

# It's integral: Replacing the trapezoidal rule to remove bias and correctly impute censored covariates with their conditional means

Sarah C. Lotspeich

*Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC, U.S.A.  
Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*

E-mail: lotspes@wfu.edu

Tanya P. Garcia

*Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*

**Summary.** Imputing censored covariates with conditional means is appealing, but existing methods saw  $>100\%$  bias. Calculating conditional means requires estimating and integrating over the survival function of the censored covariate from the censored value to infinity. Existing methods semiparametrically estimate the survival but incur bias by using the trapezoidal rule, thereby treating this indefinite integral as a definite one. We integrate with adaptive quadrature instead. Yet, the integrand is undefined beyond the data, so we identify the best extrapolation method to use with quadrature. Our approach leads to unbiased imputation in simulations and helps prioritize patients for Huntington's disease clinical trials.

**Keywords:** Adaptive quadrature; Breslow's estimator; Conditional mean imputation; Huntington's disease; Time to diagnosis.

## 1. Introduction

### 1.1. Modeling the Progression of Huntington's Disease

Prospective studies are common for genetically inherited diseases because, with genetic testing, researchers can identify at-risk subjects and follow their symptom development over time. Such studies are especially powerful for Huntington's disease, a genetically inherited neurodegenerative disease caused by unstable cytosine-adenine-guanine (CAG) repeats in the HTT gene (The Huntington's Disease Collaborative Research Group, 1993). Huntington's disease is fully penetrant, so anyone with  $> 36$  CAG is guaranteed to develop the disease. One such prospective study is the Neurobiological Predictors of Huntington's Disease (PREDICT-HD) (Paulsen et al., 2008).

Modeling the progression of Huntington's disease using data from prospective studies like PREDICT-HD is appealing, for example, as we investigate experimental treatments designed to slow or delay symptoms. Models of how impairment (i.e., in daily, motor, and cognitive function) progresses relative to the time of clinical diagnosis can help identify subjects to recruit into clinical trials. Huntington's disease symptoms are most

detectable in the few years immediately before and after a diagnosis, so subjects in this window of time would be ideal to test a new therapy in a clinical trial. However, Huntington’s disease progresses slowly, with functional, motor, and cognitive decline spanning decades, so prospective studies often end before all at-risk subjects have met the diagnosis criteria. (A diagnosis is made when motor abnormalities are unequivocal signs of Huntington’s disease (Huntington Study Group, 1996).) Therefore, the slow-moving nature of the disease leaves the key variable “time to diagnosis” right-censored among subjects who have yet to be diagnosed (i.e., their motor abnormalities will merit a diagnosis sometime after their last study visit, but exactly when is unknown). Thus, we face a pressing statistical challenge when investigating Huntington’s disease progression: how to model the association between a fully observed outcome (impairment) and a randomly right-censored covariate (time to diagnosis).

### 1.2. *Imputing a Censored Covariate*

Inspired by missing data techniques, one appealing strategy is conditional mean imputation, where we replace all right-censored times to diagnosis with their **conditional means** (Atem et al., 2017, 2019a,b). This conditional mean imputation ensures that the imputed time to diagnosis is realistic (i.e., after the last study visit) and adjusts for other variables that may influence time to diagnosis (e.g., CAG repeat length). (Conditional mean imputation could be adopted in a single or multiple imputation framework. For simplicity, we focus on single imputation in this paper; however, multiple imputation would encounter the same challenges and could be corrected in the same ways that we are about to introduce.) The conditional mean for a right-censored value is the expected time to diagnosis given that it must happen after the censored value (the last study visit) and additional covariates. In theory, this expected time to diagnosis can be anywhere from the last study visit to infinity, so computing it involves an integral over this range.

Existing approaches to conditional mean imputation use the trapezoidal rule to compute this integral (Atem et al., 2017, 2019a,b). Specifically, they define partitions based on the observed covariate values and their corresponding survival functions. However, the trapezoidal rule is intended for definite integrals (meaning those with finite lower and upper bounds), not indefinite integrals, which are needed to compute conditional means. Existing methods rely on the data to define their upper bound, ending the final partition at the largest observed covariate value. Thus, for the trapezoidal rule to hold in this indefinite integral case, the largest observed covariate value in the data must represent the variable’s true maximum (which, in theory, could be infinity); otherwise, data beyond that value will be “cut off.” Since the survival function of the censored covariate is nonnegative and decreases monotonically, this “cut-off” leads the trapezoidal rule to **underestimate** the integral and miscalculate the conditional means.

For example, if the last time to diagnosis was 10 years from study entry, the trapezoidal rule assumes that all unobserved times to diagnosis should be observed within 10 years of study entry. Yet, in reality, diagnosis could occur at any time between the last study visit and death, both of which are unique to each subject. The trapezoidal rule is thus likely to impute censored covariates with incorrect conditional means, leading to invalid statistical inference. To avoid this situation, we propose several improvements to conditional mean imputation for a censored covariate.

### 1.3. Numerical Integration of Indefinite Integrals

Many methods may come to mind that handle integrals with infinite bounds, such as Gauss–Hermite quadrature. In fact, there are many attractive methods for numerical integration already implemented in existing software that can handle indefinite bounds, for example, the `integrate` function in R, which uses adaptive quadrature (R Core Team, 2019). However, even with these methods, we can only integrate over values of the covariate where the integrand is defined.

As we will discuss in Section 2.2, calculating the conditional mean involves integrating over the conditional survival function (the integrand) of the censored covariate up to infinity. Typically, this function relies on a `step function` (in this case, Breslow’s estimator), which is only defined up to the largest uncensored covariate value. Importantly, this step function leaves the integrand undefined beyond the observed covariate values and to infinity, so more accurate quadrature alone will not improve the estimation of the conditional means. (This is “typical” because nonparametric or semiparametric estimators are often chosen because of their distribution-free robustness but they rely on step functions; a parametric estimator would already be defined up to infinity.)

To truly improve the calculation, we need the integrand to be defined up to the infinite bound in the conditional mean formula. Specifically, we need a way to extrapolate from Breslow’s estimator beyond the largest uncensored value so that we can adopt an improved approach (in our case, adaptive quadrature) to integrate over it. Extrapolation methods are well established. However, our needs are unique: we are not just interested in extending the survival curve – any of the “usual” methods like those in Klein and Moeschberger (2003) would work if so – but in further integrating over it.

In search of the best one for our purposes, we thoroughly explored various methods to extend the survival estimator and identified the best one for conditional mean imputation (Section 3.3). To our knowledge, only one paper had investigated this previously (Datta, 2005). They considered fewer methods and, in fact, we found that their recommended method could lead to `bias` even with adaptive quadrature.

Importantly, extending the survival curve for indefinite integration is not a challenge unique to imputation. Any nonparametric or semiparametric `full-likelihood approach` with a censored covariate would also need to integrate up to infinity over an integrand that is not defined over that range. Thus, our proposed improvements hold broader implications and could be adopted to improve other methods, like a maximum likelihood estimator, as well.

### 1.4. Overview

We propose an improved conditional mean calculation to impute censored covariates in statistical models: one that replaces the trapezoidal rule with adaptive quadrature with an infinite upper bound. Since `Breslow’s estimator` is not well defined for larger values than those in the data, we explore various extrapolation methods and identify the “Weibull extension” as the best one for use with quadrature. We quantify the bias introduced by the trapezoidal rule and show in extensive `simulation studies` that replacing it with adaptive quadrature and the Weibull extension corrects for that bias. We further show how imputing with biased conditional means can impact clinical trial recruitment. The rest of the paper is as follows: we describe the proposed methods in

Section 2, we evaluate those methods against existing ones through extensive simulations in Section 3, we apply both approaches to the analysis of Huntington’s disease data from the PREDICT-HD study in Section 4, and we discuss our findings in Section 5.

## 2. Methods

### 2.1. Model and Data

Consider an outcome  $Y$  and covariates  $(X, \mathbf{Z})$ , which are assumed to be related through a regression model parameterized by  $\boldsymbol{\theta}$  and denoted by  $P_{\boldsymbol{\theta}}(Y|X, \mathbf{Z})$ . For example, if  $Y$  given  $(X, \mathbf{Z})$  follows a linear regression model, we would have that  $P_{\boldsymbol{\theta}}(Y|X, \mathbf{Z}) = 1/(\sqrt{2\pi\sigma^2}) \exp\{-(Y - \alpha - \beta X - \boldsymbol{\gamma}^T \mathbf{Z})^2/(2\sigma^2)\}$ , where  $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\gamma}^T, \sigma^2)^T$ . Statistical inference of  $\boldsymbol{\theta}$  is our primary interest.

Unfortunately, estimating  $\boldsymbol{\theta}$  is difficult because the covariate  $X$  is right-censored. Rather than observe  $X$  directly, we observe  $W = \min(X, C)$  and  $\Delta = I(X \leq C)$ , where  $C$  is a random censoring value. (Having  $C$  random rather than fixed means that  $C$  changes for every subject. For Huntington’s disease studies,  $C$  is the subject-specific length of follow-up from first to last study visit.) Thus, an observation for subject  $i$  in a sample of  $n$  subjects is captured as  $(Y_i, \Delta_i, W_i, \mathbf{Z}_i)$ .

### 2.2. Conditional Mean Imputation

In missing data settings, imputation is a popular approach to obtain valid statistical inference without sacrificing the power of the full sample. Imputation is also a promising method to handle censored covariates, with one simple change. When  $X_i$  is right-censored, rather than impute any value for it, we impute a value that is larger than  $W_i$  because, by the definition of right-censoring, the true unobserved  $X$  must be larger than  $W_i$ . This partial information (i.e., that  $X > W_i$ ) is captured through a conditional mean imputation approach (Little, 1992; Richardson and Ciampi, 2003).

In conditional mean imputation, we replace right-censored covariates  $W_i$  with their corresponding conditional means

$$E(X|X > W_i, \mathbf{Z}_i) = W_i + \frac{\int_{W_i}^{\infty} S(x|\mathbf{Z}_i)dx}{S(W_i|\mathbf{Z}_i)}, \quad (1)$$

where  $S(t|\mathbf{z})$  is the conditional survival function for  $X$  given  $\mathbf{Z}$ . To our knowledge, this form was first introduced by Atem et al. (2017), with a thorough derivation set forth by Lotspeich et al. (2022). Note that we use the  $i$  subscript for  $W_i$  and  $\mathbf{Z}_i$  because these are observed values of random variables  $W$  and  $\mathbf{Z}$ , respectively, whereas  $X$  (no subscript) is still random. Importantly, deriving Equation (1) relies on the assumption of noninformative censoring, such that the censoring values  $C$  and true covariates  $X$  are assumed to be conditionally independent given the other fully observed covariates  $\mathbf{Z}$ .

Now, conditional mean imputation proceeds in two stages. First, we calculate the conditional means for all censored covariates, which requires estimating  $S(t|\mathbf{z})$  (Section 2.3) and approximating the integral over it (Sections 2.4–2.5). Then, we replace the censored covariates with these conditional means and fit the outcome model for  $Y$  given (imputed)  $X$  and  $\mathbf{Z}$  using the “usual” methods, e.g., ordinary least squares, to obtain the conditional mean imputation estimators  $\hat{\boldsymbol{\theta}}$ .

### 2.3. Estimating the Survival Function

To robustly estimate  $S(t|\mathbf{Z})$  in Equation (1) without assuming a distribution for  $X$  given  $\mathbf{Z}$  (and in doing so, bypassing some potential misspecification), existing approaches use semiparametric models (Atem et al., 2017, 2019a,b). Specifically, existing approaches use a **Cox proportional hazards model**, from which the survival function can be calculated as  $\hat{S}(t|\mathbf{z}) = S_0(t)^{\exp(\hat{\lambda}^T \mathbf{z})}$  with  $\lambda$  the log hazard ratios and  $S_0(t)$  the baseline survival function of  $X$  (i.e., at  $\mathbf{Z} = \mathbf{0}$ ).

This semiparametric model for  $S(t|\mathbf{Z})$  requires estimating two key parts: (i) the log hazard ratios  $\lambda$  and (ii) the baseline survival function  $S_0(t)$ . The log hazard ratios  $\hat{\lambda}$  are easily estimated from existing software, like the **coxph** function in the **survival** package (Therneau and Grambsch, 2000), and a common way to estimate  $S_0(t)$  is with **Breslow's estimator** (Breslow, 1972):

$$\hat{S}_0(t) = \exp \left[ - \sum_{i=1}^n \mathbf{I}(W_i \leq t) \left\{ \frac{\Delta_i}{\sum_{j=1}^n \mathbf{I}(W_j \leq W_i) \exp(\hat{\lambda}^T \mathbf{Z}_j)} \right\} \right]. \quad (2)$$

After estimating  $\hat{\lambda}$  and  $\hat{S}_0(t)$ , we will construct  $\hat{S}(t|\mathbf{z}) = \hat{S}_0(t)^{\exp(\hat{\lambda}^T \mathbf{z})}$  and use this estimated survival function to compute  $E(X|X > W_i, \mathbf{Z}_i)$  from Equation (1). Alas, computing this conditional mean still requires integrating over  $\hat{S}(t|\mathbf{z})$  from  $t = W_i$  to  $\infty$ .

### 2.4. The Problem with Using the Trapezoidal Rule to Calculate Conditional Means

Existing approaches use the trapezoidal rule to estimate this integral and compute the conditional means. That is, they estimate the integral  $\int_{W_i}^{\infty} \hat{S}_0(x)^{\exp(\hat{\lambda}^T \mathbf{Z}_i)} dx$  with

$$\frac{1}{2} \left[ \sum_{j=1}^{n-1} \mathbf{I}(W_{(j)} \geq W_i) \left\{ \hat{S}_0(W_{(j+1)})^{\exp(\hat{\lambda}^T \mathbf{Z}_i)} + \hat{S}_0(W_{(j)})^{\exp(\hat{\lambda}^T \mathbf{Z}_i)} \right\} (W_{(j+1)} - W_{(j)}) \right], \quad (3)$$

where  $W_{(1)} < \dots < W_{(n)}$  denote the  $n$  distinct, ordered values of  $W$  from the data. Going forward, let the conditional mean following the trapezoidal rule be  $\hat{E}(X|X > W_i, \mathbf{Z}_i) =$

$$W_i + \frac{1}{2} \left( \frac{\left[ \sum_{j=1}^{n-1} \mathbf{I}(W_{(j)} \geq W_i) \left\{ \hat{S}_0(W_{(j+1)})^{\exp(\hat{\lambda}^T \mathbf{Z}_i)} + \hat{S}_0(W_{(j)})^{\exp(\hat{\lambda}^T \mathbf{Z}_i)} \right\} (W_{(j+1)} - W_{(j)}) \right]}{\hat{S}_0(W_i)^{\exp(\hat{\lambda}^T \mathbf{Z}_i)}} \right).$$

This formula for the conditional mean is prominent in the current literature around imputing randomly right-censored covariates (Atem et al., 2017, 2019a,b; Lotspeich et al., 2022).

Notice that the “trapezoids” in Expression (3) are defined between the observed values  $W_{(j)} \geq W_i$  and their survival functions given the  $i$ th subject's covariates,  $\hat{S}(W_{(j)}|\mathbf{Z}_i)$ . Some  $W_{(j)}$  will be censored, so computing  $\hat{E}(X|X > W_i, \mathbf{Z}_i)$  requires evaluating  $\hat{S}_0(\cdot)$  between and beyond the uncensored data on which it is defined. Between uncensored values,  $\hat{S}_0(\cdot)$  should be carried forward (interpolated) from the last uncensored value.

Beyond the largest uncensored value,  $\hat{S}_0(\cdot)$  is not well defined; we consider multiple methods to extrapolate from it in Section 2.5.

*Remark 2.1.* Instead of using Breslow’s estimator as defined, the existing approaches (e.g., Atem et al. (2019a)) interpolate with the mean of  $\hat{S}_0(\cdot)$  from the uncensored values immediately below and above a censored  $W_{(j)}$ . Here, we will adopt **carry forward** interpolation because it is computationally simple and follows from the original formula in Breslow (1972), although we show in Section 3.3 that either mean or carry forward interpolation seems to work well.

Critically, we recognize that this use of the trapezoidal rule in Expression (3) estimates **the wrong integral**, i.e.,  $\int_{W_i}^{W_{(n)}} \hat{S}_0(x)^{\exp(\hat{\lambda}^T \mathbf{z}_i)} dx$  rather than  $\int_{W_i}^{\infty} \hat{S}_0(x)^{\exp(\hat{\lambda}^T \mathbf{z}_i)} dx$ . The **quality** of this estimate, and with it the quality of the conditional means, hinges on how well the maximum of the observed covariate  $W_{(n)}$  represents the true maximum of the covariate  $X$ . If  $W_{(n)}$  is far below the true upper bound of  $X$ , then approximating with  $\int_{W_i}^{W_{(n)}} \hat{S}_0(x)^{\exp(\hat{\lambda}^T \mathbf{z}_i)} dx$  will underestimate the integral by “cutting off” the tail of the survival function. We conclude that using the trapezoidal rule to calculate conditional means is only appropriate when  $\hat{S}_0(W_{(n)}) \approx 0$ , because in this case the survival function is entirely captured by  $W_{(1)} < \dots < W_{(n)}$ . Therefore, we set out to propose a more general approach to correctly calculate conditional means even when  $\hat{S}_0(W_{(n)}) > 0$ .

## 2.5. Replacing the Trapezoidal Rule with Adaptive Quadrature

We sought an improved calculation to capture the entirety of the indefinite integral in the conditional means by extending beyond  $W_{(n)}$  to better approximate the infinite upper bound. Conveniently, the **integrate** function in R implements “adaptive quadrature of functions ... over a finite or infinite interval” (Piessens et al., 1983; R Core Team, 2019). This function is included in the basic R functions and does not require installing any additional packages, making it an accessible and sustainable software choice. Telling the **integrate** function that we want an infinite upper bound is simple enough. In fact, as a user, it is no different than with a finite one.

Still, adopting software that can integrate up to infinity does us no good if the integrand, i.e., the survival function of the censored covariate, is not defined as such; this is a problem not just for **integrate** but for all quadrature software. Before using adaptive quadrature with an infinite upper bound, we have to “extend” Breslow’s estimator beyond the largest uncensored value  $\tilde{X} = \max(W_1 \Delta_1, \dots, W_n \Delta_n)$ . This way, we will give the **integrate** function something to integrate over on its way up to infinity and better calculate the conditional means, as desired.

### 2.5.1. Extending Breslow’s estimator beyond the largest uncensored value

We sought a method to extend Breslow’s estimator beyond the largest uncensored covariate value  $\tilde{X}$ , i.e., to extrapolate from  $\hat{S}_0(t)$  for values of  $t$  up to infinity. Extrapolating from step functions is a common challenge with censored outcomes, since popular estimators, like Kaplan–Meier, are not well defined for values of  $t > \tilde{X}$ , either. We discuss four potential methods to extend Breslow’s estimator.

*Carry forward:* Carry forward Breslow’s estimator from  $\tilde{X}$ . By estimating  $\hat{S}_0(t) = \hat{S}_0(\tilde{X})$  for all  $t > \tilde{X}$ , this asserts that all censored covariates would have had  $X = \infty$ .

*Immediate drop-off:* Do not extrapolate from Breslow’s estimator at all. Assuming that  $\hat{S}_0(t) = 0$  at all  $t > \tilde{X}$  is equivalent to assuming that the true values  $X$  for all censored covariates would have fallen just beyond their observed values  $W_i$ .

*Exponential extension:* “Tie in” an exponential survival function where Breslow’s estimator leaves off and assume that  $\hat{S}_0(t) = \exp \left( \left[ t \log \left\{ \hat{S}_0(\tilde{X}) \right\} \right] / \tilde{X} \right)$  for  $t > \tilde{X}$ .

*Weibull extension:* For added flexibility, “tie in” a Weibull survival function instead of an exponential and assume that  $\hat{S}_0(t) = \exp(-\hat{\rho}t^{\hat{\nu}})$  for  $t > \tilde{X}$ , where  $\hat{\nu}$  and  $\hat{\rho}$  are found using constrained maximum likelihood estimation (Moeschberger and Klein, 1985).

While these methods are well established for censored outcomes, to our knowledge we are the first to consider them for censored covariates. Also, our needs are unique, since we are extrapolating from the survival curve to then integrate over it. Without an extrapolation method, improving the conditional mean calculation from a step survival function like Breslow’s estimator would be impossible; no matter how well we can integrate, the integrand must be defined across the entire range, which requires extrapolation.

Either carry forward or immediate drop-off would be a valid modification if we were just modeling the survival function, since they are known to converge to the true survival functions in large samples (Klein and Moeschberger, 2003). However, neither is a good choice when we are integrating over the survival function. Carry forward makes the integral up to infinity diverge. Immediate drop-off forces the integral to cut off at  $\tilde{X}$ ; therefore, we expect it to offer little improvement over the trapezoidal rule, even with adaptive quadrature. (This is the method recommended by Datta (2005), and we show empirically in Section 3.3 that our expectation of its performance held true.) Fortunately, theoretical justification exists for both parametric extensions, so we explored them in extensive simulations before making recommendations (Section 3.2). Derivations for the parametric extensions can be found in Web Appendix A, along with an illustration of these extrapolation methods (Supplemental Fig. S1).

*Remark 2.2.* Calculating conditional means with the trapezoidal rule still involves evaluating  $\hat{S}_0(t)$  for values of  $t > \tilde{X}$ . The existing approaches (e.g., Atem et al. (2019a)) treat the largest value  $W_{(n)}$  as uncensored regardless of  $\Delta_{(n)}$  so that  $\hat{S}_0(W_{(n)}) = 0$ . This method is equivalent to immediate drop-off but its impact is subtle, since the trapezoidal rule cuts the tail off anyway.

### 3. Simulation Studies

We first show that even when the survival function is the truth, imputation using the trapezoidal rule leads to biased model estimates (Section 3.2). We then demonstrate the impact of imputing with the two conditional means in the more realistic setting when the survival function is estimated. Before we can use adaptive quadrature with an



infinite upper bound (hereafter called “adaptive quadrature”), we must decide how to extend Breslow’s estimator. We choose the Weibull extension, which we show offers low bias and high efficiency even when  $X$  given  $\mathbf{Z}$  is not truly Weibull (Section 3.2). Lastly, we highlight the improvements (i.e., substantially reduced bias and some heightened efficiency) of imputation using adaptive quadrature with an estimated survival function (Section 3.4).

### 3.1. Data Generation and Metrics for Comparison

Our simulation settings are based on those of Atem et al. (2017), who, to the best of our knowledge were the first to propose conditional mean imputation for a randomly right-censored covariate. We simulated data for samples of  $n = 100, 500$ , or  $2000$  subjects in the following way. First, a binary covariate  $Z$  was generated from a Bernoulli distribution with  $P(Z = 1) = 0.5$ . Next,  $X$  was generated from a Weibull distribution with shape = 0.75 and scale = 0.25, leading to proportional hazards in  $X$  given  $Z$ . Then, a continuous outcome was generated as  $Y = 1 + 0.5X + 0.25Z + e$ , where  $e$  was a standard normal random variable. We explored light ( $\sim 12\%$ ), moderate ( $\sim 41\%$ ), and heavy ( $\sim 78\%$ ) censoring in  $X$ , induced by generating  $C$  from an exponential distribution with rates = 0.5, 2.9, and 20, respectively. Notice that  $C$  was generated independently of all other variables, which more than satisfies our assumption of noninformative censoring. Finally,  $W = \min(X, C)$  and  $\Delta = I(X \leq C)$  were constructed.

Given a continuous outcome  $Y$ , the analysis model  $P_{\theta}(Y|X, \mathbf{Z})$  was a linear regression. We considered two imputation approaches to estimate  $\hat{\theta}$ : one using adaptive quadrature and the other using the trapezoidal rule. To assess validity, we report the empirical bias and standard errors for  $\hat{\theta}$ . To gauge statistical precision, we report the relative efficiency, which was calculated as the empirical variance of the full cohort analysis (i.e., where all  $n$  observations had uncensored  $X$ ) divided by the empirical variance of the imputation approaches. The closer the relative efficiency is to one, the more efficiency was recovered through imputation. Unless otherwise stated, all summary metrics (bias, standard errors, and relative efficiency) are based on 1000 replications.

### 3.2. Using the Gold Standard: Conditional Mean Imputation with the True Survival Function

Using the true survival function allowed us to isolate the improvements due to replacing the trapezoidal rule with adaptive quadrature for conditional mean imputation. This “gold standard” imputation approach removed the uncertainty around the survival function, since it is assumed, rather than estimated. Also, there was no need for extrapolation: the Weibull survival function  $S(t|z) = \exp\{-(t/0.25)^{0.75}\}$  was already defined for  $t$  up to infinity. Thus, we first considered imputing censored  $X$  with its conditional mean based on this true  $S(t|z)$ .

As seen in Table 1, the bias in estimating  $\hat{\beta}$  (the coefficient on censored  $X$ ) using conditional mean imputation with the trapezoidal rule was alarming: as high as 181% and 24% under heavy and moderate censoring, respectively. Meanwhile, using adaptive quadrature instead led to virtually unbiased estimates everywhere, with  $\leq 5\%$  bias for all settings. Also, even under light censoring – when the trapezoidal rule was reasonably



unbiased – imputation using adaptive quadrature could lead to more efficient inference (e.g., relative efficiency = 0.90 vs. 0.81 with  $n = 100$ ), though in some cases not by much. For either imputation approach, the relative efficiency to the full cohort analysis for  $\hat{\beta}$  decreased as censoring increased. This result was expected since we impute more and incur more uncertainty when more covariates are censored.

Inference about the other coefficients  $\hat{\alpha}$  and  $\hat{\gamma}$  was comparable between the two imputation approaches. Namely, both approaches were unbiased ( $< 3\%$ ), and while some efficiency was lost in estimating the intercept (relative efficiency  $\geq 0.35$  for  $\hat{\alpha}$ ), the efficiency for the coefficient on uncensored  $Z$  was nearly equal to the full cohort analysis (relative efficiency  $\geq 0.92$  for  $\hat{\gamma}$ ).

Recall that both imputation approaches used the same true survival function; they differed only in how they approximated the integral over it. The impact of this difference between integral approximations was evident. Replacing the trapezoidal rule with adaptive quadrature led to huge reductions in bias and some notable gains in efficiency, too.

### 3.3. Extending the Estimated Survival Function: How to Extrapolate from Breslow's Estimator

To extend Breslow's estimator, we considered three of the extrapolation methods for  $\hat{S}_0(t)$  introduced in Section 2.5.1: (i) immediate drop-off, (ii) exponential extension, and (iii) Weibull extension. (We did not consider carry forward extrapolation, since it caused the integral to diverge.) To compare them, we focused on estimating  $\hat{\beta}$ , the coefficient on  $X$ , which will be most impacted by censoring. Extrapolating  $\hat{S}_0(t)$  with the Weibull extension offered the lowest bias and best efficiency for  $\hat{\beta}$  when imputing with adaptive quadrature (Supplemental Fig. S2).

Though the “winning method” used the Weibull extension to extrapolate,  $X$  was truly generated from a Weibull distribution here. Therefore, to offer more general recommendations, we also considered an  $X$  that was generated from a log-normal distribution with mean = 0 and variance = 0.25 (on the log scale). For light ( $\sim 20\%$ ), moderate ( $\sim 35\%$ ), and heavy ( $\sim 79\%$ ) censoring, we generated  $C$  from an exponential distribution with rates = 0.2, 0.4, and 1.67, respectively. In fact, with log-normal  $X$ , the bias when using conditional mean imputation with adaptive quadrature was very low and relatively unchanged by the extrapolation methods (Supplemental Fig. S3).

We also compared mean versus carry forward interpolation between uncensored values for Breslow's estimator (Remark 2.1) and found that they performed similarly in terms of bias and efficiency (Supplemental Fig. S4). Also, as expected in Remark 2.2,

### 3.4. Constructing a More Realistic Setting: Conditional Mean Imputation with the Estimated Survival Function

Having selected the Weibull extension method for extrapolation, we compared the imputation approaches based on the adaptive quadrature and trapezoidal rule conditional mean calculations. Unlike Section 3.2, here  $S(t|z)$  was treated as unknown and instead had to be estimated; this simulation setting goes beyond the gold standard and was constructed to be more realistic.

After estimating the survival function for Weibull  $X$ , the trapezoidal rule led to an even larger bias in estimating  $\hat{\beta}$  (the coefficient on  $X$ ) with conditional mean imputation than was seen with the true survival function (Table 2). Under heavy and moderate censoring, the trapezoidal rule led to as much as 183% and 23% bias in  $\hat{\beta}$ , respectively. Meanwhile, imputation using adaptive quadrature offered no more 27% and 10% bias under heavy and moderate censoring, respectively. While larger than we would like, this residual bias with adaptive quadrature seemed to stem from the estimated survival function; recall that we saw  $\leq 5\%$  bias across these same settings when assuming the true survival function instead (Table 1). With minor exceptions (e.g., in the largest samples), adaptive quadrature continued to have efficiency gains over the trapezoidal rule even when estimating  $S(t|z)$ . Estimating the survival function for log-normal  $X$  led to similar bias when using adaptive quadrature or the trapezoidal rule (Supplemental Table S1). We were surprised, as we expected the trapezoidal rule to continue to produce bias; upon further investigation, we discovered that this was due to the symmetry of the log-normal distribution (Supplemental Fig. S6).

Having demonstrated the severe bias attributable to calculating conditional means with the trapezoidal rule with the “gold standard” imputation approach based on the true survival function, we constructed a more realistic simulation setting where the survival function had to first be estimated. The improvements to conditional mean imputation persisted with the estimated survival function, as adaptive quadrature offered much lower bias and some efficiency gains.

## 4. Application to Huntington’s Disease Data

### 4.1. *Designing Clinical Trials to Test Experimental Treatments for Huntington’s Disease*

Damage due to Huntington’s disease is irreversible, so slowing symptom progression is often the objective of experimental treatments. Clinical trials are critical to the success of potential treatments but also expensive, leading to constraints in their design and implementation, like the number of subjects recruited and length of follow-up. Thus, clinical trials seek to recruit subjects for whom the treatment could have the greatest potential impact (Paulsen et al., 2019).

Recruiting from an existing Huntington’s disease study can be a powerful first step, since more information is available than when recruiting “from scratch.” For example, we could measure symptom change leading up to potential recruitment. Information about symptom change is important, since the impact of the treatment in slowing symptom progression would be more measurable for subjects with steeply progressing symptoms. Still, an existing study only tells us how a subject’s symptoms have been changing thus far, while what we really want to know is how their symptoms would change during the trial. While this future symptom progression is not measurable, it is estimable. Specifically, we can model symptom change using data from PREDICT-HD. Then, we can estimate subjects’ symptom progression after recruitment to identify high priority subjects for a new clinical trial (i.e., those with the largest expected declines).

Time to diagnosis has been shown to be highly predictive of symptom severity, with the steepest change in symptoms seen in the years immediately before and after diagnosis

(e.g., Long et al. (2014)). Thus, time to diagnosis is an important covariate in our symptom progression model, but in a study like PREDICT-HD, where not everyone has been diagnosed, it is a randomly right-censored covariate that must first be dealt with. In the sections that follow, we discuss the details of modeling the progression of Huntington’s disease symptoms in a prospective study of diagnosed and undiagnosed subjects using data from PREDICT-HD (Section 4.2). Then, we walk through imputing censored times to diagnosis for undiagnosed subjects (Section 4.3). Finally, we discuss our strategy to recruit subjects for a new clinical trial based on these models (Section 4.4).

#### 4.2. Modeling the Progression of Huntington’s Disease Symptoms

One way to gauge symptom severity is the composite Unified Huntington Disease Rating Scale (cUHDRS), which collectively measures functional, motor, and cognitive impairments. As Huntington’s disease progresses toward diagnosis, impairment worsens and the cUHDRS is designed to decrease as it does. Following from Schobel et al. (2017),  $cUHDRS = (TFC - 10.4)/1.9 - (TMS - 29.7)/14.9 + (SDMT - 28.4)/11.3 + (SWR - 66.1)/20.1 + 10$ , where TFC is total functional capacity, TMS is total motor score, SDMT is the Symbol Digit Modality Test, and SWR is the Stroop Word Reading Test. These components measure symptom severity in different areas of life: capacity for “everyday tasks” (TFC), motor impairment (TMS), and cognitive impairment (SDMT and SWR).

We captured Huntington’s disease symptom progression over follow-up by modeling the adjusted association between the cUHDRS at the first and last study visits ( $cUHDRS_0$  and  $cUHDRS_1$ , respectively), controlling for other known covariates. Included in these covariates were (i) proximity to diagnosis, defined as  $TIME_1$  from the last visit to diagnosis, and (ii) baseline information about age, CAG repeat length, and their interaction (denoted by  $AGE_0$ ,  $CAG_0$ , and  $AGE_0 \times CAG_0$ , respectively). In addition, we included an interaction between  $cUHDRS_0$  and  $TIME_1$  because if a subject is farther from diagnosis, then their cUHDRS is not expected to be changing much, while if the subject is closer to diagnosis, it is expected to be changing noticeably. Thus, the symptom progression model of interest was captured with linear regression as  $E_\theta(cUHDRS_1 | TIME_1, cUHDRS_0, AGE_0, CAG_0) =$

$$\alpha + \beta TIME_1 + \gamma_0 cUHDRS_0 + \gamma_1 TIME_1 \times cUHDRS_0 + \gamma_2 AGE_0 + \gamma_3 CAG_0 + \gamma_4 AGE_0 \times CAG_0. \quad (4)$$

Note that the subscripts 0 and 1 delineate variables measured at or relative to the first and last visits, respectively. Covariates were rescaled, with  $AGE_0$ ,  $CAG_0$ , and  $cUHDRS_0$  centered at 18, 36, and 23.8, respectively. The remaining covariate,  $TIME_1$ , was right-censored (see Section 4.3).

To be included in our analysis, subjects needed to have (i) a CAG repeat length  $\geq 36$  on the HTT gene, (ii) not yet been diagnosed with Huntington’s disease at study entry, (iii) undergone all necessary testing to calculate the cUHDRS at the first and last visits (Supplemental Fig. S7), and (iv) returned for at least one follow-up visit. These criteria left a sample of  $n = 970$  at-risk subjects, 238 (25%) of whom were diagnosed before their last visit, leaving 75% with a censored covariate  $TIME_1$ . Since we employed single conditional mean imputation to replace censored times to diagnosis, we estimated the robust sandwich variance with the **sandwich** package (Zeileis, 2004).



### 4.3. *Imputing Censored Times to Diagnosis*

Calculating time to diagnosis was done in the following way. First, **DATE** of diagnosis was taken as the first visit where a subject met the criteria for diagnosis, i.e., a clinician assigned them to the highest rating of a 4 on the Unified Huntington’s Disease Rating Scale diagnostic confidence level (Long et al., 2014). From **DATE**, we calculated time to diagnosis from either the first or last visit. We did the former for imputation, because it was most natural to think of the symptom progression from study entry, and the latter for analysis, because time from last visit aligned better with our outcome (cUHDRS at that same time).

Since subjects who had not yet been diagnosed had no such **DATE** but would have one someday in the future,  $\text{TIME}_0$  from the first visit to diagnosis was randomly right-censored but could be imputed with conditional mean  $E(\text{TIME}_0 | \text{TIME}_0 > \text{FOLLOW\_UP}_1, \text{AGE}_0, \text{CAG}_0)$ , where  $\text{FOLLOW\_UP}_1$  was the follow-up time to the last visit. Imputation began by modeling the conditional survival function for  $\text{TIME}_0$  given other fully observed covariates ( $\text{AGE}_0, \text{CAG}_0$ ) from study entry. First, we fit the Cox proportional hazards model and calculated Breslow’s estimator (details in Web Appendix C.1). Following from our empirical findings in Section 3.3, we used the Weibull extension to extrapolate the survival estimator beyond the largest uncensored value, where  $\hat{S}(t = 11.422 | \text{AGE}_0 = 34.13, \text{CAG}_0 = 4) = 0.532$ . Also, the context of  $\text{TIME}_0$  could be used to refine the upper bound of the integral in Equation (1). Specifically,  $\text{TIME}_0$  from study entry to Huntington’s disease diagnosis could not be infinite simply because humans are not immortal. Instead, we assumed  $\text{TIME}_0$  to be within 60 years of study entry (details in Web Appendix A.3).

Now, we prepared to fit the models. Because symptoms were expected to worsen near diagnosis, time to diagnosis (in years) was a key covariate. Since cUHDRS at the last visit was our outcome, we defined time to diagnosis from the last visit, too. For uncensored subjects,  $\text{TIME}_1$  was computed by subtracting their last visit date from their **DATE** of diagnosis. For censored subjects,  $\text{TIME}_1$  was computed by subtracting their last visit date from the imputed **DATE** of diagnosis instead, where **DATE** was found by adding their conditional mean to their first visit date.

### 4.4. *Strategic Recruitment for a Clinical Trial*

Like the densities of time to diagnosis (Supplemental Fig. S8–S9), the two imputation approaches led to different models, each with its own clinical implications (Table 3). We focused on adopting the models in the following way to guide recruitment for a new clinical trial. Suppose we were recruiting 200 at-risk subjects from their last regular study visit and that the clinical trial was expected to last for 2 years. Our recruitment strategy proceeds in two steps: (i) computing the subject-specific expected change in cUHDRS between recruitment and trial end 2 years later and (ii) prioritizing subjects with the steepest expected drops in cUHDRS during that time. For demonstration, we begin by estimating one subject’s symptom progression during the trial and discussing their resulting priority (Section 4.4.1) and then outline our large-scale recruitment strategy for an entire clinical trial (Section 4.4.2).

#### 4.4.1. How to Estimate Symptom Progression and Prioritize a Subject for Recruitment

Consider a randomly selected subject whose cUHDRS was already seen to decline from  $\text{cUHDRS}_0 = 15.9$  to  $\text{cUHDRS}_1 = 13.3$  between their first to last visits in PREDICT-HD, a pre-trial change of  $\Delta_1(\text{cUHDRS}) = -2.6$ . In planning a clinical trial, the subject's symptom change during the trial was more of interest but unobservable at recruitment. Fortunately, estimating this change in cUHDRS during the trial can be a powerful alternative. Specifically, we can predict cUHDRS 2 years from recruitment, denoted by  $\widehat{\text{cUHDRS}}_2$ , using the symptom progression models and then calculate expected symptom change during the clinical trial from it as  $\widehat{\Delta}_2(\text{cUHDRS}) = \widehat{\text{cUHDRS}}_2 - \text{cUHDRS}_1$ . Thus,  $\widehat{\Delta}_2(\text{cUHDRS}) < 0$  would indicate that the subject's symptoms are expected to worsen.

For each subject, we can plug their covariates along with the estimated model parameters  $\hat{\theta}$  into Equation (4) to estimate  $\widehat{\text{cUHDRS}}_2$ . However, we wanted to predict end-of-trial cUHDRS from recruitment cUHDRS, whereas the models were fit to predict last visit cUHDRS from first visit cUHDRS. Thus, baseline covariates  $\text{AGE}_0$  and  $\text{CAG}_0$  were unchanged, but we replaced (i) time from last visit to diagnosis ( $\text{TIME}_1$ ) with time from end of trial to diagnosis ( $\text{TIME}_2$ ) and (ii) cUHDRS at first visit ( $\text{cUHDRS}_0$ ) with cUHDRS at recruitment ( $\text{cUHDRS}_1$ ). With these substitutions,  $\widehat{\text{cUHDRS}}_2$  can then be estimated from either model as  $E_{\hat{\theta}}(\text{cUHDRS}_2 | \text{TIME}_2, \text{cUHDRS}_1, \text{AGE}_0, \text{CAG}_0)$ . As a bonus,  $\widehat{\text{cUHDRS}}_2$  can be used to construct a complete trajectory of the subject's symptom severity, where  $\widehat{\Delta}_2(\text{cUHDRS})$  summarizes changes in the latter part of this trajectory (Fig. 1).

For the example subject, the model imputed using adaptive quadrature predicted their cUHDRS to be 10.8 at the end of the trial, leading to an estimated change of  $\widehat{\Delta}_2(\text{cUHDRS}) = -2.5$  during the trial. Based on this, the subject had the 43rd largest estimated decrease in cUHDRS among censored subjects, making them high priority for recruitment. In contrast, the model imputed using the trapezoidal rule predicted their cUHDRS to be 11.8 at trial end for a smaller change of  $\widehat{\Delta}_2(\text{cUHDRS}) = -1.5$  (ranking 201st and giving this subject low priority for recruitment).

Because we saw in the simulation studies (Section 3) that the trapezoidal rule estimates can be biased, particularly under heavy censoring rates like the 75% in PREDICT-HD, we have more trust in the model imputed using adaptive quadrature and believe that its expected symptom change of  $\widehat{\Delta}_2(\text{cUHDRS}) = -2.5$  would be closer to the true one. In general, misprioritizing trial candidates (e.g., by mistakenly ranking someone 201st due to a biased model when they should really have been 43rd) means that non-ideal subjects may take spots away from others with potentially more to gain.

#### 4.4.2. How to Prioritize the Entire Study for Recruitment

We used the same process outlined above for everyone and then ordered the entire study by their estimated change in symptoms,  $\widehat{\Delta}_2(\text{cUHDRS})$ , starting from the biggest decline in function (i.e., largest decrease in cUHDRS). Then, we recruited subjects ranked 1–200, prioritizing subjects expected to have the worst symptom progression and with potentially the most to gain. We call this rank-based recruitment.

Although the PREDICT-HD study is over, we demonstrated our recruitment strategy with its data. Fig. 2 summarizes the recruitment statuses based on both models for the 732 censored subjects from the study. To introduce some realistic variability, we also

created 1000 new datasets of 732 subjects each by resampling with replacement from the 732 censored subjects in PREDICT-HD. In each resampled dataset, we applied our rank-based recruitment strategy twice: once with each model. On average, the models agreed on 158 and 490 subjects to recruit and not recruit, respectively. For the other 42 subjects, the models disagreed, with the trapezoidal rule “throwing away” 42 trial spots on subjects that the adaptive quadrature model expected to have lesser changes in symptoms. For a summary across all resampled datasets, see Supplemental Fig. S10.

In an all-knowing world, we would recruit subjects for a new clinical trial who would have the steepest change in their symptoms without treatment to clearly measure the treatment effect (i.e., for a more obvious reduction in symptoms). However, we are not psychics; we cannot know which subjects will have the steepest change in symptoms, so this is not a reasonable strategy. Recruiting subjects expected to have the steepest changes in symptoms is, though. With conditional mean imputation, we modeled the progression of Huntington’s disease symptoms, despite censoring in time to diagnosis, and used these models to guide recruitment for a hypothetical trial. The models disagreed on more than a fifth of who to recruit, but given its demonstrated accuracy in the simulations, we believe that using adaptive quadrature will give statisticians confidence in their model and clinicians confidence in who they recruit based on it.

## 5. Discussion

After demonstrating that the trapezoidal rule makes existing approaches miscalculate conditional means, leading to biased statistical inference, we propose an improved calculation using adaptive quadrature with an infinite upper bound instead. We adopt the `integrate` function, which implements adaptive quadrature in R and is available as part of the “base R” packages (R Core Team, 2019). However, even though the `integrate` function can handle infinite upper bounds, we encountered an additional challenge since the integrand in the conditional means,  $S(t|\mathbf{z})$ , is only defined on the uncensored values. We provide an in-depth empirical investigation of how best to extend Breslow’s estimator for indefinite integration, offering recommendations in various real-world settings. We then demonstrate how well our method corrects for the bias attributable to the trapezoidal rule, offering valid statistical inference from censored covariates through imputation. Finally, we applied our proposed methods to model the progression of Huntington’s disease symptoms in the PREDICT-HD study relative to time of diagnosis, a censored covariate, and discussed using this model to guide recruitment for a new clinical trial.

In our simulations and real-data analysis, we focused on linear regression modeling. However, the methods apply for any outcome model that captures the associations between  $Y$ , censored  $X$ , and  $\mathbf{Z}$ . This flexibility is one of the strengths of imputation: once the censored covariates are “filled in” with their conditional means, we can apply any of the usual modeling approaches.

Our proposed recruitment strategy takes a granular approach to targeting high priority subjects. Other strategies randomly sample from strata defined by a proxy for time to diagnosis. For example, Paulsen et al. (2019) create “low” and “high” risk groups from the CAP score (Zhang et al., 2011), where the high risk group is made up of subjects

with CAP > 390.4 who are believed to be nearest to diagnosis. One potential drawback of stratified strategies like this is that creating categories loses information from the continuous CAP variable. In other words, once subjects are placed into categories, there is no way for clinicians to gauge the relative priority of subjects within a risk group; for example, a subject with a CAP of 666.4 (the largest in the study) has the same chance of being recruited as one with a CAP of 390.5 (barely qualifying as high risk). In ranking subjects from smallest to largest expected symptom change rather than categorizing, our strategy empowers clinicians to directly recruit the highest priority subjects.

Even with our improvements, there are limitations to conditional mean imputation. Namely, semiparametric imputation approaches like this one are sensitive to non-proportional hazards because they rely on the Cox model to estimate the survival function. However, we could test for this and modify the imputation model (e.g., with time-varying coefficients) to accommodate non-proportionality. Still, an entirely unspecified estimator, like the Kaplan–Meier, would be ideal. Also, standard error estimation is problematic with single imputation approaches, like the one we discuss here. However, the improvements we have proposed are needed for and could readily be adopted in a multiple imputation framework instead.

There are several interesting statistical directions for future work. The first would be to extend our framework to capture multiple censored covariates. Atem et al. (2019a) propose such an approach but use the trapezoidal rule to calculate the conditional means, so their formulas would need to be adapted. Also, to our knowledge, imputation for randomly left-censored covariates has been thus far unaddressed and should be a relatively straightforward adaptation; the formula for the appropriate conditional means,  $E(X|X < W_i, \mathbf{Z}_i)$ , would need to be derived, and then adaptive quadrature could be used to calculate them. There are also natural connections to other methods that require indefinite integration over a nonparametric or semiparametric survival estimator, for example, estimating mean residual life or maximum likelihood estimation with a censored covariate. Finally, an interesting clinical direction for future work might involve adopting our rank-based recruitment strategy for other measures of symptom progression (e.g., by ranking subjects on a proxy like CAP score).

## 6. Supplementary Material

The Web Appendices and Supplemental Material are available online through the journal. An R package **imputeCensRd** that implements the proposed methods is available at <https://github.com/sarahlotspeich/imputeCensRd>. All simulation code is available through figshare at [https://figshare.com/projects/It\\_s\\_integral/147225](https://figshare.com/projects/It_s_integral/147225).

## Acknowledgments

This research was supported by the National Institute of Environmental Health Sciences grants T32ES007018 and P30ES010126 and the National Institute of Neurological Disorders and Stroke grant K01NS099343. The authors thank PREDICT-HD for permission to present their data.

*Conflict of Interest:* None declared.



**References**

- Atem, F. D., Matsouaka, R. A. and Zimmern, V. E. (2019a) Cox regression model with randomly censored covariates. *Biometrical Journal*, **61**, 1020–1032.
- Atem, F. D., Qian, J., Maye, J. E., Johnson, K. A. and Betensky, R. A. (2017) Linear regression with a randomly censored covariate: Application to an Alzheimer’s study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 313–328.
- Atem, F. D., Sampene, E. and Greene, T. J. (2019b) Improved conditional imputation for linear regression with a randomly censored predictor. *Statistical Methods in Medical Research*, **28**, 432–444.
- Breslow, N. E. (1972) Discussion of Professor Cox’s paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 216–217.
- Datta, S. (2005) Estimating the mean life time using right censored data. *Statistical Methodology*, **2**, 65–69.
- Huntington Study Group (1996) Unified Huntington’s disease rating scale: Reliability and consistency. *Movement Disorders*, **11**, 136–142. PMID: 8684382.
- Klein, J. and Moeschberger, M. (2003) *Survival analysis: Techniques for censored and truncated data. 2nd Edition*. New York: Springer.
- Little, R. J. A. (1992) Regression with missing X’s: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.
- Long, J. D., Paulsen, J. S., Marder, K., Zhang, Y., Kim, J., Mills, J. A. and Researchers of the PREDICT-HD Huntington’s Study Group (2014) Tracking motor impairments in the progression of huntington’s disease. *Movement Disorders*, **29**, 311–319.
- Lotspeich, S. C., Grosser, K. F. and Garcia, T. P. (2022) Correcting conditional mean imputation for censored covariates and improving usability. *Biometrical Journal*, **64**, 858–862.
- Moeschberger, M. and Klein, J. (1985) A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics*, **41**, 253–259.
- Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., Duff, K., Kayson, E., Biglan, K., Shoulson, I., Oakes, D., Hayden, M. and Predict-HD Investigators and Coordinators of the Huntington Study Group (2008) Detection of Huntington’s disease decades before diagnosis: the Predict-HD study. *Journal of Neurology, Neurosurgery & Psychiatry*, **79**, 874–880.
- Paulsen, J. S., Lourens, S., Kieburtz, K. and Zhang, Y. (2019) Sample enrichment for clinical trials to show delay of onset in Huntington disease. *Movement Disorders*, **34**, 274–280.
- Piessens, R., deDoncker Kapenga, E., Uberhuber, C. and Kahaner, D. (1983) *Quadpack: a Subroutine Package for Automatic Integration*. Springer Verlag.

- R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Richardson, D. B. and Ciampi, A. (2003) Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, **157**, 355–363.
- Schobel, S., Palermo, G., Auinger, P., Long, J., Ma, S., Khwaja, O., Trundell, D., Cudkowicz, M., Hersch, S., Sampaio, C., Dorsey, E., Leavitt, B., Kiebertz, K., Sevigny, J., Langbehn, D., Tabrizi, S. and TRACK-HD, COHORT, CARE-HD, and 2CARE Huntington Study Group Investigators (2017) Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. *Neurology*, **89**, 2495–2502.
- The Huntington’s Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. *Cell*, **72**, 971–983. PMID8458085.
- Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data: Extending the Cox Model*. New York: Springer.
- Zeileis, A. (2004) Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, **11**, 1–17.
- Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., Paulsen, J. S., the PREDICT-HD Investigators and of the Huntington Study Group, C. (2011) Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, **156B**, 751–763.

**Table 1.** Simulation results for Weibull  $X$  from the full cohort analysis and imputation approaches using the true survival function and adaptive quadrature versus the trapezoidal rule. (SE: empirical standard error; RE: empirical relative efficiency to full cohort.)

Censoring	$n$	Full Cohort		Adaptive Quadrature			Trapezoidal Rule		
		Bias	SE	Bias	SE	RE	Bias	SE	RE
$\hat{\alpha}$ : Intercept									
Light	100	−0.001	0.161	0.000	0.167	0.924	−0.003	0.168	0.915
	500	0.001	0.071	−0.002	0.074	0.928	−0.003	0.074	0.921
	2000	−0.001	0.036	0.003	0.037	0.941	0.002	0.037	0.938
Moderate	100	−0.001	0.161	−0.003	0.181	0.791	−0.015	0.187	0.736
	500	0.001	0.071	−0.002	0.080	0.782	−0.009	0.082	0.754
	2000	−0.001	0.036	0.003	0.040	0.796	−0.002	0.040	0.776
Heavy	100	−0.001	0.161	−0.010	0.246	0.428	−0.033	0.272	0.350
	500	0.001	0.071	−0.002	0.107	0.440	−0.017	0.114	0.388
	2000	−0.001	0.036	−0.001	0.055	0.428	−0.012	0.058	0.385
$\hat{\beta}$ : Coefficient on Censored $X$									
Light	100	−0.001	0.275	−0.020	0.291	0.896	0.001	0.306	0.808
	500	0.001	0.116	0.007	0.123	0.887	0.017	0.126	0.847
	2000	0.002	0.057	−0.004	0.063	0.818	0.000	0.063	0.805
Moderate	100	−0.001	0.275	−0.010	0.370	0.554	0.118	0.484	0.324
	500	0.001	0.116	0.007	0.162	0.508	0.073	0.185	0.392
	2000	0.002	0.057	−0.004	0.081	0.490	0.034	0.089	0.406
Heavy	100	−0.001	0.275	0.025	0.675	0.166	0.906	1.870	0.022
	500	0.001	0.116	0.008	0.283	0.168	0.579	0.623	0.035
	2000	0.002	0.057	0.003	0.151	0.142	0.425	0.289	0.039
$\hat{\gamma}$ : Coefficient on Uncensored $Z$									
Light	100	0.005	0.208	0.006	0.201	1.066	0.006	0.201	1.068
	500	−0.002	0.090	−0.003	0.091	0.979	−0.003	0.091	0.977
	2000	0.001	0.045	−0.001	0.043	1.091	−0.001	0.043	1.091
Moderate	100	0.005	0.208	0.006	0.202	1.060	0.006	0.202	1.058
	500	−0.002	0.090	−0.003	0.091	0.965	−0.003	0.091	0.964
	2000	0.001	0.045	−0.001	0.043	1.075	−0.001	0.043	1.076
Heavy	100	0.005	0.208	0.006	0.200	1.082	0.006	0.201	1.071
	500	−0.002	0.090	0.002	0.094	0.917	0.002	0.094	0.917
	2000	0.001	0.045	0.001	0.046	0.957	0.001	0.046	0.957

Note: All entries are based on 1000 replicates.

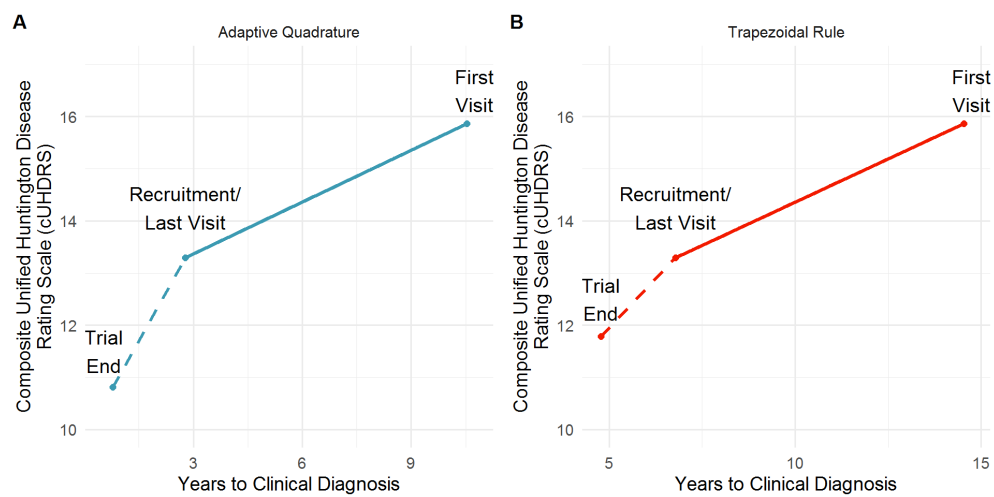
**Table 2.** Simulation results for Weibull  $X$  from the full cohort analysis and imputation approaches using the estimated survival function and adaptive quadrature versus the trapezoidal rule. (SE: empirical standard error; RE: empirical relative efficiency to full cohort.)

Censoring	$n$	Full Cohort		Adaptive Quadrature			Trapezoidal Rule		
		Bias	SE	Bias	SE	RE	Bias	SE	RE
$\hat{\alpha}$ : Intercept									
Light	100	−0.001	0.161	0.003	0.172	0.872	−0.005	0.171	0.882
	500	0.001	0.071	0.003	0.072	0.983	−0.006	0.074	0.911
	2000	−0.001	0.036	0.003	0.037	0.949	−0.002	0.036	0.965
Moderate	100	−0.001	0.161	−0.006	0.187	0.736	−0.013	0.187	0.739
	500	0.001	0.071	0.000	0.077	0.853	−0.012	0.084	0.714
	2000	−0.001	0.036	0.003	0.040	0.783	−0.005	0.041	0.771
Heavy	100	−0.001	0.161	0.009	0.266	0.365	−0.027	0.271	0.350
	500	0.001	0.071	0.002	0.111	0.405	−0.013	0.114	0.385
	2000	−0.001	0.036	0.002	0.055	0.427	−0.008	0.057	0.396
$\hat{\beta}$ : Coefficient on Censored $X$									
Light	100	−0.001	0.275	−0.011	0.312	0.778	0.018	0.317	0.756
	500	0.001	0.116	−0.012	0.125	0.857	0.012	0.129	0.805
	2000	0.002	0.057	−0.008	0.064	0.782	0.005	0.063	0.806
Moderate	100	−0.001	0.275	0.051	0.447	0.380	0.115	0.476	0.335
	500	0.001	0.116	0.017	0.179	0.416	0.073	0.199	0.338
	2000	0.002	0.057	−0.003	0.096	0.348	0.037	0.096	0.352
Heavy	100	−0.001	0.275	−0.136	0.744	0.047	0.913	1.900	0.007
	500	0.001	0.116	0.004	0.403	0.031	0.572	0.647	0.012
	2000	0.002	0.057	0.082	0.229	0.024	0.423	0.298	0.014
$\hat{\gamma}$ : Coefficient on Uncensored $Z$									
Light	100	0.005	0.208	−0.003	0.206	1.017	0.001	0.201	1.063
	500	−0.002	0.090	−0.002	0.090	0.989	0.004	0.089	1.017
	2000	0.001	0.045	−0.002	0.045	1.014	0.003	0.045	0.995
Moderate	100	0.005	0.208	−0.003	0.210	0.977	0.002	0.202	1.051
	500	−0.002	0.090	−0.001	0.091	0.978	0.004	0.090	1.002
	2000	0.001	0.045	−0.002	0.045	0.996	0.003	0.045	0.986
Heavy	100	0.005	0.208	−0.001	0.285	0.318	−0.001	0.205	0.611
	500	−0.002	0.090	0.003	0.100	0.500	−0.001	0.092	0.601
	2000	0.001	0.045	0.000	0.049	0.537	0.001	0.045	0.623

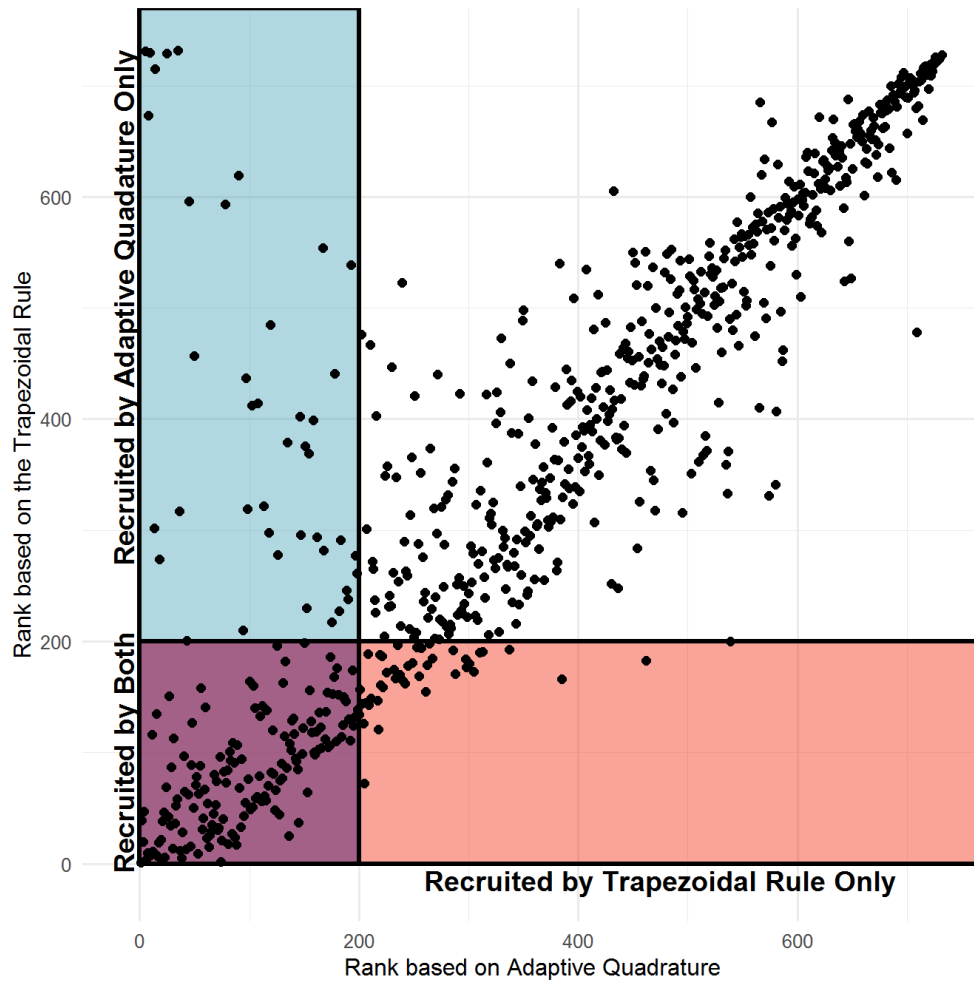
*Note:* The MLE for the Weibull extension converged in  $\geq 99.5\%$  of replicates of imputation in each setting (just 18 and 16 of 9000 total replicates did not converge with adaptive quadrature and the trapezoidal rule, respectively); all other entries are based on 1000 replicates.

**Table 3.** Huntington's disease symptom progression models in PREDICT-HD fit using normal linear regression after imputing censored  $\text{TIME}_1$  from last visit to diagnosis with conditional means. (95% CI denotes the 95% Wald-type confidence interval based on the sandwich standard errors.)

Coefficient	Adaptive Quadrature		Trapezoidal Rule	
	Estimate	95% CI	Estimate	95% CI
Intercept	21.680	(20.571, 22.790)	23.298	(22.349, 24.246)
$\text{TIME}_1$	0.084	(−0.013, 0.181)	0.117	(−0.032, 0.266)
$\text{cUHDRS}_0$	1.048	(0.941, 1.155)	0.961	(0.861, 1.061)
$\text{TIME}_1 \times \text{cUHDRS}_0$	−0.024	(−0.036, −0.011)	−0.019	(−0.036, −0.002)
$\text{AGE}_0$	−0.021	(−0.046, 0.003)	0.012	(−0.009, 0.032)
$\text{CAG}_0$	−0.089	(−0.166, −0.012)	−0.092	(−0.160, −0.025)
$\text{AGE}_0 \times \text{CAG}_0$	0.006	(0.001, 0.011)	−0.014	(−0.018, −0.010)



**Fig. 1.** For each subject, we can estimate their cUHDRS at the end of the trial using the symptom progression models and then construct a complete trajectory of their symptom severity over study follow-up (i.e., the solid line from First Visit to Recruitment/Last Visit) and the 2-year clinical trial (i.e., the dashed line from Recruitment/Last Visit to Trial End).



**Fig. 2.** Subjects were ranked by their estimated changes in symptoms based on the models, starting from the biggest decline in function (i.e., largest decrease in cUHRS), and the first 200 subjects subsequently recruited into the hypothetical clinical trial. The shaded regions capture subjects who would have been recruited based on each model, with the overlapping area in the lower left capturing subjects who would have been recruited based on either model. Points represent the  $n = 732$  censored subjects from PREDICT-HD.



## Supplementary Materials for “It’s integral: Replacing the trapezoidal rule to remove bias and correctly impute censored covariates with their conditional means”

Sarah C. Lotspeich

*Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC, U.S.A.  
Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*

E-mail: lotspes@wfu.edu

Tanya P. Garcia

*Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*

### Web Appendix A. More About the Extrapolation Methods for Breslow’s Estimator

#### Web Appendix A.1. Derivation of the Exponential Extension

Assuming that among the baseline group (i.e., with  $\mathbf{Z} = \mathbf{0}$ ),  $X$  follows an exponential distribution with rate  $\rho$ , we have  $S_0(t) = \exp\{- (t/\rho)\}$ . To connect to Breslow’s estimator,  $\hat{\rho}$  is constrained so that  $\exp\left\{-\left(\tilde{X}/\hat{\rho}\right)\right\} = \hat{S}_0(\tilde{X})$ . We can solve this constraint for  $\hat{\rho} = -\tilde{X} \log\left\{\hat{S}_0(\tilde{X})\right\}^{-1}$  and extrapolate using  $\hat{S}_0(t) = \exp\left(\left[t \log\left\{\hat{S}_0(\tilde{X})\right\}\right] / \tilde{X}\right)$  for  $t > \tilde{X}$ . This is the exponential extension introduced in Section 2.5.1, and it was originally proposed by ?).

#### Web Appendix A.2. Derivation of the Weibull Extension

Assuming that among the baseline group (i.e., with  $\mathbf{Z} = \mathbf{0}$ ),  $X$  follows a Weibull distribution with shape and scale parameters  $\nu$  and  $\rho$ , respectively, we have  $S_0(t) = \exp(-\rho t^\nu)$ . The parameters are once again constrained to ensure a clean transition from Breslow’s estimator to the extension, with  $\exp\left(-\hat{\rho}\tilde{X}^{\hat{\nu}}\right) = \hat{S}_0(\tilde{X})$ . Unlike the exponential extension, there is not a closed form solution as in Web Appendix A.1. Herein, we adopt a constrained maximum likelihood approach to find  $\hat{\nu}$  and  $\hat{\rho}$ .

The shape and scale parameters,  $\nu$  and  $\rho$ , respectively, can be estimated directly through maximum likelihood estimation. Using the probability density function and survival function of the Weibull distribution, the usual (i.e., unconstrained) log-likelihood for the shape and scale parameters can be defined as

$$l_n(\nu, \rho) = \sum_{i=1}^n \Delta_i \log\left\{\rho \nu W_i^{\nu-1} \exp(-\rho W_i^\nu)\right\} + \sum_{i=1}^n (1 - \Delta_i) \log\left\{\exp(-\rho W_i^\nu)\right\}$$

$$= -\rho \sum_{i=1}^n W_i^\nu + (\nu - 1) \sum_{i=1}^n \Delta_i \log(W_i) + n_1 \log(\rho) + n_1 \log(\nu), \quad (\text{S.1})$$

where  $n_1$  is the number of uncensored observations (i.e.,  $n_1 = \sum_{i=1}^n \Delta_i$ ).

Recall that we want this Weibull curve to “tie into” Breslow’s estimator  $\hat{S}_0(t)$  at the largest uncensored value,  $\tilde{X}$ . This constraint on the Weibull survival function can be expressed as  $\exp(-\rho \tilde{X}^\nu) \equiv \hat{S}_0(\tilde{X})$ , and it further translates into the following relationship between the shape and scale parameters:

$$\rho = -\log \left\{ \hat{S}_0(\tilde{X}) \right\} / (\tilde{X}^\nu). \quad (\text{S.2})$$

With Equation (S.2), we can modify Equation (S.1) to obtain the constrained log-likelihood in terms of just the shape parameter,

$$\begin{aligned} l_n(\nu) = & \left[ \log \left\{ \hat{S}_0(\tilde{X}) \right\} / (\tilde{X}^\nu) \right] \sum_{i=1}^n W_i^\nu + (\nu - 1) \sum_{i=1}^n \Delta_i \log(W_i) \\ & + n_1 \log \left[ \log \left\{ \hat{S}_0(\tilde{X}) \right\} / (\tilde{X}^\nu) \right] + n_1 \log(\nu). \end{aligned} \quad (\text{S.3})$$

Estimation of the maximum likelihood estimator (MLE)  $\hat{\nu}$  is done by finding the root of Equation (S.3) with a univariate Newton-type algorithm, as implemented in the `nlm` function in R (?). Our initial guess for the shape parameter (which must be  $> 0$ ) is  $\hat{\nu}^{(0)} = 1\text{E}^{-4}$ , and the algorithm is restricted to positive values for  $\hat{\nu}$ . Finally,  $\hat{\rho}$  is obtained by plugging  $\hat{\nu}$  at convergence into Equation (S.2), and with it we arrive at the parameters for the Weibull extension used to extrapolate from Breslow’s step function estimator of baseline survival,  $\hat{S}_0(t)$  for  $t > \tilde{X}$  introduced in Section 2.5.1.

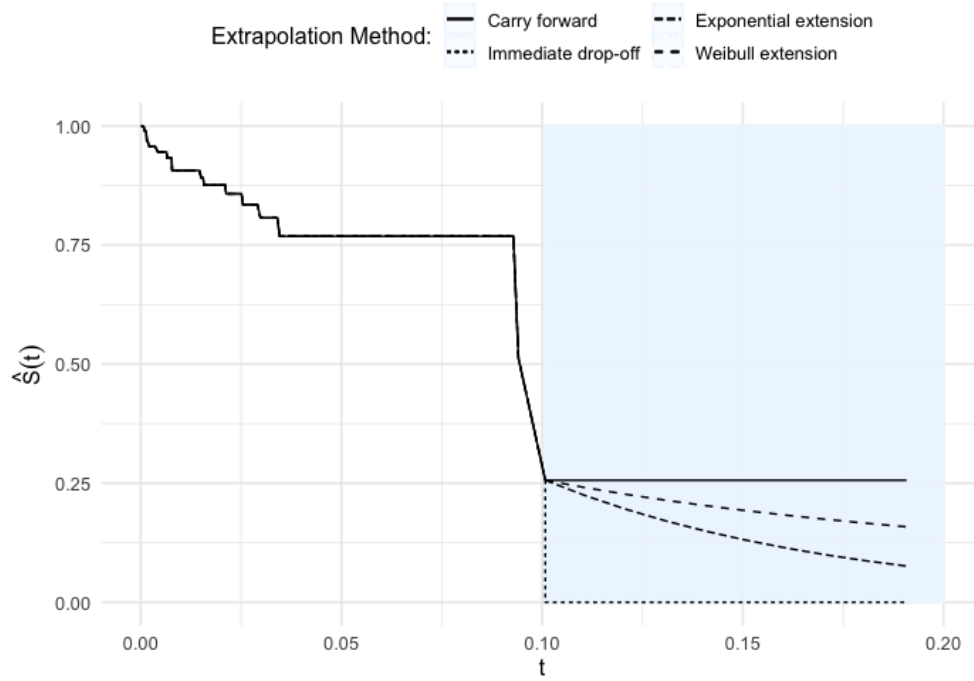
### *Web Appendix A.3. Finite Upper Limit for $X$*

In the formulas used throughout this manuscript, we integrate from  $W_i$  to  $\infty$  in calculating the conditional means. This is the most general case, and it was appropriate in our simulation studies (Section 3) because  $X$  was generated from Weibull or log-normal distributions with domains from 0 to  $\infty$ . However, in some settings we have prior information about  $X$  that allows us to replace the infinite upper bound in the integral with some known constant, denoted by  $\omega$ .

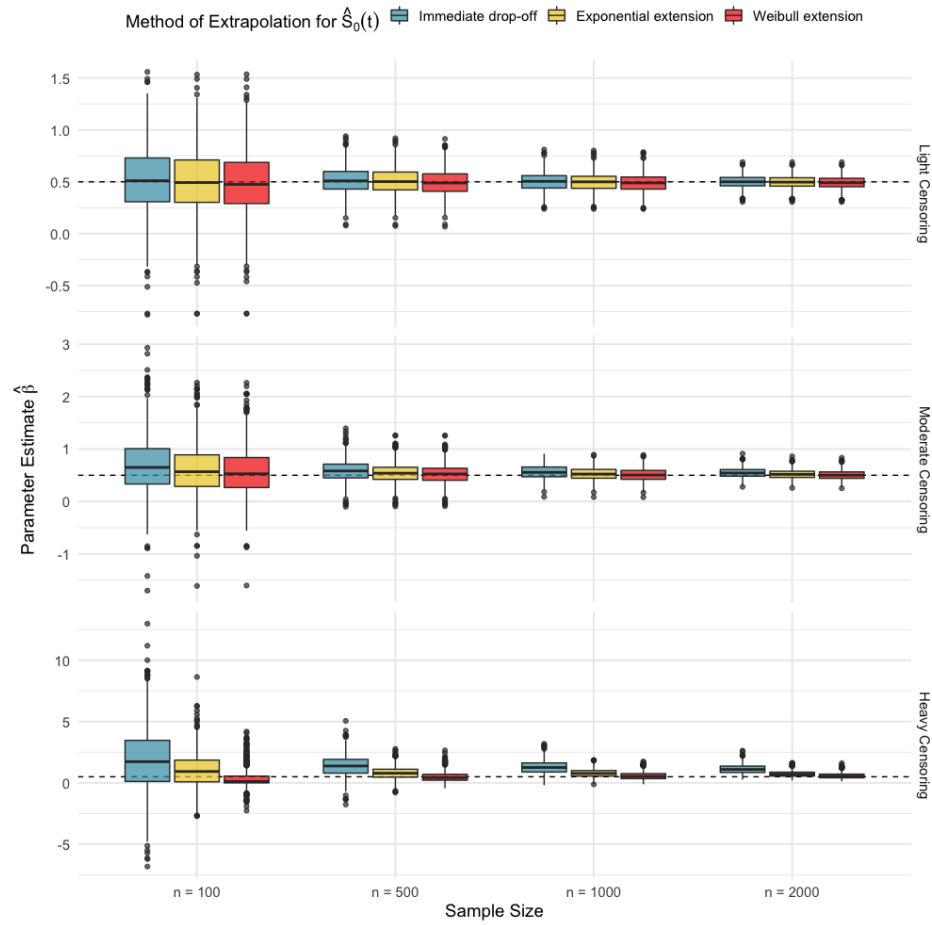
For example, in our PREDICT-HD example (Section 4), the censored covariate  $X$  was  $\text{TIME}_0$  from study entry to clinical Huntington’s disease diagnosis. Since this is an adult cohort, with all subjects having  $\text{AGE}_0 \geq 18$  years old at study entry, we set the longest time from study entry to clinical diagnosis to be  $\omega = 60$  years. We believe this is a conservative upper bound on  $\text{TIME}_0$  that is in agreement with the recent overall life expectancy estimate of 78 years in the United States (?). There is no established life expectancy estimate for people who are at-risk for Huntington’s disease; we call 78 years a “conservative” upper bound, since it is probably higher than the life expectancy in our population. Choosing this finite limit is an important consideration. While we want to extract as much information from the data as we can, we also want to avoid imputing

too far beyond the observed values of  $\text{TIME}_0$  or beyond reasonable values based on the context, leading to a trade-off between setting  $\omega$  too low or high.

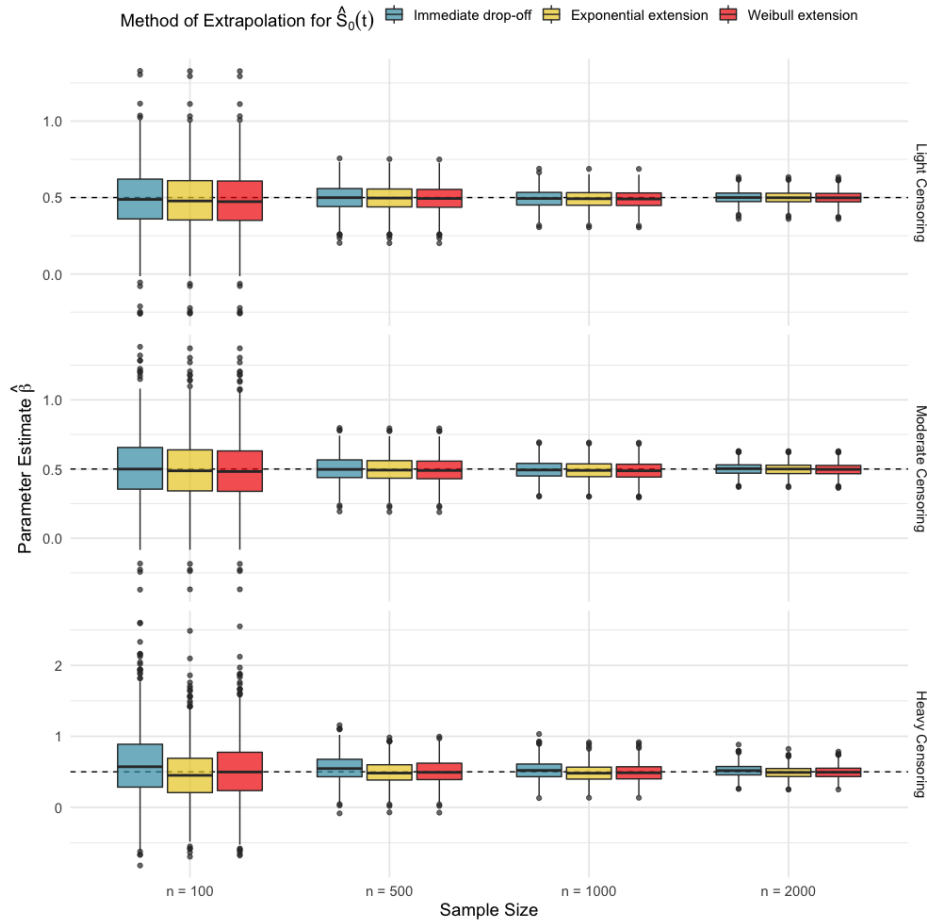
Now, our choice of finite  $\omega$  imposes an additional constraint on the Weibull extension: since  $S(\omega) \approx 0$  we further constrain  $\nu$  and  $\rho$  such that  $\exp(-\rho\omega^\nu) \approx 0$ . Thus, we can find the corresponding values of  $\hat{\rho}$  and  $\hat{\nu}$  using the `uniroot` function in R (?), since  $\rho$  can be treated as a function of  $\nu$  as in Equation (S.2).



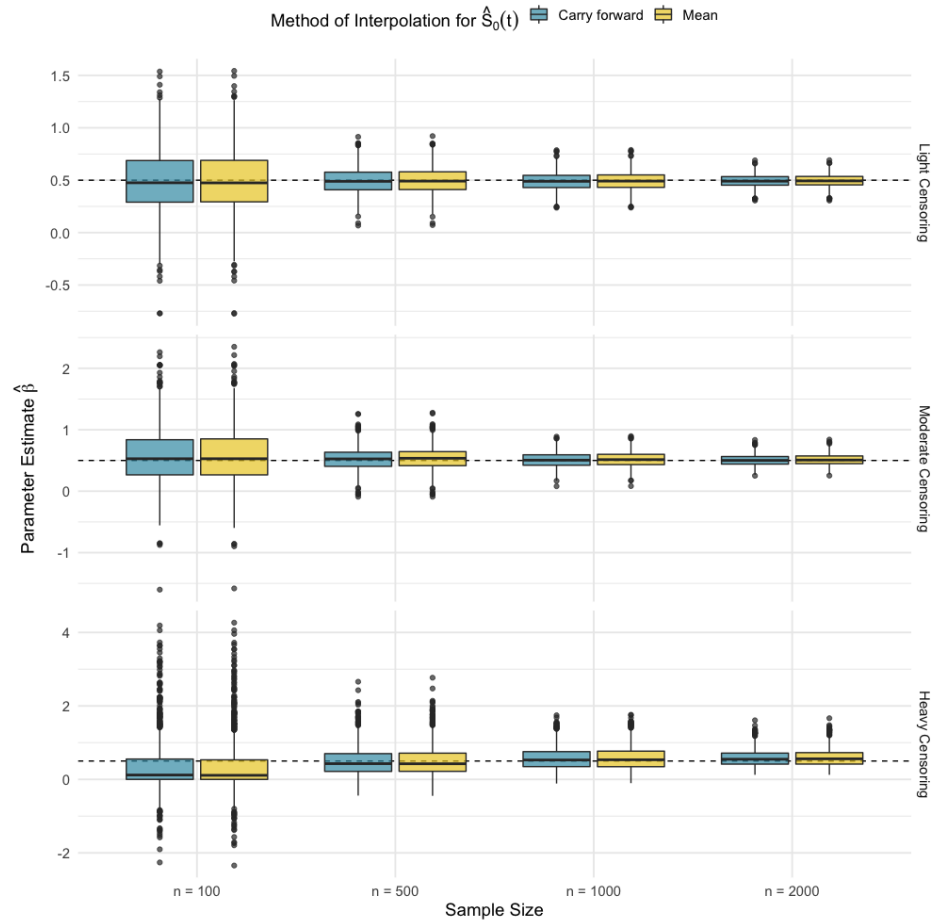
**Fig. S1.** Illustration of the four extrapolation methods for a step survival function  $\hat{S}(t)$  in simulated data. The shaded area represents values of  $t > \tilde{X}$  (the largest uncensored value), where extrapolation is needed.

**Web Appendix B. Additional Results from the Simulation Studies**

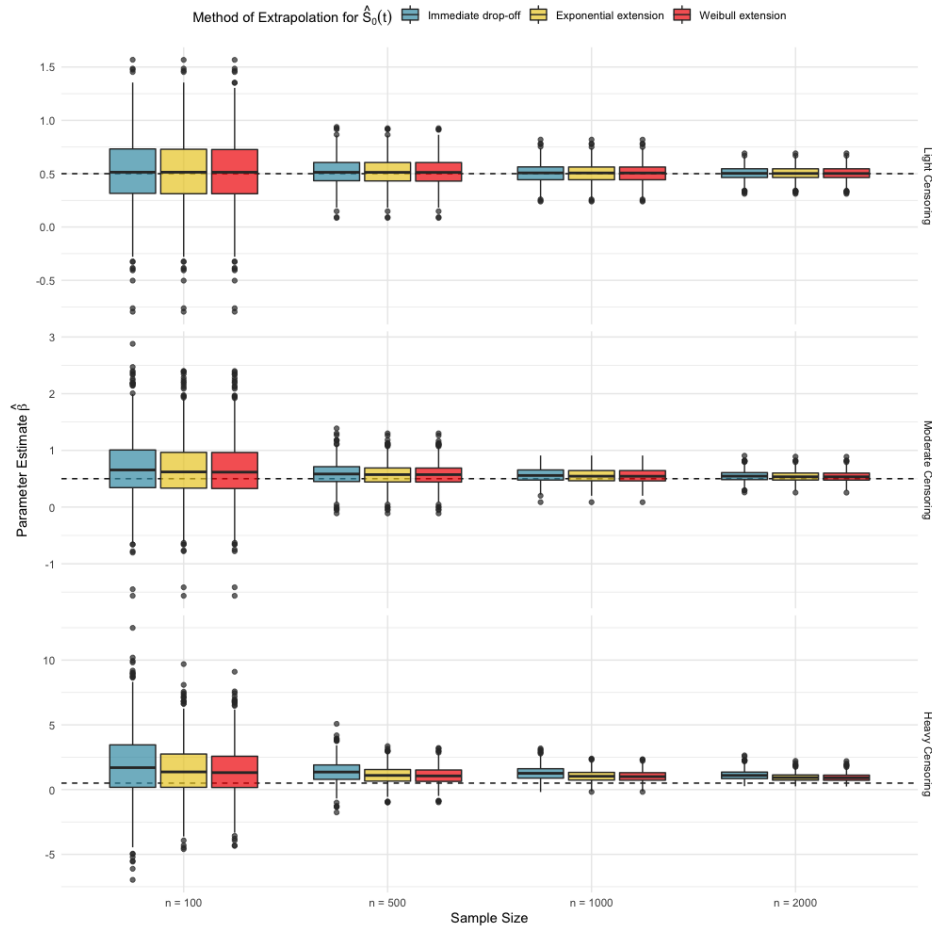
**Fig. S2.** With Weibull  $X$ , extrapolating Breslow's estimator  $\hat{S}_0(t)$  beyond the largest uncensored value  $\tilde{X}$  with any of the three extrapolation methods offered similar bias and efficiency for  $\hat{\beta}$  in conditional mean imputation with adaptive quadrature. The dashed line denotes the true parameter value,  $\beta = 0.5$ .



**Fig. S3.** With log-normal  $X$ , extrapolating Breslow's estimator  $\hat{S}_0(t)$  beyond the largest uncensored value  $\tilde{X}$  with any of the three extrapolation methods offered similar bias and efficiency for  $\hat{\beta}$  in conditional mean imputation with adaptive quadrature. The dashed line denotes the true parameter value,  $\beta = 0.5$ .



**Fig. S4.** Interpolating Breslow's estimator  $\hat{S}_0(t)$  between uncensored values with either of the two interpolation methods offered similar bias and efficiency for  $\hat{\beta}$  in conditional mean imputation with adaptive quadrature. The dashed line denotes the true parameter value,  $\beta = 0.5$ .



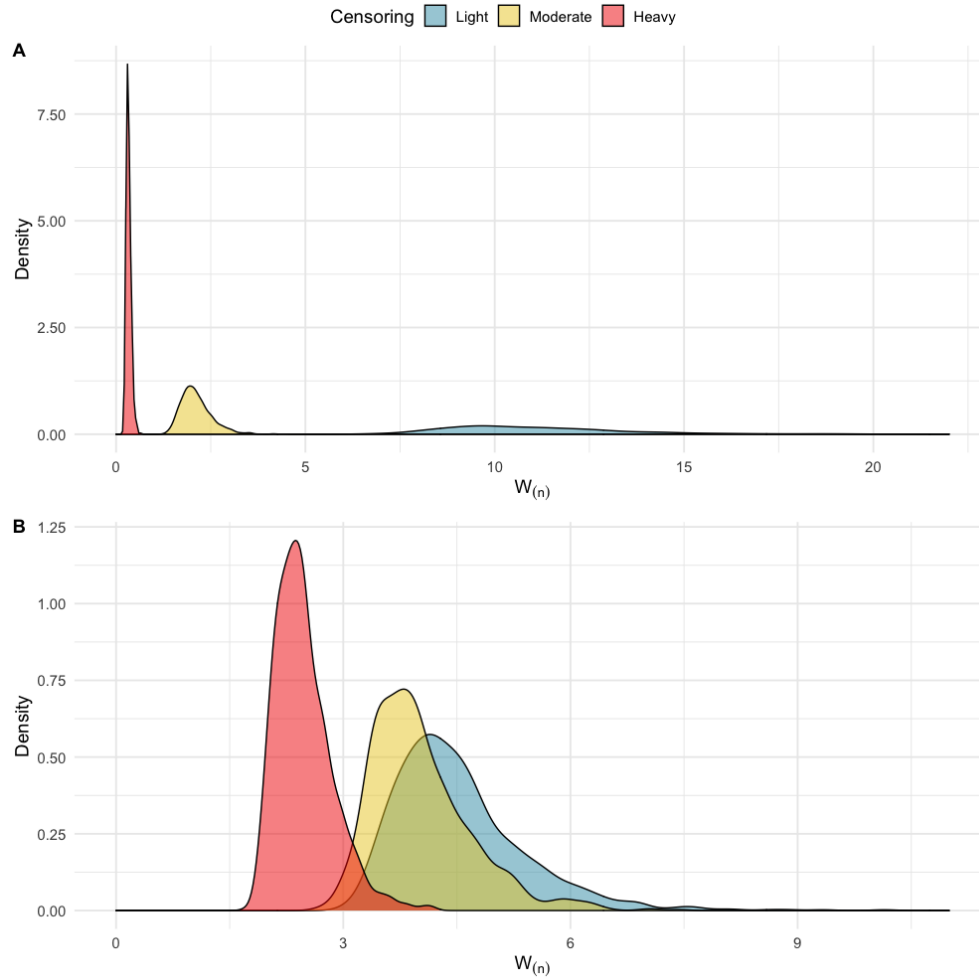
**Fig. S5.** Extrapolating Breslow's estimator  $\hat{S}_0(t)$  beyond the largest uncensored value  $\tilde{X}$  with any of the three extrapolation methods offered similar bias and efficiency for  $\hat{\beta}$  in conditional mean imputation with the trapezoidal rule. The dashed line denotes the true parameter value,  $\beta = 0.5$ .



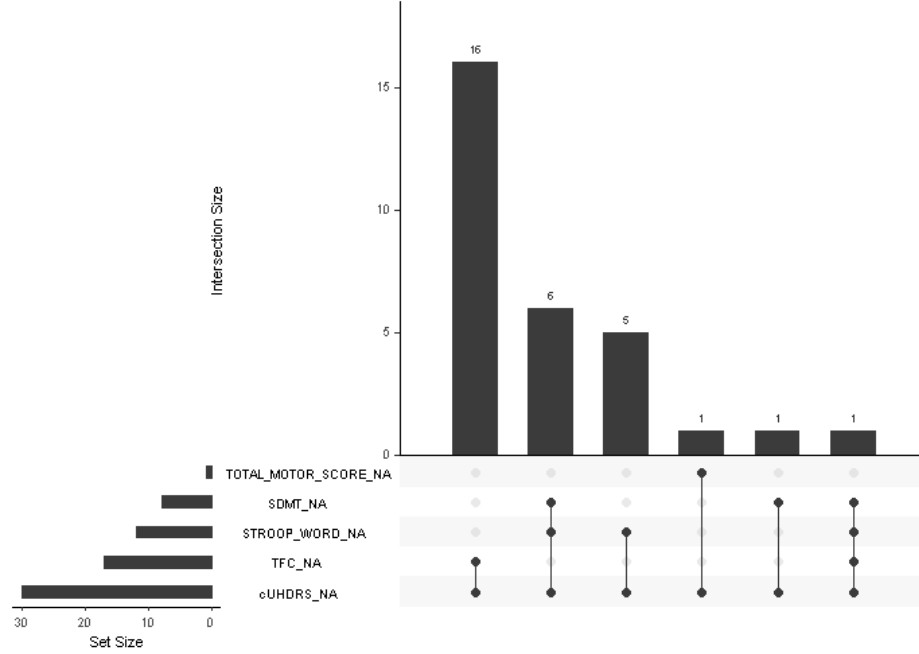
**Table S1.** Simulation results for log-normal  $X$  from the full cohort analysis and imputation approaches using the estimated survival function and adaptive quadrature versus the trapezoidal rule. (SE: empirical standard error; RE: empirical relative efficiency to full cohort.)

Censoring	$n$	Full Cohort		Adaptive Quadrature			Trapezoidal Rule		
		Bias	SE	Bias	SE	RE	Bias	SE	RE
$\hat{\alpha}$ : Intercept									
Light	100	0.005	0.242	0.001	0.263	0.852	−0.008	0.263	0.846
	500	0.005	0.109	0.003	0.115	0.899	−0.003	0.115	0.900
	2000	−0.001	0.052	0.002	0.056	0.852	0.003	0.056	0.875
Moderate	100	0.005	0.242	0.009	0.281	0.744	−0.008	0.283	0.732
	500	0.005	0.109	0.010	0.124	0.775	−0.007	0.122	0.805
	2000	−0.001	0.052	0.004	0.060	0.755	0.000	0.062	0.705
Heavy	100	0.005	0.242	−0.024	0.480	0.112	−0.060	0.496	0.105
	500	0.005	0.109	−0.000	0.189	0.141	−0.027	0.195	0.132
	2000	−0.001	0.052	0.007	0.095	0.140	−0.012	0.097	0.136
$\hat{\beta}$ : Coefficient on Censored $X$									
Light	100	−0.004	0.179	−0.001	0.201	0.792	0.011	0.204	0.775
	500	−0.001	0.077	−0.002	0.088	0.767	0.004	0.088	0.760
	2000	0.002	0.037	−0.001	0.041	0.791	−0.002	0.041	0.795
Moderate	100	−0.004	0.179	−0.012	0.219	0.670	0.011	0.225	0.635
	500	−0.001	0.077	−0.008	0.096	0.648	0.008	0.095	0.653
	2000	0.002	0.037	−0.003	0.045	0.657	0.001	0.048	0.599
Heavy	100	−0.004	0.179	0.020	0.431	0.139	0.085	0.468	0.118
	500	−0.001	0.077	0.004	0.164	0.188	0.043	0.172	0.169
	2000	0.002	0.037	−0.006	0.082	0.189	0.018	0.084	0.180
$\hat{\gamma}$ : Coefficient on Uncensored $Z$									
Light	100	−0.003	0.207	−0.005	0.212	0.958	−0.005	0.211	0.964
	500	−0.003	0.092	−0.001	0.093	0.979	−0.001	0.092	0.985
	2000	−0.002	0.044	−0.002	0.044	0.986	−0.002	0.044	0.979
Moderate	100	−0.003	0.207	−0.003	0.216	0.919	−0.005	0.213	0.946
	500	−0.003	0.092	−0.003	0.094	0.959	−0.001	0.093	0.961
	2000	−0.002	0.044	−0.002	0.045	0.960	−0.002	0.045	0.973
Heavy	100	−0.003	0.207	0.007	0.241	0.443	0.001	0.206	0.609
	500	−0.003	0.092	−0.009	0.104	0.467	−0.008	0.095	0.553
	2000	−0.002	0.044	−0.002	0.052	0.479	−0.003	0.046	0.591

*Note:* The MLE for the Weibull extension converged in  $\geq 99.8\%$  of replicates of imputation in each setting (just 8 and 6 of 9000 total replicates did not converge with adaptive quadrature and the trapezoidal rule, respectively); all other entries are based on 1000 replicates.



**Fig. S6.** Due to the Weibull distribution's skewness, higher censoring rates led to smaller values of  $W_{(n)}$ , which led to worse performance (i.e., higher bias) when calculating the conditional mean with the trapezoidal rule. **A** and **B** are the empirical densities of  $W_{(n)}$  when  $X$  was generated from a Weibull and a log-normal distribution, respectively, under light, moderate, or heavy censoring.

**Web Appendix C. Additional Results from the PREDICT-HD Analysis**

**Fig. S7.** Patterns of missing data in the outcome `cUHDRS` (composite Unified Huntington Disease Rating Scale) and its component variables total functional capacity (TFC), total motor score (`TOTAL_MOTOR_SCORE`), Symbol Digit Modality Test (SDMT), and Stroop Word Reading Test (`STROOP_WORD`) at study entry. This plot was created using the **nanianr** package (?).

### Web Appendix C.1. Details About Imputing Censored Times to Diagnosis

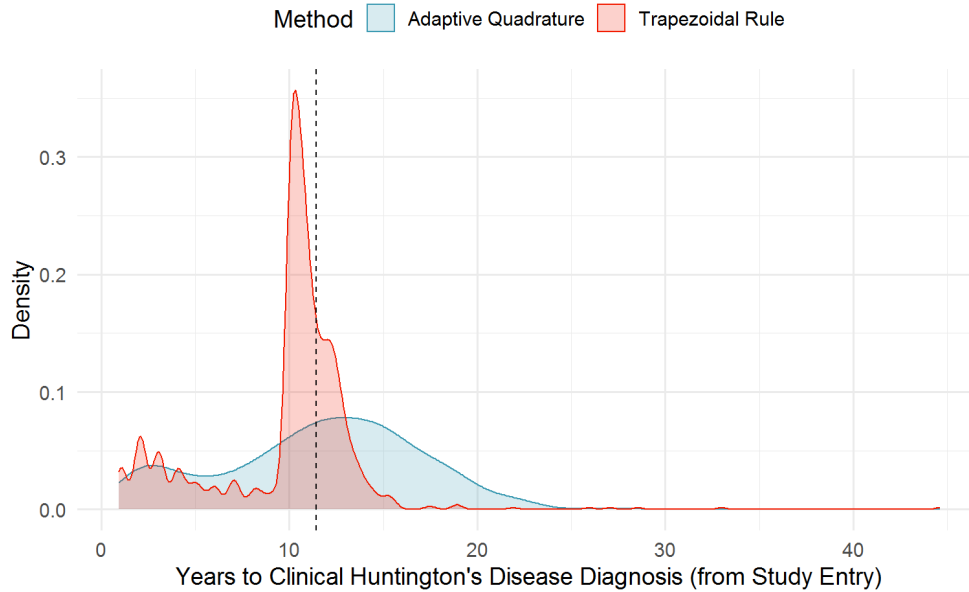
Imputation began by modeling the conditional survival function for  $\text{TIME}_0$  given other fully observed covariates from study entry. First, we fit the Cox proportional hazards model for  $h_{\lambda}(\text{TIME}_0 | \text{AGE}_0, \text{CAG}_0) = \lambda_0(\text{TIME}_0) \exp(\lambda_1 \text{AGE}_0 + \lambda_2 \text{AGE}_0 \times \text{CAG}_0)$ , and tested for proportional hazards using the `coxph` and `cox.zph` functions, respectively, from the **survival** package (?). (There was no evidence that the assumption was violated, with both  $p$ -values  $> 0.1$ .) The covariates  $\text{AGE}_0$  and  $\text{AGE}_0 \times \text{CAG}_0$ , were chosen to align with the CAP model proxy for time to diagnosis from (?). Then, we calculated Breslow's estimator  $\hat{S}_0(\text{TIME}_0)$  based on the estimated log hazard ratios  $\hat{\lambda}_1 = -0.038$  and  $\hat{\lambda}_2 = 0.022$ .

With this, we had an estimator  $\hat{S}(\text{TIME}_0 | \text{AGE}_0, \text{CAG}_0)$  for values of  $\text{TIME}_0$  up to  $\tilde{X} = 11.422$ , the longest observed time from study entry to diagnosis in PREDICT-HD. Following from our empirical findings in Section 3.3, we used the Weibull extension to extrapolate the survival estimator beyond the largest uncensored value, where  $\hat{S}(t = 11.422 | \text{AGE}_0 = 34.13, \text{CAG}_0 = 4) = 0.532$ . While we cannot guarantee that these data follow a Weibull distribution, the added flexibility of this extension over the exponential

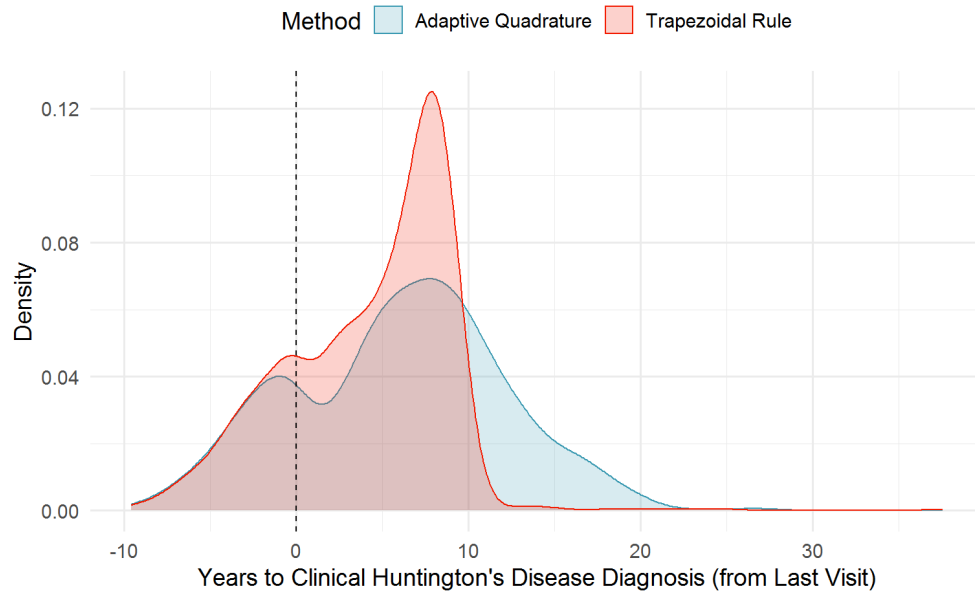
was appealing. Also, unlike our simulations, the context of  $\text{TIME}_0$  could be used to refine the upper bound of the integral in Equation (2.1). Specifically,  $\text{TIME}_0$  from study entry to clinical Huntington’s disease diagnosis could not be infinite for the simple reason that humans are not immortal. Instead, we assumed  $\text{TIME}_0$  of diagnosis would be no longer than 60 years from study entry. Additional details are in Web Appendix A.3.

### Web Appendix C.2. Comparing Imputed Times to Diagnosis

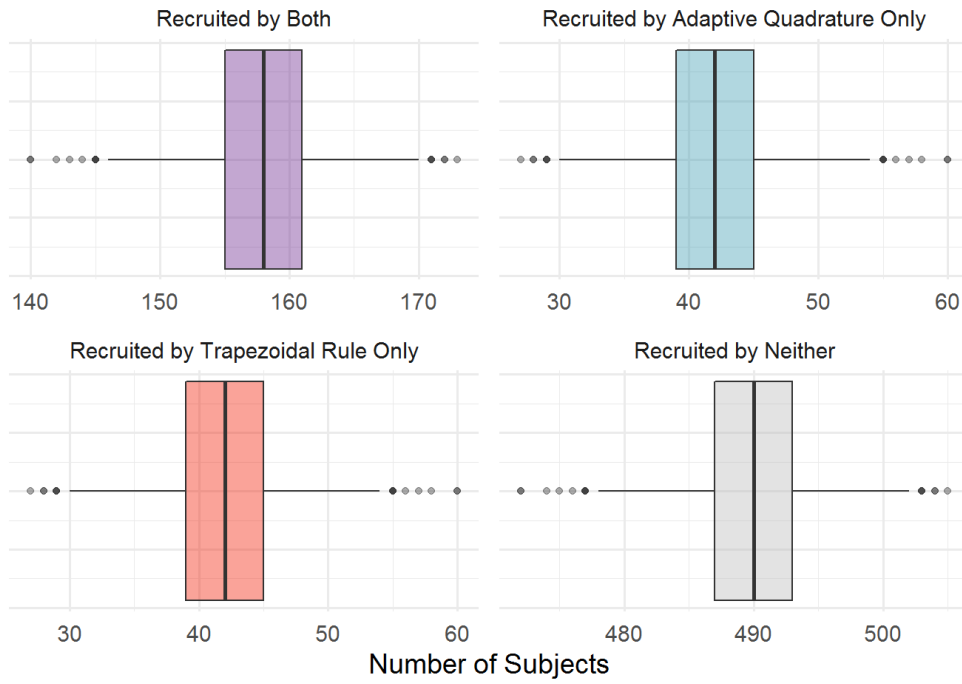
Empirical densities of observed and imputed  $\text{TIME}_0$  from study entry to clinical Huntington’s disease diagnosis for the two imputation approaches exhibited some distinct differences (Figure S8). Using adaptive quadrature for imputation led to a smooth, uni-modal density, with a peak not long after the largest uncensored value of  $\tilde{X} = 11.422$  years from study entry to diagnosis. Imputing using the trapezoidal rule instead led to a more volatile density that peaked earlier, at around 10 years to diagnosis. Interestingly, the trapezoidal rule led to a higher maximum of 45 years to diagnosis versus 29 with adaptive quadrature, but other quantiles were similar (e.g., within 4 years). We also noted differences between the densities of  $\text{TIME}_1$  from the last visit to clinical Huntington’s disease diagnosis (Figure S9), with adaptive quadrature still leading to more support for more larger values of  $\text{TIME}_1$ , representing longer pre-diagnosis follow-up.



**Fig. S8.** Histograms of observed and imputed times from study entry to Huntington’s disease diagnosis in the PREDICT-HD study. The dashed line denotes the longest uncensored value observed in the data,  $\tilde{X} = 11.4$  years from study entry to diagnosis.



**Fig. S9.** Histograms of observed and imputed times from last visit to Huntington's disease diagnosis in the PREDICT-HD study. The dashed line denotes the time of diagnosis.



**Fig. S10.** Statuses of  $n = 732$  resampled subjects considered for recruitment into a hypothetical clinical trial based on Huntington’s disease symptom progression models using the two imputation approaches in PREDICT-HD. New datasets of  $n = 732$  subjects were created by resampling from censored subjects in PREDICT-HD with replacement 1000 times.