day = union(quest_day, 0)) |>
    ungroup()

# set up imputation
md.pattern(data)

init <- mice(data, maxit = 0)
pred_matrix <- init$predictorMatrix
meth <- init$method

# don't use survival time because we use na_est.
pred_matrix[, "time"] <- 0
pred_matrix["time", ] <- 0
meth["time"] <- ""

# Use patient ID as clustering variable
pred_matrix[, "id"] <- -2

# Impute individual q30 at the questionnaire level
meth["q30"] <- "2l.pmm"


# use mean matching at the cluster level for baseline observations
# this ensures only one ecog per patient per imputation
meth["ecog_fstcnt"] <- "2lonly.pmm"
meth["plan_fstcnt_coded"] <- "2lonly.pmm"
meth["diagnosis"] <- "2lonly.pmm"

# run imputation
imputed_data <- mice(data,
                     m = 50,
                     maxit = 50,
                     predictorMatrix = pred_matrix,
                     method = meth,
                     seed = 1234)
```

10.2 Appendix B: Model Diagnostics

The regression model output, chain mixing plots, and density plots for all models will be shown, as exemplified below for the analysis of the primary estimand, here based on simulated data.

```
Bayesian Constrained Partial Proportional Odds Ordinal Logistic Model

Dirichlet Priors With Concentration Parameter 0.308 for Intercepts

blrm(formula = y ~ tx + rcs(time, 4) + yprev * gap + ecog_fstcnt +
    diagnosis, ppo = ~time, cppo = function(y) y, data = data_for_model,
    iter = 2000, chains = 4, refresh = 5, method = "sampling")


Frequencies of Responses

  1   2   3   4   5   6   7
126 277 248 111  75  22  10

                   Mixed Calibration/             Discrimination               Rank
Discrim.
               Discrimination Indexes                   Indexes
Indexes
  Obs 869    LOO log L-1351.6+/-20.67    g  1.013 [0.864, 1.154]    C   0.668 [0.659,
0.674]
Draws4000      LOO IC 2703.19+/-41.34    gp  0.215 [0.189, 0.24]    Dxy 0.336 [0.319,
0.348]
  Chains4     Effective p31.37+/-1.44    EV  0.144 [0.11,
0.175]
Time40.4s      B 0.219 [0.216, 0.222]    v  0.843 [0.595,
1.088]
  p    23                               vp 0.036 [0.029,
0.045]

                  Mean Beta Median Beta S.E.   Lower   Upper   Pr(Beta>0)
y>=2               3.5307    3.5178     0.4918  2.5871  4.5012 1.0000
y>=3               1.8046    1.7993     0.4561  0.9278  2.7261 1.0000
y>=4               0.4311    0.4234     0.4411 -0.4308  1.3124 0.8348
y>=5              -0.4887   -0.4925     0.4496 -1.3998  0.4001 0.1338
y>=6              -1.8581   -1.8541     0.4757 -2.8363 -0.9723 0.0003
y>=7              -3.0931   -3.0865     0.5513 -4.2002 -2.0457 0.0000
tx                -0.2649   -0.2632     0.1329 -0.5194 -0.0030 0.0213
time              -0.0875   -0.0874     0.0545 -0.1975  0.0178 0.0562
time'              0.1028    0.1050     0.1540 -0.1979  0.4123 0.7510
time''            -0.2173   -0.2217     0.3983 -1.0344  0.5402 0.2928
yprev=2           -0.1957   -0.1943     0.3674 -0.9386  0.4922 0.2915
yprev=3           -0.0729   -0.0709     0.3570 -0.7576  0.6461 0.4125
yprev=4            0.2382    0.2359     0.3915 -0.5966  0.9349 0.7410
yprev=5           -0.2832   -0.2801     0.4239 -1.0814  0.5750 0.2448
yprev=6            0.0675    0.0684     0.4648 -0.8143  0.9988 0.5628
yprev=7            0.3186    0.3172     0.7269 -1.1128  1.6695 0.6620
gap                0.0186    0.0186     0.0492 -0.0810  0.1094 0.6480
ecog_fstcnt=2      1.5078    1.5072     0.1932  1.1483  1.9038 1.0000
ecog_fstcnt=3plus  2.4779    2.4674     0.6196  1.2242  3.6425 1.0000
diagnosis=2       -0.5802   -0.5766     0.2561 -1.0662 -0.0789 0.0105
diagnosis=3       -0.9750   -0.9743     0.2574 -1.4747 -0.4831 0.0000
```

```
diagnosis=4       -0.9269   -0.9224     0.2883 -1.4733 -0.3578 0.0005
diagnosis=5       -0.5739   -0.5756     0.1924 -0.9356 -0.1911 0.0015
yprev=2 * gap      0.0284    0.0291     0.0608 -0.0806  0.1575 0.6832
yprev=3 * gap     -0.0300   -0.0295     0.0562 -0.1465  0.0756 0.3018
yprev=4 * gap     -0.0386   -0.0385     0.0617 -0.1561  0.0862 0.2638
yprev=5 * gap     -0.0040   -0.0037     0.0629 -0.1255  0.1183 0.4752
yprev=6 * gap     -0.0515   -0.0523     0.0650 -0.1761  0.0721 0.2240
yprev=7 * gap     -0.1084   -0.1066     0.1102 -0.3296  0.1039 0.1615
time x f(y)       -0.0062   -0.0062     0.0062 -0.0174  0.0071 0.1578
                  Symmetry
y>=2              1.04
y>=3              1.02
y>=4              1.02
y>=5              1.00
y>=6              1.01
y>=7              0.91
tx                0.99
time              0.99
time'             0.98
time''            1.00
yprev=2           1.01
yprev=3           1.01
yprev=4           1.00
yprev=5           1.01
yprev=6           1.01
yprev=7           0.99
gap               1.01
ecog_fstcnt=2     1.03
ecog_fstcnt=3plus 1.04
diagnosis=2       1.01
diagnosis=3       0.99
diagnosis=4       1.02
diagnosis=5       1.01
yprev=2 * gap     0.99
yprev=3 * gap     1.01
yprev=4 * gap     1.03
yprev=5 * gap     1.01
yprev=6 * gap     1.04
yprev=7 * gap     0.97
time x f(y)       1.01
```

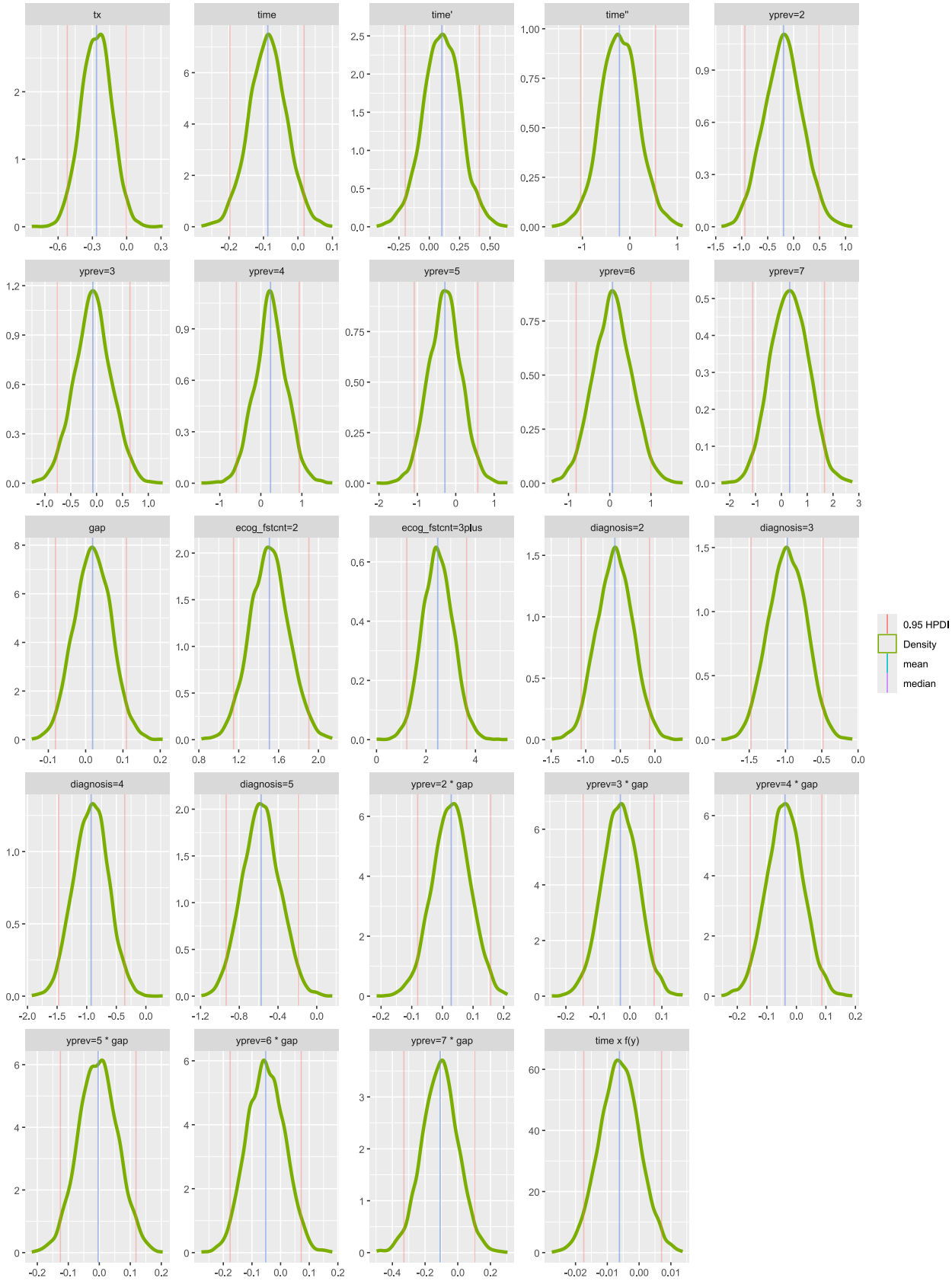

Figure 5: Posterior density plots of each model parameter.

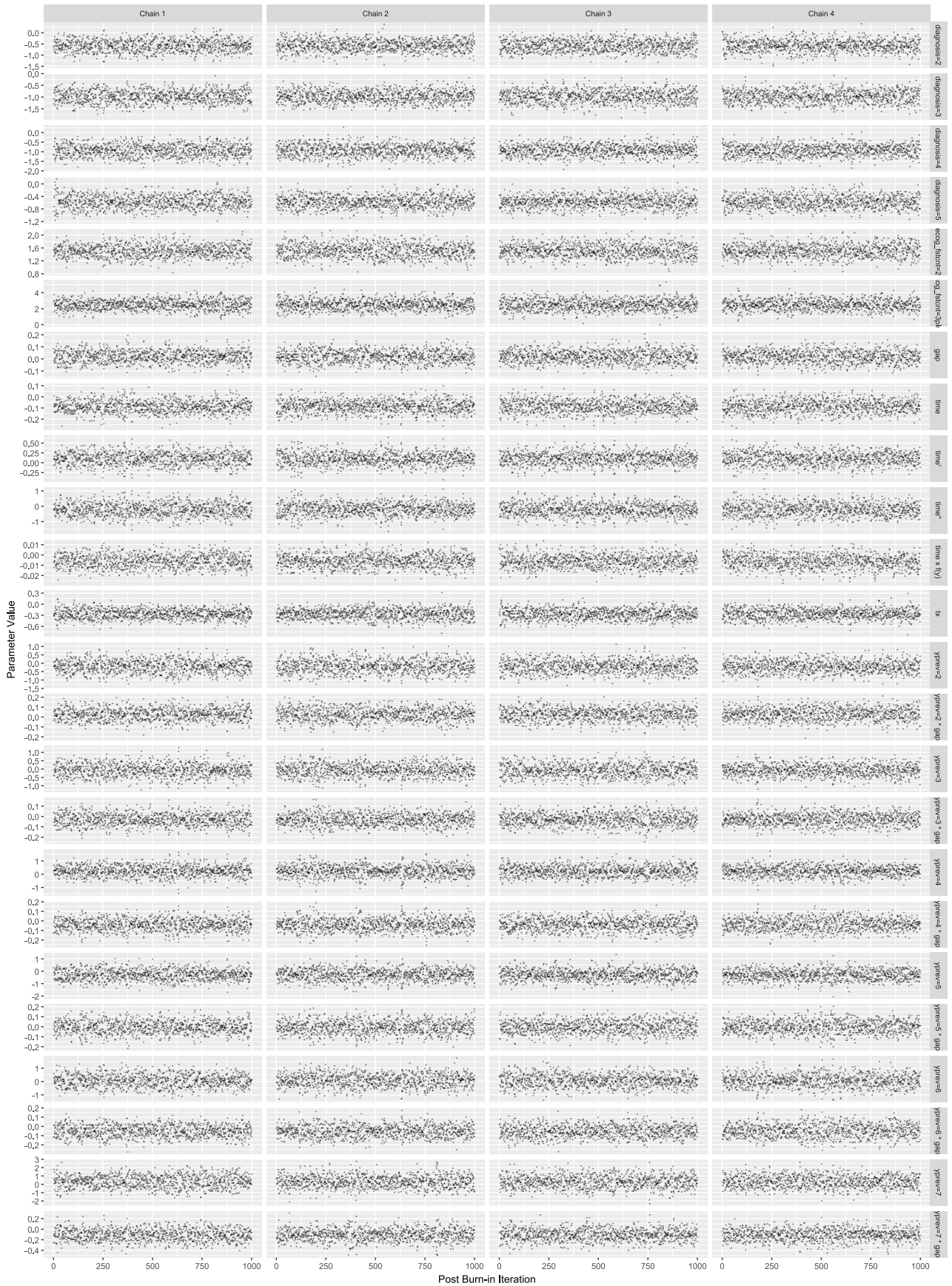

Figure 6: Chain mixing plots of each model parameter.

10.3 Appendix C: Simulating a Large Dataset of QoL Trajectories

10.3.1 Overview

To prospectively evaluate the operating characteristics of our proposed analysis method and compare it to common alternatives, we conducted a simulation study. The goal was to generate a large dataset with known properties that realistically mimics the longitudinal QoL data structure of patients with advanced cancer. This process involved two main stages: (1) fitting a model to a historical dataset to obtain realistic parameters, and (2) using these parameters to simulate large, complete longitudinal data under different treatment effect scenarios.

10.3.2 Stage 1: Deriving Realistic Parameters from Historical Data

The basis for our simulation was a historical dataset of patients with advanced cancer from the University Hospital Basel. This dataset contained 618 measurements (including death) from 259 patients, with a median gap between measurements of 9 weeks (IQR = 6-14). Our ability to estimate the serial correlation accurately is thus very limited, but likely still a better representation of reality than completely simulating data from scratch.

- **Defining an 8-level ordinal outcome**: The 7-level EORTC QoL scale was inverted so that 1 represented "Excellent" and 7 represented "Very Poor". Death was added as the 8th, absorbing state.
- **Discretizing time**: Follow-up time, originally in days, was converted to weeks. If multiple measurements occurred in the same week, the one corresponding to the worst health state was retained.
- **Creating lagged variables**: The previous QoL state (`yprev`) and the time gap in weeks since the last measurement (`gap`) were computed for each observation.

A Bayesian first-order Markov model was then fitted to this prepared historical dataset to extract plausible parameter values for the simulation. The model was specified in R using the `rmsb` package as follows:

```
model_for_simulation <- blrm(
  formula = Ocens(y.a, y.b) ~ tx + pol(week, 2) + yprev * gap +
            ecog_fstcnt + diagnosis + pat_age + gender + plan_fstcnt_coded,
  data = prepared_historical_data,
  ppo = ~week,
  cppo = function(y) y,
  iter = 2000, chains = 4
)
```

This model treats unobserved weeks as interval-censored (`Ocens(y.a, y.b)`), where the state is known to be between 1 and 7. It models time with a second-degree polynomial (`pol(week, 2)`), as this makes the output simpler to use for the generation of a new dataset compared to splines. SOPs were very similar when using a restricted cubic spline with 4-knots (not shown). The effect of `week` is allowed to be non-proportional, with the non-proportionality constrained to be linear across the outcome categories (`ppo = ~week`, `cppo = function(y) y`). The posterior medians of the coefficients from this model were used as the "true" parameters for the next stage. As shown in Table Table 4, all MCMC convergence diagnostics were well within acceptable limits.

```
Iterations: 2000 on each of 4 chains, with 4000 posterior distribution samples saved

For each parameter, n_eff is a crude measure of effective sample size
and Rhat is the potential scale reduction factor on split chains
(at convergence, Rhat=1)
```

	n_eff	Rhat
y>=2	5675	1.000
y>=3	5247	1.000
y>=4	4736	0.999
y>=5	4935	0.999
y>=6	4462	1.000
y>=7	3702	1.000
y>=8	4399	1.000
tx=1	6621	0.999
week	2952	1.000
week^2	6494	1.000
yprev=2	6120	1.000
yprev=3	4905	0.999
yprev=4	5900	1.000
yprev=5	4990	1.000
yprev=6	5155	0.999
yprev=7	6233	0.999
gap	5310	1.000
ecog_fstcnt=2	7536	1.000
ecog_fstcnt=3plus	6903	0.999
diagnosis=2	8565	1.000
diagnosis=3	6956	1.000
diagnosis=4	6745	0.999
diagnosis=5	6900	1.000
pat_age	6506	0.999
gender=male	5847	1.000
plan_fstcnt_coded=2	6996	0.999
plan_fstcnt_coded=3	6982	1.000
plan_fstcnt_coded=4	5556	1.000
plan_fstcnt_coded=5	7530	1.000
yprev=2 * gap	6281	0.999
yprev=3 * gap	5288	1.000
yprev=4 * gap	6740	0.999
yprev=5 * gap	5462	1.000
yprev=6 * gap	5727	0.999
yprev=7 * gap	6292	0.999
week x f(y)	3132	1.000

Table 4: Convergence diagnostics for the first-order Markov model fitted to the historical data and serving as a basis for the simulated datasets.

10.3.3 Stage 2: Simulating the Dataset

Using the parameters from the model (`model_for_simulation`), we simulated a large ($N = 100,000$) population with complete weekly QoL data for 26 weeks. This was accomplished using the `rmsb::simMarkovOrd` function, which generates data from a first-order Markov process. The baseline covariate distributions were preserved by sampling with replacement from the historical dataset's baseline characteristics. Details and code can be found on **[github link placeholder].**

We generated two datasets under a primary alternative hypothesis where the treatment had a beneficial effect, corresponding to a true transitional odds ratio (OR) of 0.8. We explored two scenarios:

1. **Proportional Treatment Effect**: The treatment effect (OR = 0.8 for a transition to a worse state) was applied equally to all transitions, including the transition to death.
2. **Non-Proportional Treatment Effect**: The treatment effect was applied only to transitions between living QoL states (1-7), with no effect on the risk of death.

This process resulted in two large, complete datasets where the true QoL trajectory for every patient under their assigned treatment was known. These datasets served as the basis for planning the analyses described in the main body of the SAP. The overall cumulative incidence of death in the simulated dataset is an overestimate. This is a direct consequence of the data-generating model (`model_for_simulation`), which used the same interval-censoring approach that, as we demonstrate in Section 10.5, tends to overestimate the probability of the absorbing death state when intermediate QoL states are sparse. However, in LIQPLAT the enrolled patients have on-average more severe disease than in the historical dataset used to fit `model_for_simulation`. The cumulative incidence of the simulated data of about 18% over 26 weeks is thus likely a good estimate for the LIQPLAT population. Figure 7 and Figure 8 show the SOPs over time for both simulated datasets.

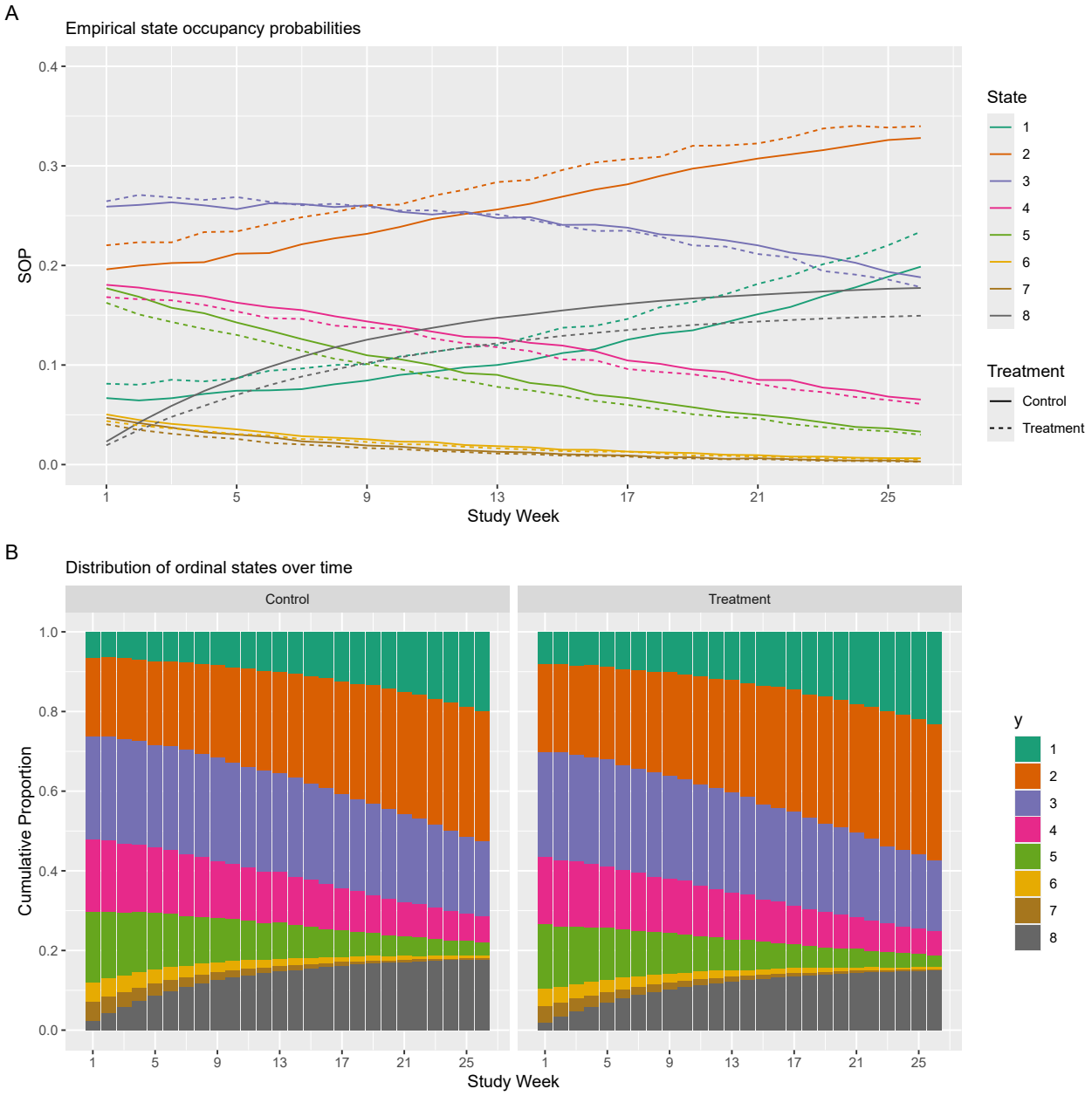

Figure 7: Empirical state occupancy probabilities (SOPs) for each ordinal state over time from the simulated dataset with a proportional treatment effect across all thresholds (OR = 0.8, N = 100′000). Panel A uses a line chart to highlight time trends for each state, panel B shows the data as a stacked bar chart to visualize the change in distribution of states over time.

In the dataset with a proportional treatment effect the difference in weeks spent with good or better quality of life is 1.1. The difference in weeks spent with bad QoL or death is −0.9.

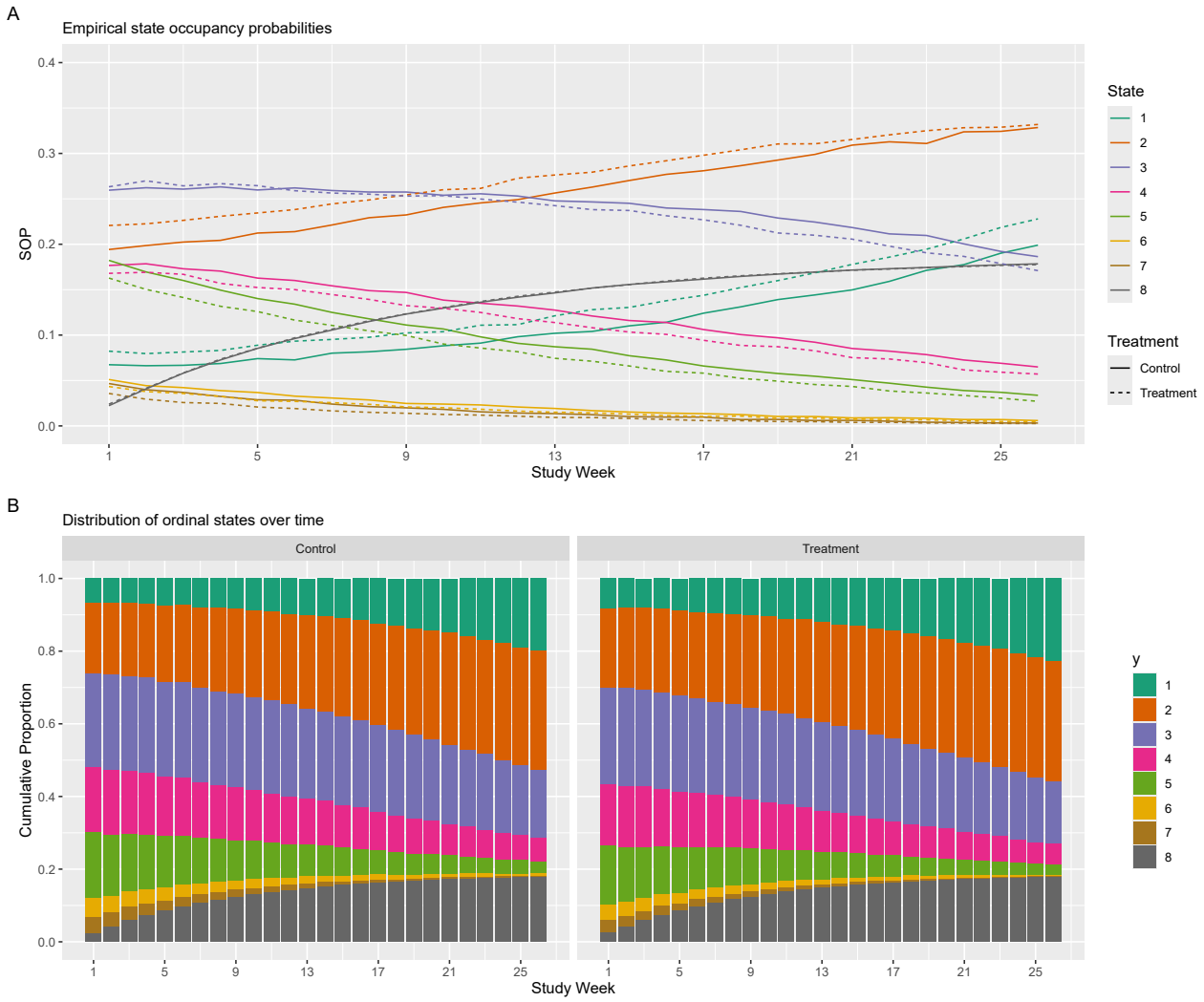

Figure 8: Empirical state occupancy probabilities (SOPs) for each ordinal state over time from the simulated dataset with a non-proportional treatment (OR = 0.8 for thresholds 1-6, and OR = 1.0 for threshold 7, N = 100′000). Panel A uses a line chart to highlight time trends for each state, panel B shows the data as a stacked bar chart to visualize the change in distribution of states over time.

In the dataset with a non-proportional treatment effect (no effect on mortality) the difference in weeks spent with good or better quality of life is 0.8. The difference in weeks spent with bad quality of life or dead is −0.5, and, as expected, also −0.5 for the difference in weeks spent with bad quality of life, when excluding participants who died.

When plotting the trajectories of a random sample of individual patients, we can see that the serial correlation appears to be very weak, i.e., patients appear to more often transition to a different state than the previous day than not.

> **Note**
>
> This very weak serial correlation might explain why the multilevel model with a random intercept performs equally well in our simulations. It's possible that our simulated dataset actually underestimates the serial correlation, because the historical data has so few observations per patient with mostly very large gaps. We have not explored this further.

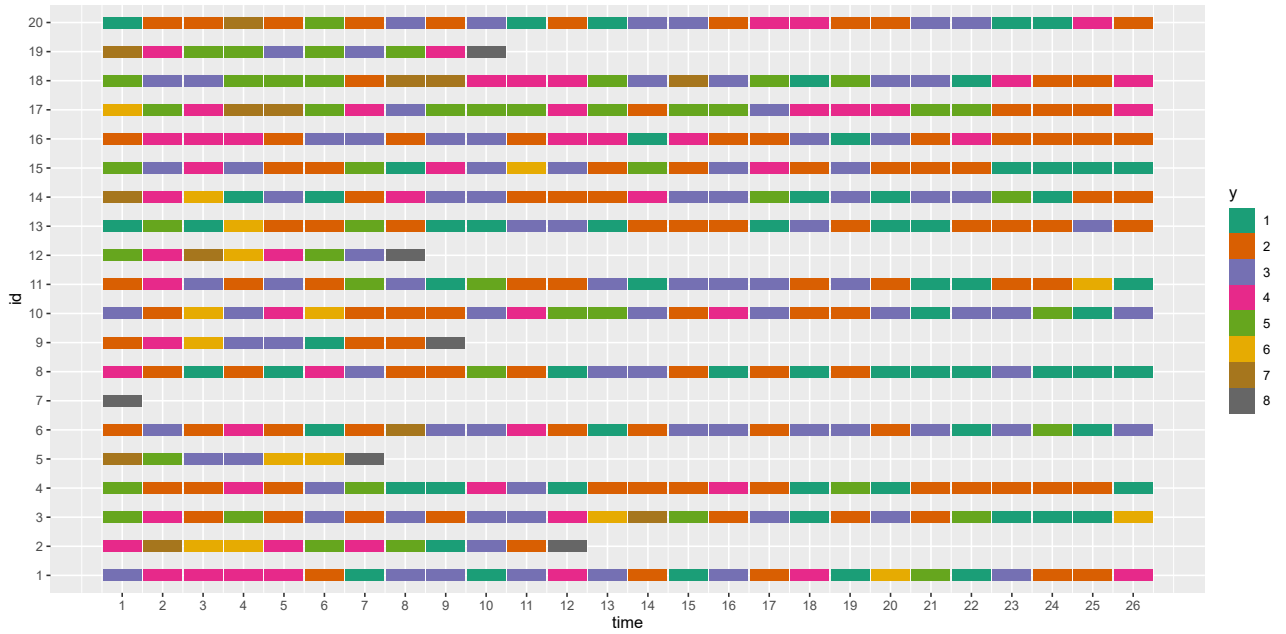

Figure 9: Daily ordinal status for a random sample of 20 patients from the simulated large dataset.

10.4 Appendix D - Simulating Operating Characteristics for Different Models

10.4.1 Aims

To evaluate and compare the operating characteristics of the first-order Markov model to other common alternative endpoints, namely an ordinal (cumulative logit) model at month 6 after baseline, a Cox proportional-hazards model for time-to-deterioration, a linear model for change from baseline, and a multilevel model with a random intercept at the patient level. We are evaluating the probability of rejecting H_0 when $\beta_{\text{treatment}} = 0$ holds based on the posterior probability of any effect larger than 95% ($P(\text{any benefit}) > 0.95$), i.e., a Bayesian type I error, and when $\beta_{\text{treatment}} \neq 0$, i.e., Bayesian power. We also investigate bias and coverage of the 95% quantile-based credible interval.

10.4.2 Research question

Which method of analyzing QoL trajectories under sparse follow-up yields the highest Bayesian power at a constant type I error rate?

10.4.2.1 Estimands

$\beta_{\text{treatment}}$ does not necessarily target an ATE and may have different posterior distributions and thus type I error and power than the estimate for the ATE from the same model. However, for feasibility reasons we did not formulate and estimate ATE for each statistical model, but used each model's treatment parameter to evaluate the type I error rate and power. The computational burden of marginalization, particularly for the multilevel and Markov models would have been too high. While the power and type I error for the ATE of each model might slightly differ for $\beta_{\text{treatment}}$, we expect this difference to be small and to not meaningfully alter our conclusions.

10.4.3 Methods

10.4.3.1 Data generating mechanisms

For each simulation iteration, we sampled patients from the large simulated dataset with a non-proportional treatment effect without replacement. For details on the data generation see Section 10.3. Under H_A, we sampled 180 patients from the treatment group and 90 patients from the control group to match the $2:1$ allocation ratio of LIQPLAT. Under H_0, 270 patients were randomly drawn only from the control population. 180 of these patients were then labeled as treated, and 90 patients were labeled as controls, thereby enforcing the null condition. Patients who died within the follow-up period were removed from the dataset. To reflect sparse follow-up, we then randomly sampled just 15% of the longitudinal QoL observations after baseline (see Figure 10).

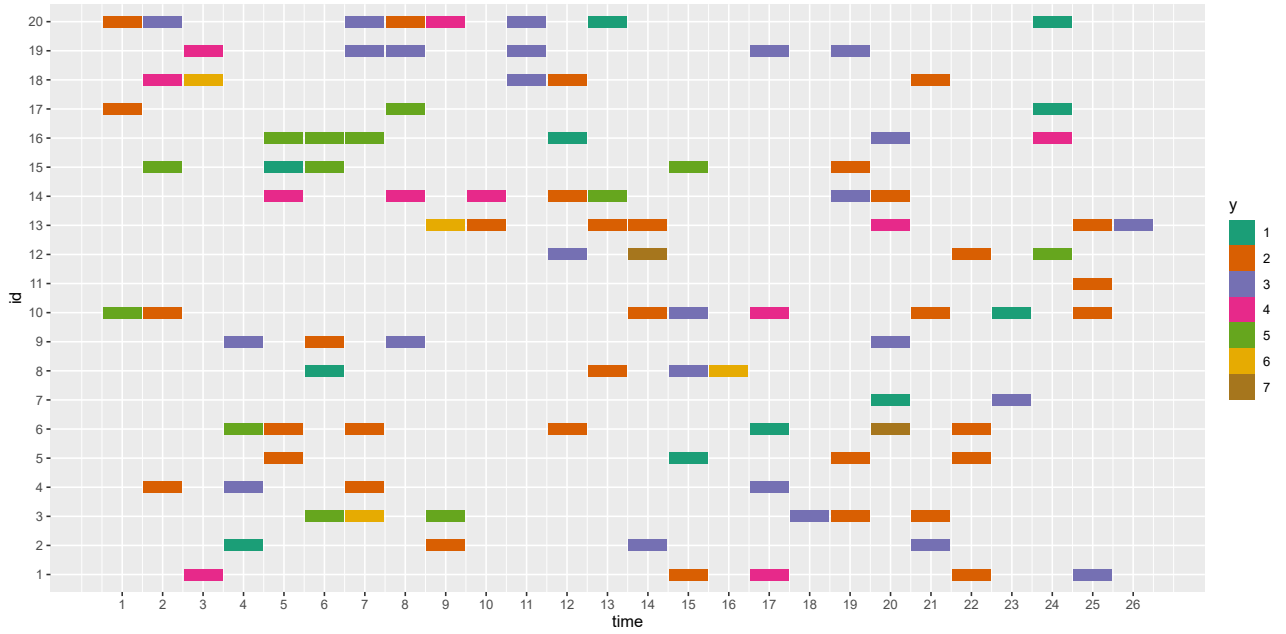

Figure 10: Tile-plot visualizing the sparse follow-up data (85% missing) for a random sample of 20 patients from the Principal Stratum of the simulated dataset with a non-proportional treatment effect.

10.4.3.2 Simulation and Analysis Pipeline

All models were fitted to the same sparsely populated datasets. Models which considered longitudinal data modeled time using a second degree polynomial as used in the data generation (see Section 10.3), and all models included the predictive baseline variables functional status at baseline (`ecog_fstcnt`) and diagnosis category (`diagnosis`). Patient age, gender, and treatment plan at baseline, which were used in generating the simulated dataset, were not included to reflect residual outcome heterogeneity.

Before the main study, a set of pilot simulations ($n_{sim} = 20$) was conducted to verify the end-to-end analysis pipeline and assess initial MCMC convergence for each of the five competing models under both H_0 and H_A scenarios. By inspecting trace plots and $\hat{R}$, these preliminary runs confirmed the stability of all models.

We use bootstrap-after-jackknife resampling to compute the Monte Carlo standard error for the performance measures of interest [14]:

$$\mathrm{MCE}_{\mathrm{JK}}\left(\hat{\theta}_{\mathrm{MC}}(\mathbf{Y})\right) = \sqrt{\frac{n_{\mathrm{sim}}-1}{n_{\mathrm{sim}}}\sum_{i=1}^{n_{\mathrm{sim}}}\left(\hat{\theta}_{\mathrm{MC}}^{n_{\mathrm{sim}}-1}(\mathbf{Y}_{-i}) - \overline{\hat{\theta}_{\mathrm{MC}}^{n_{\mathrm{sim}}-1}(\mathbf{Y})}\right)^2}$$

where

$$\overline{\hat{\theta}_{\mathrm{MC}}^{n_{\mathrm{sim}}-1}(\mathbf{Y})} := \frac{1}{n_{\mathrm{sim}}}\sum_{i=1}^{n_{\mathrm{sim}}}\hat{\theta}_{\mathrm{MC}}^{n_{\mathrm{sim}}-1}(\mathbf{Y}_{-i})$$

$\mathrm{MCE}_{\mathrm{JK}}$ was programatically implemented as shown below:

```r
jackknife_mcse <- function(estimates, statistic = mean) {

  # Number of simulation repetitions
  nsim <- length(estimates)
  # Vector to store the "leave-one-out" estimates
  leave_one_out_estimates <- numeric(nsim)
  # Step 1: Create the "leave-one-out" estimates
```

```r
  # Loop through each simulation run, remove it, and recalculate the statistic
  for (i in 1:nsim) {
    leave_one_out_estimates[i] <- statistic(estimates[-i])
  }
  # Step 2: Calculate the average of all the leave-one-out estimates
  mean_of_estimates <- mean(leave_one_out_estimates)
  # Step 3: Calculate the sum of squared differences
  sum_sq_diff <- sum((leave_one_out_estimates - mean_of_estimates)^2)
  # Step 4: Apply the final jackknife formula
  mcse <- sqrt(((nsim - 1) / nsim) * sum_sq_diff)
  return(mcse)
}
```

For the purpose of this statistical analysis plan we accept more uncertainty, a $\text{MCE}_{\text{JK}} < 0.02$ for the type I error rate and power of each method. The final number of repetitions ($n_{sim} = 1000$) for the main simulation was determined by iteratively increasing the number of iterations until the precision was reached (see Table 5).

n_{sim}	Monte Carlo SE for power based on posterior probability
200	0.0354
400	0.025
600	0.0203
800	0.0176
1000	0.0157

Table 5: Monte Carlo standard errors for different n_{sim} of the Bayesian power for the first-order Markov model-derived from the posterior distribution ($P(\text{any benefit}) > 0.95$).

10.4.3.3 Software

We use the R language and environment for statistical computing [11]. All models use the probabilistic programming language stan through either `rmsb` [8] version 1.1-2 or `brms` [12] version 2.22.11.

10.4.3.4 MCMC Settings and Convergence

To monitor convergence to the posterior, we use 4 chains for inference, with a chain length of 2000 per chain and a warmup of 1000 iterations. We evaluate the converge of each iteration based on the potential-scale-reduction factor $\hat{R}$, where we accept values smaller than 1.05.

10.4.4 Model Specifications & Results

10.4.4.1 First-order Markov Model

10.4.4.1.1 Conceptual Framework & Data Preparation

We use the same model as intended for the primary QoL analysis of the LIQPLAT trial, but use a second degree polynomial to model time to match the data-generating mechanism.

10.4.4.1.2 Statistical Model

Let y_{it} be the ordinal QoL state (from 1 to 7, where 1 is best) for patient i at week t. Let $y_{it'}$ be their last observed state at a prior week t', and let the time gap be $\Delta t = t - t'$. The model is specified as a cumulative logit model for the transition probabilities:

$$
\begin{aligned}
\text{logit}\ \left(P(y_{it} \geq j \mid y_{it'})\right) &= \alpha_j - \left(\eta_{it} + \gamma_{it,j}\right) \\
\eta_{it} &= \beta_{\text{tx}} \cdot \text{Treatment}_i + f(t) + \sum_{k=2}^{7} \beta_{\text{yprev}\,=k} \cdot \mathbb{I}(y_{it'} = k) \\
&\quad + \beta_{\text{gap}} \cdot \Delta t + \sum_{k=2}^{7} \beta_{\text{yprev}\,=k\times\ \text{gap}} \cdot \mathbb{I}(y_{it'} = k) \cdot \Delta t + \mathbf{X}_i \beta_{\text{covars}} \\
\gamma_{it,j} &= (\tau \cdot t) \cdot j \\
\alpha &\sim \text{Dirichlet}\ (0.308) \\
\beta_k, &\sim \text{Normal}\ (0, 100) \\
\tau &\sim \text{Normal}\ (0, 100)
\end{aligned}
$$

Where:

- α_j are the category-specific intercepts (cutpoint) for $j = 1, \ldots, 6$. The prior is induced by a Dirichlet distribution on the baseline cell probabilities with a concentration parameter 0.308.
- η_{it} is the main linear predictor for effects assumed to satisfy the proportional odds assumption (i.e., their effect is constant across the $j - 1$ cumulative logits).
- $\gamma_{it,j}$ models a deviation from the proportional odds assumption. We allow the effect of time to be non-proportional, but constrain the effect to be linear in the outcome category j.
- β_{tx} is the effect for treatment
- $f(t)$ is a polynomial function of time (`poly(time, 2)`).
- $\sum_{k=2}^{7} \beta_{\text{yprev}\,=k} \cdot I(y_{it'} = k)$ is the effect of the previous QoL state, modeled as a categorical variable with 6 parameters relative to state 1 as the reference category. $I(\cdot)$ is an indicator function.
- β_{gap}: The linear effect of the time gap since the last measurement.
- $\sum_{k=2}^{7} \beta_{\text{yprev}\,=k\times\ \text{gap}} \cdot I(y_{it'} = k) \cdot \Delta t$ is an interaction term allowing the effect of the patient's previous QoL state to differ depending on the gap.
- $\mathbf{X}_i \beta_{\text{covars}}$ represent the effect for baseline covariates (`ecog_fstcnt`, `diagnosis`)
- $\gamma_{it,j}$ models a deviation from the proportional odds assumption. We allow the effect of time to be non-proportional, but constrain the effect to be linear in the outcome category j. This means the odds ratio for t can change linearly across the different cutpoints j of the QoL scale.

10.4.4.1.3 R Code

```
 model <- blrm(
                formula = y ~ tx + pol(time, 2) + yprev * gap + ecog_fstcnt +
diagnosis,
                data = data_for_model,
                ppo = ~time,
                cppo = function(y) y,
                iter = 2000,
                chains = 4,
                method = "sampling",
            )
```

10.4.4.1.4 Convergence Diagnostics

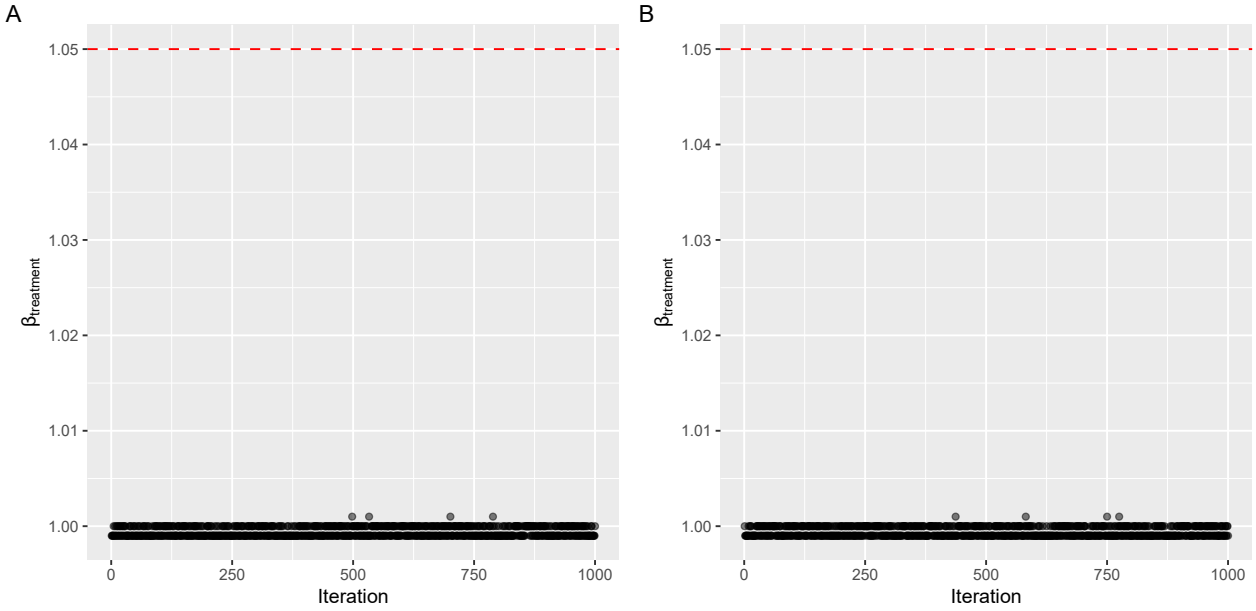

Figure 11: $\hat{R}$ values for $\beta_{treatment}$ across all iterations of the first-order Markov model along with the $\hat{R} = 1.05$ threshold. Under H_0 (left) and H_A (right).

10.4.4.1.5 Operating Characteristics

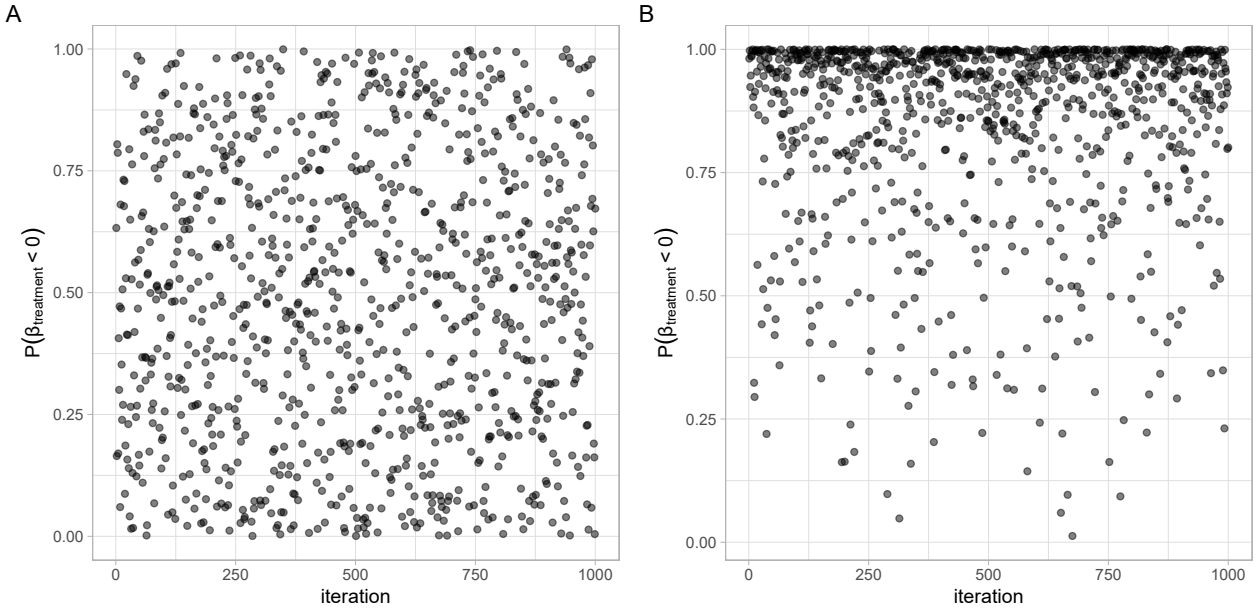

Figure 12: $P(\text{benefit})$ for each iteration of the first-order Markov model, under no effect ($H_0; OR = 1$) (left) and a treatment effect ($H_A; OR = 0.8$) (right).

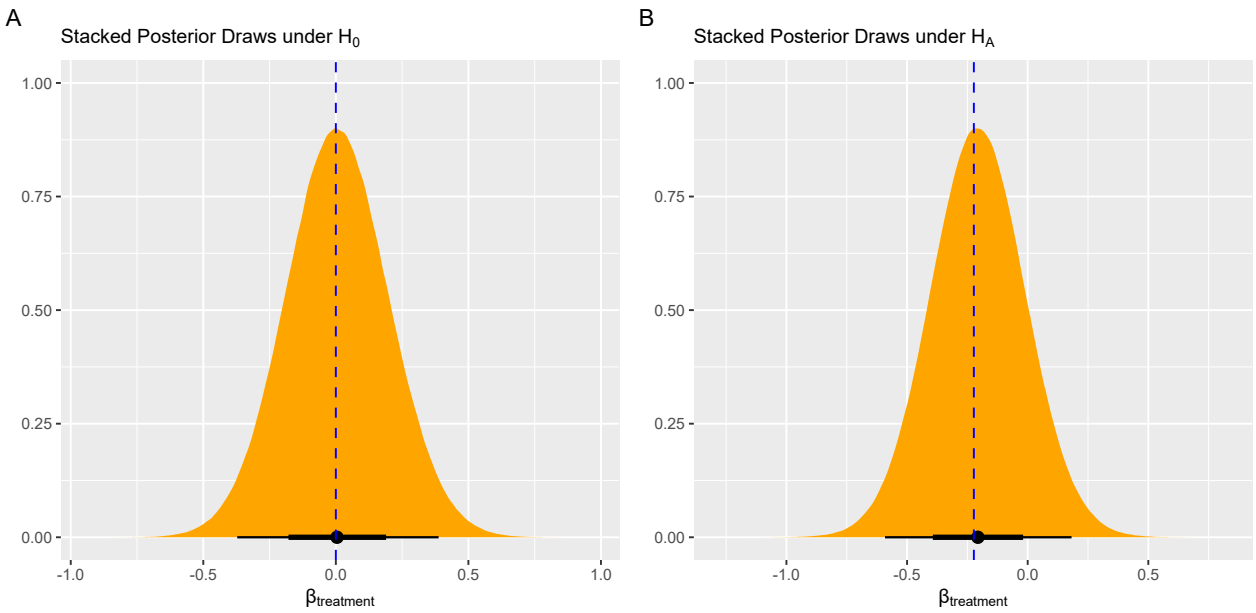

Figure 13: Stacked posterior draws for the treatment parameter (η_{tx}) from all iterations of the first-order Markov model, under no effect ($H_0; OR = 1$) (left) and a treatment effect ($H_A; OR = 0.8$) (right). The blue dotted line indicates the true parameter value ($\log(0.8) \approx -0.223$).

10.4.4.2 Cumulative Logit Multilevel Model with a Random Intercept

> **Note**
>
> I have not yet run all 1000 simulation for this model.

10.4.4.2.1 Conceptual Framework & Data Preparation

This approach models the longitudinal ordinal QoL data using a cumulative logit model that includes a patient-specific random intercept. This accounts for within-patient correlation by assuming each patient has a unique baseline QoL state around which their states fluctuate over time. This implies a compound symmetry correlation structure, where the correlation between any two measurements on the same patient is assumed to be constant, regardless of the time gap.

10.4.4.2.2 Statistical Model

Let y_{it} be the ordinal QoL state (from 1 to 7) for patient i at week t. The model is a Bayesian multilevel cumulative logit model. Following the `brms` parameterization, the probability of being in category j or lower is modeled as:

$$
\begin{aligned}
\text{logit}\ (P(y_{it} \leq j)) &= \alpha_j - \eta_{it} \\
\eta_{it} &= u_{0i} + \beta_{\text{tx}} \cdot \text{Treatment}_i + f(t) + \mathbf{X}_i \beta_{\text{covars}} \\
\alpha_1 &\sim \text{Normal}\ (-1.0675705, 1) \\
\alpha_2 &\sim \text{Normal}\ (-0.5659488, 1) \\
\alpha_3 &\sim \text{Normal}\ (-0.1800124, 1) \\
\alpha_4 &\sim \text{Normal}\ (0.1800124, 1) \\
\alpha_5 &\sim \text{Normal}\ (0.5659488, 1) \\
\alpha_6 &\sim \text{Normal}\ (1.0675705, 1) \\
\beta_k &\sim \text{Uniform}\ (-\infty, \infty) \\
u_{0i} &\sim \text{Normal}\ (0, \sigma_u^2) \\
\sigma_u &\sim \text{Student-t}^+(3, 0, 2.5)
\end{aligned}
$$

Where:

- α_j are the category-specific intercepts (cutpoints) for $j = 1, ..., 6$.
- η_{it} is the linear predictor for patient i at week t.
- u_{0i} is the random intercept for patient i
- β_{tx} is the fixed effect for treatment.
- $f(t)$ is a polynomial function of time (`poly(time, 2)`).
- $\mathbf{X}_i \beta_{\text{covars}}$ represents the fixed effects for baseline covariates (`ecog_fstcnt`, `diagnosis`).

The priors for a_j correspond to the probit-transformed cumulative probabilities of a uniform distribution over the 7 QoL categories. We chose this specification as we are not aware of a method to implement the Dirichlet prior on baseline cell probabilities within the `brms` framework in a manner analogous to the `rmsb` package. This approach provides an alternative way to specify weakly informative, equispaced priors for the model's cutpoints.

10.4.4.2.3 R Code

The model was implemented using the `brms` package:

```
model_ri <- brm(
  y ~ 1 + tx + poly(time, 2) + ecog_fstcnt + diagnosis + (1|id),
  data = data_for_model,
  family = cumulative(logit),
  prior = c(prior(normal(-1.0675705, 1), class = Intercept, coef = 1),
            prior(normal(-0.5659488, 1), class = Intercept, coef = 2),
            prior(normal(-0.1800124, 1), class = Intercept, coef = 3),
            prior(normal( 0.1800124, 1), class = Intercept, coef = 4),
            prior(normal( 0.5659488, 1), class = Intercept, coef = 5),
            prior(normal( 1.0675705, 1), class = Intercept, coef = 6)),
  cores = 4,
  seed = 123,
  init_r = 0.2,
  iter = 2000
)
```

10.4.4.2.4 Convergence Diagnostics

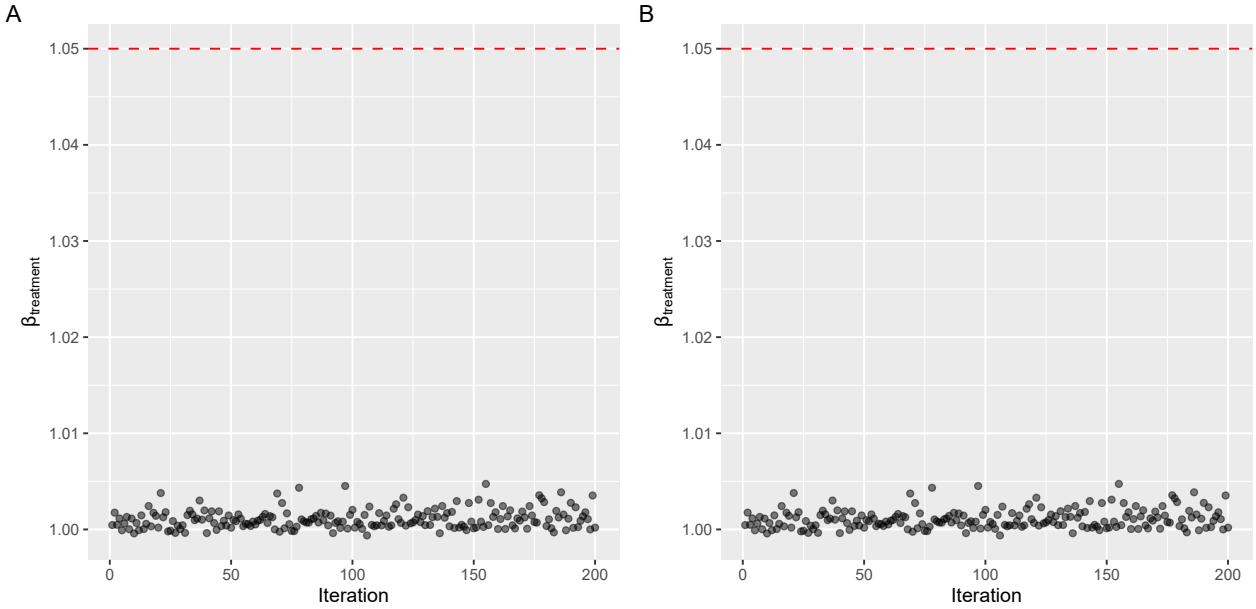

Figure 14: $\hat{R}$ values for $\beta_{treatment}$ across all iterations of the multilevel model with a random intercept at the patient level along with the threshold $\hat{R} = 1.05$. Under H_0 (left) and H_A (right).

10.4.4.2.5 Operating Characteristics

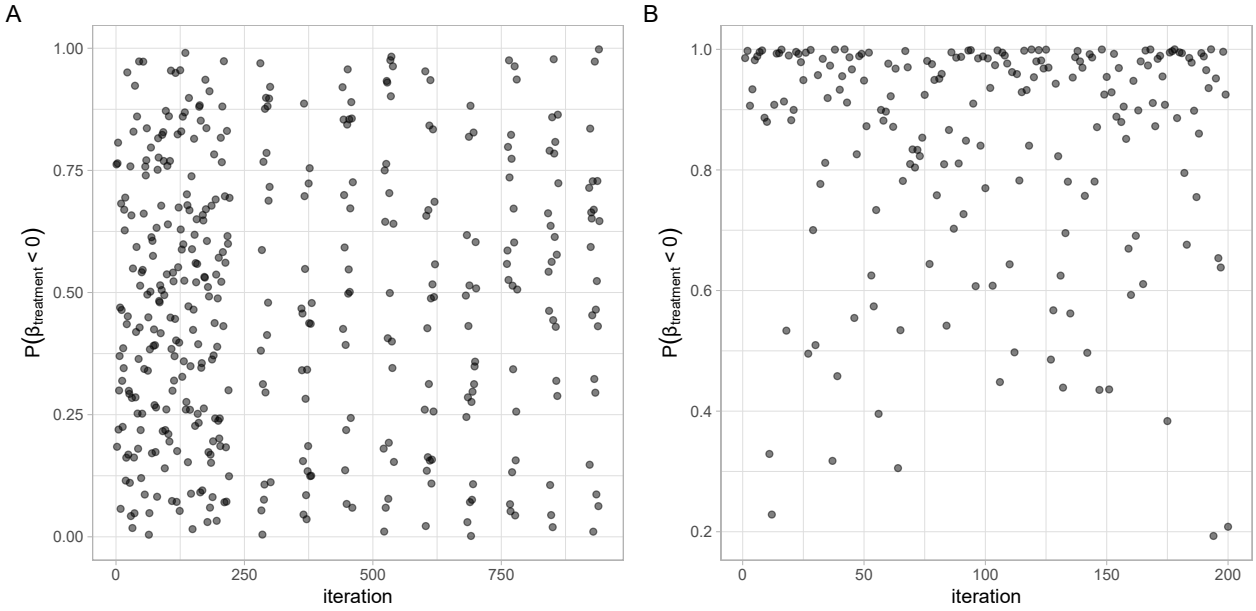

Figure 15: $P(\text{benefit})$ for each iteration of the multilevel model with a random intercept at the patient level, under no effect ($H_0; OR = 1$) (left) and a treatment effect ($H_A; OR = 0.8$) (right).

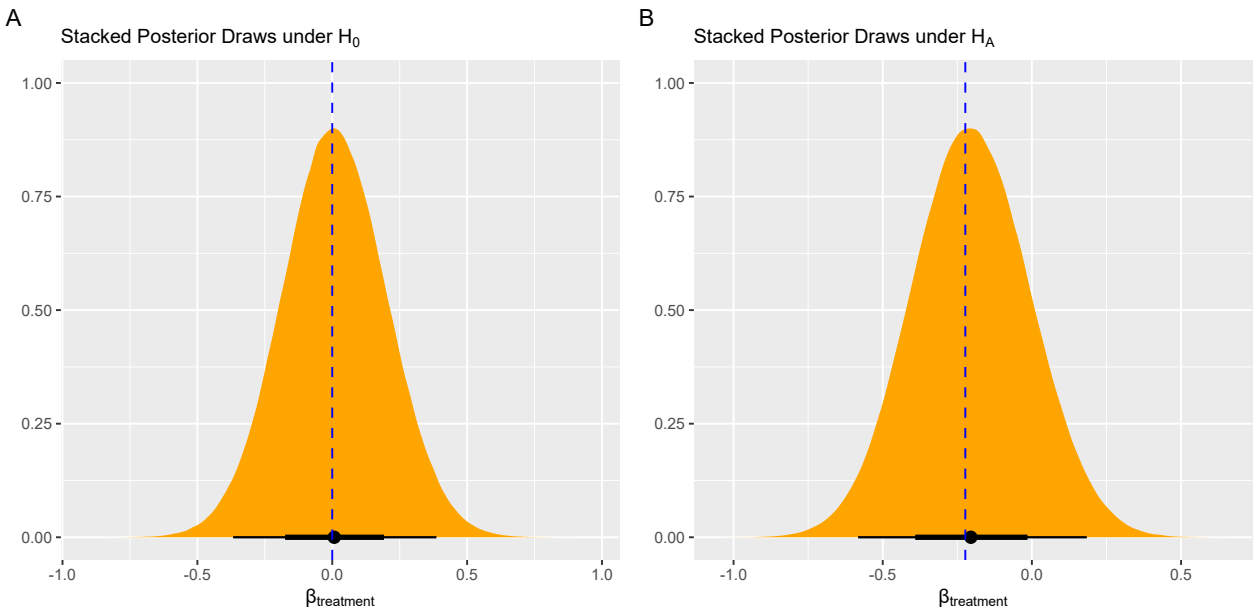

Figure 16: Stacked posterior draws for the treatment parameter (η_{tx}) from all iterations with the multilevel model with a random intercept at the patient level, under no effect ($H_0; OR = 1$) (left) and a treatment effect ($H_A; OR = 0.8$) (right). The blue dotted line indicates the true parameter value ($\log(0.8) \approx -0.223$).

10.4.4.3 Time-to-Deterioration Model

10.4.4.3.1 Conceptual Framework & Data Preparation

This model reframes the QoL analysis as a time-to-event problem. The outcome is the time until the first clinically significant deterioration in QoL.

The data was prepared by identifying, for each patient, their baseline QoL state. "Deterioration" was defined as an increase of 1 or more points on the 7-point QoL scale compared to baseline. The event time was the week of the first such deterioration. Patients who did not experience deterioration by week 26 were right-censored at that time.

10.4.4.3.2 Statistical Model

A Cox proportional-hazards model was used to model the hazard of deterioration.

$$
\begin{aligned}
h(t \mid \mathbf{X}_i) &= h_0(t) \exp\big(\beta_0 + \beta_{\text{tx}} \cdot \text{Treatment}_i + \mathbf{X}_{i,\ \text{covars}} \beta_{\text{covars}}\big) \\
\beta_0 &\sim \text{Student-t}\ (3, 3.3, 2.5) \\
\beta_k &\sim \text{Uniform}\ (-\infty, \infty) \quad \text{for} \quad k > 0 \\
h_0(t) &\sim \text{Dirichlet}\ (1) \quad (\text{on baseline hazard increments})
\end{aligned}
$$

Where:

- $h(t \mid \mathbf{X}_i)$ is the hazard of deterioration for patient i at week t.
- $h_0(t)$ is the baseline hazard function.
- β_0 is the model intercept.
- β_{tx} is the log-hazard ratio for the treatment effect.
- $\mathbf{X}_{i,\ \text{covars}} \beta_{\text{covars}}$ represents the effects of baseline covariates (`ecog_fstcnt`, `diagnosis`).

10.4.4.3.3 R Code

The model was implemented using the `brms` package with a `cox` family:

```
model_ttd <- brm(
  formula = time | cens(censoring) ~ tx + ecog_fstcnt + diagnosis,
  data = data_for_model,
  family = cox(),
  iter = 2000, chains = 4, seed = 123
)
```

10.4.4.3.4 Convergence Diagnostics

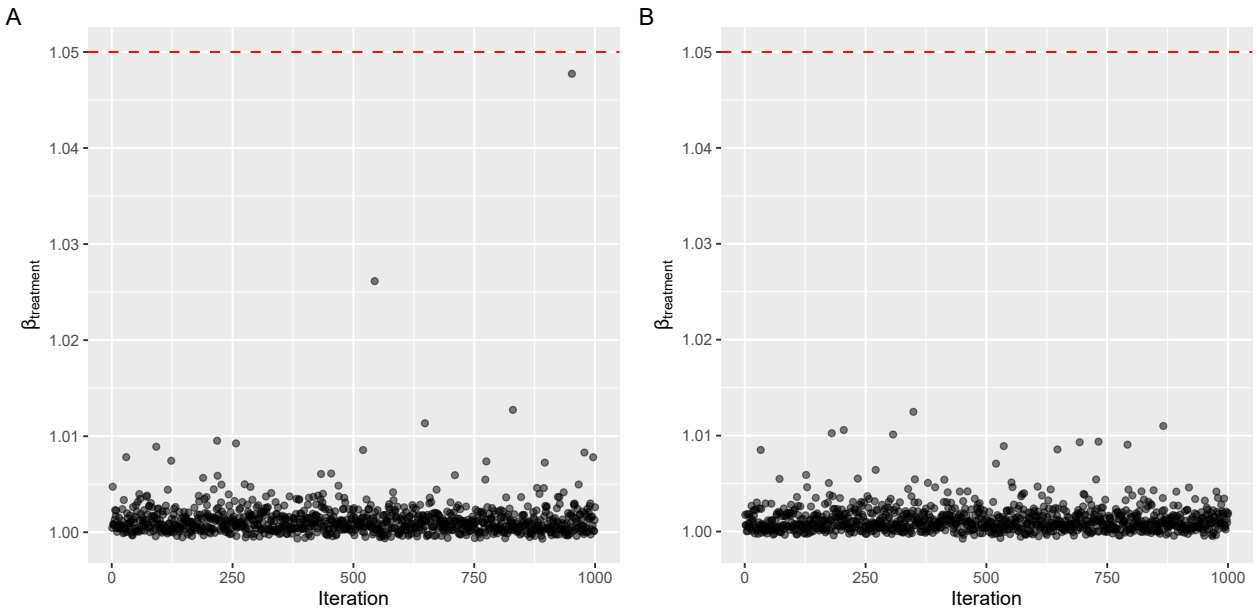

Figure 17: $\hat{R}$ values for $\beta_{treatment}$ across all iterations of the Cox proportional-hazards model for time-to-deterioration, under no effect (H_0; $HR = 1$) (left) and a treatment effect (H_A; $HR \sim 0.9$) (right).

10.4.4.3.5 Operating Characteristics

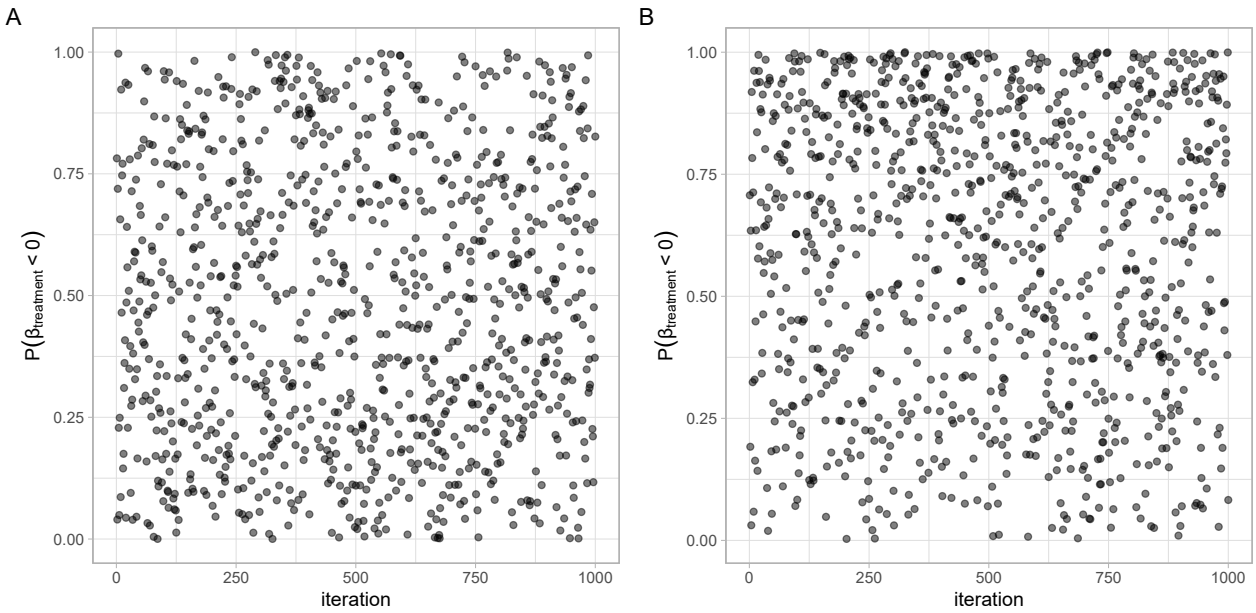

Figure 18: $P(\text{benefit})$ for each iteration of the Cox proportional-hazards model for time-to-deterioration, under no effect (H_0; $HR = 1$) (left) and a treatment effect (H_A; $HR \sim 0.9$) (right).

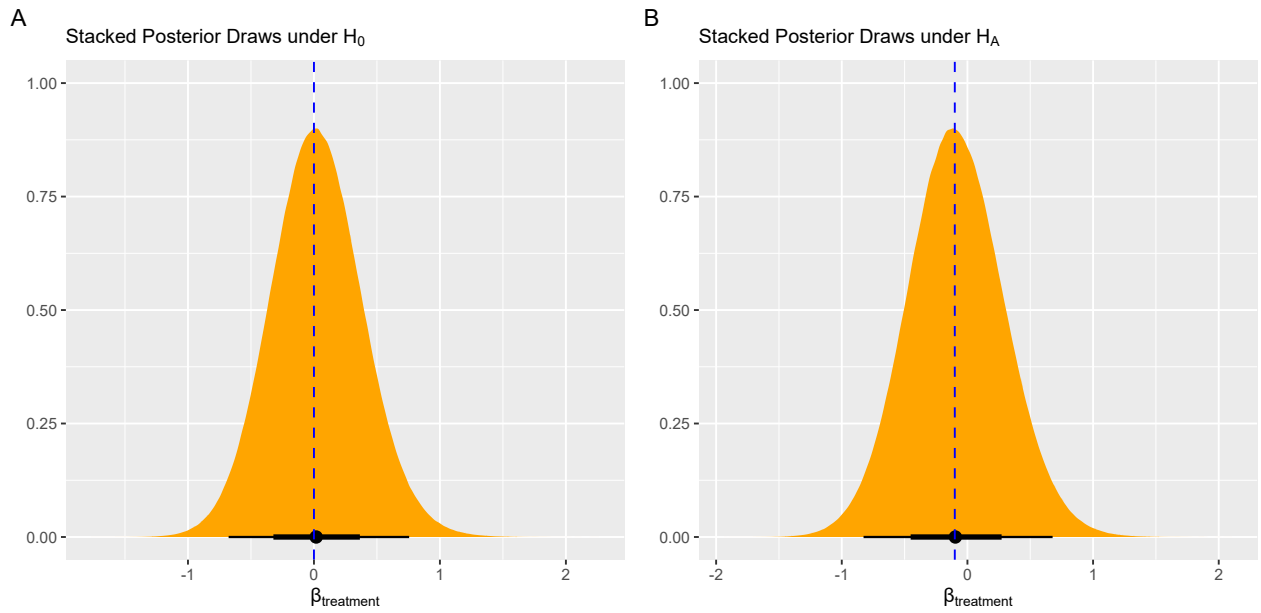

Figure 19: Stacked posterior draws for the treatment parameter (β_{tx}) from all simulations using a Cox proportional-hazards model for time-to-deterioration, under no effect ($H_0; HR = 1$) (left) and a treatment effect ($H_A; HR \sim 0.9$) (right). The blue dotted line indicates the true parameter value. The 'true' HR under H_A was approximated by fitting the same model to the full dataset of 100, 000 patients, restricted to 'always-survivors.'

10.4.4.4 Change from Baseline Model

10.4.4.4.1 Conceptual Framework & Data Preparation

This approach treats the ordinal QoL state as a continuous variable and analyzes the change from baseline to a single follow-up timepoint.

For each patient, the QoL state ($y_{i,\text{ baseline}}$) at baseline was recorded. The outcome was the QoL state from the single observation closest to 26 weeks and after week 14 $y_{i,\text{ w-26}}$. Consequently, patients without a measurement between week 14 and 26 were dropped. The analysis model then used the change score ($\Delta y_i = y_{i,\text{ week=26}} - y_{i,\text{ baseline}}$) as the dependent variable in a linear regression, adjusting for the baseline state.

10.4.4.4.2 Statistical Model

A Bayesian linear regression model was fitted to the change score.

$$
\begin{aligned}
\Delta y_i &\sim \text{Normal}\ (\mu_i, \sigma_\epsilon^2) \\
\mu_i &= \beta_0 + \beta_{\text{tx}} \cdot \text{Treatment}_i + \beta_{\text{baseline}} \cdot y_{i,\text{ baseline}} + \mathbf{X}_{i,\text{ covars}} \beta_{\text{covars}} \\
\beta_0 &\sim \text{Student-t}\ (3, 2.5, 2.5) \\
\beta_k &\sim \text{Uniform}\ (-\infty, \infty) \quad \text{for} \quad k > 0 \\
\sigma_\epsilon &\sim \text{Student-t}^+(3, 0, 2.5)
\end{aligned}
$$

Where:

- Δy_i is the change in QoL score for patient i.
- μ_i is the expected change.
- $y_{i,\text{ baseline}}$ is the numeric QoL score at baseline.
- $\mathbf{X}_{i,\text{ covars}}$ includes `ecog_fstcnt` and `diagnosis`.

10.4.4.4.3 R Code

The model was implemented using the `brms` package:

```
model_cfb <- brm(
  formula = y_delta ~ tx + ecog_fstcnt + diagnosis + ybaseline,
  data = data_for_model,
  family = gaussian(),
  iter = 2000, chains = 4, seed = 123
)
```

10.4.4.4.4 Convergence Diagnostics

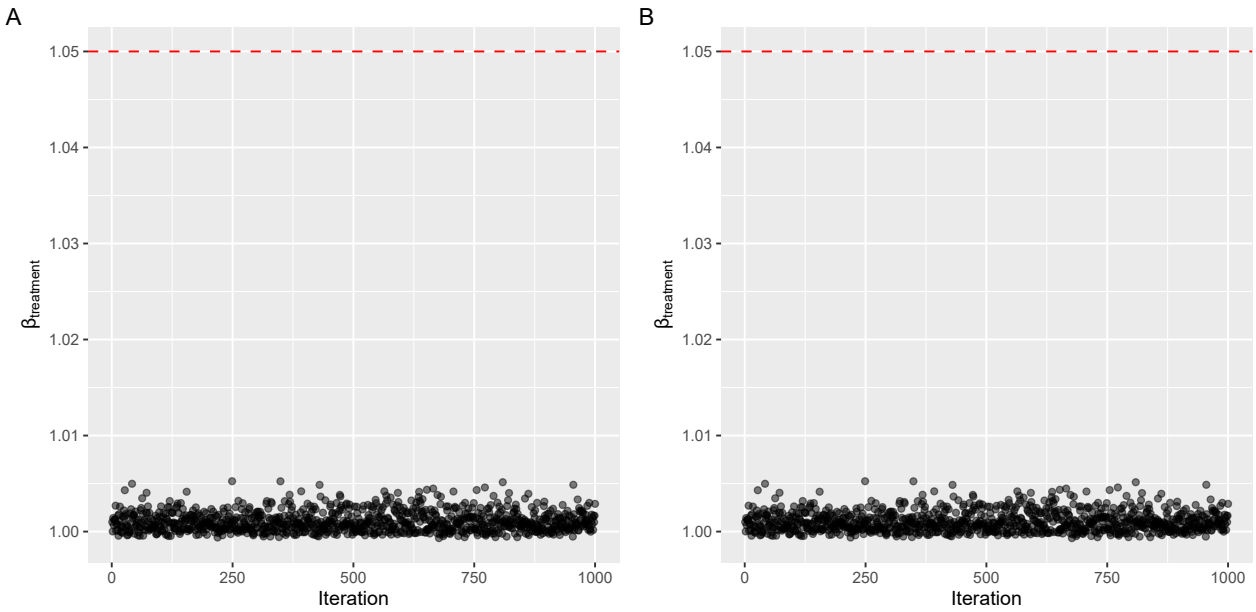

Figure 20: $\hat{R}$ values for $\beta_{treatment}$ across all iterations of the linear model for change from baseline, under no effect ($H_0; \beta_{tx} = 0$) (panel A) and a treatment effect ($H_A; \beta_{tx} \sim -0.09$) (panel B).

10.4.4.4.5 Operating Characteristics

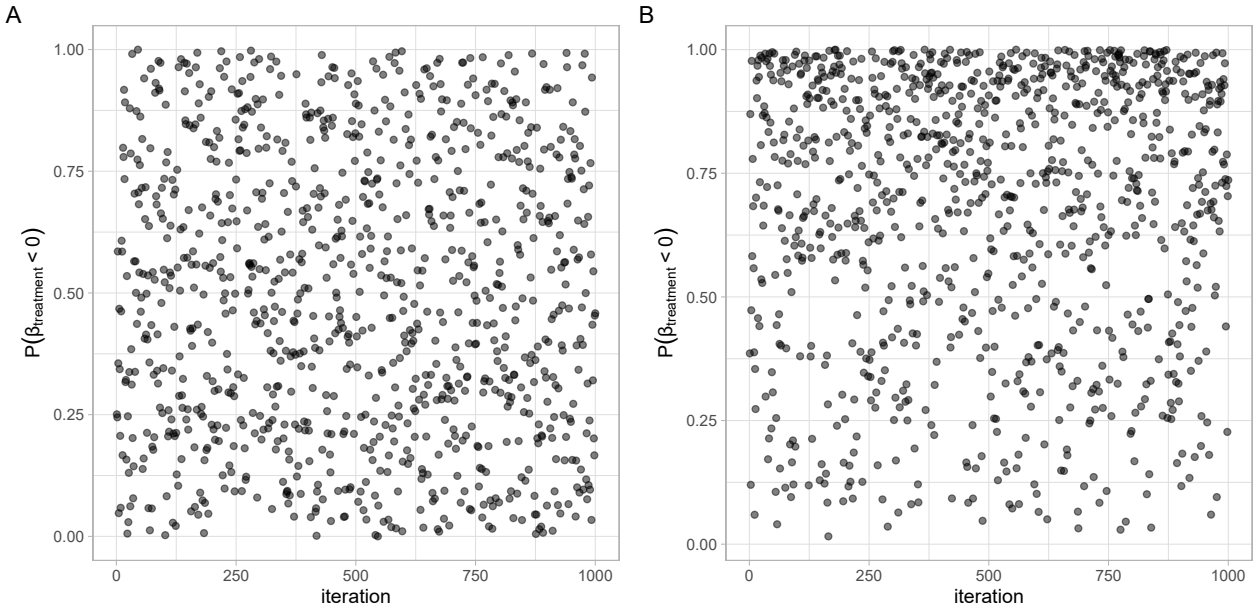

Figure 21: $P(\text{benefit})$ for each iteration of the linear model for change from baseline, under no effect $(H_0; \beta_{tx} = 0)$ (panel A) and a treatment effect $(H_A; \beta_{tx} \sim -0.09)$ (panel B).

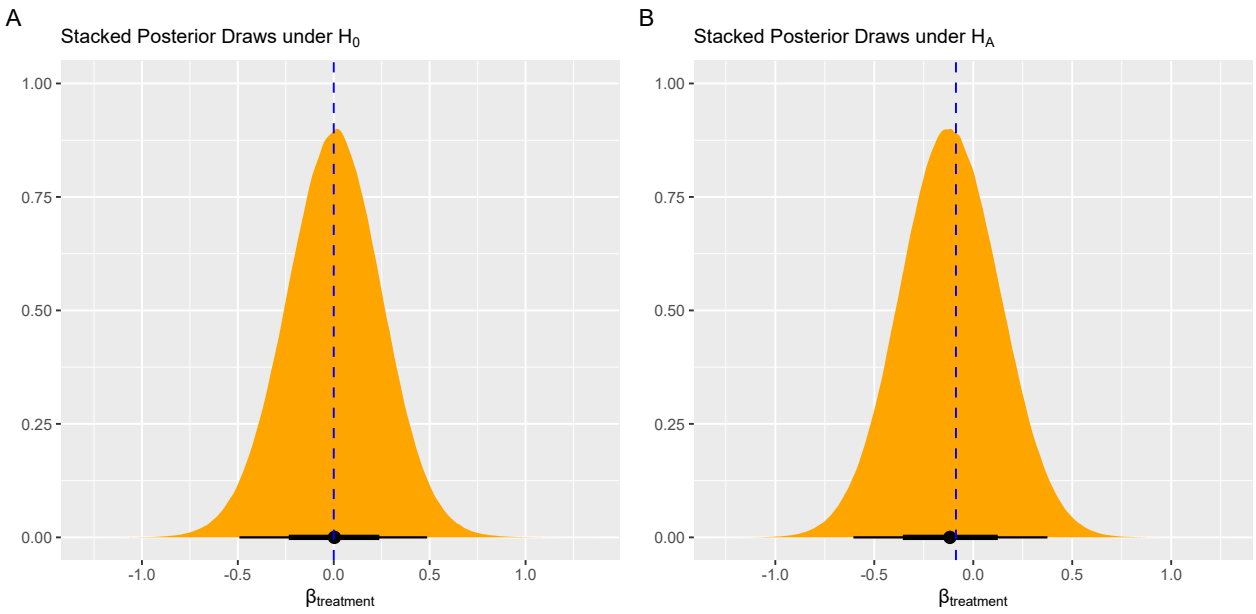

Figure 22: Stacked posterior draws for the treatment parameter (β_{tx}) from all simulations using the linear model for change from baseline, under no effect $(H_0; \beta_{tx} = 0)$ (panel A) and a treatment effect $(H_A; \beta_{tx} \sim -0.09)$ (panel B). The blue dotted line indicates the true parameter value. The 'true' difference in change from baseline under H_A was approximated by manual calculation using the full dataset of 100,000 patients, restricted to 'always-survivors.'

10.4.4.5 Cumulative Logit Model at Month 6

10.4.4.5.1 Conceptual Framework & Data Preparation

This model ignores the longitudinal nature of the data and performs a cumulative logit regression using only the QoL data from the 6-month timepoint.

For each patient, the single QoL state closest to 26 weeks but after week 14 was selected. Consequently, patients without a measurement between week 14 and 26 were dropped. In our simulation this does not bias the estimate of β_{tx}, as the treatment effect in the simulated datasets is constant over time. However, in LIQPLAT and many trials we would not expect a constant treatment effect (if it exists), which would lead to bias under H_A when treating observations from before time t as if they had occurred at time t. On the other, if the chosen accepted range is small, more participants will be dropped and the power to detect a difference between groups will decrease. Simulations to illustrate the impact are warranted. The baseline QoL state and functional status at baseline and diagnosis category were included as predictors.

10.4.4.5.2 Statistical Model

A Bayesian proportional odds logistic regression model was fitted to the 6-month QoL outcome.

$$\begin{aligned}
\text{logit}\ (P(y_i \geq j)) &= \alpha_j - \left(\beta_{\text{tx}} \cdot \text{Treatment}_i + \beta_{\text{baseline}} \cdot y_{i,\ \text{baseline}} + \mathbf{X}_{i,\ \text{covars}} \beta_{\text{covars}}\right) \\
\alpha &\sim \text{Induced by Dirichlet}\ (0.308) \\
\beta_k &\sim \text{Normal}\ (0, 100)
\end{aligned}$$

Where:

- y_i is the ordinal QoL state for patient i at the 6-month timepoint.

- α is the vector of ordered intercepts (cutpoints) for $j = 2, \ldots, 7$. The prior is induced by a Dirichlet distribution on the baseline cell probabilities with a concentration parameter 0.308.
- The model adjust for baseline QoL (`ybaseline`), and other covariates (`ecog_fstcnt`, `diagnosis`).

10.4.4.5.3 R Code

The model was implemented using the `blrm` function from the `rmsb` package:

```
model_6m <- blrm(
  formula = y ~ tx + ybaseline + ecog_fstcnt + diagnosis,
  data = data_for_model,
  iter = 2000, chains = 4, seed = 123
)
```

10.4.4.5.4 Convergence Diagnostics

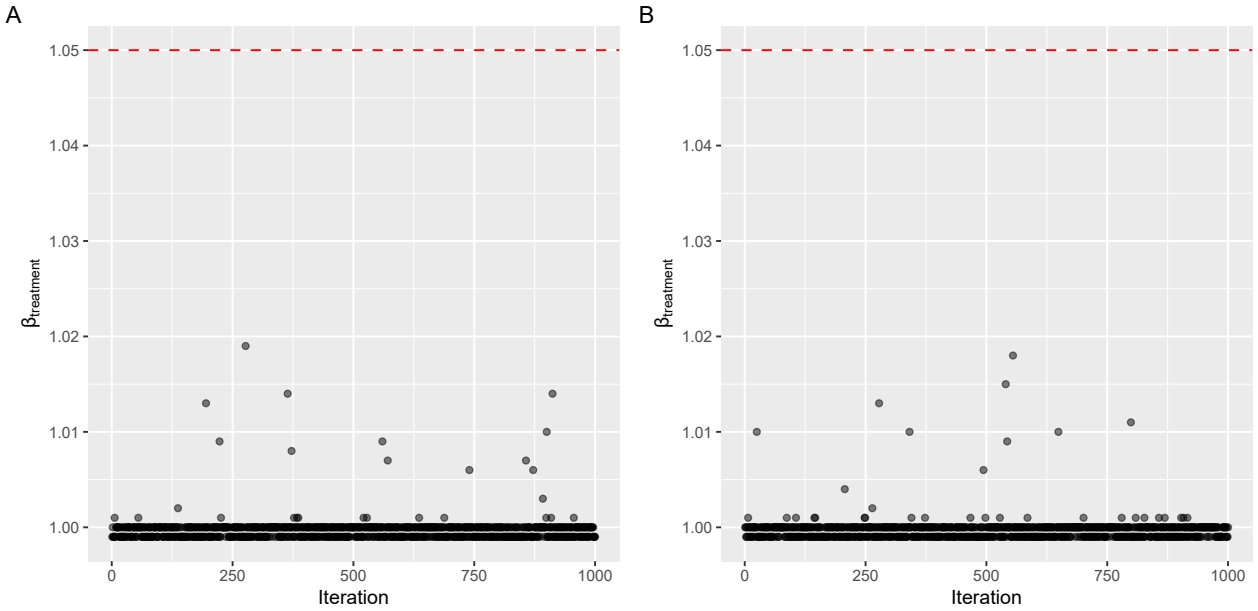

Figure 23: $\hat{R}$ values for $\beta_{treatment}$ across all iterations of the cumulative logit model at month 6 after baseline, under no effect (H_0; OR = 1) (left) and a treatment effect (H_A; OR = 0.8) (right).

10.4.4.5.5 Operating Characteristics

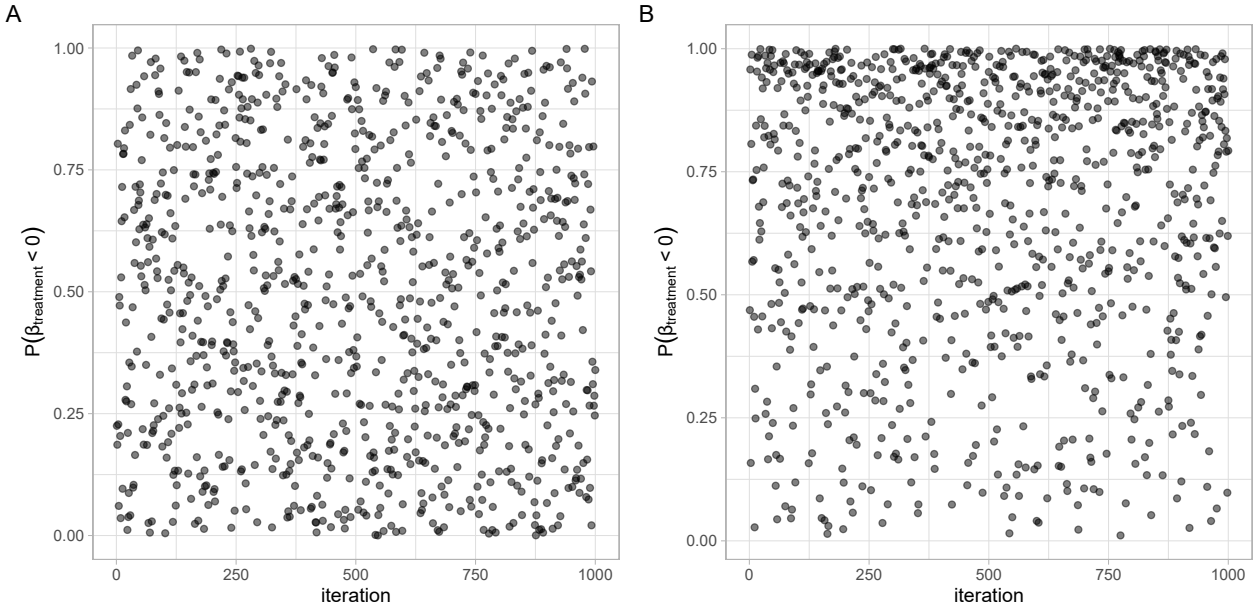

Figure 24: $P(\text{benefit})$ for each iteration of the cumulative logit model at month 6 after baseline, under no effect (H_0; OR = 1) (left) and a treatment effect (H_A; OR = 0.8) (right).

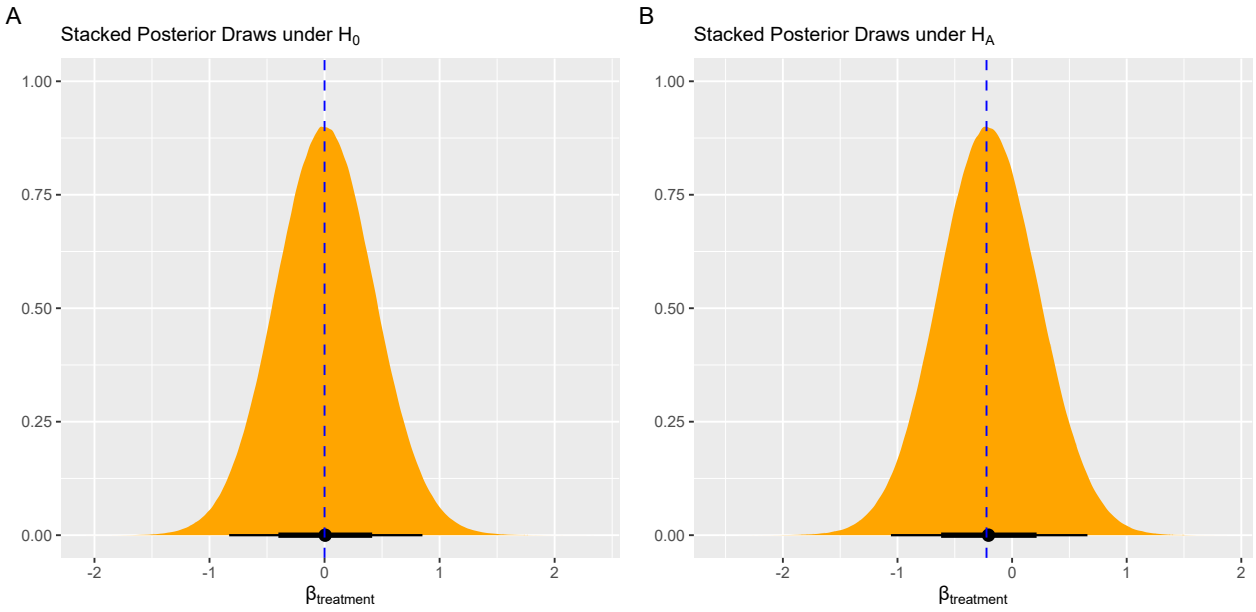

Figure 25: Stacked posterior draws for the treatment parameter (β_{tx}) from all iterations using the cumulative logit model at month 6 after baseline, under no effect (H_0; OR = 1) (left) and a treatment effect (H_A; OR = 0.8) (right). The blue dotted line indicates the true parameter value ($\log(0.8) \approx -0.223$).

10.4.4.6 Convergence Summary

Across all $n_{sim} = 1000$ repetitions, every model successfully converged for every run. There were not instances where $\hat{R}$ for the treatment parameter exceeded the pre-specified threshold of 1.05.

10.4.4.7 Key Findings

Model	Absolute Bias (H_A)	Coverage (H_A)	Absolute Bias (H_0)	Coverage (H_0)	Power	Type I Error
Markov Model	0.017 (0.0045)	0.934 (0.0079)	0.005 (0.0044)	0.938 (0.0076)	0.450 (0.0157)	0.054 (0.0072)
Random Intercept	0.020 (0.0097)	0.975 (0.0111)	0.008 (0.0065)	0.963 (0.0093)	0.445 (0.0352)	0.044 (0.0101)
Time-to-Deterioration	0.008 (0.0088)	0.934 (0.0079)	0.019 (0.0081)	0.954 (0.0066)	0.112 (0.0100)	0.047 (0.0067)
6-Month Ordinal	0.020 (0.0101)	0.927 (0.0082)	0.006 (0.0097)	0.942 (0.0074)	0.197 (0.0126)	0.050 (0.0069)
Change from Baseline	−0.029 (0.0057)	0.944 (0.0073)	0.002 (0.0055)	0.954 (0.0066)	0.176 (0.0120)	0.050 (0.0069)

Table 6: Bias and Coverage under H_0 and H_A, Bayesian Power and Type I error when $P(\text{benefit}) > 0.95$ is chosen as a cut-off for benefit for the five endpoints (Monte Carlo SE in parenthesis).

- **Power and Type I Error**: The first-order Markov model and the multilevel random intercept model had substantially higher Bayesian power (both approximately 45%) to detect the true treatment effect. The other three methods (Time-to-Deterioration, Change from Baseline, and the 6-month Cumulative Logit) all had power below 20%. All five models as specified had a Bayesian Type I error rate close to the 5% level under the null hypothesis.
- **Bias and Coverage**:For most models, the estimated bias was within range of Monte Carlo error. However, the Change from Baseline model showed a bias under the alternative hypothesis (H_A). This is an expected consequence of its design, which pools observations from a wide time window (weeks 14-26) and treats them as a single timepoint, even though QoL changes during the period.
- **Monte Carlo Standard Error (MCSE)**: As shown in the table (values in parentheses), the MCE_{JK} for all reported performance measures was well below our tolerance of 0.02. This confirms that the simulation results are stable and that $n_{sim} = 1000$ was a sufficient number of repetitions.

10.4.4.8 Conclusion

The **first-order Markov model** and the **multilevel model with a random intercept** provide the highest statistical power of all models evaluated to detect a clinically relevant treatment effect in the context of sparse longitudinal QoL data, while controlling the Type I error rate and providing accurate, well-calibrated estimates.

10.5 Appendix E - Investigation of the Interval Censoring Approach

10.5.1 Background and Method

An intuitive method for handling intermittently missing QoL data is to treat the unobserved weeks as interval-censored. For a patient known to be alive, their QoL state for an unobserved week is somewhere in the range of `[1, 7]`. The `rmsb` package facilitates this through the `Ocens()` function in the model formula, which models an outcome known to fall within a lower (`y.a`) and upper (`y.b`) bound.

We investigated the validity of this approach using our simulated dataset. The investigation proceeded as follows:

1. **Create a sparse, realistic dataset**: From the complete simulated population of 100,000 we sampled 2500 patients and created a sparse dataset. We retained only the baseline measurement, the measurement at death (if applicable), and a random 15% of all other intermediate QoL measurements. This resulted in a dataset with heavy missingness ($\sim 85\%$) for the non-absorbing QoL states.
2. **Add intervalcensoring**: For weeks where a patient is known to be alive but QoL is missing, `y.a` was set to 1 and `y.b` was set to 7. For weeks where Y is not missing `y.a` = `y.b`.
3. **Fit the model**:

```
model <-
  blrm(
    formula = Ocens(y.a, y.b)  ~ tx + pol(time, 2) + yprev + ecog_fstcnt + diagnosis,
    data = df_censored,
    ppo = ~ time,
    cppo = function(y) y,
    refresh = 5,
    iter = 2000,
    chains = 4,
    method = "sampling"
  )
```

4. **Compare results**: We then compared the marginalized, model-derived SOPs with the true, empirical SOPs calculated from the complete dataset of 2500 patients.

10.5.2 Findings and Conclusion

We found a substantial overestimate of the cumulative incidence of the absorbing state (death), as shown in Figure 26.

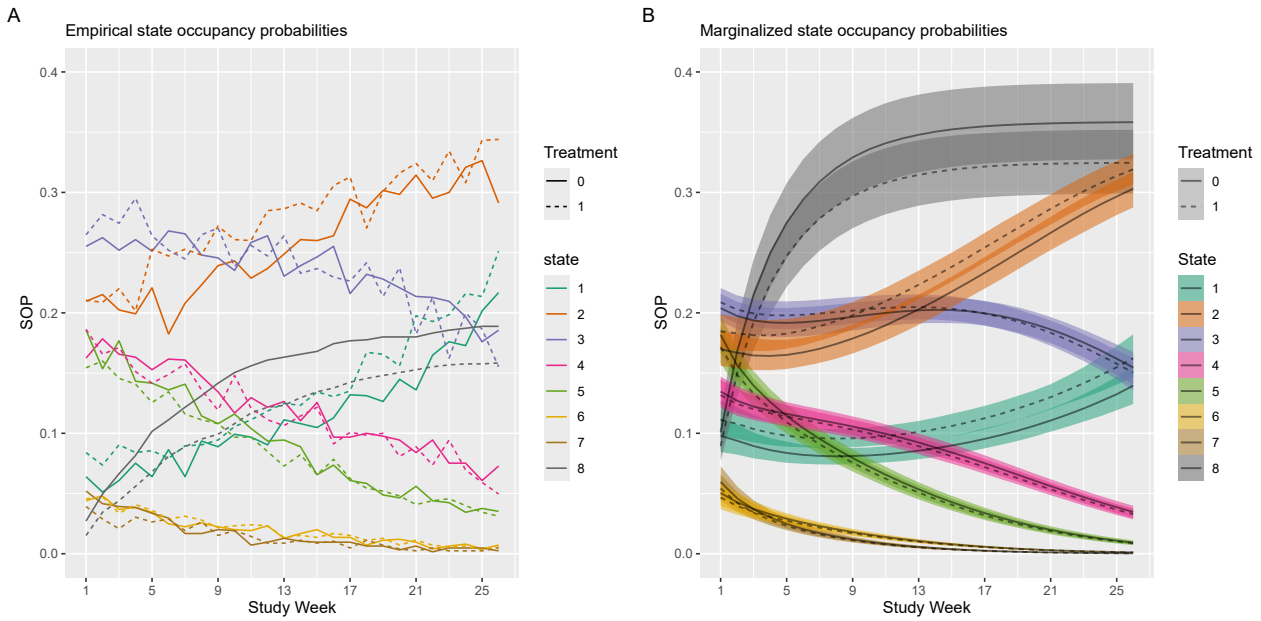

Figure 26: Comparison of empirical state occupancy probabilities (SOPs) from a complete dataset (A) vs. model-derived marginalized SOPs from a dataset with heavy censoring (85%) for non-absorbing states (B). $N = 1250$ per treatment group, the population average odds-ratio from which the 2500 patients were sampled is 0.8. The model-derived SOP for state 8 (death) is severely overestimated, leading to underestimates of SOPs for non-absorbing states.

We hypothesize this bias arises from the differential missingness between the absorbing and non-absorbing states. With the time of death being known precisely while the intermediate QoL states are overwhelmingly missing, the model appears to incorrectly attribute too much probability mass to the known absorbing state over time. As the SOPs must sum to one at every timepoint, this overestimation for the death state necessarily leads to an underestimation of the time spent in the living QoL states, which would directly bias our primary estimand.

Given this demonstrated potential for bias, we concluded that the interval censoring approach was not sufficiently robust for our primary analysis. This finding was a key motivation for adopting the **Principal Stratum analysis**, which analyzes the treatment effect conditional on survival and thus circumvents the challenge of modeling the transition to death with sparse intermediate data, but relies on strong assumptions.

Bibliography

[1] F. Harrell, "Statistical Errors in the Medical Literature." Apr. 2017.

[2] M. Di Maio, "Reading and interpreting quality-of-life results in cancer trials," *NEJM Evid.*, vol. 4, no. 6, p. EVIDra2400340, Jun. 2025.

[3] M. D. Rohde, B. French, T. G. Stewart, and F. E. Harrell, "Bayesian transition models for ordinal longitudinal outcomes," *Statistics in Medicine*, vol. 43, no. 18, pp. 3539–3561, Jun. 2024, doi: 10.1002/sim.10133.

[4] P. Fayers, A. Bottomley, EORTC Quality of Life Group, and Quality of Life Unit, "Quality of life research within the EORTC-the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer," *Eur. J. Cancer*, pp. S125–33, Mar. 2002.

[5] P. M. Fayers, Aaronson, NK: Bjordal, K, M. Groenvold, D. Curran, and A. Bottomley, "The EORTC QLQ-C30 Scoring Manual (3 rd Edition)." European Organisation for Research, Treatment of Cancer, 2001.

[6] T. M. Liddell and J. K. Kruschke, "Analyzing ordinal data with metric models: What could possibly go wrong?," *J. Exp. Soc. Psychol.*, vol. 79, pp. 328–348, Nov. 2018.

[7] F. E. Harrell Jr, "Modeling longitudinal responses using generalized least squares," *Regression Modeling Strategies*. in Springer series in statistics. Springer International Publishing, Cham, pp. 143–160, 2015.

[8] F. Harrell, "rmsb: Bayesian Regression Modeling Strategies." 2025. [Online]. Available: https://cran.r-project.org/package=rmsb

[9] B. C. Kahan, J. Hindley, M. Edwards, S. Cro, and T. P. Morris, "The estimands framework: a primer on the ICH E9(R1) addendum," *BMJ*, vol. 384, p. e76316, Jan. 2024.

[10] X. Zhou and J. P. Reiter, "A note on Bayesian inference after multiple imputation," *Am. Stat.*, vol. 64, no. 2, pp. 159–163, May 2010.

[11] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2025. [Online]. Available: https://www.r-project.org/

[12] P.-C. Bürkner, "brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017, doi: 10.18637/jss.v080.i01.

[13] I. R. White and P. Royston, "Imputing missing covariate values for the Cox model: IMPUTING MISSING COVARIATE VALUES FOR THE COX MODEL," *Stat. Med.*, vol. 28, no. 15, pp. 1982–1998, Jul. 2009.

[14] R. Kelter, "The Bayesian simulation study (BASIS) framework for simulation studies in statistical and methodological research," *Biom. J.*, vol. 66, no. 1, p. e2200095, Jan. 2024.