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., Bridgewater, 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 the composite endpoint of 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 patients alive at a pre-specified time, some may fail to be evaluated due to missed visits or withdrawal, yielding missing data. There is a distinction between the two types of unobserved data. Data pre-empted due to death are generally considered not existing and undefined, whereas missing data are considered existing but not collected. The question addressed in this paper is how to draw inference about the effect of treatment when clinical evaluation data may be pre-empted by death or missing.

Ignoring the complication of missing data, the so-called issue of "truncation due to death" is a thorny one. 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 (Tsiatis and Davidian, 2004). In this approach, the model for the clinical evaluation endpoints often 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 endpoint 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 two randomized trials.

1.1 *HT-ANAM 302 Study*

In this study, patients with non-small cell lung cancer-cachexia were randomized 2:1 to receive either Anamorelin ($n = 330$) or Placebo ($n = 165$) (Garcia et al., 2015). 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; p = 0.79 based on Fisher's exact test, p=0.66 based on Logrank test).

1.2 *ABC Trial*

The Awakening and Breathing Controlled (ABC) trial randomized critically ill patients receiving mechanical ventilation 1:1 within each study site to management with a paired sedation plus ventilator weaning protocol involving daily interruption of sedatives through spontaneous awakening trials (SATs) and spontaneous breathing trials (SBTs) ($n = 168$) or sedation per usual care (UC) and SBTs ($n = 168$) (Girard et al., 2008). In a single site substudy (Jackson et al., 2010) ($n = 94$ for UC+SBT, $n = 93$ for SAT+SBT), the researchers assessed differences in cognitive, psychological and functional outcomes at 3 and 12 months after randomization. Our focus is cognitive function at 12 months after randomization, which is derived from the results of nine cognitive tests. Each test score was converted to a T-score and cognitive function was represented by a cognition score, the mean of the available nine T-scores. In Table 1, we present treatment-specific summaries of death prior to 12 months and missingness of cognition scores among survivors. In this substudy, there was a statistically significant difference with respect to death prior to 12 months (62% vs. 41% for UC + SBT vs. SAT + SBT, respectively; p = 0.005 based on Fisher's exact test, p = 0.005 based on Logrank test).

[Table 1 about here.]

2. Problem Formulation

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. Let X denote baseline covariates, excluding treatment assignment T. Let L denote the survival time and $\Delta_k = I(L > t_k)$. Let $Z = g(Y_0, \ldots, Y_K)$ be the study's primary functional endpoint, which is only defined if $\Delta_K = 1$. 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 $Z = (Y_1 + Y_2)/2 - Y_0$. In the ABC substudy, Y_0 is not available (as is commonly the case in study of patients with critical illness), $K = 2$, Y_k is the cognition score and $Z = Y_2$.

In the absence of missing data among patients alive at the post-baseline assessment time points, the data for an individual are

$$F = (T, X, L, Y_0, \Delta_1 Y_1, \ldots, \Delta_K Y_K)$$

We use the subscripts i and j to denote data for the ith and jth patients, respectively.

2.1 *Ranking*

In the absence of missing data, we propose to rank patients as follows:

- If $\Delta_{K,i} = \Delta_{K,j} = 1$, then patient i (j) is ranked better than patient j (i) if $Z_i > Z_j$ $(Z_j < Z_i)$ and ranked the same if $Z_i = Z_j$.
- If $\Delta_{K,j} = 0$ $(\Delta_{K,i} = 0)$ and $\Delta_{K,i} = 1$ $(\Delta_{K,j} = 1)$, then patient i (j) is ranked better than patient j (i).
- If $\Delta_{K,i} = \Delta_{K,j} = 0$, then patient i (j) is ranked better than patient j (i) if $L_i > L_j$ $(L_j < L_i)$ and ranked the same if $L_i = L_j$.

We let R denote the rank for an individual.

In this ranking, patients who die prior to time t_K are ranked according to their survival time, with shorter survival times assigned worse ranks. Patients who survive past time t_K are then assigned ranks (higher than those died prior to time t_K), according to the value of Z, with lower values of Z assigned worse ranks.

2.2 *Treatment Effect Quantification*

Let θ be the probability that the rank for a random individual randomized to treatment $T = 0$ is less than the rank of a random individual randomized to treatment $T = 1$ minus

the probability that the rank for a random individual randomized to treatment $T = 0$ is greater than the rank of a random individual randomized to treatment $T = 1$. Values of $\theta > 0$ (< 0) favor $T = 1$ ($T = 0$). Under the null hypothesis of no treatment effect, there will be no difference between the ranks of the two treatment groups and θ will be zero. The goal is to draw inference about θ.

In the absence of missing data, we estimate θ by

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

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 can be calculated to further characterize the treatment effect.

2.3 *Missing Data and Benchmark Imputation Assumptions*

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

$$O = (T, X, L, Y_0, \Delta_1\tau_1, \Delta_1\tau_1 Y_1, \ldots, \Delta_K\tau_K, \Delta_K\tau_K Y_K).$$

We have assumed that T, X, L and Y_0 are always observed.

For subjects alive at t_K, let $Y_{obs} = \{Y_k : \tau_k = 1, k \geqslant 1\}$ and $Y_{mis} = \{Y_k : \tau_k = 0, k \geqslant 1\}$ denote the observed and missing post-baseline functional outcomes. Let $S = (\tau_1, \ldots, \tau_K)$ be the missing pattern. In order to rank subjects in the presence of missing data, we need to know how to impute Z for patients alive at τ_K. It is sufficient to impute Y_{mis} for these patients.

Assumptions are required in order to perform this imputation. We make the following untestable benchmark assumptions:

$$f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = s) = f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = \mathbf{1}) \qquad (1)$$

for all $s \neq \mathbf{1}$, where $\mathbf{1}$ is a K-dimensional vector of 1's. These assumptions are the complete

case missing value (CCMV) restrictions (Little, 1993) applied to the missing data patterns for patients alive at t_K.

To understand the CCMV assumptions, consider the special case where $K = 2$. In this setting, (1) reduces to the following three assumptions:

Assumption 1:

$$f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = (1,0)) = f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = \mathbf{1}) \tag{2}$$

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 (Y_0, X, T), the distribution of Y_2 is the same for those whose functional measure at t_2 is missing and those whose measure is observed.

Assumption 2:

$$f(Y_1|\Delta_2 = 1, Y_2, Y_0, X, T, S = (0,1)) = f(Y_1|\Delta_2 = 1, Y_2, Y_0, X, T, S = \mathbf{1}) \tag{3}$$

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 is the same for those whose functional measure at t_1 is missing and those whose measure is observed.

Assumption 3:

$$f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = (0,0)) = f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = \mathbf{1}) \tag{4}$$

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 is the same for those whose functional measures at t_1 and t_2 are missing and those whose measures are fully observed.

2.4 *Sensitivity Analysis*

The CCMV benchmark assumptions are untestable. Thus, as noted in NRC (2010), it is essential to evaluate the robustness of inferences to deviations from the benchmark assump-

tions. Exponential tilting is one method that has been employed to construct a neighborhood of assumptions that is centered around the benchmark assumptions. The neighborhood is indexed by sensitivity analysis parameters, where typically sensitivity analysis parameters set to zero reduce to the benchmark assumptions.

One of the challenges with any sensitivity analysis is the dimension of the sensitivity analysis parameters. While the size of the neighborhood grows with the dimension of the sensitivity analysis parameters, it becomes more complex to communicate results. In our analyses, we consider a two-dimensional sensitivity analysis parameter. Specifically, we consider an exponential tilting class of assumptions of the following form:

$$f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = s) \propto \exp(\beta_T Z) f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = \mathbf{1}) \quad (5)$$

for all $s \neq \mathbf{1}$, where β_T is a treatment-specific sensitivity parameter. Note that setting $\beta_T = 0$, reduces to the CCMV benchmark assumptions.

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, (5) reduces to the following three assumptions (where $\beta_T' = 2\beta_T$):

Assumption 1':

$$f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = (1,0)) \propto \exp(\beta_T' Y_2) \underbrace{f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (6)$$

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':

$$f(Y_1|\Delta_2 = 1, Y_2, Y_0, X, T, S = (0,1)) \propto \exp(\beta_T' Y_1) \underbrace{f(Y_1|\Delta_2 = 1, Y_2, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (7)$$

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':

$$f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = (0,0)) \propto \exp\left(\beta'_T(Y_1 + Y_2)\right) \underbrace{f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = \mathbf{1})}_{\text{Reference Distribution}} \quad (8)$$

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.

Importantly, 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 $f(Y_1, Y_2|\Delta_2 = 1, T, S = \mathbf{1})$ 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, $f(Y_2|\Delta_2 = 1, Y_1, T, S = (1,0))$ 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 $(1-\rho^2_T)\sigma^2_{T,2}$; $f(Y_1|\Delta_2 = 1, Y_2, T, S = (0,1))$ 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 $f(Y_1, Y_2|\Delta_2 = 1, T, S = (0,0))$ 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 (6), (7) and (8) have more (less) mass at higher values than their reference distributions.

For the ABC substudy, $K = 2$ and $Z = Y_2$. In this case, there is no need to impute Y_1 in order to rank patients. Thus, (5) reduces to the following two assumptions: Assumption 1', and

Assumption 3":

$$f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = (0,0)) \propto \exp(\beta_T Y_2) f(Y_1, Y_2|\Delta_2 = 1, Y_0, X, T, S = \mathbf{1}) \quad (9)$$

Assumption 3" is equivalent to the following two assumptions:

Assumption 3.1":

$$f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = (0,0)) \propto \exp(\beta_T Y_2) f(Y_2|\Delta_2 = 1, Y_1, Y_0, X, T, S = \mathbf{1}) \quad (10)$$

and

Assumption 3.2":

$$f(Y_1|\Delta_2 = 1, Y_0, X, T, S = (0,0)) = f(Y_1|\Delta_2 = 1, Y_0, X, T, S = \mathbf{1}) \quad (11)$$

Assumption 3.1" says that for subjects alive at t_2, 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 measures at t_1 and t_2 are missing is, when $\beta_T > 0$ (< 0), more heavily weighted toward higher (lower) values of Y_2 than those whose measures are fully observed. Assumption 3.2" says that for subjects alive at t_2 and who share the same baseline factors, the distribution of Y_1 for those whose functional measures at t_1 and t_2 are missing is the same as those whose measures are fully observed.

2.5 *Modeling and Imputation Inference*

Our imputation approach will require specification of a model for $f(\overline{Y}_K|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$. 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. In the HT-ANAM 302 study population, experts expect that LBM will be between 24kg to 140kg. In the ABC trial, the cognitive score was constructed to be between 0 and 100.

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 $Y_k^\dagger = \phi(Y_k)$ and $\overline{Y}_k^\dagger = (Y_1^\dagger, \ldots, Y_k^\dagger)$. Importantly, there is a one-to-one mapping between the conditional distributions $h(\overline{Y}_K^\dagger|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$ and $f(\overline{Y}_K|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$. In particular,

$$f(\overline{Y}_K|\Delta_K = 1, Y_0, X, T, S = \mathbf{1}) = h(\overline{Y}_K^\dagger|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})\left|\prod_{k=1}^{K}\frac{d\phi(Y_k)}{dY_k}\right|. \quad (12)$$

We will construct a model for $f(\overline{Y}_K|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$ by positing a model for $h(\overline{Y}_K^\dagger|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$ and using the above formula.

To proceed, we write

$$h(\overline{Y}_K^\dagger|\Delta_K = 1, Y_0, X, T, S = \mathbf{1}) = \prod_{k=1}^{K} h(Y_k^\dagger|\Delta_K = 1, \overline{Y}_{k-1}^\dagger, Y_0, X, T, S = \mathbf{1}) \quad (13)$$

and posit a model for each component of the product. In our examples, we consider models of the form:

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

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

$$\sum_{i=1}^{n} I(T_i = t)\Delta_{K,i}\left(\prod_{k=1}^{K}\tau_{k,i}\right)\{Y_{k,i}^\dagger - \mu_{k,t}(\overline{Y}_{k-1}^\dagger, Y_0, X; \boldsymbol{\alpha}_{k,t})\}^2$$

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 $\{Y^{\dagger}_{k,i} - \mu_{k,t}(\overline{Y}^{\dagger}_{k-1,i}, Y_{0,i}, X_i; \widehat{\boldsymbol{\alpha}}_{k,t}) : T_i = t, \Delta_{K,i} = 1, \tau_{1,i} = \ldots, \tau_{K,i} = 1, i = 1, \ldots, n\}$. Let $\widehat{h}_{k,t}$ denote the kernel density estimator of $h_{k,t}$. We then estimate $f(\overline{Y}_K|\Delta_K = 1, Y_0, X, T, S = \mathbf{1})$ by

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

In Section 2.6, we show how to draw Y_{mis} under Assumption (2.5) using the Metroplis-Hastings algorithm. For each individual i alive at t_K and who is in a stratum $s \neq \mathbf{1}$, we impute the missing functional outcomes by drawing from the estimated density that is proportional to $\exp(\beta_T Z) f(Y_{mis}|\Delta_K = 1, Y_{obs} = Y_{obs,i}, Y_0 = Y_{0,i}, X = X_i, T = T_i, S = s)$. For each such individual, we draw M copies of the missing functional outcomes. This is then used 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.

2.6 *Imputation*

The goal is to sample from (5) for all $s \neq \mathbf{1}$. That is, we want to sample from the conditional density that is proportional to

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

The closed form of $f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = s)$ is in general not available. Thus, numerical sampling techniques need to be applied to draw the samples (Robert and Casella, 1999).

We provide the following detailed steps of a random-walk Metropolis-Hastings algorithm to sample from $f(Y_{mis}|\Delta_K = 1, Y_{obs}, Y_0, X, T, S = s)$:

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

(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

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

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 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. Data Analysis

3.1 *HT-ANAM 302 Study*

For the analysis of the HT-ANAM 302 Study, the imputation 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 set $B_L = 24$ and $B_U = 140$. We specify 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, 10 imputed datasets were generated. A total of 500 bootstrap samples were used to characterize uncertainty (i.e., percentile confidence intervals, standard errors). The bootstrap process takes into account variation due to model fitting. Under the benchmark assumptions, $\widehat{\theta} = 0.30$ (95% CI: 0.18 to 0.37, $p < 0.0001$) (Table 2), which indicates that patients treated with Anamorelin have a significantly higher probability of having a better clinical outcome, as described by the composite of LBM and survival, than patients treated with placebo. The left panel of Figure 1 displays the treatment-specific cumulative distribution functions of the composite endpoint for this study. 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: -1.53 kg to -0.51 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: 0.33 kg to 0.87 kg) (Table 2).

For the sensitivity analysis, we ranged 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). The top left panel 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 including the "worst" scenario, the lower bound of the 95% CI for θ is always greater than 0 suggesting that the conclusions from the benchmark analysis are robust. The top right panel 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. The left panel of Figure 4 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.

3.2 *ABC Trial*

In the ABC Trial, the imputation incorporated patient age (AGE) and years of education (EDU). In this example, there is no Y_0 and we set $B_L = 0$ and $B_U = 100$. We specify the following models for $\mu_{k,t}(\overline{Y}^{\dagger}_{k-1}, X; \boldsymbol{\alpha}_{k,t})$:

$$\mu_{1,t}(X, \boldsymbol{\alpha}_{1,t}) = \alpha_{1,t,1} + \alpha_{1,t,2}EDU + \alpha_{1,t,3}AGE + \alpha_{1,t,4}EDU * AGE$$

$$\mu_{2,t}(\overline{Y}^{\dagger}_{1}, X; \boldsymbol{\alpha}_{2,t}) = \alpha_{2,t,1} + \alpha_{2,t,2}EDU + \alpha_{2,t,3}AGE + \alpha_{2,t,4}EDU * AGE + \alpha_{2,t,5}Y^{\dagger}_{1}$$

To estimate θ, 10 imputed datasets were generated. Under the benchmark assumptions, $\widehat{\theta} = 0.18$ (95% CI: 0.03 to 0.33, $p = 0.023$), indicating that a randomly selected SAT+SBT patient has a greater probability of being ranked higher than a randomly selected UC+SBT patient (Table 2). The right panel of Figure 1 displays the treatment-specific cumulative distribution functions of the composite endpoint for this study. For the UC+SBT group, we

estimate that 50% of the subjects will survive past 72 days (95% CI: survive past 30 days to survive to 1-year). In the SAT+SBT group, we estimate that 50% of subjects will survive to 12 months with cognitive scores of 30 or greater (95% CI: survive past 357 days to survive with cognitive score of 37) (Table 2).

We allowed the sensitivity parameter β_T to vary between -0.2 and 0.2 corresponding to a shift the imputed mean 12 month cognition scores, relative to the benchmark imputation, of 5-15 units (Figure 2), a difference that is clinically significant. The sensitivity analysis reveals that the SAT+SBT group is favored over the control group in all of the scenarios (Figure 3 and 4). However, the difference between the two arms is not statistically significant for scenarios when β_0 for the UC+SBT arm is greater than 0 and β_1 for the SAT+SBT arm is less than -0.04. That is, there must be differential missing data mechanisms in the two arms in order to "lose statistical significance".

Based on the primary and sensitivity analysis results, we conclude that there is robust evidence that a difference exists between the control and the intervention arms in the composite endpoints of survival and cognitive performance, favoring the intervention arm.

[Table 2 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

4. Discussion

In this paper, we proposed a global sensitivity analysis approach for randomized clinical trials with death and intermittent missing data. Our method is based on the construction of a composite endpoint that mixes both the survival and the functional outcome data.

Treatment effect estimation and comparison are based on ranks. Complete case missing value constraints are considered as the benchmark assumption for missing data imputation. Sensitivity analysis is further conducted to evaluate the robustness of the findings through exponential tilting.

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.

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 ($\Delta_K = 0$) the same and is also commonly used. The proposed method can easily incorporate alternatives to death such as "unable to complete" the functional evaluation as may occur in studies similar to the ABC Trial. 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.

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.

5. Software

A web-based software is developed for the proposed method. The software is available at `http://sow.familyds.com/shiny/composite/`. Source code in the form of R code, together with a sample data set is available on request from the authors.

Acknowledgments and Conflicts

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 and NIH grant R24HL111895.

References

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.

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.

Garcia, J. M., Boccia, R. V., Graham, C. D., Yan, Y., Duus, E. M., Allen, S., and Friend, J. (2015). Anamorelin for patients with cancer cachexia: an integrated analysis of two phase 2, randomised, placebo-controlled, double-blind trials. *Lancet Oncology* **16,** 108–116.

Girard, T. D., Kress, J. P., Fuchs, B. D., Thomason, J. W. W., Schweickert, W. D., Pun, B. T., Taichman, D. B., Dunn, J. G., Pohlman, A. S., Kinniry, P. A., Jackson, J. C., Canonico, A. E., Light, R. W., Shintani, A. K., Thompson, J. L., Gordon, S. M., Hall, J. B., Dittus, R. S., Bernard, G. R., and Ely, E. W. (2008). Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (awakening and breathing controlled trial): a randomised controlled trial. *Lancet* **371,** 126–134.

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

Jackson, J. C., Girard, T. D., Gordon, S. M., Thompson, J. L., Shintani, A. K., Thomason, J. W. W., Pun, B. T., Canonico, A. E., Dunn, J. G., Bernard, G. R., Dittus, R. S., and Ely, E. W. (2010). Long term cognitive and psychological outcomes in the awakening and breathing controlled trial. *American Journal of Respiratory and Critical Care Medicine* **182,** 183–191.

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.

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

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

NRC (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press.

Robert, C. P. (1997). Discussion of richardson and green's paper. *Journal of Royal Statistics (B)* **59,** 758–764.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer.

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.

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.

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

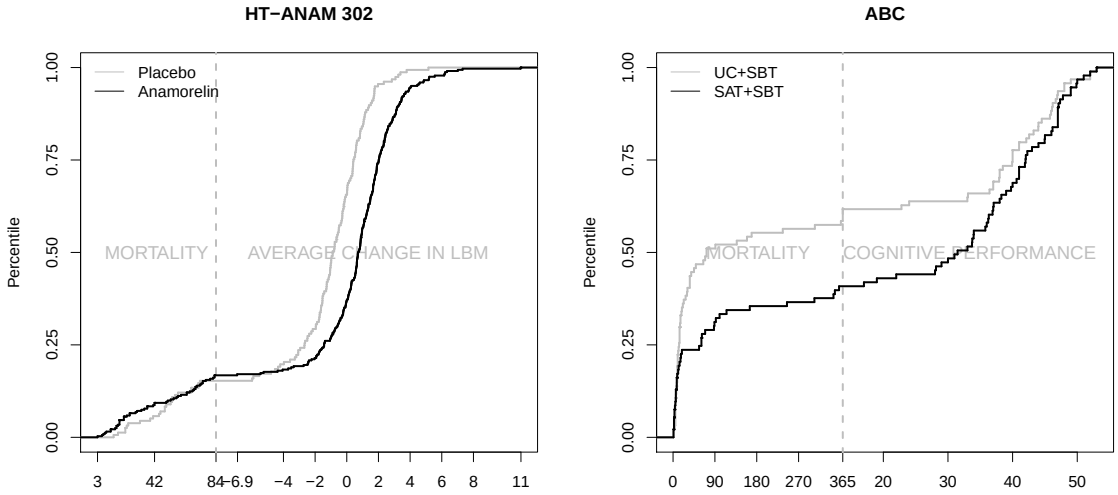

Figure 1. Cumulative distribution function of the composite endpoint under benchmark assumptions. On the x-axis, the average change in LBM (Z) segment for the HT-ANAM 302 Study is in kg. The cognitive performance (Z) segment for the ABC Trial is the performance score. The mortality (L) segments for both studies are in days

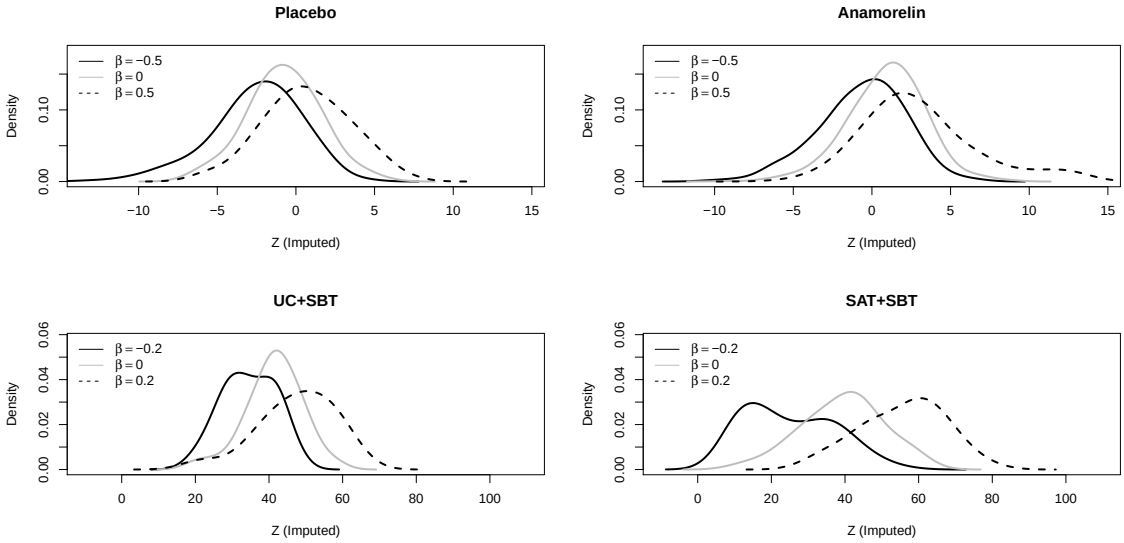

Figure 2. Treatment-specific densities of the imputed Z for different choices of the sensitivity parameters β for HT-ANAM 302 Study (first row) and ABC Trial (second row)

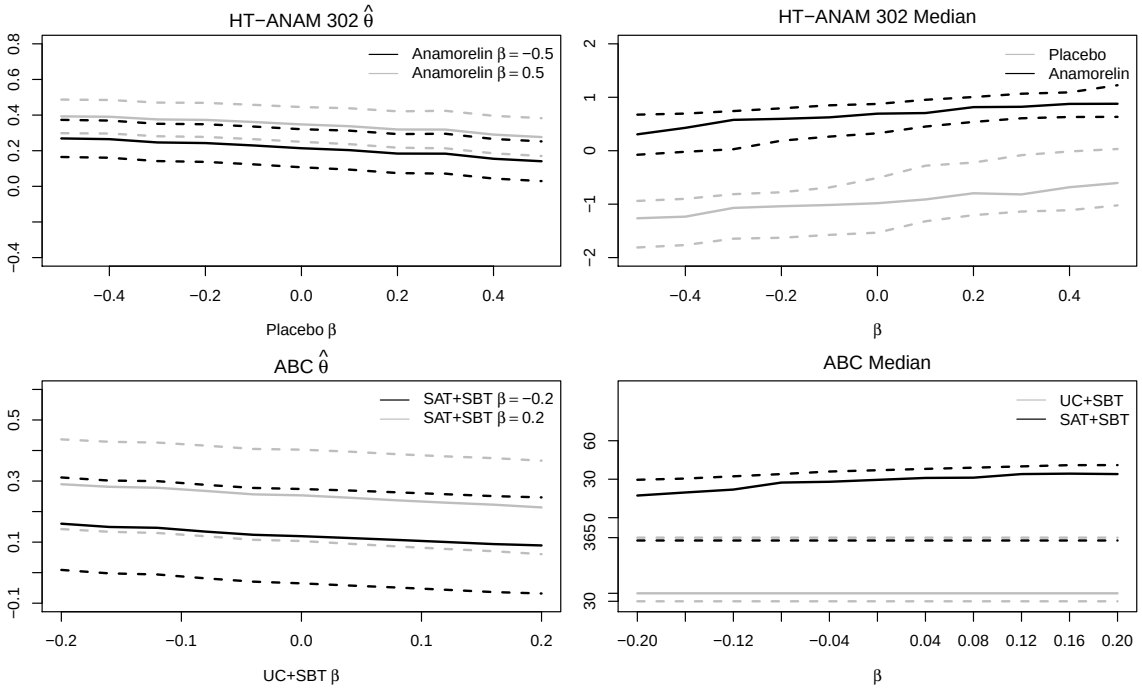

Figure 3. Sensitivity analysis: First column presents study-specific estimates of θ (with 95% confidence intervals) for various choices of sensitivity analysis parameters. Second column presents study- and treatment-specific estimates of the median (with 95% confidence intervals) for various choices of sensitivity analysis parameters; for the UC+SBT arm and the lower bound of the SAT+SBT arm in the ABC trial, the median is survival in days

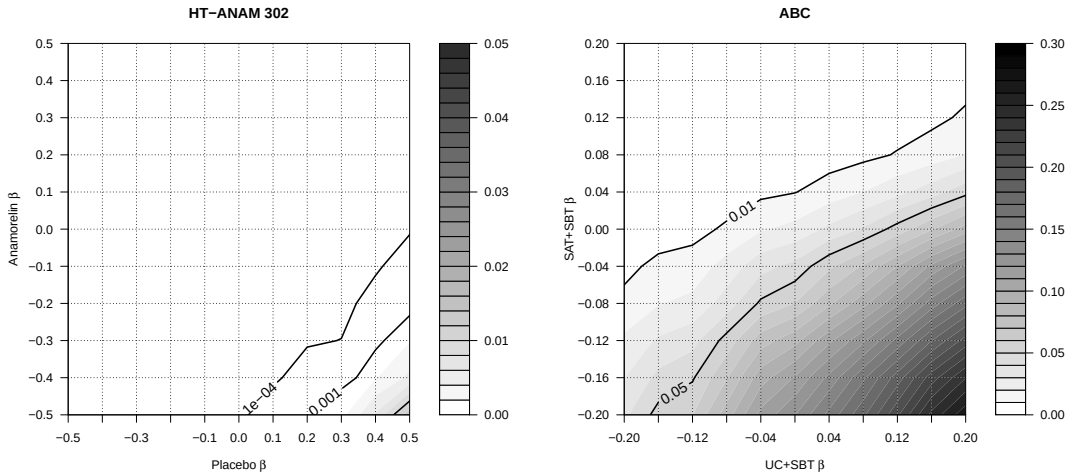

Figure 4. Sensitivity analysis: Contour plot of the primary hypothesis testing p-values as function of treatment-specific sensitivity analysis parameters β

Table 1
Treatment-specific summaries of death prior to week 12 (month 12), and missingness of LBM (cognition scores) among survivors from the HT-ANAM 302 Study (ABC Trial).

HT-ANAM 302 Study	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%)
ABC Trial	UC + SBT $n = 94$	SAT + SBT $n = 93$
Died Prior to Month 12	58 (61.7%)	38 (40.9%)
Survivors with complete data	18 (19.1%)	32 (34.4%)
Survivors missing only Month 3	1 (1.1%)	0 (0.0%)
Survivors missing only Month 12	8 (8.5%)	8 (8.6%)
Survivors missing both Months 3 and 12	9 (9.6%)	15 (16.1%)

Table 2
Hypothesis testing, estimation of θ and median (p_{50}) of the distribution of the composite endpoint under benchmark assumptions. For estimation of the median, t_x indicates a survival time of x days

	$\widehat{\theta}$ (95% CI)	p-value
HT-ANAM 302 Study	0.30(0.18,0.37)	< 0.0001
ABC Trial	0.18(0.03,0.33)	0.023
		$\widehat{p}_{50}$ (95% CI)
HT-ANAM 302 Study	Placebo	-0.98(-1.53,-0.51)
	Anamorelin	0.69(0.33, 0.87)
ABC Trial	UC + SBT	$t_{72}(t_{30}, t_{365})$
	SAT + SBT	30 (t_{357}, 37)