Health Utility Survival for Randomized Clinical Trials: Extensions and Statistical Properties

Yangqing Deng[1,*], Meiling Hao[2,*], Shao Hui Huang[3], Geoffrey Liu[4,5,6], John R. de Almeida[7,8], Wei Xu[1,6#]

[1]Department of Biostatistics, University Health Network, Toronto, ON, Canada

[2]School of Statistics, University of International Business and Economics, Beijing, China

[3]Department of Radiation Oncology, University Health Network, Toronto, ON, Canada

[4]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[5]Medical Oncology and Hematology, Princess Margaret Cancer Centre, Toronto, ON, Canada

[6]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[7]Department of Otolaryngology—H&N Surgery, University Health Network, Toronto, ON, Canada

[8]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[*]Co-first-authors

[#]Corresponding author: Wei Xu, Email: wei.xu@uhn.ca

SUMMARY

Overall survival has been used as the primary endpoint for many randomized trials that aim to examine whether a new treatment is non-inferior to the standard treatment or placebo control. When a new treatment is indeed non-inferior in terms of survival, it may be important to assess other outcomes including health utility. However, analyzing health utility scores in a secondary analysis may have limited power since the primary objectives of the original study design may not include health utility. To comprehensively consider both survival and health utility, we developed a composite endpoint, HUS (Health Utility-adjusted Survival), which combines both

survival and utility. HUS has been shown to be able to increase statistical power and potentially reduce the required sample size compared to the standard overall survival endpoint. Nevertheless, the asymptotic properties of the test statistics of HUS endpoint have yet to be fully established. Besides that, the standard version of HUS cannot be applied to or have limited performance in certain scenarios, where extensions are needed. In this manuscript, we propose various methodological extensions of HUS and derive the asymptotic distributions of the test statistics. By comprehensive simulation studies and a data application using retrospective data based on a translational patient cohort in Princess Margaret Cancer Centre, we demonstrate the better efficiency and feasibility of HUS comparing to different methods.

Keywords: Health utility; Overall survival; Time-to-event data; Hazard ratio; Proportional hazards; Randomized controlled trials

1. INTRODUCTION

In clinical studies, including superiority and non-inferiority trials, overall survival (OS) is commonly used as the primary endpoint to compare a new treatment with a standard or controlled treatment. In some scenarios, non-inferiority trials are preferred due to some specific potential benefits of the new treatment (e.g., lower costs, fewer side effects, better quality of life, or less difficulty to implement), and it would suffice to show that the new treatment is not worse than control with respect to OS. After establishing non-inferiority, the next step is usually to demonstrate that the new treatment can benefit patients in some other clinical endpoints beyond OS, and one of such endpoints that clinicians have interest is health utility.[1] Health utility is a value ascribed to individuals' preference for a specific health state. As a measurement usually ranging from 0 to 1, health utility can quantify the health state of a patient at a certain time-point, and a higher value usually corresponds to a healthier state. Usually, a heath utility of 0 is akin to death and, in some instances, negative utilities can indicate a state worse than death. Health utilities are typically elicited either indirectly from patients by means of patient-reported outcomes with instruments such as the EQ-5D or they can be directly measured with utility elicitation techniques such as time-trade-off or standard gamble, whereby an individual is presented with a hypothetical health state and asked to select between the hypothetical health state and a lesser time or lower probability of a healthier state, respectively. Various methods for performing statistical analysis using health utility scores at different time points have been proposed in the literature.[2–4] However, the health utility analysis may be underpowered since the primary objective of the study design is usually based on OS, without taking health utility into consideration, especially if the compared treatments differ substantially in survival but only

moderately in utility. In some trials, utility is only evaluated if OS demonstrates a significant improvement. Meanwhile, it may not be desirable to perform statistical testing on OS and health utility separately since it will result in multiplicity and require multiple testing adjustment, which will lead to potential loss of statistical power. Hence, using a composite endpoint that combines survival and health utility to perform a single test may be preferred, since it may increase the statistical power and reduce the required sample size. Furthermore, if utility is analyzed separately, the absence of utility scores following a patient's death constitutes informative missingness. One might assume that utility is zero at and after death; however, this assumption may not always hold and can lead to biased results. Therefore, incorporating both survival and utility into a composite endpoint may naturally offer a more appropriate approach for handling such a missing mechanism.

While methods like Q-TWiST (Quality-adjusted Time Without Symptoms of disease or Toxicity) that can combine survival and utility have been proposed and used to analyze clinical trial data, they have their own drawbacks.[5–12] For instance, though researchers have derived some statistical properties as well as sample size formulas for Q-TWiST,[8] the implementation of this approach is limited to scenarios where each patient's status can be divided into three states (toxicity, time without symptoms and toxicity, and relapse), and the weights for different states are artificially pre-selected. In many other scenarios, especially with utility scores measured on a continuous scale, clinicians may be more interested in analyzing them in the original scale rather than forcing them into three categories, since some information is lost in the categorization process, which will make the statistical test have lower power.

A more general approach, usually referred to as QALY (Quality-Adjusted Life Years), offers an intuitive way to combine survival and health utility,[5,13–17] and a similar concept called quality-adjusted progression-free survival has also been used in some randomized trials.[18–20] Nevertheless, the formal statistical frameworks of these methods have not been established, and the potential advantages and feasibility of these methods compared to the traditional survival endpoint have not been fully evaluated through comprehensive simulation studies.

A composite endpoint called HUS (Health Utility adjusted Survival) was proposed in recent literature,[21] which can combine longitudinal health utility and survival performance to assess treatment effects. The authors also provided a detailed statistical testing framework and procedures for power analysis and sample size calculations. Although they demonstrated through simulations that HUS may increase the statistical power over classic tests based on OS only and thus reduce the required sample size, the asymptotic properties of HUS have yet to be thoroughly explored. Establishing detailed theoretical properties may make HUS a more solid approach that can benefit future clinical trials.

Meanwhile, on some special occasions, it may be beneficial to modify the test statistics in order to better capture treatment effects. For example, the utility scores recorded at later time-points may be more important than those recorded at earlier time-points, since they are better indicators of how well a patient has recovered, or ultimately how much the treatment has helped the patients improve their quality of life. The early utility scores may indicate a patient's discomfort level during or right after going through a treatment such as surgery, but they may be less important if the clinicians care more about the patient's quality of life after the recovering

process. In some other scenarios, with multiple measurements of utility recorded at each time-point, it may be useful to consider giving them different weights before combining them into a single utility score, and the weights may be pre-determined based on the clinicians' knowledge.

With the above considerations, we propose some important extensions of HUS, including a time-weighted version that allows assigning different weights to different time-points, denoted by twHUS, and a natural way to combine different utility measurements using weights. We also provide a simple process to apply HUS to clinical observational data with covariates, which may greatly extend the range of studies that HUS can be implemented on. Most importantly, we derive the asymptotic theories for HUS and the proposed extensions.

This manuscript is structured as follows. In Section 2, we present the methodology of the HUS endpoint as well as its extensions, and then we establish the theoretical properties of HUS. In Section 3, we use comprehensive simulation studies, including different scenarios with a single utility score or multiple quality of life (QoL) scores, without covariates or with covariates, to demonstrate the effectiveness of the HUS endpoint. At last, we provide a thorough discussion regarding the strengths and drawbacks of HUS as well as potential future directions in Section 4.

2. METHODS

2.1. HEALTH UTILITY SURVIVAL (HUS)

In this section, we give a brief review of health utility survival (HUS). Suppose the total length of the study is T, and $S_g(t)$ and $\overline{U}_g(t)$ are the survival function (proportion of patients alive at t) and average utility score of those alive at t for treatment group g ($g = 1,2$). Intuitively, we can

define a composite endpoint using $\int_0^T S_1(t)\bar{U}_1(t)\mathrm{d}t$ and $\int_0^T S_2(t)\bar{U}_2(t)\mathrm{d}t$. To allow survival and utility to be weighted differently, we propose a highly general class of tests analogous to the two-sample test proposed in [22], which is defined as

$$Q_{\mathrm{HUS},1} = \int_0^T [S_1(t)]^{\lambda_1}[\bar{U}_1(t)]^{\lambda_2}\mathrm{d}t, \tag{1}$$

$$Q_{\mathrm{HUS},2} = \int_0^T [S_2(t)]^{\lambda_1}[\bar{U}_2(t)]^{\lambda_2}\mathrm{d}t, \tag{2}$$

where λ_1 and λ_2 are preselected weights to reflect the importance of survival and utility. Larger weights correspond to higher importance, and the standard HUS uses $\lambda_1 = \lambda_2 = 1$. Since true $S_1(t)$ and $S_2(t)$ are unknown, we need to estimate them based on observed data. The simplest way is to obtain Kaplan-Meier (KM) estimates for the two groups separately.[23] Then we can substitute $S_1(t)$, $S_2(t)$ with $\hat{S}_1(t)$, $\hat{S}_2(t)$.

We may also use the Cox proportional hazards model,[24] treating the treatment assignment as a covariate, to obtain estimates of the survival functions. This requires the proportional hazards assumption for the two groups, which may have benefits when this assumption is not violated.

The test statistic to examine whether the two treatment groups differ in HUS is defined as

$$\mathcal{T} = Q_{\mathrm{HUS},1} - Q_{\mathrm{HUS},2}. \tag{3}$$

We can use the bootstrap or the permutation algorithm to get the empirical confidence intervals or p-values. With our derived asymptotic properties of HUS, we may also use an alternative approach to simulate the distribution of $\mathcal{T}$ and obtain its p-value. More details are provided in Appendix A of the Supplementary Materials.

2.2. Time-Weighted HUS

In some clinical studies, the importance of health utility at different time-points may vary. For example, the utility scores recorded in the later stage of the study may be more important than that recorded in the earlier stage (e.g., around surgery time), as the later scores may show how well a patient has recovered. Besides, some treatments may show less effect in the beginning, but benefit the patients' quality of life significantly more after a period of time. As a result, it may make more sense to give different weights to utility scores at different time points. We propose a time-weighted version of HUS (twHUS), with

$$Q_{\text{tHUS},1} = \int_0^T [S_1(t)]^{\lambda_1} [\overline{U}_1(t) w(t)]^{\lambda_2} \mathrm{d}t, \tag{4}$$

$$Q_{\text{tHUS},2} = \int_0^T [S_2(t)]^{\lambda_1} [\overline{U}_2(t) w(t)]^{\lambda_2} \mathrm{d}t, \tag{5}$$

where $w(t)$ is a function of weight across time. For example, we may let $w(t)$ linearly increase from 0 at baseline to 1 at the end of surgery, and then it may stay at 1 until the end of study. We use this setting by default unless otherwise specified.

2.3. Special Case with Multiple Quality of Life Scores

Note that our previous framework only focuses on one utility score, while sometimes we may have different QoL scores measuring different aspects of the wealth and comfort of patients. Suppose we have M different QoL scores, and $U_{gi,m}(t)$ is the mth QoL score for subject i at

time t in treatment group g ($g = 1,2$). To combine different scores, we may define a new score U_{gi}^{*}, with

$$U_{gi}^{*}(t) = \sum_{m=1}^{M} v_m U_{gi,m}(t), \tag{6}$$

where v_m's are the weights for different scores, and $\sum_{m=1}^{M} v_m = 1$. In some cases, if we are interested in the worst score at each time point, we may also consider $U_{gi}^{*}(t) = \min_m(U_{gi,m}(t))$. Once the combined score is defined and calculated, we can apply our previously described procedures.

2.4. HUS for Observational Data

The original framework of HUS was intended for data from randomized trials, where covariates are balanced in the two treatment groups. However, we may also be interested in analyzing observational studies that the study subjects were not randomized by treatments. In this case, due to the potential confounders that are likely unbalanced, directly applying HUS to the full data may be problematic. A simple but effective approach to alleviate the confounding issue is to conduct propensity matching first, and then analyze the matched pairs.[25] While there are more sophisticated techniques regarding propensity scores in survival analyses,[26–28] to maintain simplicity and focus, we will only demonstrate propensity score matching in our simulations.

2.5. Theoretical Properties

2.5.1 Asymptotic Distribution of HUS

In this section, we investigate the asymptotic properties of the HUS test statistic. The observed test statistic is $\hat{\mathcal{T}} = \hat{Q}_{\mathrm{HUS},1} - \hat{Q}_{\mathrm{HUS},2}$, with

$$\hat{Q}_{\mathrm{HUS},1} = \int_0^T \left[\hat{S}_1(t)\right]^{\lambda_1} [\bar{U}_1(t)]^{\lambda_2} \mathrm{d}t, \tag{7}$$

$$\hat{Q}_{\mathrm{HUS},2} = \int_0^T \left[\hat{S}_2(t)\right]^{\lambda_1} [\bar{U}_2(t)]^{\lambda_2} \mathrm{d}t. \tag{8}$$

For the simplest case without weights, we have $\lambda_1 = \lambda_2 = 1$, and

$$\hat{Q}_{\mathrm{HUS},1} = \int_0^T \hat{S}_1(t)\bar{U}_1(t)\mathrm{d}t, \tag{9}$$

$$\hat{Q}_{\mathrm{HUS},2} = \int_0^T \hat{S}_2(t)\bar{U}_2(t)\mathrm{d}t. \tag{10}$$

Note that in the above formulas, we have already replaced the unknown survival functions with KM estimates $\hat{S}_1(t)$ and $\hat{S}_2(t)$. Denote the survival time, censoring time and observed time of the ith subject in group g by T_{gi}, C_{gi} and X_{gi} respectively, and their utility score at time t by $U_{gi}(t)$, of which the expectation is $U_g(t)$. Define $N_{gi}(t) = I\{X_{gi} \le t, \Delta_{gi} = 1\}, \Delta_{gi} = I\{X_{gi} \le C_{gi}\}, Y_{gi}(t) = I\{X_{gi} \ge t\}, \bar{\mathbb{Y}}_g(t) = \sum_{i=1}^{n_g} Y_{gi}(t)$, $g = 1,2$.

The following assumptions are needed to establish the asymptotic distribution of $\mathcal{T}$ under the null hypothesis.

A.1 $\mathrm{P}(T_{gi} > T) > 0$.

A.2 $U_1(t), U_2(t)$ are of bounded variation on $[0, T]$.

Assumption A.1 is very common in survival analysis.[29,30] Assumption A.2 is to make sure the weak convergence of $\sqrt{n_g}\{\overline{U}_g(t) - U_{gi}(t)\}$, $g = 1,2$, which is very common in counting process theory.

Lemma 1 Under assumptions A.1-A.2, we have $\sqrt{n_1}\{\hat{S}_1(t)\overline{U}_1(t) - S_1(t)U_1(t)\}$ converges to a mean zero Gaussian process $\mathcal{G}_1(t)U_1(t) + S_1(t)\mathcal{U}_1(t)$. Here, $\mathcal{G}_1(t)$ is a mean zero Gaussian process with variance being $S_1^2(t)\int_0^t \lambda(u)/p(X_{1i} > u)\mathrm{d}u$, $\mathcal{U}_1(t)$ is a mean zero Gaussian process with variance being $\mathrm{E}[U_{1i}(t) - U_1(t)]^2$, and $\lambda(t)$ is the hazard function for X_{1i}.

Theorem 1 Under assumptions A.1-A.2 and the null hypothesis, if $\frac{n_1}{n_2} \to c > 0$, we have

$$\sqrt{n_1}\hat{\mathcal{T}} \to_d \sqrt{1+c}\int_0^T \mathcal{G}_1(t)U_1(t) + S_1(t)\mathcal{U}_1(t)\mathrm{d}t. \tag{11}$$

Here, $\to_d$ means converges in distribution. More details and the proofs of Lemma 1 and Theorem 1 are provided in Appendix A of the Supplementary Materials.

We can derive asymptotic properties using similar techniques for HUS with weights, where λ_1, λ_2 are pre-selected values.

Lemma 2 Under assumptions A.1-A.2, we have $\sqrt{n_1}\{\left[\hat{S}_1(t)\right]^{\lambda_1}[\overline{U}_1(t)]^{\lambda_2} - [S_1(t)]^{\lambda_1}[U_1(t)]^{\lambda_2}\}$ converges to a mean zero Gaussian process:

$\lambda_1\mathcal{G}_1(t)[S_1(t)]^{\lambda_1-1}[U_1(t)]^{\lambda_2} + \lambda_2[S_1(t)]^{\lambda_1}[U_1(t)]^{\lambda_2-1}\mathcal{U}_1(t)$.

Theorem 2 Under assumptions A.1-A.2 and the null hypothesis, if $\frac{n_1}{n_2} \to c$, we have

$$\sqrt{n_1}\hat{\mathcal{T}} \to_d \sqrt{1+c}\int_0^T \lambda_1\mathcal{G}_1(t)[S_1(t)]^{\lambda_1-1}[U_1(t)]^{\lambda_2} + \lambda_2[S_1(t)]^{\lambda_1}[U_1(t)]^{\lambda_2-1}\mathcal{U}_1(t)\mathrm{d}t. \tag{12}$$

The proofs for Lemma 2 and Theorem 2 are similar to the proofs for Lemma 1 and Theorem 1. They are available in Appendix A of the Supplementary Materials. Under the null hypothesis, the distribution of $\sqrt{n_1}\hat{\mathcal{T}}$ can be approximated well via a perturbation-resampling method.[31–33] Let Z_{1i} $(i = 1, \dots, n_1)$ and $Z_{2i} (i = 1, \dots, n_2)$ be independent random samples from $N(0,1)$. For HUS with weights λ_1, λ_2, following from the proof of Lemma 2, we can approximate the distribution of $\sqrt{n_1}\hat{\mathcal{T}}$ by

$$\begin{aligned}\Gamma^* = \sqrt{n_1}\int_0^T \lambda_1 \hat{S}_1^{\lambda_1-1}(t)\hat{S}_1(t)\sum_{i=1}^{n_1} Z_{1i}\int_0^t \frac{I\{X_{1i}\le t, \Delta_{1i}=1\}}{\overline{\mathbb{Y}}_1(X_{1i})}\overline{U}_1^{\lambda_2}(t)\mathrm{d}t \\ + \sqrt{n_1^{-1}}\int_0^T \lambda_2 \hat{S}_1^{\lambda_1}(t)\overline{U}_1^{\lambda_2-1}(t)\sum_{i=1}^{n_1} Z_{1i}\{U_{1i}(t)-\overline{U}_1(t)\}\mathrm{d}t \\ - \sqrt{\frac{n_1}{n_2}}\Bigg\{\sqrt{n_2}\int_0^T \lambda_1 \hat{S}_2^{\lambda_1-1}(t)\hat{S}_2(t)\sum_{i=1}^{n_1} Z_{2i}\int_0^t \frac{I\{X_{2i}\le t, \Delta_{2i}=1\}}{\overline{\mathbb{Y}}_2(X_{2i})}\overline{U}_2^{\lambda_2}(t)\mathrm{d}t \\ + \sqrt{n_2^{-1}}\int_0^T \lambda_2 \hat{S}_2^{\lambda_1}(t)\overline{U}_2^{\lambda_2-1}(t)\sum_{i=1}^{n_1} Z_{2i}\{U_{2i}(t)-\overline{U}_2(t)\}\mathrm{d}t\Bigg\}. \end{aligned} \quad (13)$$

According to our experience, this approximation approach has similar performance compared to the bootstrap method. As the bootstrap method is more straightforward and more commonly used, we recommend using it by default. The results in this manuscript are based on bootstrap, unless otherwise specified. Some comparison of the two methods can be found in Appendix B of the Supplementary Materials.

2.5.2 Asymptotic Distribution of Time-Weighted HUS

For twHUS, the test statistics are calculated using

$$\hat{Q}_{\text{tHUS},1} = \int_0^T \left[\hat{S}_1(t)\right]^{\lambda_1} [\overline{U}_1(t) w(t)]^{\lambda_2} \mathrm{d}t, \tag{14}$$

$$\hat{Q}_{\text{tHUS},2} = \int_0^T \left[\hat{S}_2(t)\right]^{\lambda_1} [\overline{U}_2(t) w(t)]^{\lambda_2} \mathrm{d}t. \tag{15}$$

To compare two treatment arms, we look at $\hat{\mathcal{T}}_t = \hat{Q}_{\text{tHUS},1} - \hat{Q}_{\text{tHUS},2}$. Here $\hat{\mathcal{T}}_t$ is defined as the test statistic, and we want to give the asymptotic distribution under the null hypothesis.

Lemma 3 Under assumptions A.1-A.2, we have $\sqrt{n_1}\{\left[\hat{S}_1(t)\right]^{\lambda_1} [w(t)\overline{U}_1(t)]^{\lambda_2} - [S_1(t)]^{\lambda_1} [w(t)U_1(t)]^{\lambda_2}\}$ converges to a mean zero Gaussian process, $\lambda_1 \mathcal{G}_1(t)[S_1(t)]^{\lambda_1 - 1}[w(t)U_1(t)]^{\lambda_2} + \lambda_2 [S_1(t)]^{\lambda_1}[w(t)]^{\lambda_2}[U_1(t)]^{\lambda_2 - 1}\mathcal{U}_1(t)$.

Theorem 3 Under assumptions A.1-A.2 and the null hypothesis, if $\frac{n_1}{n_2} \to c$, we have

$$\sqrt{n}\hat{\mathcal{T}}_t \to_d \sqrt{1+c} \int_0^T \lambda_1 \mathcal{G}_1(t)[S_1(t)]^{\lambda_1 - 1}[w(t)U_1(t)]^{\lambda_2} + \lambda_2 [S_1(t)]^{\lambda_1}[w(t)]^{\lambda_2}[U_1(t)]^{\lambda_2 - 1}\mathcal{U}_1(t)\mathrm{d}t. \tag{16}$$

The proofs for Lemma 3 and Theorem 3 are provided in Appendix A of the Supplementary Materials. For twHUS, by perturbation-resampling,[31–33] it is also possible to approximate the distribution of $\sqrt{n_1}\hat{\mathcal{T}}_t$ by

$$
\begin{aligned}
&\Gamma_t^* \\
&= \sqrt{n_1} \int_0^T \lambda_1 \hat{S}_1^{\lambda_1 - 1}(t) \hat{S}_1(t) \sum_{i=1}^{n_1} Z_{1i} \int_0^t \frac{I\{X_{1i} \le t, \Delta_{1i} = 1\}}{\bar{\mathbb{Y}}_1(X_{1i})} [w(t)\bar{U}_1(t)]^{\lambda_2} \mathrm{d}t \\
&+ \sqrt{n_1^{-1}} \int_0^T \lambda_2 \hat{S}_1^{\lambda_1}(t) [w(t)]^{\lambda_2} \bar{U}_1^{\lambda_2 - 1}(t) \sum_{i=1}^{n_1} Z_{1i} \{U_{1i}(t) - \bar{U}_1(t)\} \mathrm{d}t \\
&- \sqrt{\frac{n_1}{n_2}} \Bigg\{ \sqrt{n_2} \int_0^T \lambda_1 \hat{S}_2^{\lambda_1 - 1}(t) \hat{S}_2(t) \sum_{i=1}^{n_1} Z_{2i} \int_0^t \frac{I\{X_{2i} \le t, \Delta_{2i} = 1\}}{\bar{\mathbb{Y}}_2(X_{2i})} [w(t)\bar{U}_2(t)]^{\lambda_2} \mathrm{d}t \\
&+ \sqrt{n_2^{-1}} \int_0^T \lambda_2 \hat{S}_2^{\lambda_1}(t) [w(t)]^{\lambda_2} \bar{U}_2^{\lambda_2 - 1}(t) \sum_{i=1}^{n_1} Z_{2i} \{U_{2i}(t) - \bar{U}_2(t)\} \mathrm{d}t \Bigg\}. \qquad (17)
\end{aligned}
$$

2.6. Handling Missing Utility Scores

In most of the clinical studies, it is difficult to collect utility scores at every time point for all subjects. As previous literature has shown[21], in scenarios with missing utility scores, a simple but efficient way is to use linear functions to fill in the utility scores for each subject. However, using this approach may result in very unstable estimates when the number of recorded scores for a subject is too small. In this manuscript, we apply an alternative method, where we use the average score of the treatment group at a time point plus a small variation to impute the missing scores at that time point. The small variation follows a normal distribution with its mean equal to zero and variance equal to the sample variance of the non-missing utility scores at that timepoint. We only implement this procedure on time points where at least 80% of the subjects have utility scores recorded. Next, we fill in the rest of the missing scores using linear functions for each subject separately. Following this procedure, a single imputation is conducted. Our experience shows that the new imputation method yields higher statistical power compared to the previous approach. Some examples are provided in Appendix B of the Supplementary Materials.

3. RESULTS

3.1. SIMULATIONS WITH SINGLE UTILITY SCORE

We conduct simulations under various scenarios focusing on comparing the performances of different versions of HUS, as the advantages of the standard version of HUS has already been thoroughly studied.[21] By default, we simulate a randomized clinical trial data with two treatment arms mimicking the PET-NECK trial, a randomized phase III non-inferiority trial that compares Positron emission tomography-computerized tomography-guided watch-and-wait policy (PET-CT) with planned neck dissection (planned ND) for head and neck cancer patients.[1] We assume the length of study to be 36 months ($T = 36$) with each patient receiving surgery at 3 months ($C = 3$). Let T_{gi}, X_{gi} and δ_{gi} be the true survival time, observed survival time and survival status for patient i from treatment group g respectively, and n_1, n_2 be the sample sizes for group 1 and group 2. Similar to what was done in prior work,[21] we simulate the survival data using

$$T_{gi} \sim \text{Exp}(h_g),$$

$$C_{gi} \sim \text{Unif}(0, \zeta),$$

$$X_{gi} = \min(T_{gi}, C_{gi}, T),$$

$$\delta_{gi} = \begin{cases} 1 & (X_{gi} < C_{gi} \text{ and } X_{gi} < T) \\ 0 & (\text{otherwise}) \end{cases},$$

where ζ is chosen to control the censoring rate, denoted by $p_{\text{censoring}}$. Under this setting, it is easy to see that the hazard ratio of treatment 1 against treatment 2 is h_1/h_2. Note that our focus is to compare the performances of different HUS methods with different patterns of the utility difference given that the two treatments do not differ in OS, we let $h_1/h_2 = 1$ unless otherwise specified.

When simulating the health utility score, we first define the base utility at time t for group g using functions $U_{g0}(t)$. For example, $U_{g0}(t)$ can be defined as

$$U_{g0}(t) = \begin{cases} A_{g1} + \dfrac{A_{g2} - A_{g1}}{C} t\,, & (0 \le t \le C); \\ \dfrac{TA_{g2} - CA_{g3}}{T - C} + \dfrac{A_{g3} - A_{g2}}{T - C} t\,, & (C < t \le T); \end{cases}$$

which represents the average utility for group g starts from A_{g1} at baseline, linearly changes to A_{g2} at 3 months, and then linearly changes to A_{g3} at the end of the study. This setting mimics typical clinical trials where a patient's health utility reaches the lowest at the end of surgery and gradually recovers afterwards. We use $\mathrm{N}(U_{g0}(t), 0.01)$ to generate $U_{gi}(t)$, the health utility score of patient i from group g at time-point t. Since it rarely happens in practice that utility scores are fully collected at all time-points, we assume that the scores are only collected at $t = 1, C$ and T unless otherwise specified. Also, we assume there is a p_{missingU} chance that the score is missing for a subject when $t = C$ or T. We choose $p_{\mathrm{missingU}} = 0.3$ by default. Figure 1 shows the base utility functions for different scenarios. Note that in Scenarios B1 and B2, the utility changes are piece-wise smooth but not piece-wise linear. In Scenario B2, treatment 1 reaches the lowest point earlier than treatment 2. We conducted 200 iterations to assess statistical power in scenarios where the null hypothesis is false and increase the number of iterations to 500 when estimating type I error rates in scenarios where the null hypothesis is true.

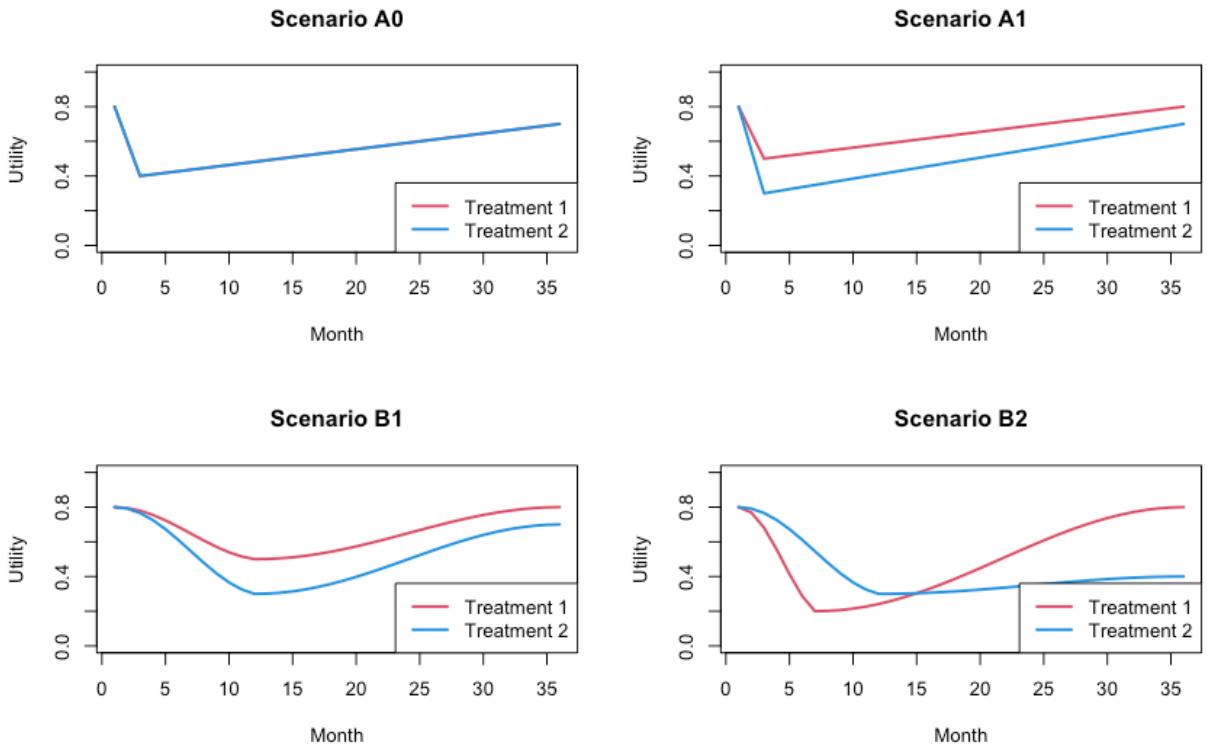

Figure 1. Utility plots for different scenarios with a single utility score.

As shown in Table 1, after applying different methods to our simulated data, we find that all methods can control type I errors in Scenario A0. Since the difference between the two treatment groups lies in utility but not in survival, assigning more weight to utility results in higher power in Scenarios A1 and B1-B2. Meanwhile, twHUS has similar performance to the standard HUS except in Scenario B2, which makes sense because in the other scenarios, the difference between the treatments is relatively consistent throughout the study, whereas in Scenario B2, treatment 1 performs worse in the earlier stage but becomes much better in the later stage. In such a scenario, using the time weights with more weights given to the later time-points, it is easier to detect the advantage of treatment 1 over that of treatment 2, and thus twHUS can obtain higher power compared to the standard HUS. For both HUS and twHUS, since the survival data was generated under the proportional hazards assumption, the Cox model is able to get slightly better results,

and thus yields slightly higher power compared to the KM method. These simulation results demonstrate the flexibility of HUS and that it is important to choose appropriate weights given their potential impact on the testing results. Note that in our main simulations, the missingness of the utility score is independent of its value. In Appendix B of the Supplementary Materials, we demonstrate that HUS can work relatively well when there is moderate informative missing, where lower utility scores are more likely to be missing.

3.2. Simulations with Multiple Quality of Life Scores

To examine the performance of our method for combing multiple measures with weights, we simulate multiple QoL scores using similar procedures as we used in section 3.1. Figure 2 shows the base QoL functions for different scores in 4 different scenarios. As shown in Table 2, in Scenario M0, where there is no difference between the two treatments in either measurement, HUS can control the type I error regardless of the choice of weights. In Scenario M1, where the difference is the same in both scores, the choice of weights does not affect the power. In Scenario M2, treatment 1 performs better than treatment 2 in terms of score 1, but it is worse in score 2. As a result, assigning more weight to score 1 (i.e., choosing a larger v_1) leads to higher power. In Scenario M3, with four QoL scores having the same difference, different choices of weights yield very similar results. This shows that choosing different weights may have a significant impact on the testing result when the differences in scores are not the same. We recommend choosing the same weight for all measurements if there is no prior knowledge of which measurements are more important.

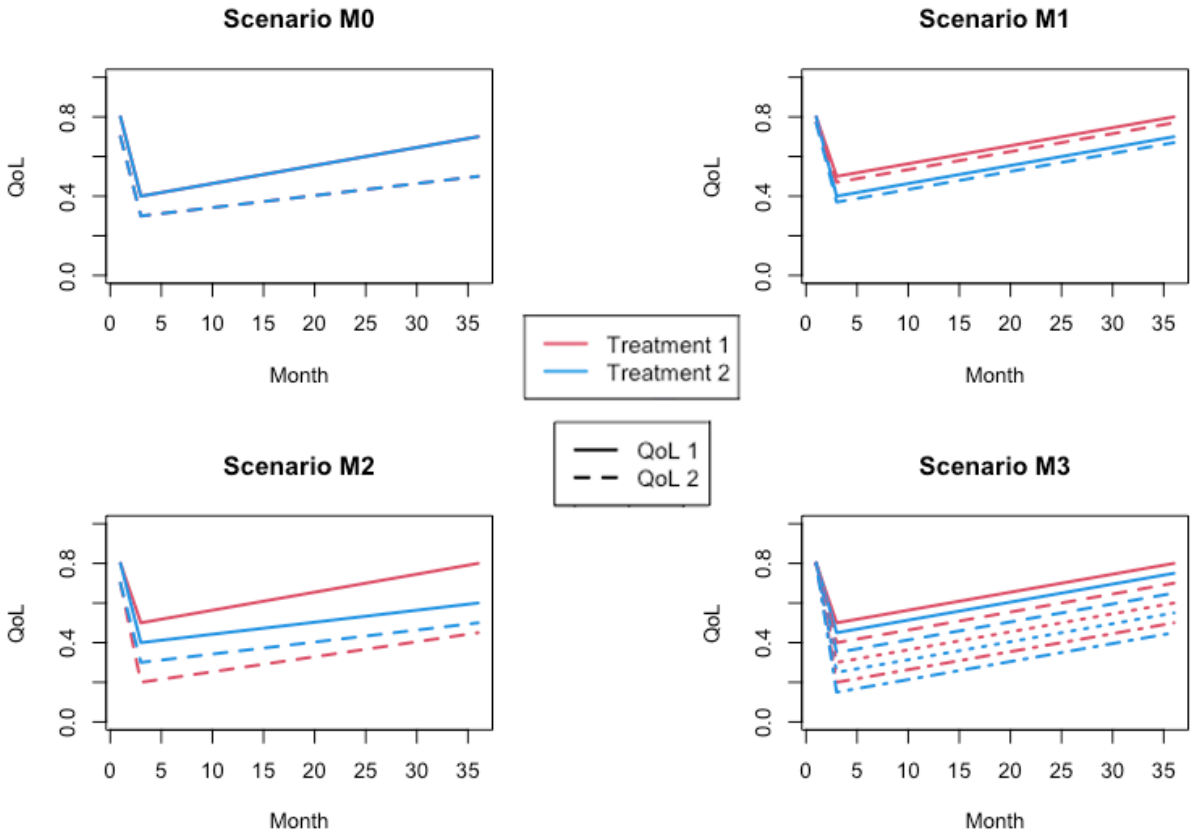

Figure 2. Quality of life plots for different scenarios with multiple measures of quality of life.

3.3. Simulation with Covariates

In this section, we consider scenarios where we have observational data and need to be adjusted for covariates. We simulate 10000 samples with covariates:

$$Age_i \sim \text{Unif}(35, 80),$$

$$Sex_i \sim \text{Bernoulli}(0.5).$$

Sex is defined as 0 for female and 1 for male. Each subject is assigned to a treatment group (1 or 2) with a probability that is a function of covariates. The probability of receiving treatment 2 is

$$p_i = 0.5 + \eta_{age}(Age_i - 57.5) + \eta_{sex}(Sex_i - 0.5).$$

We choose $\eta_{age} = 0.01$ and $\eta_{sex} = 0.2$, meaning that older patients and male patients are more likely to receive treatment 2 (control). As a result, treatment 1 has 60% female and 40% male, while treatment 2 has 40% female and 60% male. The average age is 48.8 for treatment 1 and 56.0 for treatment 2.

For each replication, we randomly select n_1 subjects from treatment group 1 and n_2 subjects from treatment group 2. For the ith subject in group g, denote their age and sex by Age_{gi} and Sex_{gi} respectively. We simulate the survival data using

$$T_{gi} \sim \text{Exp}\left(h_g e^{\beta_{age}(Age_{gi}-57.5)+\beta_{sex}(Sex_{gi}-0.5)}\right),$$

$$C_{gi} \sim \text{Unif}(0, \zeta),$$

$$X_{gi} = \min\left(T_{gi}, \xi_{gi}, T\right),$$

$$\delta_{gi} = \begin{cases} 1 & (X_{gi} < C_{gi} \text{ and } X_{gi} < T) \\ 0 & (\text{otherwise}) \end{cases},$$

where ζ is chosen to control the censoring rate, denoted by $p_{\text{censoring}}$. Given the distributions used to generate age and sex in the full sample, their mathematical expectations are 57.5 and 0.5, respectively. h_g represents the baseline hazard rate for group g when age and sex are centered at 57.5 and 0.5. The hazard ratio of treatment 1 against treatment 2, given the same age and sex, is h_1/h_2. The hazard ratio of male vs female given the same age and treatment is $e^{\beta_{sex}}$ (e.g., this hazard ratio is 1.22 if we choose $\beta_{sex} = 0.2$). The hazard ratio of age 80 vs age 35 given the same sex and treatment is $e^{45\beta_{age}}$ (e.g., this hazard ratio is 6.05 if we choose $\beta_{age} = 0.04$).

For health utility, we use the same baseline functions $U_{g0}(t)$ as before. For the ith subject in group g, we simulate its utility scores as

$$U_{gi}(t) = U_{g0}(t) + \beta_{g,age}(Age_{gi} - 57.5) + \beta_{g,sex}(Sex_{gi} - 0.5) + e_{gi},$$

which means we allow the age and sex's effects on health utility to be different depending on the treatment. For example, if we let $\beta_{1,age} = \beta_{2,age} = -0.01$ and $\beta_{1,sex} = \beta_{2,sex} = -0.1$, it will mean that older patients and male patients tend to have lower utility scores. We consider 5 scenarios: Scenario C0 uses the same base utility function as in Scenario A0, while Scenarios C1-C4 use the same base utility function as in Scenario A1. The effect sizes of the covariates are included in Table 3.

As Table 3 suggests, in Scenario C0, where there is no survival or utility difference between the two treatments, while there are covariates that are affecting the treatment assignment as well as utility scores, the naïve approach that does not adjust for covariates may have highly inflated type I errors. Meanwhile, the propensity matching technique can control type I errors. In Scenarios C1, the power of the naïve method may seem much higher, mostly due to its high inflation. In Scenarios C3-C4, the two methods have similar power, though propensity matching may slightly lose power since it uses less samples. However, in Scenario C2, where the signs of the covariate effects are modified, the naïve method has very low power, while propensity matching still has decent power. This demonstrates that the confounding issue may also result in loss of power. In conclusion, it is safer to use propensity matching when dealing with observational data, since it can reduce the bias brought by confounders.

3.4. Application to HUI3 Data

To further demonstrate the feasibility of HUS, we apply it to the HUI3 (Health Utilities Index 3) data on a translational patient cohort in Princess Margaret Cancer Centre.[34] It is a retrospective

dataset that records the patients' utility scores throughout the study as well as many baseline variables. We are interested in comparing the survival and health utility performance in patients who only received surgery treatments (Sx alone) and patients who received surgery plus combination chemotherapy and radiotherapy (Sx + POCRT). The survival time for patients was recorded from baseline up to 82.4 months, while the utility data was only recorded from baseline to 24 months. The median follow-up time is 43.8 months. We consider 4 covariates when conducting propensity matching: age, gender, stage (early or late) and HPV status. Figure 3 shows the sample sizes before and after matching as well as estimated curves based on survival, utility and the product of survival and utility. We conduct different tests to the data and compare their results.

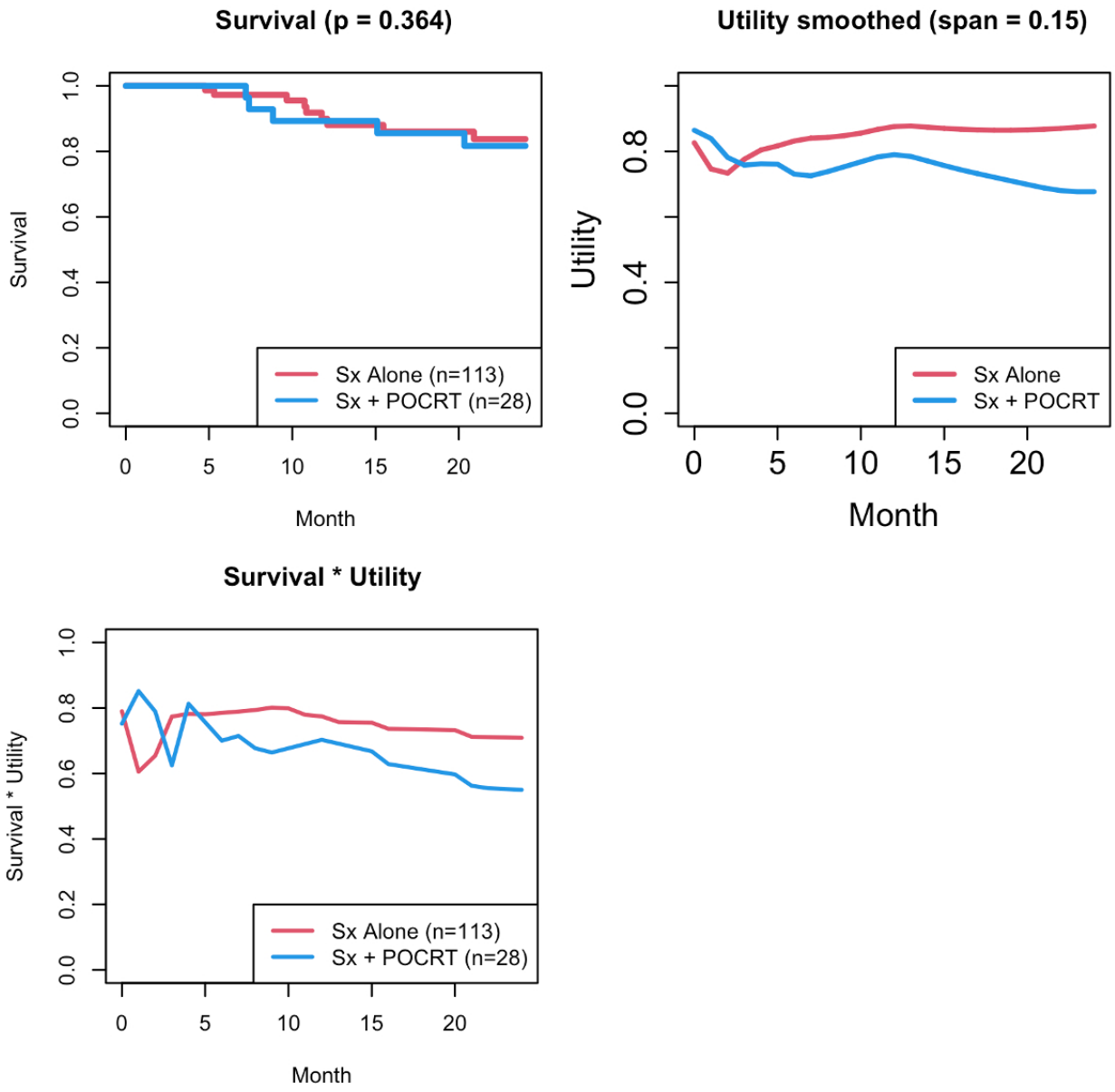

Figure 3. Curves (survival, utility, product of survival and utility) using the group average for the two treatment groups from baseline to 24 months.

As shown in Table 4, in this application, the OS-based log-rank test has insignificant results, which agrees with the survival plots in Figure 3, where there is no significant difference between the two treatment groups. Note that the p-values in those plots are results using 24 months' survival data only, whereas the p-values of OS in Table 4 use all time-points (up to 82.4 months).

Meanwhile, HUS (except $\lambda_2 = 0.5$) and the test based on utility only are able to obtain significant results, which also agrees with Figure 3. A larger λ_2 leads to a smaller p-value, which makes sense because the difference appears to be in health utility. Giving it a larger weight will lead to more significant results. These test results suggest that the treatment group that receives surgery alone tends to have better utility than the group that receives surgery and postoperative combination chemotherapy and radiotherapy. We also conducted the analysis with propensity matching, however, the sample size left after matching is too small to give meaningful conclusions.

4. DISCUSSION

We have presented the extensions of HUS to compare treatment effects with a composite endpoint combining survival and health utility with different focuses, and we have established the theoretical properties of HUS. As demonstrated by our comprehensive simulation studies and HUI3 data application, HUS can be applied to not only randomized trial data but also to observational study data, and different versions of HUS show different advantages in various scenarios. The time-weighted version, twHUS, may yield better results when we care more about the patients' eventual recovery. Using propensity matching is important for observational data as it helps reduce the inflation caused by confounders.

Note that when dealing with multiple quality of life measures, the proposed weighted-average method in this manuscript is straightforward, but there may be other techniques that can boost power in certain scenarios, especially considering there are various choices of measures with different emphases.[35–37] Exploring other options may be a potential future direction. For instance,

we may consider taking a powered sum and applying a technique similar to the SPU (Sum of Powered U-score) test,[38,39] or building different models with different assumptions and then taking the model average.[40] To better handle sparsely measured utility scores and dependent censoring, we may also consider jointly modeling the scores using a mixed model.[41] Another point worth mentioning is that the choice of weights is important in practice. The main purpose of this manuscript is to further develop the statistical framework of HUS and demonstrate the potentials of its different versions. In practice, we recommend choosing equal weights by default. Based on the clinicians' input, more weights may be given to the measures or timepoints that are considered as more important. On the other hand, if pilot data are available, different choices of weights may be applied based on the prior information and compared before finalizing the choice for the new study. In the analysis stage of the new study, we also recommend conducting sensitivity analyses.

Meanwhile, we have proposed the Cox version of HUS and showed that it has similar performance to using the KM estimates, which is likely due to the fact that our survival data are generated by proportional hazard models. In the future, we may also explore other situations where the proportional hazard assumption is violated and compare the performances of using the Cox estimates and the KM estimates. For example, when the proportional hazards assumption does not hold, the Cox model can be interpreted as estimating a time-averaged hazard ratio, which may still be quite useful.[42,43] Besides, it is also possible to perform HUS analysis with other survival models like the flexible parametric model,[44] which may show certain benefits in some situations.

Regarding observational study data, there are other options we may compare in terms of handling the confounding issue besides the propensity score matching approach.[25,45,46] For example, we may use regression models to estimate and take out the covariate effects, and then use HUS analysis on the residuals. The major challenge about this approach is that it may require very large sample sizes to get good estimates, since when we have many covariates to consider, the number of parameters to estimate in the regression models may be very large. Also, there may be unmeasured confounders that we cannot directly adjust for. Hence, it will be worthwhile to find more efficient and robust ways to analyze observational data with HUS. Besides, though our proposed approach to deal with missing data works well in our simulation settings, where the missingness is independent of or moderately associated with the utility score, we may need to explore better options, given that the missing mechanism in real data may be much less ideal. For example, death may be more related to the utility score. It is important to incorporate more advanced techniques to make HUS more robust in more complicated scenarios.[47,48] Since this manuscript focuses on the development of statistical properties, the currently used imputation approach is intuitive but relatively simple, which may introduce more bias than some more sophisticated methods. In the future, we may implement more advanced imputation techniques and compare their results.[49,50]

Finally, the current HUS framework is limited to comparing two treatment groups. In certain scenarios, we may have more than two groups of interest, and it may be desirable to have a single test to detect whether there is a difference in multiple groups instead of comparing two groups at a time, which will result in multiple testing and loss of power. For example, in our HUI3 application, we may have interests in comparing three treatment groups: patients who only

went through radiation therapy, patients who only received surgery, and patients who received combination treatments. We may develop a multi-group comparison test based on HUS that can be very useful in these situations.

5. SOFTWARE

R code for our simulation studies is available at https://github.com/yangq001/HUS.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of Dr. Hisham Mehanna (Institute of Head and Neck Studies and Education, University of Birmingham) and Dr. Sue Yom (Department of Radiation Oncology, University of California) for clinical insights and discussion.

FUNDING

This work was supported by the Alan Brown Chair in Molecular Genomics, the Lusi Wong Family Fund, and the Posluns Family Fund, all through the Princess Margaret Cancer Foundation, the Fundamental Research Funds for the Central Universities in UIBE(CXTD14-05), and the National Natural Science Foundation of China(NSFC)(Grant Nos. 12371264 and 12171329).

1

2

REFERENCES

1. Mehanna H, McConkey CC, Rahman JK, et al. PET-NECK: a multicentre randomised Phase III non-inferiority trial comparing a positron emission tomography–computerised tomography-guided watch-and-wait policy with planned neck dissection in the management of locally advanced (N2/N3) nodal metastases in patients with squamous cell head and neck cancer. *Health Technol Assess*. 2017;21(17):1-122. doi:10.3310/hta21170

2. Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL. Use of the Health Utilities Index With Stroke Patients and Their Caregivers. *Stroke*. 1997;28(10):1888-1894. doi:10.1161/01.STR.28.10.1888

3. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health Qual Life Outcomes*. 2003;1(1):54. doi:10.1186/1477-7525-1-54

4. Jewell EL, Smrtka M, Broadwater G, et al. Utility Scores and Treatment Preferences for Clinical Early-Stage Cervical Cancer. *Value Health*. 2011;14(4):582-586. doi:10.1016/j.jval.2010.11.017

5. Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. *Stat Med*. 1990;9(11):1259-1276. doi:10.1002/sim.4780091106

6. Gelber RD. Quality-of-Life-Adjusted Evaluation of Adjuvant Therapies for Operable Breast Cancer. *Ann Intern Med*. 1991;114(8):621. doi:10.7326/0003-4819-114-8-621

7. Gelber RD, Goldhirsch A, Cole BF, Wieand HS, Schroeder G, Krook JE. A Quality-Adjusted Time Without Symptoms or Toxicity (Q-TWiST) Analysis of Adjuvant Radiation Therapy and Chemotherapy for Resectable Rectal Cancer. *JNCI J Natl Cancer Inst*. 1996;88(15):1039-1045. doi:10.1093/jnci/88.15.1039

8. Murray S, Cole B. Variance and Sample Size Calculations in Quality-of-Life-Adjusted Survival Analysis (Q-TWiST). *Biometrics*. 2000;56(1):173-182. doi:10.1111/j.0006-341X.2000.00173.x

9. Konski AA, Winter K, Cole BF, Ang KK, Fu KK. Quality-adjusted survival analysis of Radiation Therapy Oncology Group (RTOG) 90-03: Phase III randomized study comparing altered fractionation to standard fractionation radiotherapy for locally advanced head and neck squamous cell carcinoma. *Head Neck*. 2009;31(2):207-212. doi:10.1002/hed.20949

10. Zbrozek AS, Hudes G, Levy D, et al. Q-TWiST Analysis of Patients Receiving Temsirolimus or Interferon Alpha for Treatment of Advanced Renal Cell Carcinoma. *PharmacoEconomics*. 2010;28(7):577-584. doi:10.2165/11535290-000000000-00000

11. Seymour JF, Gaitonde P, Emeribe U, Cai L, Mato AR. A Quality-Adjusted Survival (Q-TWiST) Analysis to Assess Benefit-Risk of Acalabrutinib Versus Idelalisib/Bendamustine

Plus Rituximab or Ibrutinib Among Relapsed/Refractory (R/R) Chronic Lymphocytic Leukemia (CLL) Patients. *Blood*. 2021;138(Supplement 1):3722-3722. doi:10.1182/blood-2021-147112

12. Jerusalem G, Delea TE, Martin M, et al. Quality-Adjusted Survival with Ribociclib Plus Fulvestrant Versus Placebo Plus Fulvestrant in Postmenopausal Women with HR±HER2− Advanced Breast Cancer in the MONALEESA-3 Trial. *Clin Breast Cancer*. 2022;22(4):326-335. doi:10.1016/j.clbc.2021.12.008

13. Glasziou PP, Cole BF, Gelber RD, Hilden J, Simes RJ. Quality adjusted survival analysis with repeated quality of life measures. *Stat Med*. 1998;17(11):1215-1229. doi:10.1002/(sici)1097-0258(19980615)17:11<1215::aid-sim844>3.0.co;2-y

14. Prieto L, Sacristán JA. Problems and solutions in calculating quality-adjusted life years (QALYs). *Health Qual Life Outcomes*. 2003;1(1):80. doi:10.1186/1477-7525-1-80

15. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull*. 2010;96(1):5-21. doi:10.1093/bmb/ldq033

16. Touray MML. Estimation of Quality-adjusted Life Years alongside clinical trials: the impact of 'time-effects' on trial results. *J Pharm Health Serv Res*. 2018;9(2):109-114. doi:10.1111/jphs.12218

17. Chung CH, Hu TH, Wang JD, Hwang JS. Estimation of Quality-Adjusted Life Expectancy of Patients With Oral Cancer: Integration of Lifetime Survival With Repeated Quality-of-Life Measurements. *Value Health Reg Issues*. 2020;21:59-65. doi:10.1016/j.vhri.2019.07.005

18. Billingham LJ, Abrams KR, Jones DR. Methods for the analysis of quality-of-life and survival data in health technology assessment. *Health Technol Assess Winch Engl*. 1999;3(10):1-152.

19. Diaby V, Adunlin G, Ali AA, Tawk R. Using quality-adjusted progression-free survival as an outcome measure to assess the benefits of cancer drugs in randomized-controlled trials: case of the BOLERO-2 trial. *Breast Cancer Res Treat*. 2014;146(3):669-673. doi:10.1007/s10549-014-3047-y

20. Oza AM, Lorusso D, Aghajanian C, et al. Patient-Centered Outcomes in ARIEL3, a Phase III, Randomized, Placebo-Controlled Trial of Rucaparib Maintenance Treatment in Patients With Recurrent Ovarian Carcinoma. *J Clin Oncol*. 2020;38(30):3494-3505. doi:10.1200/JCO.19.03107

21. Deng Y, De Almeida JR, Xu W. Health Utility Adjusted Survival: a Composite Endpoint for Clinical Trial Designs. Published online April 9, 2024. doi:10.1101/2024.04.08.24305511

22. Fleming TR, Harrington DP. A class of hypothesis tests for one and two sample censored survival data. *Commun Stat - Theory Methods*. 1981;10(8):763-794. doi:10.1080/03610928108828073

23. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452

24. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187-202. doi:10.1111/j.2517-6161.1972.tb00899.x

25. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786

26. Cheng C, Li F, Thomas LE, Li F (Frank). Addressing Extreme Propensity Scores in Estimating Counterfactual Survival Functions via the Overlap Weights. *Am J Epidemiol*. 2022;191(6):1140-1151. doi:10.1093/aje/kwac043

27. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res*. 2017;26(4):1654-1670. doi:10.1177/0962280215584401

28. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J*. 2022;15(1):14-20. doi:10.1093/ckj/sfab158

29. Hao M, Song X, Sun L. Reweighting estimators for the additive hazards model with missing covariates: REWEIGHTING ESTIMATORS FOR THE ADDITIVE HAZARDS MODEL. *Can J Stat*. 2014;42(2):285-307. doi:10.1002/cjs.11210

30. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis: Fleming/Counting*. John Wiley & Sons, Inc.; 2005. doi:10.1002/9781118150672

31. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. 1993;80(3):557-572. doi:10.1093/biomet/80.3.557

32. Parzen MI, Wei LJ, Ying Z. Simultaneous Confidence Intervals for the Difference of Two Survival Functions. *Scand J Stat*. 1997;24(3):309-314. doi:10.1111/1467-9469.t01-1-00065

33. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clin Trials*. 2012;9(5):570-577. doi:10.1177/1740774512455464

34. Ren J, Pang W, Hueniken K, et al. Longitudinal health utility and symptom-toxicity trajectories in patients with head and neck cancers. *Cancer*. 2022;128(3):497-508. doi:10.1002/cncr.33936

35. Fisk JD. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. *J Neurol Neurosurg Psychiatry*. 2005;76(1):58-63. doi:10.1136/jnnp.2003.017897

36. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Ann Med*. 2001;33(5):358-370. doi:10.3109/07853890109002090

37. Pickard AS, Ray S, Ganguli A, Cella D. Comparison of FACT- and EQ-5D–Based Utility Scores in Cancer. *Value Health*. 2012;15(2):305-311. doi:10.1016/j.jval.2011.11.029

38. Kim J, Bai Y, Pan W. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. *Genet Epidemiol*. 2015;39(8):651-663. doi:10.1002/gepi.21931

39. Pan W, Kim J, Zhang Y, Shen X, Wei P. A Powerful and Adaptive Association Test for Rare Variants. *Genetics*. 2014;197(4):1081-1095. doi:10.1534/genetics.114.165035

40. BIOS consortium, Social Science Genetic Association Consortium, Baselmans BML, et al. Multivariate genome-wide analyses of the well-being spectrum. *Nat Genet*. 2019;51(3):445-451. doi:10.1038/s41588-018-0320-8

41. Li N, Elashoff RM, Li G. Robust Joint Modeling of Longitudinal Measurements and Competing Risks Failure Time Data. *Biom J*. 2009;51(1):19-30. doi:10.1002/bimj.200810491

42. Rauch G, Brannath W, Brückner M, Kieser M. The Average Hazard Ratio – A Good Effect Measure for Time-to-event Endpoints when the Proportional Hazard Assumption is Violated? *Methods Inf Med*. 2018;57(03):089-100. doi:10.3414/ME17-01-0058

43. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika*. 1981;68(1):105-112. doi:10.1093/biomet/68.1.105

44. Lambert PC, Royston P. Further Development of Flexible Parametric Models for Survival Analysis. *Stata J Promot Commun Stat Stata*. 2009;9(2):265-290. doi:10.1177/1536867X0900900206

45. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ*. Published online October 23, 2019:l5657. doi:10.1136/bmj.l5657

46. Wunsch H, Linde-Zwirble WT, C. Angus D. Methods to adjust for bias and confounding in critical care health services research involving observational data. *J Crit Care*. 2006;21(1):1-7. doi:10.1016/j.jcrc.2006.01.004

47. Graham JW. *Missing Data*. Springer New York; 2012. doi:10.1007/978-1-4614-4018-5

48. Naeim A, Keeler EB, Mangione CM. Options for Handling Missing Data in the Health Utilities Index Mark 3. *Med Decis Making*. 2005;25(2):186-198. doi:10.1177/0272989X05275153

49. Jahangiri M, Kazemnejad A, Goldfeld KS, et al. A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis. *BMC Med Res Methodol*. 2023;23(1):161. doi:10.1186/s12874-023-01968-8

50. Cao Y, Allore H, Vander Wyk B, Gutman R. Review and evaluation of imputation methods for multivariate longitudinal data with mixed-type incomplete variables. *Stat Med*. 2022;41(30):5844-5876. doi:10.1002/sim.9592

Table 1. Simulation results for Scenarios A0-A1, B1-B2. Type I errors for Scenario A0, power for Scenarios A1, B1-B2.

Scenario A0								
n_1, n_2	KM				Cox			
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)
50	0.048	0.047	0.053	0.048	0.047	0.046	0.052	0.046
100	0.046	0.045	0.051	0.046	0.051	0.048	0.053	0.052
150	0.054	0.054	0.052	0.050	0.052	0.051	0.054	0.052
Scenario A1								
n_1, n_2	KM				Cox			
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)
50	0.81	0.41	0.99	0.80	0.85	0.46	1	0.85
100	0.95	0.62	1	0.95	0.98	0.70	1	0.98
150	0.99	0.78	1	0.99	1	0.86	1	1
Scenario B1								
n_1, n_2	KM				Cox			
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)
50	0.74	0.32	0.97	0.72	0.80	0.39	0.98	0.78
100	0.90	0.57	1	0.9	0.92	0.60	1	0.92
150	0.99	0.67	1	0.98	0.99	0.76	1	0.98
Scenario B2								
n_1, n_2	KM				Cox			
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$ (twHUS)
50	0.24	0.11	0.71	0.58	0.30	0.12	0.75	0.62
100	0.44	0.17	0.90	0.82	0.49	0.18	0.92	0.85
150	0.54	0.21	0.99	0.92	0.60	0.22	0.98	0.95

Table 2. Simulation results for Scenarios M0-M3. Type I errors for Scenario M0, power for Scenarios M1-M3.

Scenario M0			
n_1, n_2	$v_1 = 0.5$ $v_2 = 0.5$	$v_1 = 0.8$ $v_2 = 0.2$	$v_1 = 0.2$ $v_2 = 0.8$
50	0.050	0.053	0.052
100	0.050	0.051	0.049
150	0.052	0.045	0.051
Scenario M1			
n_1, n_2	$v_1 = 0.5$ $v_2 = 0.5$	$v_1 = 0.8$ $v_2 = 0.2$	$v_1 = 0.2$ $v_2 = 0.8$
50	0.35	0.33	0.32
100	0.59	0.54	0.54
150	0.69	0.67	0.69
Scenario M2			
n_1, n_2	$v_1 = 0.5$ $v_2 = 0.5$	$v_1 = 0.8$ $v_2 = 0.2$	$v_1 = 0.2$ $v_2 = 0.8$
50	0.06	0.25	0.02
100	0.12	0.47	0
150	0.14	0.60	0.02

Scenario M3		
n_1, n_2	$v_1 = 0.25$ $v_2, v_3, v_4 = 0.25$	$v_1 = 0.4$ $v_2, v_3, v_4 = 0.2$
50	0.16	0.14
100	0.32	0.28
150	0.34	0.33

Table 3. Effect sizes of covariates and testing results for Scenarios C0-C4.

Scenario C0						
Effect size	Assign	Survival	Utility (group 1)	Utility (group 2)		
Age	0.01	0.04	-0.01	-0.01		
Sex	0.2	0.2	-0.1	-0.1		
Type I error						
n_1, n_2	Naïve			Propensity matching		
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$
50	0.55	0.42	0.62	0.04	0.04	0.03
100	0.84	0.71	0.89	0.04	0.04	0.04
Scenario C1						
Effect size	Assign	Survival	Utility (group 1)	Utility (group 2)		
Age	0.01	0.04	-0.01	-0.01		
Sex	0.2	0.2	-0.1	-0.1		
Power						
n_1, n_2	Naïve			Propensity matching		
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$
50	0.97	0.80	1	0.32	0.21	0.52
Scenario C2						
Effect size	Assign	Survival	Utility (group 1)	Utility (group 2)		
Age	0.01	-0.02	0.01	0.01		
Sex	0.2	-0.1	0.1	0.1		
Power						
n_1, n_2	Naïve			Propensity matching		
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$
50	0.07	0.06	0.10	0.38	0.20	0.52
100	0.07	0.04	0.14	0.63	0.38	0.84
150	0.06	0.03	0.14	0.76	0.44	0.92
Scenario C3						
Effect size	Assign	Survival	Utility (group 1)	Utility (group 2)		
Age	0	0.04	-0.01	-0.01		
Sex	0	0.2	-0.1	-0.1		
Power						
n_1, n_2	Naïve			Propensity matching		
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$
50	0.46	0.24	0.68	0.41	0.21	0.62
100	0.65	0.35	0.85	0.68	0.36	0.86
150	0.80	0.47	0.94	0.84	0.44	0.97

Scenario C4						
Effect size		Assign	Survival	Utility (group 1)		Utility (group 2)
Age		0	0	0		0
Sex		0	0	0		0
Power						
n_1, n_2	Naïve			Propensity matching		
	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$	$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 2$
50	0.59	0.28	0.84	0.56	0.28	0.76
100	0.79	0.49	0.97	0.73	0.44	0.94
150	0.92	0.56	0.99	0.90	0.56	0.98

Table 4. P-values of different tests. OS stands for the log-rank test to test the difference in overall survival. HU only means compare the health utility without comparing survival (i.e., set $\lambda_1 = 0, \lambda_2 = 1$ in HUS).

OS	HU only	HUS			
		$\lambda_2 = 1$	$\lambda_2 = 0.5$	$\lambda_2 = 1.5$	$\lambda_2 = 1$ (twHUS)
0.513	< 0.0001*	0.048*	0.186	0.016*	0.052