Biometrics 000, 000–000 DOI: 000
000 0000

Inference in Randomized Trials with Death and Missingness

Chenguang Wang[1,*], Daniel O. Scharfstein[2], Elizabeth Colantuoni[2], Timothy D. Girard[3], and Ying Yan[4]

[1]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[3]Center for Health Services Research, Vanderbilt University School of Medicine, Nashville, TN

[4]Helsinn Therapeutics (U.S.), Inc., Iselin, NJ

**email:* cwang68@jhmi.edu

Summary: In randomized studies involving severely ill patients, functional outcomes are often unobserved due to missed clinic visits, premature withdrawal or death. It is well known that if these unobserved functional outcomes are not handled properly, biased treatment comparisons can be produced. In this paper, we propose a procedure for comparing treatments that is based on a composite endpoint that combines information on both the functional outcome and survival. We further propose a missing data imputation scheme and sensitivity analysis strategy to handle the unobserved functional outcomes not due to death. Illustrations of the proposed method are given by analyzing data from a recent non-small cell lung cancer clinical trial and a recent trial of sedation interruption among mechanically ventilated patients.

Key words: Composite endpoint; Death-truncated data; Missing data; Sensitivity analysis.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Consider a randomized trial in which patients at high risk of death are scheduled to be clinically evaluated at pre-specified points in time after randomization. These clinical evaluations may be pre-empted due to death. Among living patients, clinical evaluations may be missing due to skipped visits or premature withdrawal from the study. There is a distinction between the two types of unobserved data. Data pre-empted due to death are generally considered undefined, whereas missing data are considered defined but uncollected. The question addressed in this paper is how to draw inference about the effect of treatment when clinical evaluation data may be unobserved due to death or missingness.

The issue of "truncation due to death" is challenging even in the absence of missing data. A number of methods have been proposed for analyzing such data (Kurland et al., 2009). Broadly speaking, the methods can be categorized into four main groups: (1) conditional, (2) joint, (3) causal and (4) composite. In the conditional approach, treatment effects are evaluated by conditioning on survival at each follow-up time (Kurland and Heagerty, 2005; Shardell and Miller, 2008). This approach is problematic because survival is a post-randomization factor and conditioning on a factor that may be affected by treatment can introduce bias (Rosenbaum, 1984). The joint approach introduces a common set of latent random effects for modeling both clinical evaluation endpoints and survival (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004; Ibrahim et al., 2010; Rizopoulos, 2012). In this approach, the model for the clinical evaluation endpoints allows trajectories of the functional endpoint after death, which is not scientifically meaningful. The causal inference approach frames the problem in terms of counterfactuals and seeks to estimate the "principal stratum" causal effect (Frangakis and Rubin, 2002; Hayden et al., 2005; Chiba and VanderWeele, 2011). The issue with this approach is that the principal stratum is the cohort of patients who would survive to a particular point in time regardless of treatment assignment and a clinician

cannot, at the time of the treatment decision, readily identify whether a patient is a member of this stratum or not. Nonetheless, this approach is useful for understanding the mechanistic effect of treatment on clinical outcomes. The fourth approach creates a composite outcome that mixes both the survival and functional evaluation endpoints (Diehr et al., 2001; Lachin, 1999; Joshua Chen et al., 2005). The problem with this approach is that it requires that the outcomes for patients be ordered. Further, the composite outcome approach does not allow one to separately tease out the effect of treatment on survival and on the functional outcome. *If* patients can be ordered in a way that makes scientific sense, the simplicity of the composite outcome approach can be a useful way of globally assessing treatment effects that are causally interpretable.

In this paper, we consider the composite outcome approach and address how to handle missing clinical evaluation data among those alive at the assessment times. We develop and illustrate our methodology in the context of the Study HT-ANAM-302 (also known as ROMANA 2), a randomized trial among advance lung cancer patients with cachexia. A second example with a trial of sedation interruption among mechanically ventilated patients is included in Web Appendix A.

In Study HT-ANAM-302, patients were randomized 2:1 to receive either anamorelin ($n = 330$) or Placebo ($n = 165$) (Temel et al., 2016). Patients were scheduled to have their lean body mass (LBM) evaluated at baseline and at 6 and 12 weeks after randomization. Eight survivors from each treatment group were missing LBM at baseline and are excluded from our analysis. In Table 1, we present treatment-specific summaries of death prior to week 12 and missingness of LBM among survivors. In this study, there was no statistically significant differences with respect to death prior to week 12 (15% vs. 17% for Placebo vs. anamorelin).

[Table 1 about here.]

2. Problem Formulation

2.1 *Notation*

We consider a two-arm randomized study design in which continuous functional measures are scheduled to be collected at baseline and K post-baseline assessment times $t_1, \ldots, t_K$. Let Y_0 denote the baseline measure and Y_k $(k = 1, \ldots, K)$ denote the post-baseline measure scheduled to be collected at time t_k. We use $\overline{Y}_k$ to denote $(Y_1, Y_2, \ldots, Y_k)$. Let X denote baseline covariates, excluding treatment assignment T. Let L denote the survival time and $A_k = I(L > t_k)$. Let $Z = g(Y_0, \ldots, Y_K)$ be the study's functional endpoint (i.e., an outcome measured on a living patient). We assume that Z is coded so that higher values denote better function. In the HT-ANAM-302 study, $K = 2$, Y_k is LBM and $Z = (Y_1 + Y_2)/2 - Y_0$ (the clinically meaningful endpoint defined in the protocol).

We consider the primary endpoint to be a finite-valued random variable U which assigns a score to each patient such that (1) each patient who dies prior to t_K is assigned a score according to their survival time (L), with shorter survival times assigned lower scores and (2) each patient who survives past t_K is assigned a score (higher than those who died prior to t_K) according to their functional status (Z), with lower functional status assigned lower scores. More formally, U is a function of (A_K, W) where $W = L$ if $A_K = 0$ and $W = Z$ if $A_K = 1$ and is defined such for all $\omega \in \Omega$ (sample space), $U(\omega) < c$ (an arbitrary constant) when $A_K(\omega) = 0$ and for all $\omega, \omega' \in \Omega$

$$
\begin{aligned}
&U(\omega) < U(\omega') \quad \text{if } A_K(\omega) = A_K(\omega'), W(\omega) < W(\omega') \\
&U(\omega) > U(\omega') \quad \text{if } A_K(\omega) = A_K(\omega'), W(\omega) > W(\omega') \\
&U(\omega) = U(\omega') \quad \text{if } A_K(\omega) = A_K(\omega'), W(\omega) = W(\omega') \\
&U(\omega) < U(\omega') \quad \text{if } A_K(\omega) = 0, A_K(\omega') = 1 \\
&U(\omega) > U(\omega') \quad \text{if } A_K(\omega) = 1, A_K(\omega') = 0.
\end{aligned}
$$

For our methods, only the ordering of U is important, not the actual score assignments. That is, given a sample of (A_K, W)'s, the above conditions are sufficient for ranking the

sample from worst to best; for any two subjects i and j, the conditions allow us to know whether $U_i < U_j$, $U_i = U_j$ or $U_i > U_j$. This endpoint is a composite outcome in the sense that it is univariate and contains information on survival and, when measurable, functional status.

For a patient alive at assessment k ($k \geqslant 1$), their outcome may be missing. When $A_k = 1$, define R_k to be the indicator that Y_k is observed. Thus, the observed data are:

$$O = (T, X, Y_0, L, A_1R_1, A_1R_1Y_1, \ldots, A_KR_K, A_KR_KY_K).$$

We have assumed that T, X, and Y_0 are always observed and that L is observed when $L < t_K$ (i.e., no censoring before t_K as is typical in well-designed clinical trials). For patients alive at t_K (i.e., $A_K = 1$), let $S = (R_1, \ldots, R_K)$ denote the missing data pattern; further, let $Y_{obs}^{(s)} = \{Y_k : R_k = 1, k \geqslant 1, S = s\}$ and $Y_{mis}^{(s)} = \{Y_k : R_k = 0, k \geqslant 1, S = s\}$ denote the observed and missing post-baseline functional outcomes. Note that Z is unobserved when $S \neq \mathbf{1}$, where $\mathbf{1}$ is a K-dimensional vector of 1's. We assume that we observe n i.i.d. copies of O. When necessary, we will subscript random variables by i and j to denote data specific to individual i and j, respectively.

2.2 *Treatment Effect Quantification*

In the classic two-sample Mann-Whitney test (Mann and Whitney, 1947), the population distributions from which the independent samples are drawn are assumed to be absolutely continuous. This assumption obviates tied observations. The samples are used to estimate the probability (η) that the outcome for a random individual drawn from the first population is less than the outcome for a random individual drawn from the second population. Under the null hypothesis of equality of the population distributions, $\eta = 0.5$. If, however, the population distributions are not absolutely continuous, η may not be distribution-free under the null.

In our setting, we want to allow for treatment-specific population distributions of the

composite outcome that may not be absolutely continuous. To address this issue, we define the treatment effect parameter θ to be the probability that the outcome for a random individual randomized to treatment $T = 0$ is less than the outcome of a random individual randomized to treatment $T = 1$ minus the probability that the outcome for a random individual randomized to treatment $T = 0$ is greater than the outcome of a random individual randomized to treatment $T = 1$. Values of $\theta > 0$ and $\theta < 0$ favor $T = 1$ and $T = 0$, respectively. Under the null hypothesis of no treatment effect, θ will be zero. Our goal is to draw inference about θ.

In the absence of missing data, we estimate θ by

$$\widehat{\theta} = \frac{1}{n_0 n_1} \sum_{i:T_i=0} \sum_{j:T_j=1} \{\mathrm{I}(U_i < U_j) - \mathrm{I}(U_i > U_j)\}$$

where $n_0 = \sum_i (1 - T_i)$ and $n_1 = \sum_i T_i$. In addition to estimating θ, quantiles of the treatment-specific distribution of the composite endpoint U can be calculated to help further characterize the treatment effect.

2.3 *Missing Data and Imputation Assumptions*

In order to estimate θ in the presence of missing data, we need to know how to impute Z for patients alive at t_K with $s \neq \mathbf{1}$. It is sufficient to impute $Y_{mis}^{(s)}$ for these patients.

Assumptions are required in order to perform this imputation. We introduce the following class of untestable assumptions:

$$f(Y_{mis}^{(s)} | A_K = 1, Y_{obs}^{(s)}, Y_0, X, T, S = s) \propto \exp(\beta_T Z) f(Y_{mis}^{(s)} | A_K = 1, Y_{obs}^{(s)}, Y_0, X, T, S = \mathbf{1}) \quad (1)$$

for all $s \neq \mathbf{1}$, where β_T is a treatment-specific sensitivity parameter. Note that setting $\beta_T = 0$ (i.e., benchmark assumption in the class) reduces to the complete case missing value (CCMV) restrictions (Little, 1993) applied to the missing data patterns for patients alive at t_K. It can be shown that CCMV is different from the missing at random (MAR) assumption. Because of the difficulty and subtlety of the MAR assumption in the presence

of non-monotone missing data (Robins and Gill, 1997; Tsiatis, 2007), we have anchored the class of assumptions around CCMV.

To understand this class of assumptions, consider the case where $K = 2$ and, as in the HT-ANAM-302 study, $Z = (Y_1 + Y_2)/2 - Y_0$. In this case, (1) reduces to the following three assumptions, where $\beta_T' = \beta_T/2$ (due to the definition of the functional endpoint Z):

Assumption 1:

$$f(Y_2|A_2 = 1, Y_1, Y_0, X, T, S = (1,0)) \propto \exp(\beta_T' Y_2) \underbrace{f(Y_2|A_2 = 1, Y_1, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (2)$$

This assumption says that for subjects alive at t_2, who are observed at time t_1, who share the same functional measure at t_1 and who share the same baseline factors, the distribution of Y_2 for those whose functional measure at t_2 is missing is, when $\beta_T' > 0$ (< 0), more heavily weighted toward higher (lower) values of Y_2 than those whose functional measure at t_2 is observed.

Assumption 2:

$$f(Y_1|A_2 = 1, Y_2, Y_0, X, T, S = (0,1)) \propto \exp(\beta_T' Y_1) \underbrace{f(Y_1|A_2 = 1, Y_2, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (3)$$

This assumption says that for subjects alive at t_2, who are observed at time t_2, who share the same functional measure at t_2 and who share the same baseline factors, the distribution of Y_1 for those whose functional measure at t_1 is missing is, when $\beta_T' > 0$ (< 0), more heavily weighted toward higher (lower) values of Y_1 than those whose functional measure at t_1 is observed.

Assumption 3:

$$f(Y_1, Y_2|A_2 = 1, Y_0, X, T, S = (0,0)) \propto \exp\left(\beta_T'(Y_1 + Y_2)\right) \underbrace{f(Y_1, Y_2|A_2 = 1, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (4)$$

This assumption says that for subjects alive at t_2 and who share the same baseline factors, the joint distribution of Y_1 and Y_2 for those whose functional measures at t_1 and t_2 are

missing is, when $\beta'_T > 0$ (< 0), more heavily weighted toward higher (lower) values of Y_1 and Y_2 than those whose measures are fully observed.

When $\beta'_T = 0$ in above assumptions, there is no differential weighting. The differences between the distributions being contrasted in the above assumptions increases with $|\beta'_T|$. To better illustrate these assumptions, ignore conditioning on Y_0 and X and suppose $f(Y_1, Y_2|A_2 = 1, T, S = \mathbf{1})$ is multivariate normal with mean $(\mu_{T,1}, \mu_{T,2})$ and variance-covariance matrix

$$\Sigma_T = \begin{bmatrix} \sigma^2_{T,1} & \rho_T \sigma_{T,1} \sigma_{T,2} \\ \rho_T \sigma_{T,1} \sigma_{T,2} & \sigma^2_{T,2} \end{bmatrix}$$

Then, $f(Y_2|A_2 = 1, Y_1, T, S = (1,0))$ is normal with mean $\mu_{T,2} + \beta'_T(1-\rho^2_T)\sigma^2_{T,2} + \rho_T \frac{\sigma_{T,2}}{\sigma_{T,1}}(Y_1 - \mu_{T,1})$ and variance $(1-\rho^2_T)\sigma^2_{T,2}$; $f(Y_1|A_2 = 1, Y_2, T, S = (0,1))$ is normal with mean $\mu_{T,1} + \beta'_T(1-\rho^2_T)\sigma^2_{T,1} + \rho_T \frac{\sigma_{T,1}}{\sigma_{T,2}}(Y_2 - \mu_{T,2})$ and variance $(1-\rho^2_T)\sigma^2_{T,1}$; and $f(Y_1, Y_2|A_2 = 1, T, S = (0,0))$ is multivariate normal with mean $(\mu_{T,1} + \beta'_T\sigma^2_{T,1} + \beta'_T\rho_T\sigma_{T,1}\sigma_{T,2}, \mu_{T,2} + \beta'_T\sigma^2_{T,2} + \beta'_T\rho_T\sigma_{T,1}\sigma_{T,2})$ and variance-covariance matrix Σ_T. If $\rho_T > 0$, then the above means increase linearly in β'_T; β'_T has no impact on the above variances and covariances. Thus, $\beta'_T > 0$ ($\beta'_T < 0$) implies that the distributions on the left hand sides of Equations (2), (3) and (4) have more (less) mass at higher values than their reference distributions.

2.4 *Modeling and Inference*

Our imputation approach requires specification of a model for $f(\overline{Y}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$. In specifying this model, it is important to utilize an approach that respects bounds (possibly population-specific) on the functional outcomes; failure to do so can result in nonsensical imputations.

To address this issue, we consider a data transformation of Y_k ($k = 1, \ldots, K$) by a transformation function

$$\phi(y_k) = \log\left\{\frac{y_k - B_L}{B_U - y_k}\right\},$$

where (B_L, B_U) denote the lower and upper bound. Let $Y^\dagger_k = \phi(Y_k)$ and $\overline{Y}^\dagger_k = (Y^\dagger_1, \ldots, Y^\dagger_k)$.

Importantly, there is a one-to-one mapping between the conditional distributions $h(\overline{Y}^{\dagger}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$ and $f(\overline{Y}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$. We construct a model for $f(\overline{Y}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$ by positing a model for $h(\overline{Y}^{\dagger}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$. To proceed, we write

$$h(\overline{Y}^{\dagger}_K|A_K = 1, Y_0, X, T, S = \mathbf{1}) = \prod_{k=1}^{K} h(Y^{\dagger}_k|A_K = 1, \overline{Y}^{\dagger}_{k-1}, Y_0, X, T, S = \mathbf{1}) \tag{5}$$

and posit a model for each component of the product.

In our examples, we consider models of the form:

$$h(Y^{\dagger}_k|A_K = 1, \overline{Y}^{\dagger}_{k-1}, Y_0, X, T = t, S = \mathbf{1}) = h_{k,t}(Y^{\dagger}_k - \mu_{k,t}(\overline{Y}^{\dagger}_{k-1}, Y_0, X; \boldsymbol{\alpha}_{k,t}))$$

where $\mu_{k,t}(\overline{Y}^{\dagger}_{k-1}, Y_0, X; \boldsymbol{\alpha}_{k,t})$ is a specified conditional mean function (depending on time k and treatment t) of $\overline{Y}^{\dagger}_{k-1}$, Y_0, X and $\boldsymbol{\alpha}_{k,t}$, $\boldsymbol{\alpha}_{k,t}$ is an unknown parameter vector and $h_{k,t}$ is an unspecified time/treatment-specific mean zero density function. The parameter vectors $\boldsymbol{\alpha}_{k,t}$ can be estimated by minimizing the least squares objective function

$$\sum_{i=1}^{n} I(T_i = t) A_{K,i} \left(\prod_{k=1}^{K} R_{k,i} \right) \{Y^{\dagger}_{k,i} - \mu_{k,t}(\overline{Y}^{\dagger}_{k-1,i}, Y_{0,i}, X_i; \boldsymbol{\alpha}_{k,t})\}^2$$

Let $\widehat{\boldsymbol{\alpha}}_{k,t}$ denote the least squares estimator of $\boldsymbol{\alpha}_{k,t}$. The density function $h_{k,t}$ can be estimated by kernel density estimation based on the residuals $\{Y^{\dagger}_{k,i} - \mu_{k,t}(\overline{Y}^{\dagger}_{k-1,i}, Y_{0,i}, X_i; \widehat{\boldsymbol{\alpha}}_{k,t}) : T_i = t, A_{K,i} = 1, R_{1,i} = \ldots, R_{K,i} = 1, i = 1, \ldots, n\}$ or estimated with parametric assumptions (e.g. normality) if the sample size is small. Let $\widehat{h}_{k,t}$ denote the kernel density estimator of $h_{k,t}$. We then estimate $f(\overline{Y}_K|A_K = 1, Y_0, X, T, S = \mathbf{1})$ by

$$\widehat{f}(\overline{Y}_K|A_K = 1, Y_0, X, T, S = \mathbf{1}) = \prod_{k=1}^{K} \widehat{h}_{k,t}(Y^{\dagger}_k - \mu_{k,t}(\overline{Y}^{\dagger}_{k-1}, Y_0, X; \widehat{\boldsymbol{\alpha}}_{k,t})) \left| \frac{d\phi(Y_k)}{dY_k} \right|.$$

For each individual i alive at t_K and who is in a stratum $s \neq \mathbf{1}$, we impute (see next section) the missing functional outcomes by drawing from the density that is proportional to $\exp(\beta_T Z)\widehat{f}(Y^{(s)}_{mis}|A_K = 1, Y^{(s)}_{obs} = Y_{obs,i}, Y_0 = Y_{0,i}, X = X_i, T = T_i, S = \mathbf{1})$. For each such individual, we draw M copies of the missing functional outcomes to create M complete datasets. For each complete dataset m, we estimate θ by $\widehat{\theta}_m$. Our overall estimator of θ

is $\tilde{\theta} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_m$. Confidence intervals can be constructed by non-parametric bootstrap, where individuals are sampled with replacement within each treatment group.

2.5 *Imputation*

We propose the following Metropolis-Hastings algorithm to draw from

$$\exp(\beta_T Z)\widehat{f}(Y_{mis}^{(s)}|A_K = 1, Y_{obs}^{(s)}, Y_0, X, T, S = \mathbf{1}).$$

For ease of notation, we suppress the superscript s for Y_{mis} and Y_{obs} in the following steps.

(1) Set $l = 0$. Choose arbitrary initial values for Y_{mis}, denoted by $Y_{mis}^{(0)}$. Let $Z^{(0)}$ be the primary functional endpoint with data $(Y_{obs}, Y_{mis}^{(0)})$.

(2) Set $l = l + 1$.

(3) Generate Y'_{mis} from a (multivariate) Gaussian distribution with mean $Y_{mis}^{(l-1)}$ and variance Λ.

(4) Calculate the acceptance ratio as

$$a = \frac{\exp\{\beta_T Z'\}\widehat{f}(Y'_{mis}, Y_{obs}|A_K = 1, Y_0, X, T, S = \mathbf{1})}{\exp\{\beta_T Z^{(l-1)}\}\widehat{f}(Y_{mis}^{(l-1)}, Y_{obs}|A_K = 1, Y_0, X, T, S = \mathbf{1})}$$

where Z' and $Z^{(l-1)}$ are the primary functional endpoints with data (Y_{obs}, Y'_{mis}) and $(Y_{obs}, Y_{mis}^{(l-1)})$, respectively.

(5) Accept $Y_{mis}^{(l)} = Y'_{mis}$ with probability $\min(1, a)$ and $Y_{mis}^{(l)} = Y_{mis}^{(l-1)}$ with probability $1 - \min(1, a)$.

(6) Repeat Steps 2-5 until the Markov chain converges.

(7) Draw random samples from the set $\{Y_{mis}^{(l_0)}, Y_{mis}^{(l_0+1)}, \ldots\}$ as the imputed missing values, where l_0 corresponds to the burn-in number.

Note that out-of-boundary candidates Y'_{mis} are rejected at Step 5 since the acceptance ratio will be 0. The tuning parameter Λ in Step 3 affects the acceptance rate. In practice, calibration of Λ may be applied to achieve a desirable acceptance rate. Note a higher acceptance rate often corresponds to a slower convergence. Robert (1997) suggested an

acceptance rate of 1/4 for models of high dimension and 1/2 for models of dimension 1 or 2. As an example of calibration, Muller (1991) proposed to successively modify Λ as the product of a scale factor and the variance of the available samples. The calibration process continues until the acceptance rate is close to 1/4 and the variance of the available samples stabilizes. Furthermore, various diagnostics such as Geweke diagnostic may be applied to evaluate the convergence of the Markov chain (Cowles and Carlin, 1996).

3. Simulation Study

We considered a study design in which two post-baseline functional assessments are scheduled (i.e., $K = 2$) to be collected at t_1 and t_2. We defined $Z = (Y_1+Y_2)/2-Y_0$. For each simulation, we generate a dataset with n individuals - half assigned $T = 0$ and half assigned $T = 1$. For each individual, we simulated data according to the following algorithm:

- Draw Y_0 from standard normal distribution.
- Given T and Y_0, draw L_1 from an exponential distribution with mean $1/\exp(\lambda_{T,0}+\lambda_{T,1}Y_0)$. If $L_1 < t_1$, set $L = L_1$ and stop.
- Given T and Y_0, draw Y_1 from a normal distribution with mean $\mu_T + \gamma_T Y_0$, and variance 1.
- Given T and $\overline{Y}_1$, draw L_2 from an exponential distribution with mean $1/\exp(\lambda_{T,0}+\lambda_{T,1}Y_1)$. If $L_2 < t_2 - t_1$, set $L = L_2 + t_1$ and stop.
- Given T and $\overline{Y}_1$, draw Y_2 from a normal distribution with mean $\mu_T + \gamma_T Y_1$ and variance 1.
- Given T and $\overline{Y}_2$, draw S from multinomial distribution with

$$P[S = s|T, \overline{Y}_2] = \frac{\exp(\mu'_{T,s} + \beta_T Z)}{1 + \sum_{s' \neq \mathbf{1}} \exp(\mu'_{T,s'} + \beta_T Z)}, \quad s \neq \mathbf{1}$$

and

$$P[S = \mathbf{1}|T, \overline{Y}_2] = \frac{1}{1 + \sum_{s' \neq \mathbf{1}} \exp(\mu'_{T,s'} + \beta_T Z)}.$$

Note that the data generation mechanism for S is equivalent to the exponential tilting class of assumptions posited by (1).

We considered two major scenarios in the simulation study. Scenario I is focused on evaluating the impact of survival and functional status among survivors on the treatment effect evaluation. In this scenario, we assumed there is no missing data among survivors in either arm (i.e., $\mu'_{T,s} \equiv -\infty$ for all T and s). Scenario II is focused on evaluating the impact of missing data and the proposed sensitivity analysis strategy on the treatment effect evaluation. In this scenario, we assumed there were no deaths in either arm (i.e., $\lambda_{T,0} = -\infty$ for all T). For all simulations, we set $\gamma_0 = \gamma_1 \equiv 1$ and $\mu_0 \equiv 0$. For assessing the performance of our estimation procedure for θ, we report mean squared error, coverage rate of 95% percentile-based bootstrap confidence intervals (1000 re-samples) and the null hypothesis rejection rate. In our simulation study, we considered sample sizes of 200 and 500. Each simulation was based on 500 replications. For missing data imputation, we set $M = 1$ and used a burn-in of 1000 iterations for each MCMC chain.

3.1 *Scenario I*

We set $\lambda_{0,0} = \lambda_{1,0} = -0.5$ and $\lambda_{0,1} = 1$. In varying the survival rate at the end of the study, we considered study lengths (i.e., t_2) of 0.2 and 0.5 and set $t_1 = t_2/2$. Table 2 shows that, for all settings, θ is well estimated and the 95% bootstrap confidence interval covers the true value of θ with probability close to the nominal level. Under the null hypothesis $H_0 : \theta = 0$ (i.e. $\lambda_{1,1} = 1$ and $\mu_1 = 0$), the type I error rate is well controlled and as expected the power to detect a treatment effect increases as the size of the study or $|\theta|$ increases.

[Table 2 about here.]

3.2 *Scenario II*

We set $\mu'_{1,s} = -2.5$ and $\mu'_{0,s} = -\infty$ (i.e., no missing data in arm $T = 0$) for all $s \neq \mathbf{1}$. We set $\beta_1 = -2.0$. We considered $\mu_1 = -0.25$, 0.00 and 0.25, yielding missing data rates in arm $T = 1$ of 21%, 15% and 10%, respectively. Table 3 shows that when β_1 is correctly specified (i.e., $\beta_1^* = -2.0$), the multiple imputation procedure produces unbiased estimates of θ with nominal coverage probabilities. However, when β_1 is mis-specified (i.e., $\beta_1^* = 0$), there is, as expected, bias in estimation of θ, poor confidence interval coverage and inflated type I error.

[Table 3 about here.]

4. Data Analysis

For the analysis of the HT-ANAM-302 Study, the imputation of week 6 and week 12 LBM incorporated the following baseline covariates: Eastern Cooperative Oncology Group (ECOG) performance status (0 or 1 vs. 2), age ($\leqslant 65$ vs. > 65), sex, body mass index (BMI) (underweight, < 18.5, or not), and weight loss over the prior 6 months (WL) ($\leqslant 10\%$ vs. $> 10\%$). In this example, we applied a data transformation setting the lower (B_L) and upper (B_U) bound of LBM to be 24 and 140, respectively. We specified the following models for $\mu_{k,t}(\overline{Y}^{\dagger}_{k-1}, Y_0, X; \boldsymbol{\alpha}_{k,t})$:

$$\begin{aligned}\mu_{1,t}(Y_0, X, \boldsymbol{\alpha}_{1,t}) &= \alpha_{1,t,1} + \alpha_{1,t,2}Y_0 + \alpha_{1,t,3}ECOG + \alpha_{1,t,4}AGE \\ &\quad + \alpha_{1,t,5}SEX + \alpha_{1,t,6}BMI + \alpha_{1,t,7}WL \\ \mu_{2,t}(\overline{Y}^{\dagger}_{1}, Y_0, X; \boldsymbol{\alpha}_{2,t}) &= \alpha_{2,t,1} + \alpha_{2,t,2}Y_0 + \alpha_{2,t,3}ECOG + \alpha_{2,t,4}AGE \\ &\quad + \alpha_{2,t,5}SEX + \alpha_{2,t,6}BMI + \alpha_{2,t,7}WL + \alpha_{2,t,8}Y^{\dagger}_1\end{aligned}$$

To estimate θ, $M = 10$ imputed datasets were generated. A burn-in of 2000 iterations was used for each MCMC chain. After the burn-in, imputed data were drawn every 50 iterations. Trace plots of the MCMC chains for 5 randomly selected patients are reported in Web Appendix B. A total of 1000 bootstrap samples were used to compute standard

errors and 95% percentile-based confidence intervals; two-sided p-values were computed using a standard normal approximation to the Wald statistic (estimator divided by bootstrap standard error).

Under the benchmark assumptions, $\widehat{\theta} = 0.30$ (95% CI: 0.16 to 0.37, $p < 0.0001$), which indicates that patients treated with anamorelin have a significantly higher probability of having a better clinical outcome, as described by the composite of survival and average change in LBM from baseline, than patients treated with Placebo. Figure 1 displays the treatment-specific cumulative distribution functions of the composite endpoint U, where we have labeled the values of the composite endpoint according to the survival time L and functional endpoint Z among survivors. Note that the distribution of survival is similar across the treatment groups and differences in the distribution of the composite endpoint are being driven by differences in the functional endpoint among survivors. In the Placebo group, we estimate that more than half the patients will survive and have an average change in LBM from baseline greater than -0.98 kg (95% CI: -1.27 kg to -0.28 kg). In the anamorelin group, we estimate that more than half the patients will survive and have an average change in LBM from baseline greater than 0.69 kg (95% CI: 0.43 kg to 0.93 kg).

For the sensitivity analysis, we varied β_T from -0.5 to 0.5. This range corresponds to an induced shift, relative to the benchmark imputation, of about 1.5 kg in the mean of the imputed average LBM change, which represents a clinically important change (Figure 2). Panel (A) of Figure 3 presents estimates of θ and its associated 95% confidence interval as a function of β_0 (i.e., sensitivity analysis parameter in the Placebo arm), for two extreme values of $\beta_1 = -0.5, 0.5$ (i.e., sensitivity analysis parameter in the anamorelin arm). For all the sensitivity scenarios, the lower bound of the 95% CI for θ is always greater than 0 suggesting that the conclusions from the benchmark analysis are robust. Panel (B) of Figure 3 presents the treatment-specific estimates (along with 95% confidence intervals) of the median

of the composite endpoint and its 95% confidence interval as a function of β_T for this study. Panel (C) of Figure 3 presents a contour plot of the p-values associated with testing the null hypothesis $\theta = 0$ for each combination of β_0 and β_1 for this study. The figures shows that, for all combinations, the null hypothesis is rejected in favor of anamorelin. We conclude that anamorelin is superior to Placebo in terms of improving the composite endpoint, driven by improvements in LBM (as depicted in Figure 1).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

5. Discussion

In this paper, we proposed a global sensitivity analysis approach for randomized controlled clinical trials with death and intermittent missing data. Our method is based on the construction of a composite endpoint that combines both the survival and the functional outcome data. Complete case missing value constraints are considered as the benchmark assumption for intermittent missing data imputation. Sensitivity analysis is further conducted to evaluate the robustness of the findings through exponential tilting. The sensitivity analysis strategy differs from previous work in two important ways. First, it handles non-monotone missing data anchored at CCMV benchmark restrictions. With the exception of Minini and Chavance (2004) and Vansteelandt et al. (2007), previous work has focused on monotone missing data anchored at MAR-type assumptions (see, for example, Rotnitzky et al. (2001b); Little (1994); Rotnitzky et al. (1998); Scharfstein et al. (1999); Robins et al. (2000); Rotnitzky et al. (2001a); Birmingham et al. (2003); Daniels and Hogan (2008); National Research Council Panel on Handling Missing Data in Clinical Trials (2010); Scharfstein et al. (2014)). Furthermore, previous work has not been imputation-based. Second, our proposal uses a

parsimonious way to introduce sensitivity parameters, which is directly connected to the functional outcome.

The CCMV benchmark restrictions are untestable and may be considered unreasonable in some settings. Thus, the proposed sensitivity analysis strategy is critically important. Our proposal will fail in settings where there are very few survivors with complete functional outcome data, on which Z is defined.

We emphasize that there exists multiple approaches to address the "truncation by death" issue. Which approach is the "best" depends on the target of inference (Kurland et al., 2009). Provided that death and the functional outcome can be ordered in a scientifically meaningful way, the composite endpoint approach is desirable when the goal is to globally evaluate the efficacy and safety of a medical intervention under the intention to treat paradigm. Before utilizing the proposed method, researchers should employ mixed methods to confirm that the ordering is consistent with the health preferences of the patient population under investigation.

The ranking scheme we proposed is similar to the "untied worst-rank score analysis" in Lachin (1999). An alternative approach, the "worst-rank score analysis", ranks all the patients who died ($A_K = 0$) the same and is also commonly used. The proposed method can be easily extended to incorporate alternatives to death such as "unable to complete" the functional evaluation as may occur in studies similar to the trial of sedation interruption among mechanically ventilated patients. The principle for choosing the ranking scheme, nonetheless, is that the ranking orders should be clinically meaningful and closely related to the goal of evaluating the efficacy and safety of the treatment.

In the proposed approach, we assume that the survival status is always known and there is no censoring. Such an assumption is generally reasonable for well-controlled clinical trials with relatively short study durations. When this assumption does not hold, we need to extend

the imputation strategy to first impute the survival time for censored subjects. Depending on the imputed survival length, missing data may nor may not need to be imputed for these subjects.

In this paper, we have assumed complete information on baseline covariates, including the functional measure. The missing baseline functional measures could be imputed by extending the patterns of missingness to include this measure. With regards to other missing baseline covariates, we recommend, prior to implementation of our proposed methods, imputation using readily available software.

We proposed numerical sampling techniques, specifically the random-walk Metroplis Hastings algorithm, for sampling the missing outcomes. Alternatively, the slice sampling algorithm (Neal, 2003) can be applied to take into account the restricted ranges of the missing outcomes. The computation load may be reduced for special cases in which there is a closed form expression for the target distributions.

The proposed approach can handle randomized studies with more than two treatment arms by conducting pairwise treatment comparisons and adjusting for multiplicity. The proposed missing data imputation strategy may be applied without change since it is conducted separately for each arm.

6. Software

A web-based software is developed for the proposed method. The software is demonstrated at `http://www.olssol.com/app/composite/`. Source code in the form of R code is available on request from the authors.

Acknowledgments and Conflicts

The authors are grateful to the editor and referees for their constructive comments. The methods developed in this paper were motivated by a consulting project between the first

two authors (CW and DS) and Helsinn Therapeutics. CW and DS were compensated for their consultation services. CW and DS were not paid for preparation of this manuscript. This research was also partially supported by contracts from FDA and PCORI, NIH grant R24HL111895 and NCI grant CA183854.

References

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society: Series B* **65,** 275–297.

Chiba, Y. and VanderWeele, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology* **173,** 745–751.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91,** 883–904.

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. CRC Press.

Diehr, P., Patrick, D. L., Spertus, J., Kiefe, C. I., Donell, M., and Fihn, S. D. (2001). Transforming self-rated health and the SF-36 scales to include death to improve interpretability. *Medical Care* **39,** 670–680.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Hayden, D., Pauler, D. K., and Schoenfeld, D. (2005). An esitmator for treatment comparisons among survivors in randomized trials. *Biometrics* **61,** 305–310.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28,** 2796–2801.

Joshua Chen, Y., Gould, A. L., and Nessly, M. L. (2005). Treatment comparisons for a

partially categorical outcome applied to a biomarker with assay limit. *Statistics in Medicine* **24,** 211–228.

Kurland, B. F. and Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* **6,** 241–258.

Kurland, B. F., Johnson, L. L., Egleston, B. L., and Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical Science* **24,** 211–222.

Lachin, J. M. (1999). Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials* **20,** 408–422.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88,** 125–134.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81,** 471–483.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* pages 50–60.

Minini, P. and Chavance, M. (2004). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics* **5,** 531–544.

Muller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical report, Purdue University.

National Research Council Panel on Handling Missing Data in Clinical Trials (2010). *The Prevention and Treatment of Missing Data in Clinical Trials.* The National Academies Press.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics* **31,** 705–741.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R.* CRC Press.

Robert, C. P. (1997). Discussion of Richardson and Green's paper. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59,** 758–764.

Robins, J., Rotnitzky, A., and Scharfstein, D. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran, E., editor, *Statistical Models for Epidemiology*, pages 1–94. Springer-Verlag.

Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* **16,** 39–56.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147,** 656–666.

Rotnitzky, A., Robins, J., and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93,** 1321–1339.

Rotnitzky, A., Scharfstein, D., Su, T., and Robins, J. (2001a). A sensitivity analysis methodology for randomized trials with potentially non-ignorable cause-specific censoring. *Biometrics* **57,** 103–113.

Rotnitzky, A., Scharfstein, D., Su, T.-L., and Robins, J. (2001b). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57,** 103–113.

Scharfstein, D., McDermott, A., Olson, W., and F, W. (2014). Global sensitivity analysis for repeated measures studies with informative drop-out. *Statistics in Biopharmaceutical Research* **6,** 338–348.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). *Journal of the American Statistical Association* **94,** 1096–1146.

Shardell, M. and Miller, R. R. (2008). Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine* **27,** 1008–1025.

Temel, J. S., Abernethy, A. P., Currow, D. C., Friend, J., Duus, E. M., Yan, Y., and Fearon, K. C. (2016). Anamorelin in patients with non-small-cell lung cancer and cachexia (romana 1 and romana 2): results from two randomised, double-blind, phase 3 trials. *The Lancet Oncology* **17,** 519–531.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14,** 809–834.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94,** 841–860.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53,** 330–339.

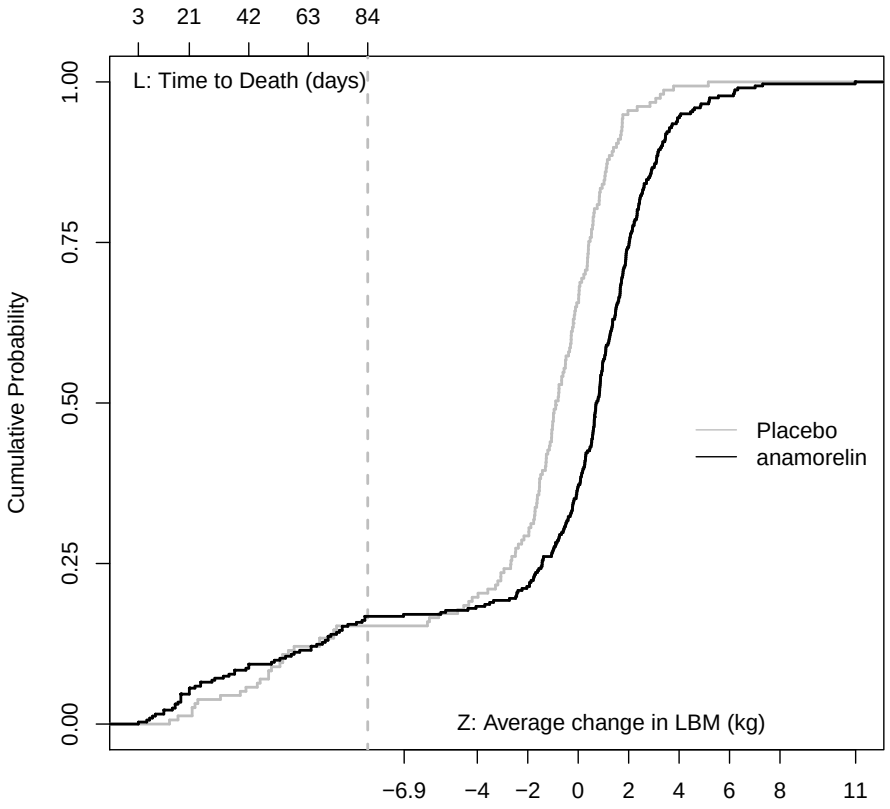

Figure 1: Cumulative distribution function of the composite endpoint for each treatment group based on the multiple imputation algorithm with the benchmark assumptions. The composite endpoint is labeled according to the survival time L among patients that die and the functional endpoint Z among patients that survive to 12 weeks.

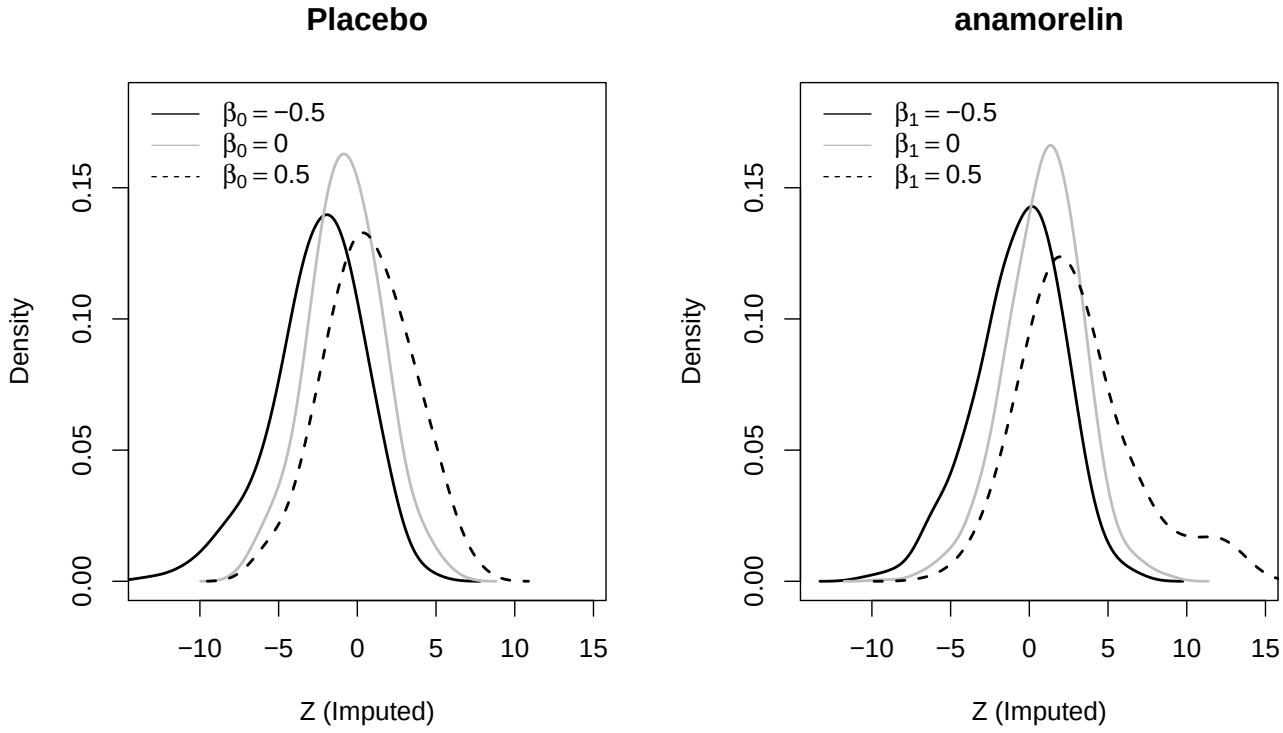

Figure 2: Treatment-specific densities of the imputed Z (average change in LBM from baseline) for different choices of the sensitivity parameters β_T

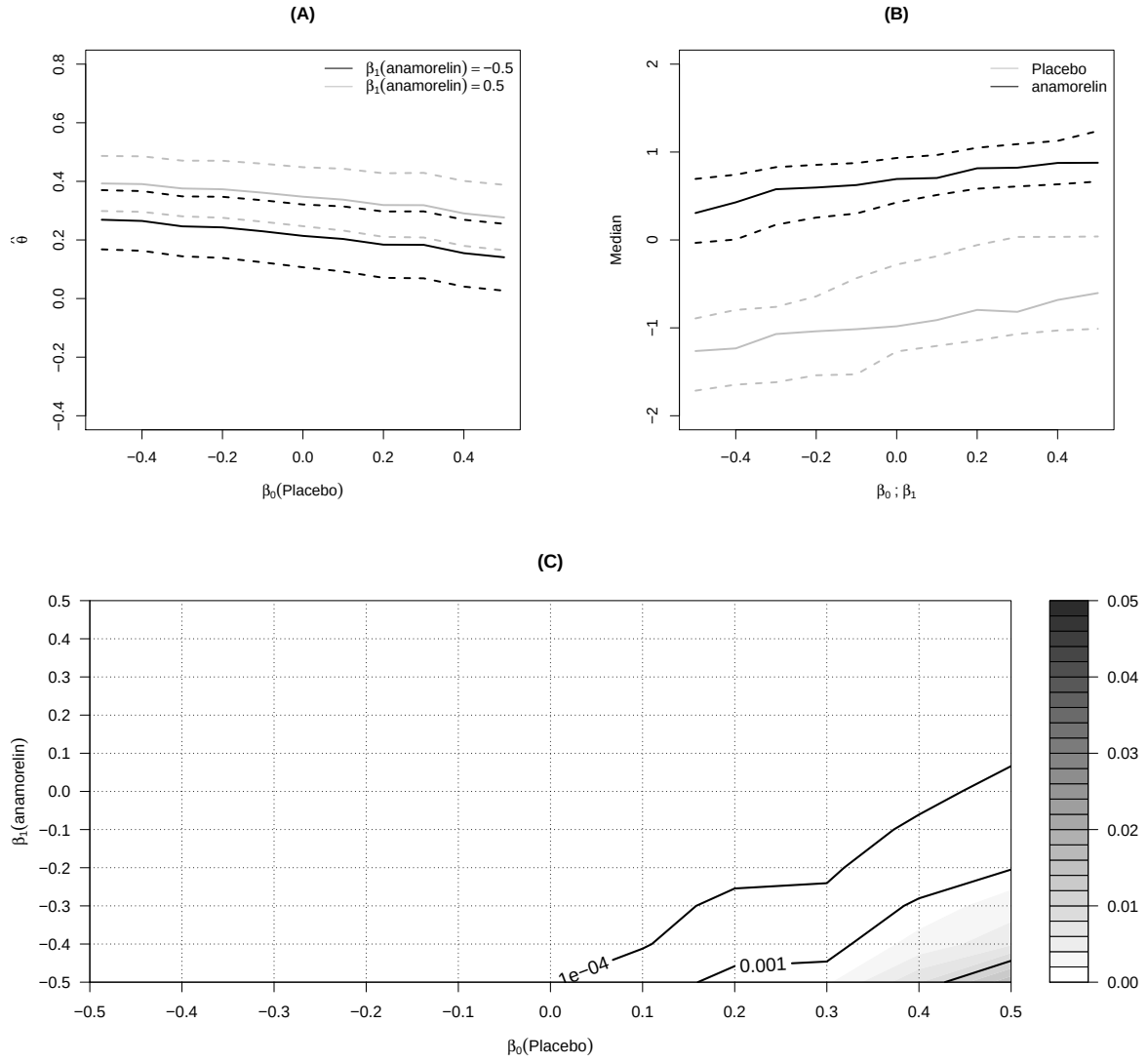

Figure 3: Sensitivity analysis: Panel (A) presents estimates of θ (with 95% confidence intervals) for various choices of the sensitivity analysis parameters. Note that β_1 and β_0 are the sensitivity analysis parameters for the anamorelin and Placebo groups, respectively. Panel (B) presents the treatment-specific estimates of the median (with 95% confidence intervals) of the composite endpoint for various choices of sensitivity analysis parameters. Panel (C) presents the contour plot of the p-values obtained by testing the null hypothesis of $\theta = 0$ as function of treatment-specific sensitivity analysis parameters.

Table 1: Treatment-specific summaries of death prior to week 12, and missingness of LBM among survivors from the HT-ANAM-302 Study. †: Patients with $A_K = 0$.

	Placebo ($n = 157$)	anamorelin ($n = 322$)
Died Prior to Week 12†	24 (15.3%)	54 (16.8%)
Survivors with complete data	93 (59.2%)	185 (57.5%)
Survivors missing only Week 6	3 (1.9%)	17 (5.3%)
Survivors missing only Week 12	17 (10.8%)	31 (9.6%)
Survivors missing both Weeks 6 and 12	20 (12.7%)	35 (10.9%)

Table 2: Scenario I Simulation Study Results. MSE*: mean squared error $\times 1000$. Rej*: rejection rate for $H_0 : \theta = 0$. Cov*: bootstrap 95% confidence interval coverage rate. The Death Rates for $T = 0$ are 0.188 or 0.354 corresponding to the study length (t_2) of 0.2 and 0.5, respectively.

	Death Rate			True	Sample	Estimation		Rate	
$\lambda_{1,1}$	$T=0$	$T=1$	μ_1	θ	Size	$\widehat{\theta}$	MSE*	Rej*	Cov*
1.3	0.188	0.230	0.0	-0.056	200	-0.060	5.5	0.092	0.978
					500	-0.054	2.9	0.186	0.938
		0.293	0.5	0.088	200	0.085	7.1	0.198	0.944
					500	0.086	2.5	0.358	0.958
	0.354	0.388	0.0	-0.051	200	-0.053	6.7	0.104	0.936
					500	-0.046	2.7	0.154	0.956
		0.463	0.5	0.007	200	0.007	7.6	0.072	0.928
					500	0.006	2.6	0.042	0.960
1.0	0.188	0.188	0.0	-0.001	200	0.002	6.9	0.050	0.952
					500	0.004	2.7	0.048	0.958
		0.236	0.5	0.178	200	0.181	7.5	0.602	0.932
					500	0.177	2.7	0.934	0.946
	0.354	0.354	0.0	0.000	200	-0.003	6.1	0.032	0.974
					500	0.000	2.7	0.058	0.944
		0.418	0.5	0.080	200	0.079	7.2	0.180	0.946
					500	0.084	2.7	0.352	0.948
0.7	0.188	0.151	0.0	0.051	200	0.047	6.4	0.090	0.960
					500	0.053	2.4	0.174	0.952
		0.180	0.5	0.265	200	0.269	5.8	0.924	0.954
					500	0.262	2.7	0.996	0.944
	0.354	0.315	0.0	0.054	200	0.051	6.3	0.096	0.958
					500	0.053	2.5	0.174	0.964
		0.362	0.5	0.163	200	0.160	6.0	0.518	0.950
					500	0.165	2.7	0.884	0.954

Table 3: Scenario II Simulation Study Results. MSE*: mean squared error ×1000. Rej*: rejection rate for $H_0 : \theta = 0$. Cov*: bootstrap 95% confidence interval coverage rate. β_1^*: sensitivity parameter for $T = 1$. Missing rate*: overall functional endpoint missing rate.

	Missing		True	Sample	Estimation		Rate	
β_1^*	Rate*	μ_1	θ	Size	$\widehat{\theta}$	MSE*	Rej*	Cov*
0	0.21	-0.25	-0.186	200	-0.049	26.8	0.090	0.640
				500	-0.045	23.5	0.146	0.268
	0.15	0.00	0.000	200	0.104	18.4	0.236	0.780
				500	0.110	15.1	0.516	0.476
	0.10	0.25	0.186	200	0.275	14.4	0.906	0.810
				500	0.271	9.5	1.000	0.614
-2	0.21	-0.25	-0.186	200	-0.192	7.1	0.612	0.952
				500	-0.189	2.9	0.928	0.950
	0.15	0.00	0.000	200	-0.014	7.6	0.054	0.952
				500	-0.011	3.1	0.050	0.952
	0.10	0.25	0.186	200	0.180	7.5	0.572	0.950
				500	0.178	2.7	0.928	0.948