Supplementary Materials for "Extrapolation before imputation reduces bias when imputing censored covariates"

Lotspeich and Garcia

November 27, 2023

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 $\boldsymbol{Z} = \mathbf{0}$), X follows an exponential distribution with rate ρ, we have $S_0(t) = \exp\{-(t/\rho)\}$. To allow the parametric segment to extend seamlessly from Breslow's estimator, the chosen rate $\widehat{\rho}$ is constrained so that $\exp\left\{-\left(\widetilde{X}/\widehat{\rho}\right)\right\} = \widehat{S}_0(\widetilde{X})$. We can solve this constraint for $\widehat{\rho} = -\widetilde{X}\log\left\{\widehat{S}_0(\widetilde{X})\right\}^{-1}$ and then extrapolate using $\widehat{S}_0(t) = \exp\left(\left[t\log\left\{\widehat{S}_0(\widetilde{X})\right\}\right]/\widetilde{X}\right)$ for $t > \widetilde{X}$. This is the *exponential extension* described in Section 2.5.1 and originally proposed by Brown et al. (1974).

Web Appendix A.2 Derivation of the Weibull Extension

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

In general, the shape and scale parameters, ν and ρ, 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

$$\begin{aligned} l_n(\nu,\rho) &= \sum_{i=1}^{n}\Delta_i\log\left\{\rho\nu W_i^{\nu-1}\exp\left(-\rho W_i^{\nu}\right)\right\} + \sum_{i=1}^{n}(1-\Delta_i)\log\left\{\exp\left(-\rho W_i^{\nu}\right)\right\} \\ &= -\rho\sum_{i=1}^{n}W_i^{\nu} + (\nu-1)\sum_{i=1}^{n}\Delta_i\log\left(W_i\right) + n_1\log\left(\rho\right) + n_1\log\left(\nu\right), \end{aligned} \tag{S.1}$$

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 connect with Breslow's estimator $\widehat{S}_0(t)$ at the largest uncensored covariate value, $\widetilde{X}$. This constraint on the Weibull survival function can be expressed as $\exp(-\rho\widetilde{X}^{\nu}) \equiv \widehat{S}_0(\widetilde{X})$, and it further translates into the following relationship between the shape and scale parameters:

$$\rho = -\log\left\{\widehat{S}_0(\widetilde{X})\right\}/(\widetilde{X}^{\nu}). \tag{S.2}$$

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

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

The maximum likelihood estimator (MLE) $\widehat{\nu}$ is obtained by finding the root of Equation (S.3) with a univariate Newton-type algorithm, as implemented in the `nlm` function in R (R Core Team 2019). Our initial guess for the shape parameter, which must be > 0, is $\hat{\nu}^{(0)} = 1\mathrm{E}^{-4}$, and the algorithm is restricted to positive values for $\widehat{\nu}$. Finally, $\widehat{\rho}$ is obtained by plugging $\widehat{\nu}$ 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, $\widehat{S}_0(t)$ for $t > \widetilde{X}$ 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 infinity 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 infinity. 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 ω.

For example, in our PREDICT-HD example (Section 4), the censored covariate X being imputed was `TIME_start` from study entry to clinical Huntington's disease diagnosis. Since this is an adult cohort, with all subjects having `AGE` ≥ 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 `TIME_start` that is in agreement with the recent overall life expectancy estimate of 78 years in the United States (U.S. Census Bureau 2017). 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 `TIME_start` or beyond reasonable values based on the context, leading to a trade-off between setting ω too low or too high.

Now, our choice of finite ω imposes an additional constraint on the Weibull extension. Since $S(\omega) \approx 0$ we further constrain ν and ρ such that $\exp\left(-\rho\omega^{\nu}\right) \approx 0$. Thus, we can find the corresponding values of $\widehat{\rho}$ and $\widehat{\nu}$ using the `uniroot` function in R (R Core Team 2019), since ρ can be treated as a function of ν as in Equation (S.2).

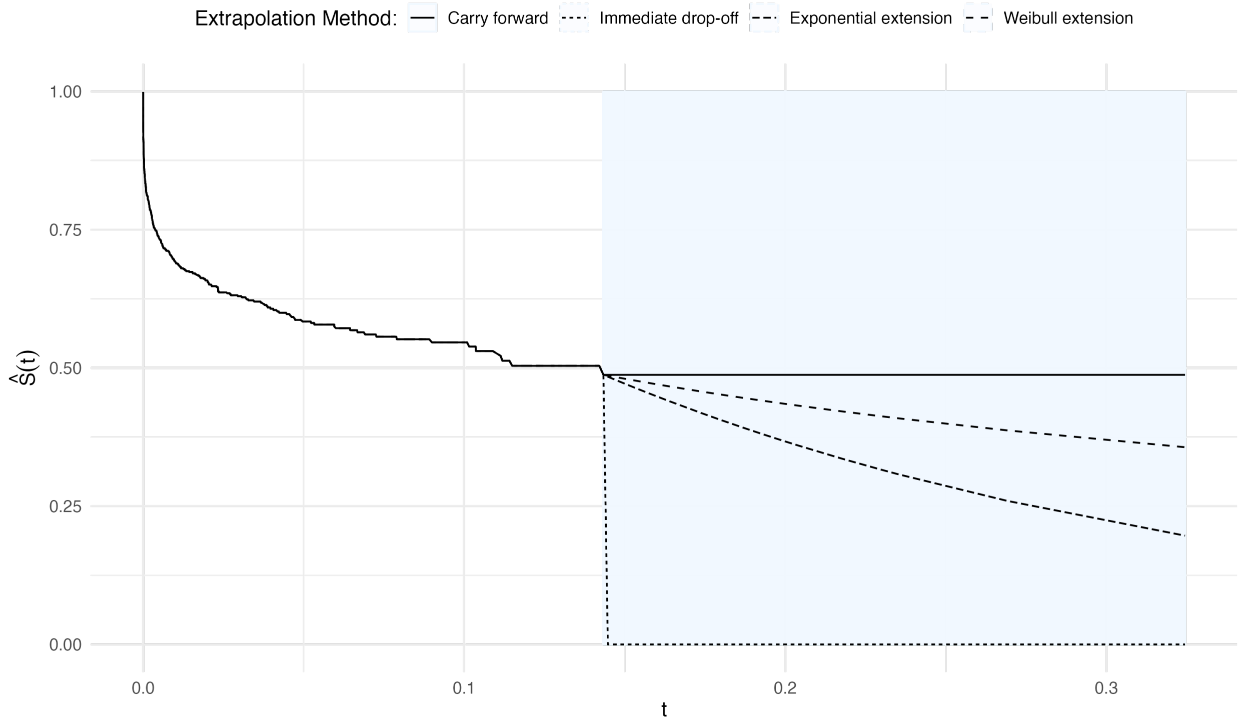

Figure S1: Illustration of the four extrapolation methods for a step survival function $\widehat{S}(t)$ in simulated data. The shaded area represents values of $t > \widetilde{X}$ (the largest uncensored value), where extrapolation is needed.

Web Appendix B Additional Results from the Simulation Studies

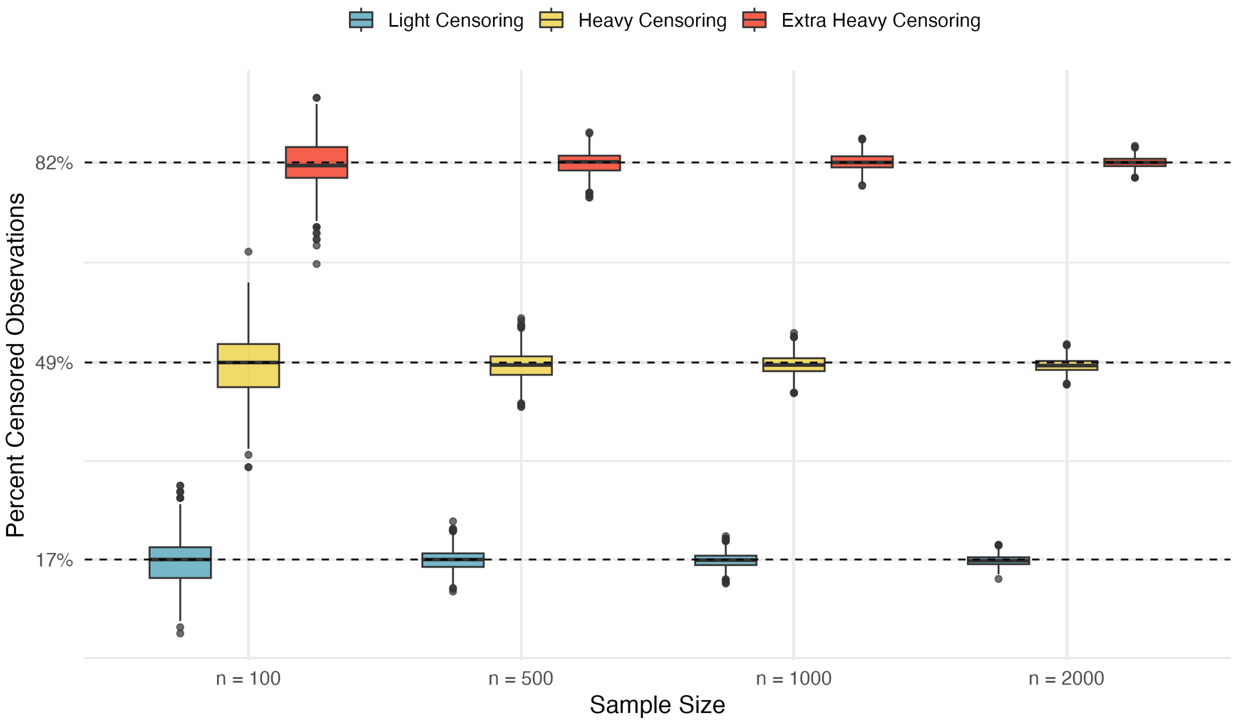

Figure S2: We explored light ($\sim 17\%$), heavy ($\sim 49\%$), and extra heavy ($\sim 82\%$) censoring in Weibull X, induced by generating C from an exponential distribution with rates $= 0.5$, 2.9, and 20, respectively.

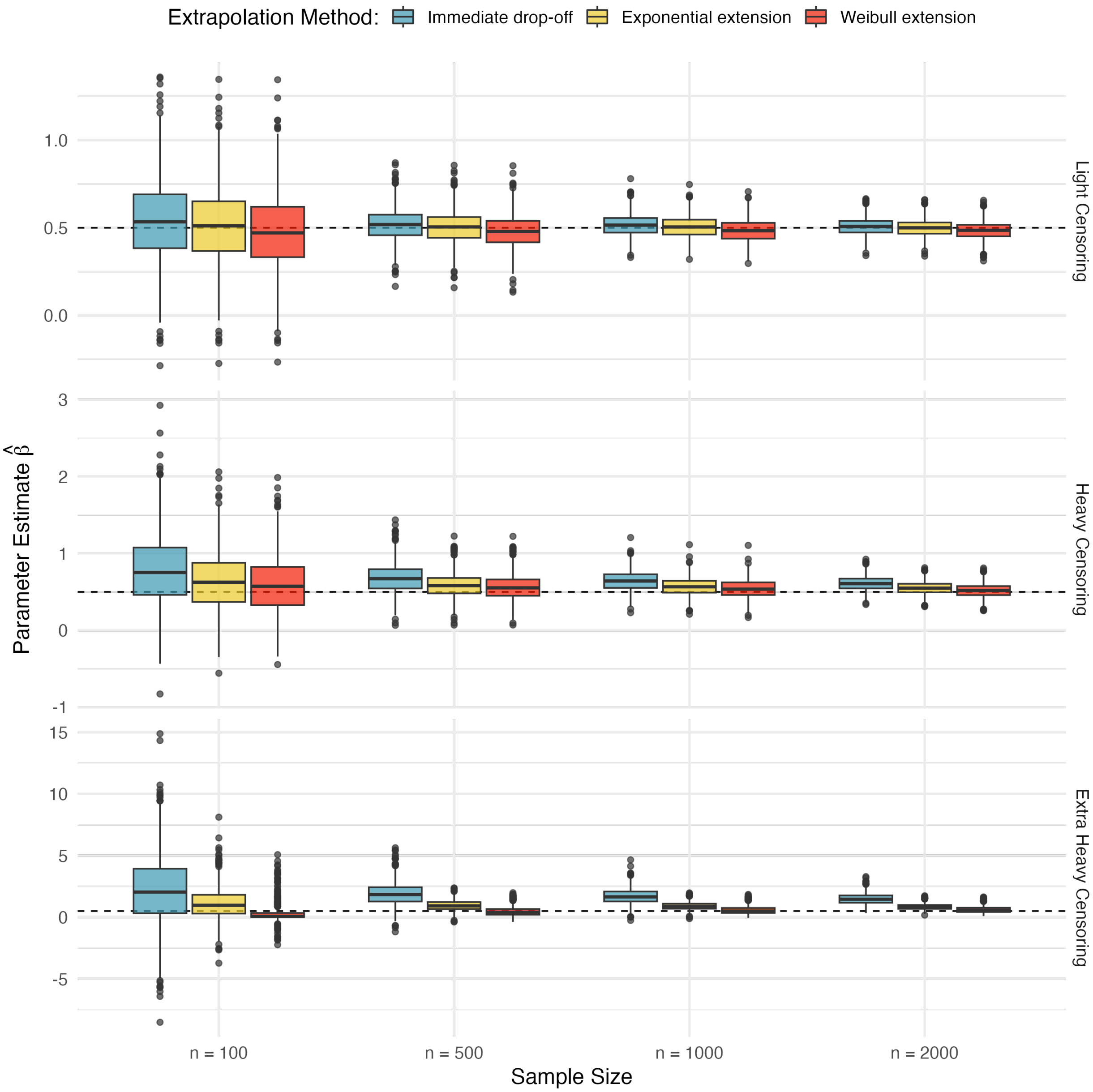

Figure S3: With Weibull X, extrapolating Breslow's estimator $\widehat{S}_0(t)$ beyond the largest uncensored value $\widetilde{X}$ with the Weibull extension offered the lowest bias and best efficiency for $\hat{\beta}$ in extrapolated conditional mean imputation. The dashed line denotes the true parameter value, $\beta = 0.5$.

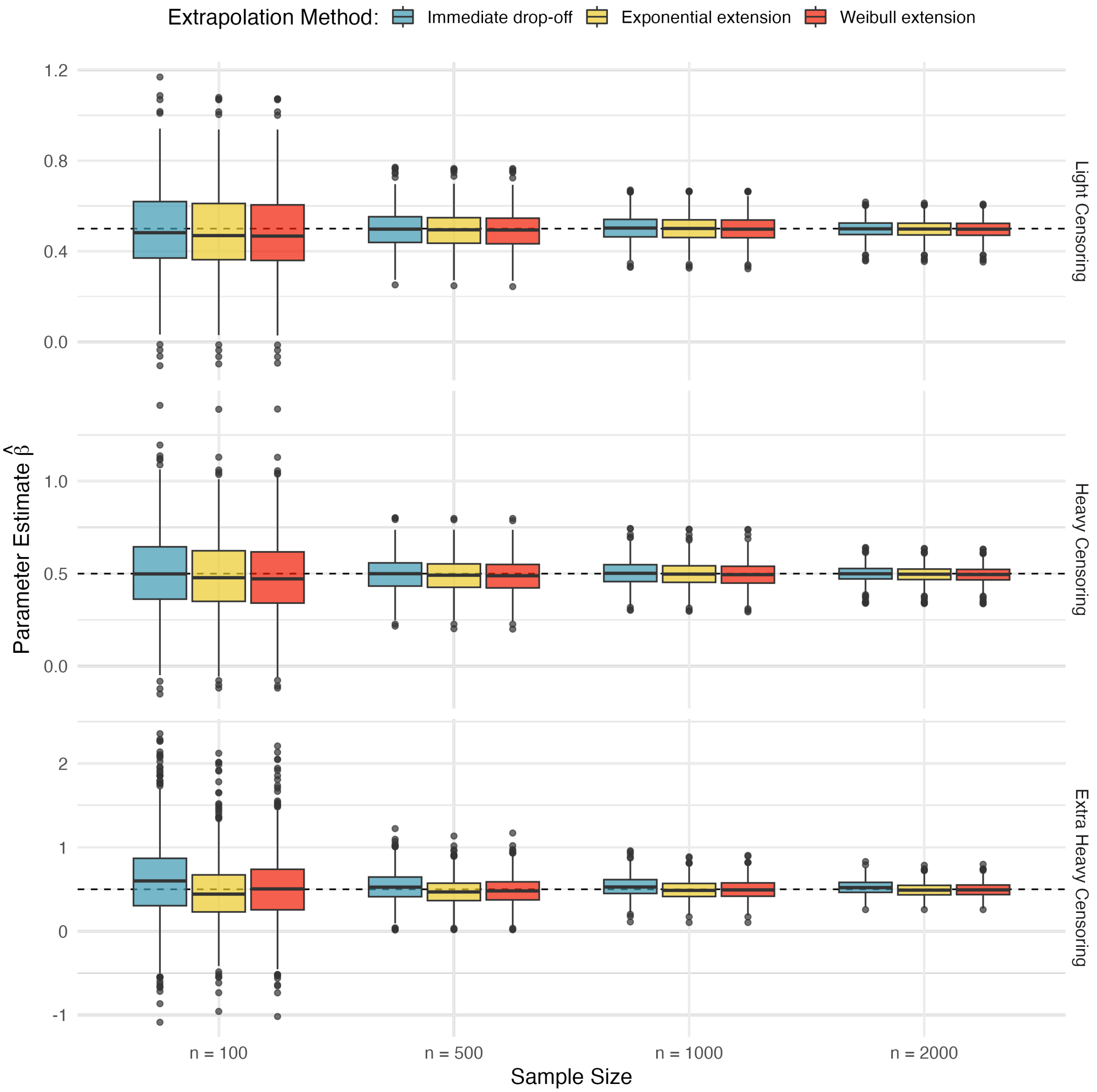

Figure S4: With log-normal X, extrapolating Breslow's estimator $\widehat{S}_0(t)$ beyond the largest uncensored value $\widetilde{X}$ with any of the three extrapolation methods offered similar bias and efficiency for $\hat{\beta}$ in extrapolated conditional mean imputation. The dashed line denotes the true parameter value, $\beta = 0.5$.

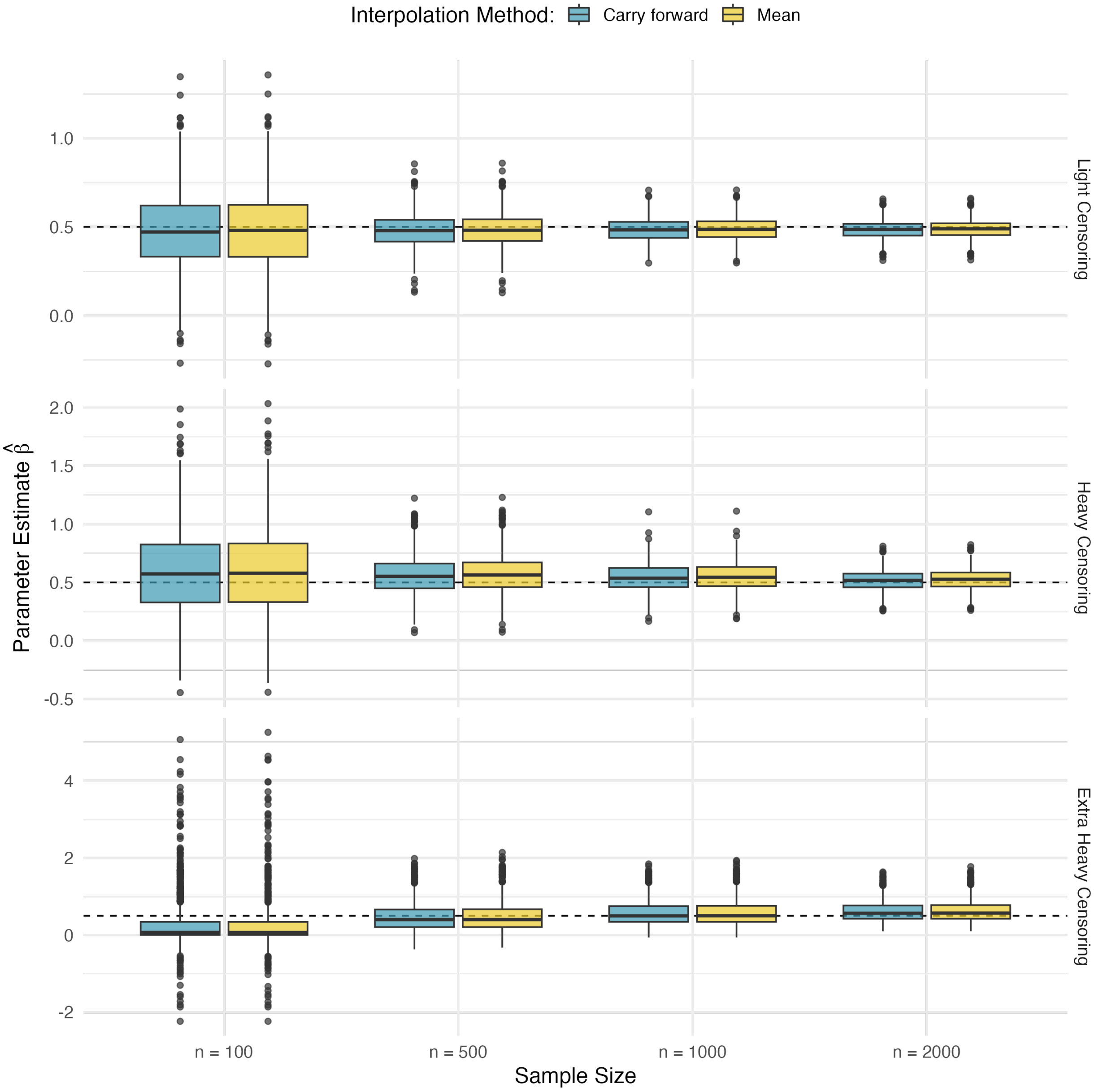

Figure S5: Interpolating Breslow's estimator $\widehat{S}_0(t)$ between uncensored values with either of the two interpolation methods offered similar bias and efficiency for $\hat{\beta}$ in extrapolated conditional mean imputation. Between uncensored values, $\widehat{S}_0(\cdot)$ was either be carried forward from the last uncensored value or taken as the mean of the uncensored values immediately before and after. The dashed line denotes the true parameter value, $\beta = 0.5$.

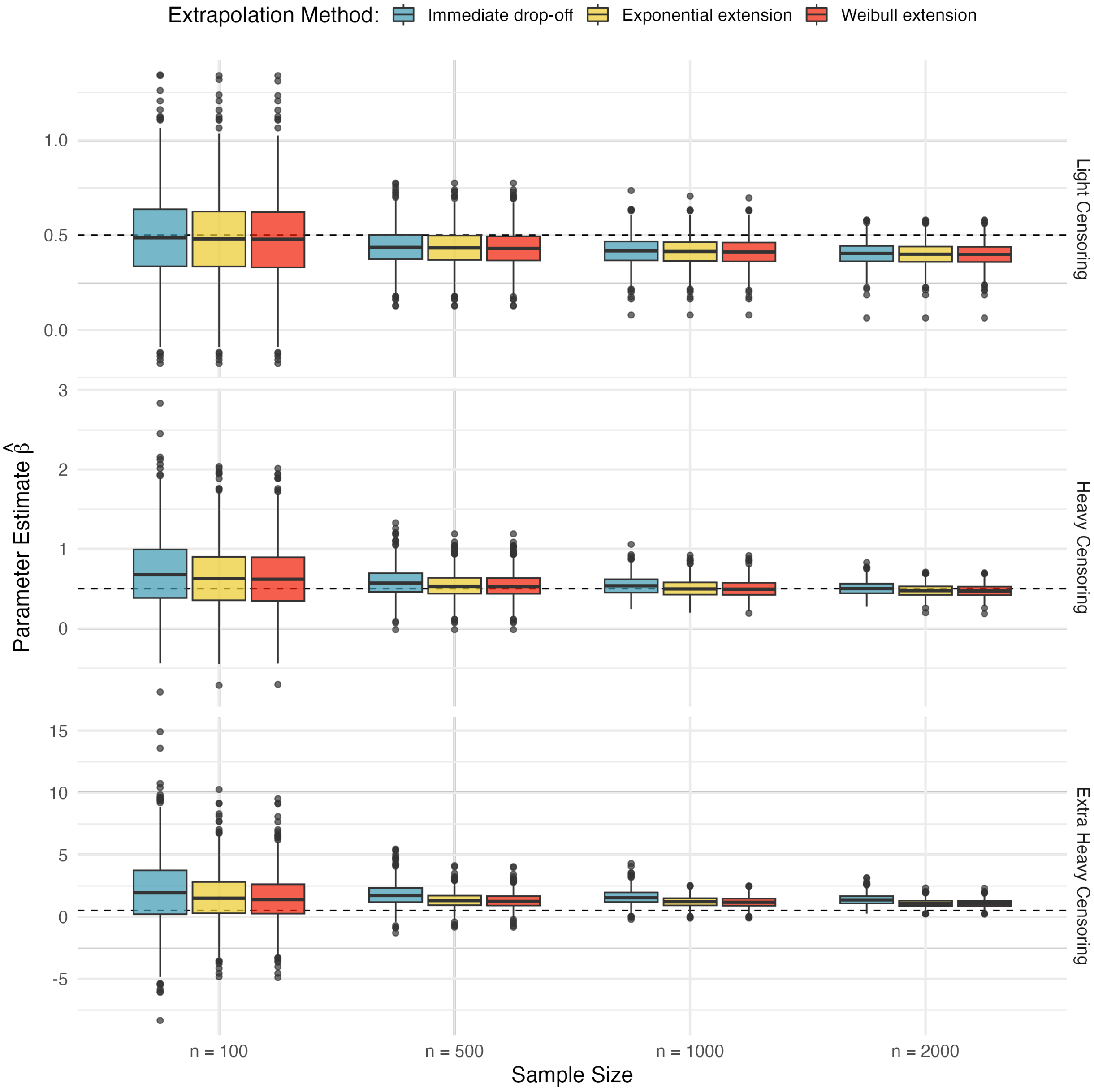

Figure S6: Extrapolating Breslow's estimator $\widehat{S}_0(t)$ beyond the largest uncensored value $\widetilde{X}$ with any of the three extrapolation methods offered similar bias and efficiency for $\hat{\beta}$ in non-extrapolated conditional mean imputation. The dashed line denotes the true parameter value, $\beta = 0.5$.

Table S1: Simulation results for Weibull X independent of Z from the full cohort analysis (i.e., where all n observations had uncensored X) and imputation approaches.

		Full Cohort			**Extrapolated Conditional Mean Imputation**				**Non-Extrapolated Conditional Mean Imputation**			
Censoring	n	**Bias**	**(%)**	**SE**	**Bias**	**(%)**	**SE**	**RE**	**Bias**	**(%)**	**SE**	**RE**
					$\hat{\alpha}$: Intercept							
Light	100	0.002	(0.22)	0.164	0.003	(0.33)	0.168	0.948	−0.001	(−0.11)	0.169	0.936
	500	−0.003	(−0.32)	0.068	−0.001	(−0.09)	0.070	0.945	−0.005	(−0.47)	0.071	0.939
	1000	0.000	(0.00)	0.051	0.002	(0.20)	0.051	0.976	−0.001	(−0.10)	0.052	0.963
	2000	0.001	(0.07)	0.037	0.002	(0.22)	0.038	0.948	0.000	(0.00)	0.038	0.946
Heavy	100	0.002	(0.22)	0.164	−0.003	(−0.34)	0.187	0.763	−0.010	(−0.99)	0.190	0.743
	500	−0.003	(−0.32)	0.068	−0.007	(−0.72)	0.078	0.770	−0.012	(−1.23)	0.079	0.754
	1000	0.000	(0.00)	0.051	−0.002	(−0.22)	0.055	0.842	−0.007	(−0.69)	0.056	0.811
	2000	0.001	(0.07)	0.037	0.000	(−0.04)	0.041	0.809	−0.005	(−0.46)	0.041	0.806
Extra Heavy	100	0.002	(0.22)	0.164	0.020	(1.97)	0.247	0.440	−0.009	(−0.94)	0.269	0.370
	500	−0.003	(−0.32)	0.068	−0.006	(−0.64)	0.107	0.406	−0.022	(−2.17)	0.112	0.373
	1000	0.000	(0.00)	0.051	−0.005	(−0.48)	0.077	0.438	−0.017	(−1.69)	0.080	0.403
	2000	0.001	(0.07)	0.037	−0.003	(−0.29)	0.056	0.427	−0.012	(−1.24)	0.058	0.404
					$\hat{\beta}$: Coefficient on Censored X							
Light	100	−0.009	(−1.85)	0.274	−0.022	(−4.39)	0.303	0.821	0.007	(1.33)	0.318	0.747
	500	0.004	(0.74)	0.109	−0.010	(−1.91)	0.122	0.797	0.010	(2.08)	0.125	0.768
	1000	0.005	(1.00)	0.078	−0.005	(−1.04)	0.087	0.809	0.010	(1.94)	0.088	0.791
	2000	0.000	(−0.05)	0.056	−0.008	(−1.66)	0.063	0.783	0.002	(0.43)	0.063	0.776
Heavy	100	−0.009	(−1.85)	0.274	0.033	(6.61)	0.448	0.376	0.114	(22.78)	0.508	0.292
	500	0.004	(0.74)	0.109	0.031	(6.24)	0.183	0.356	0.079	(15.88)	0.192	0.322
	1000	0.005	(1.00)	0.078	0.022	(4.34)	0.125	0.393	0.061	(12.27)	0.130	0.365
	2000	0.000	(−0.05)	0.056	0.005	(1.03)	0.094	0.357	0.040	(8.06)	0.094	0.358
Extra Heavy	100	−0.009	(−1.85)	0.274	−0.165	(−32.95)	0.741	0.137	0.745	(148.99)	1.905	0.021
	500	0.004	(0.74)	0.109	0.024	(4.73)	0.398	0.075	0.600	(119.91)	0.626	0.031
	1000	0.005	(1.00)	0.078	0.085	(16.95)	0.306	0.066	0.533	(106.57)	0.420	0.035
	2000	0.000	(−0.05)	0.056	0.100	(19.92)	0.239	0.055	0.445	(89.00)	0.303	0.034
					$\hat{\gamma}$: Coefficient on Uncensored Z							
Light	100	0.000	(−0.12)	0.205	0.001	(0.32)	0.206	0.991	0.000	(0.14)	0.205	1.000
	500	0.002	(0.90)	0.090	0.002	(0.73)	0.090	0.987	0.002	(0.73)	0.090	0.991
	1000	−0.002	(−0.88)	0.062	−0.002	(−0.91)	0.062	1.004	−0.002	(−0.84)	0.063	0.993
	2000	−0.001	(−0.34)	0.045	−0.001	(−0.34)	0.045	0.988	−0.001	(−0.32)	0.045	0.991
Heavy	100	0.000	(−0.12)	0.205	0.002	(0.65)	0.208	0.978	0.001	(0.24)	0.206	0.996
	500	0.002	(0.90)	0.090	0.002	(0.70)	0.091	0.967	0.002	(0.74)	0.091	0.977
	1000	−0.002	(−0.88)	0.062	−0.002	(−0.97)	0.063	0.974	−0.002	(−0.73)	0.064	0.956
	2000	−0.001	(−0.34)	0.045	−0.001	(−0.30)	0.046	0.955	−0.001	(−0.28)	0.046	0.974
Extra Heavy	100	0.000	(−0.12)	0.205	−0.005	(−1.83)	0.257	0.639	0.000	(−0.04)	0.206	0.990
	500	0.002	(0.90)	0.090	0.004	(1.70)	0.099	0.823	0.002	(0.94)	0.091	0.970
	1000	−0.002	(−0.88)	0.062	−0.001	(−0.39)	0.067	0.869	−0.002	(−0.63)	0.064	0.942
	2000	−0.001	(−0.34)	0.045	0.000	(−0.20)	0.049	0.855	−0.001	(−0.25)	0.046	0.963

Note: **Bias (%)**: empirical bias (empirical percent bias); **SE**: empirical standard error; **RE**: empirical relative efficiency to the full-cohort analysis. The censored covariate X was generated from a Weibull distribution with shape = 0.75 and scale = 0.25. All other variables were generated as in Section 3.1. True parameter values were $(\alpha, \beta, \gamma) = (1, 0.5, 0.25)$. All entries are based on 1000 replicates.

Table S2: Simulation results for log-normal X from the full cohort analysis (i.e., where all n observations had uncensored X) and imputation approaches.

		Full Cohort			**Extrapolated Conditional Mean Imputation**				**Non-Extrapolated Conditional Mean Imputation**			
Censoring	n	**Bias**	**(%)**	**SE**	**Bias**	**(%)**	**SE**	**RE**	**Bias**	**(%)**	**SE**	**RE**
					$\hat{\alpha}$: Intercept							
Light	100	−0.001	(−0.14)	0.233	0.009	(0.94)	0.251	0.862	−0.006	(−0.58)	0.252	0.850
	500	0.002	(0.17)	0.103	0.010	(1.01)	0.112	0.839	0.001	(0.13)	0.112	0.835
	1000	−0.001	(−0.10)	0.074	0.003	(0.25)	0.080	0.864	−0.005	(−0.46)	0.080	0.851
	2000	0.000	(0.02)	0.051	0.004	(0.37)	0.054	0.891	−0.003	(−0.25)	0.054	0.885
Heavy	100	−0.001	(−0.14)	0.233	0.011	(1.08)	0.270	0.744	−0.014	(−1.37)	0.277	0.706
	500	0.002	(0.17)	0.103	0.015	(1.49)	0.121	0.715	−0.001	(−0.07)	0.123	0.696
	1000	−0.001	(−0.10)	0.074	0.006	(0.62)	0.086	0.732	−0.007	(−0.68)	0.087	0.718
	2000	0.000	(0.02)	0.051	0.007	(0.67)	0.058	0.775	−0.005	(−0.47)	0.058	0.766
Extra Heavy	100	−0.001	(−0.14)	0.233	−0.017	(−1.72)	0.449	0.269	−0.072	(−7.18)	0.480	0.235
	500	0.002	(0.17)	0.103	0.017	(1.66)	0.190	0.292	−0.023	(−2.28)	0.199	0.266
	1000	−0.001	(−0.10)	0.074	0.003	(0.26)	0.138	0.287	−0.030	(−3.03)	0.143	0.267
	2000	0.000	(0.02)	0.051	0.010	(0.98)	0.094	0.289	−0.023	(−2.35)	0.097	0.276
					$\hat{\beta}$: Coefficient on Censored X							
Light	100	−0.005	(−1.09)	0.165	−0.018	(−3.67)	0.185	0.795	−0.005	(−1.02)	0.187	0.777
	500	−0.001	(−0.22)	0.073	−0.009	(−1.87)	0.082	0.778	−0.004	(−0.80)	0.082	0.777
	1000	0.001	(0.23)	0.052	−0.002	(−0.39)	0.059	0.799	0.001	(0.27)	0.059	0.794
	2000	0.000	(0.07)	0.036	−0.003	(−0.53)	0.040	0.817	0.000	(−0.10)	0.040	0.816
Heavy	100	−0.005	(−1.09)	0.165	−0.021	(−4.11)	0.206	0.642	0.002	(0.36)	0.213	0.602
	500	−0.001	(−0.22)	0.073	−0.014	(−2.84)	0.091	0.631	−0.004	(−0.87)	0.092	0.629
	1000	0.001	(0.23)	0.052	−0.006	(−1.16)	0.065	0.642	0.001	(0.22)	0.066	0.637
	2000	0.000	(0.07)	0.036	−0.005	(−1.04)	0.044	0.658	−0.001	(−0.17)	0.044	0.652
Extra Heavy	100	−0.005	(−1.09)	0.165	0.014	(2.79)	0.403	0.168	0.082	(16.34)	0.442	0.139
	500	−0.001	(−0.22)	0.073	−0.015	(−2.92)	0.162	0.201	0.022	(4.42)	0.170	0.182
	1000	0.001	(0.23)	0.052	0.000	(−0.07)	0.121	0.187	0.026	(5.28)	0.125	0.176
	2000	0.000	(0.07)	0.036	−0.006	(−1.28)	0.083	0.186	0.016	(3.30)	0.084	0.184
					$\hat{\gamma}$: Coefficient on Uncensored Z							
Light	100	0.006	(2.22)	0.203	0.004	(1.52)	0.207	0.959	0.012	(4.83)	0.206	0.971
	500	0.000	(0.05)	0.090	0.000	(−0.18)	0.091	0.963	0.008	(3.00)	0.091	0.974
	1000	−0.002	(−0.95)	0.064	−0.003	(−1.35)	0.065	0.971	0.005	(1.93)	0.065	0.971
	2000	−0.002	(−0.79)	0.045	−0.003	(−1.30)	0.045	0.972	0.005	(2.00)	0.045	0.992
Heavy	100	0.006	(2.22)	0.203	0.002	(0.89)	0.210	0.935	0.016	(6.53)	0.207	0.957
	500	0.000	(0.05)	0.090	−0.002	(−0.67)	0.092	0.939	0.012	(4.90)	0.091	0.964
	1000	−0.002	(−0.95)	0.064	−0.003	(−1.38)	0.065	0.960	0.010	(4.11)	0.065	0.962
	2000	−0.002	(−0.79)	0.045	−0.004	(−1.74)	0.046	0.924	0.010	(4.00)	0.045	0.978
Extra Heavy	100	0.006	(2.22)	0.203	−0.001	(−0.34)	0.248	0.669	0.028	(11.25)	0.212	0.917
	500	0.000	(0.05)	0.090	−0.005	(−1.91)	0.107	0.701	0.024	(9.79)	0.093	0.926
	1000	−0.002	(−0.95)	0.064	−0.008	(−3.02)	0.075	0.713	0.022	(8.75)	0.067	0.907
	2000	−0.002	(−0.79)	0.045	−0.010	(−3.85)	0.053	0.703	0.022	(8.96)	0.045	0.964

Note: **Bias (%)**: empirical bias (empirical percent bias); **SE**: empirical standard error; **RE**: empirical relative efficiency to the full-cohort analysis. The censored covariate X was generated from a log-normal distribution with mean $= 0.05Z$ and variance $= 0.25$ (on the log scale). All other variables were generated as in Section 3.1. True parameter values were $(\alpha, \beta, \gamma) = (1, 0.5, 0.25)$. The MLE for the Weibull extension converged in $\geq 99.8\%$ of replicates of imputation in each setting (just 3 of 12,000 total replicates did not converge); all other entries are based on 1000 replicates.

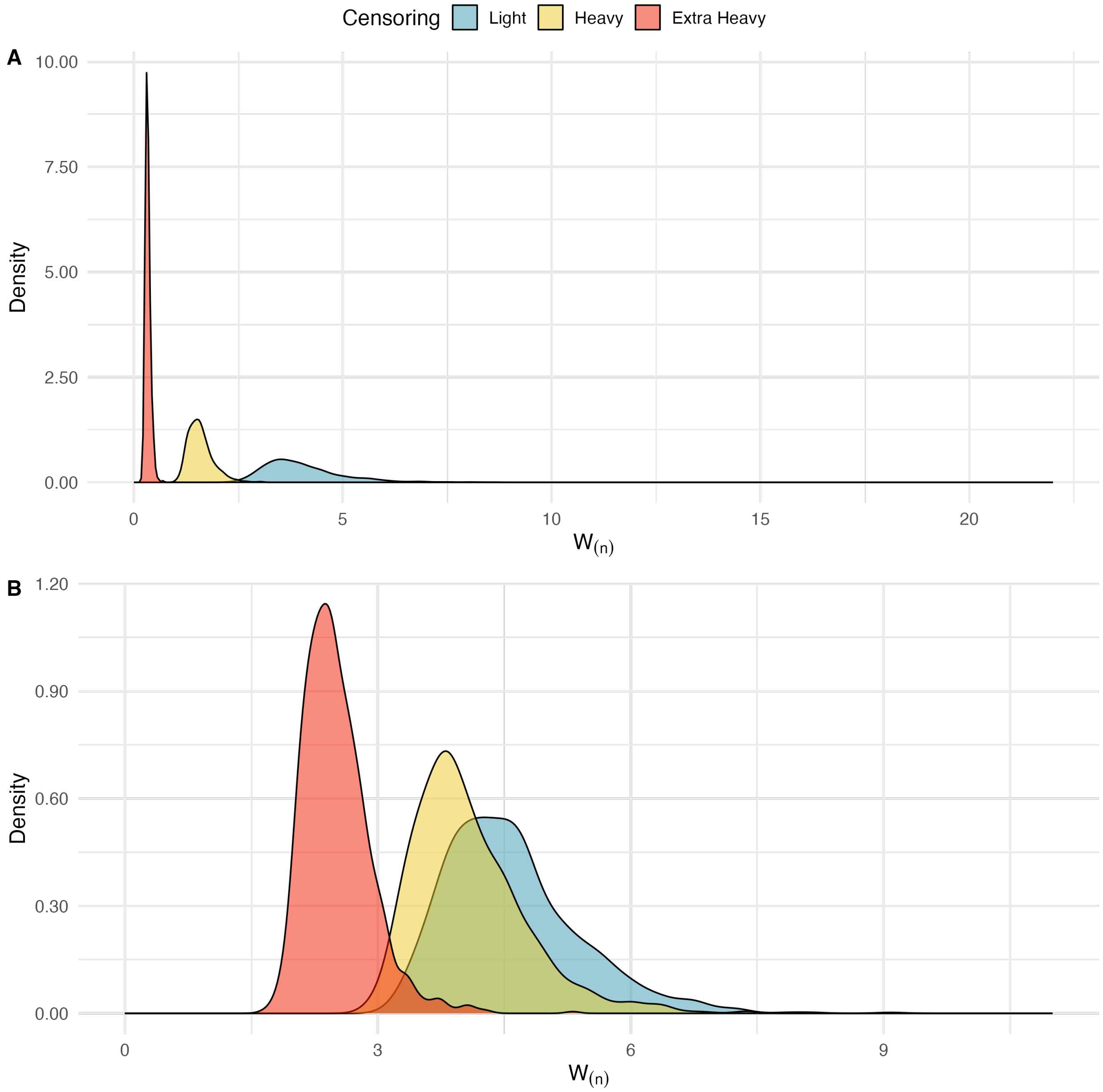

Figure S7: Due to the Weibull distribution's skewness, higher censoring rates led to smaller values of $W_{(n)}$ (the maximum of the observed covariate used as the integral's upper bound by the trapezoidal rule). With smaller values of $W_{(n)}$, the trapezoidal rule was cutting off more of the survival function, leading to worse performance (i.e., higher bias) with non-extrapolated conditional mean imputation. **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, heavy, or extra heavy censoring.

Web Appendix C Additional Results from the PREDICT-HD Analysis

To be included in our analysis, subjects needed to have (i) a $\texttt{CAG}$ repeat length ≥ 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, 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 time from study entry to diagnosis.

Of the subjects who were excluded, 30 were excluded specifically due to missing data in one or more of the component tests needed to calculate the cUHDRS at their first and last visits. The patterns of missing data in these variables for these subjects are displayed in Figure S8. From this figure, it can be seen that most subjects excluded for missing data (16 out of 30) were missing their total functional capacity score. There were also six subjects excluded for missing both assessments of cognitive impairment (Symbol Digit Modality Test and Stroop Word Reading Test), and five subjects excluded for missing just the Stroop Word Reading Test. The remaining three excluded subjects were the only ones with their patterns.

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

Imputation began by modeling the conditional survival function for $\texttt{TIME\_start}$ given other fully observed covariates from study entry. First, we fit the Cox proportional hazards model for

$$h_{\boldsymbol{\lambda}}(\texttt{TIME\_start}|\texttt{AGE}, \texttt{CAG}) = \lambda_0(\texttt{TIME\_start}) \exp\left(\lambda_1 \texttt{AGE} + \lambda_2 \texttt{AGE} \times \texttt{CAG}\right),$$

and tested for proportional hazards using the `coxph` and `cox.zph` functions, respectively, from the **survival** package (Therneau & Grambsch 2000). There was no evidence that the assumption was violated, with both p-values > 0.1. The covariates $\texttt{AGE}$ and $\texttt{AGE} \times \texttt{CAG}$, were chosen to align with the CAP model proxy for time to diagnosis from Zhang et al. (2011). Then, we calculated Breslow's estimator $\widehat{S}_0(\texttt{TIME\_start})$ based on the estimated log hazard ratios $\hat{\lambda}_1 = -0.038$ and $\hat{\lambda}_2 = 0.022$.

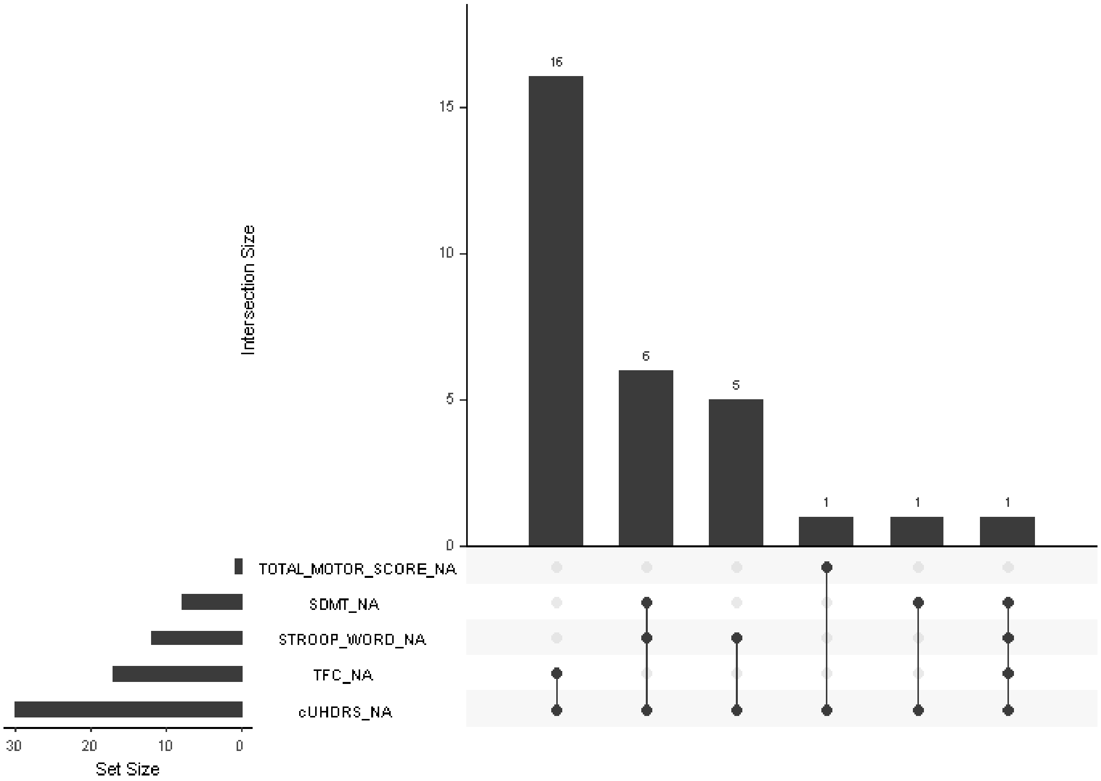

Figure S8: Patterns of missing data in the outcome `cUHDRS` (composite Unified Huntington Disease Rating Scale) and its component variables total functional capacity (`TFC`), total motor score (`TOTAL_MOTOR_SCORE`), Symbol Digit Modality Test (`SDMT`), and Stroop Word Reading Test (`STROOP_WORD`) at study entry. This plot was created using the **naniar** package (Tierney et al. 2021).

With this, we had an estimator $\widehat{S}(\texttt{TIME\_start}|\texttt{AGE}, \texttt{CAG})$ for values of `TIME_start` up to $\widetilde{X} = 11.42$, 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 $\widehat{S}_0(t = 11.42) = 0.89$. 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 `TIME_start` could be used to refine the upper bound of the integral in Equation (1). Specifically, `TIME_start` 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 `TIME_start` 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 `TIME_start` from study entry to clinical Huntington's disease diagnosis for the two imputation approaches exhibited some distinct differences (Figure S9). Using extrapolated conditional mean imputation led to a smooth, unimodal density, with a peak not long after the largest uncensored value of $\widetilde{X} = 11.42$ years from study entry to diagnosis. Non-extrapolated conditional mean imputation instead led to a more volatile density that peaked earlier, at around 10 years to diagnosis. Interestingly, the latter approach led to a higher maximum of 45 years to diagnosis versus 29 years with the former, but other quantiles were similar. We also noted differences between the densities of `TIME_end` from the last visit to clinical Huntington's disease diagnosis (Figure S10), with extrapolated conditional mean imputation still leading to more support for larger values of `TIME_end`, representing longer pre-diagnosis follow-up.

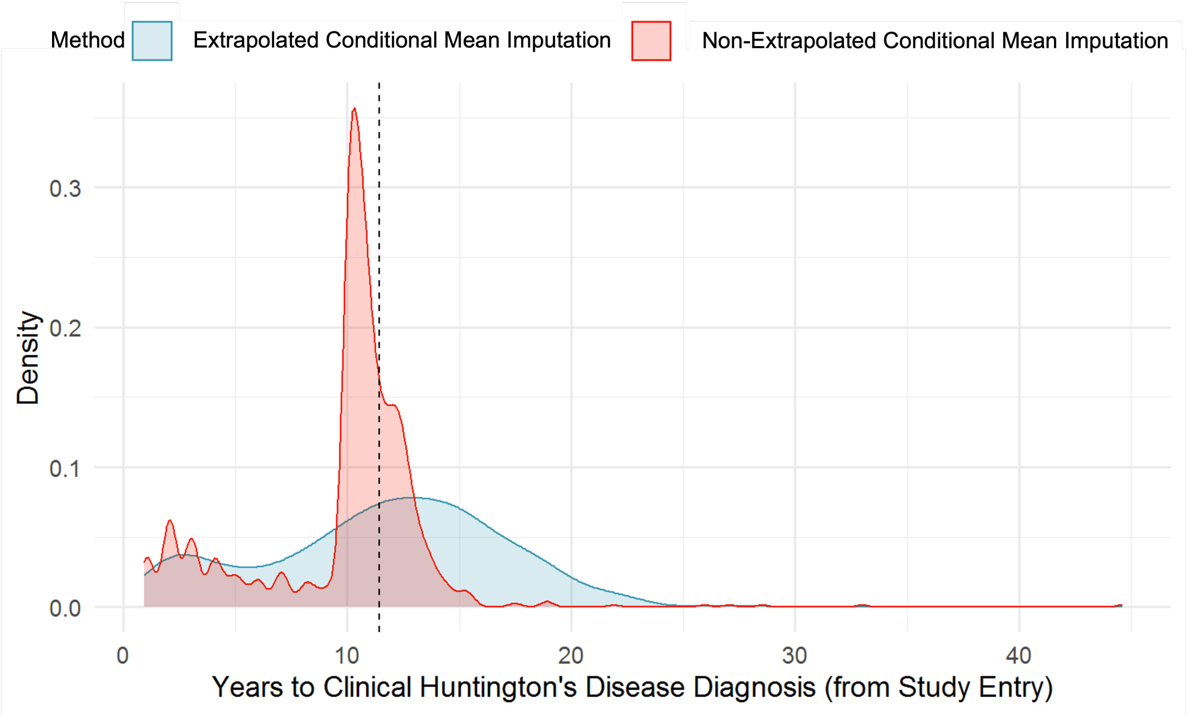

Figure S9: Histograms of observed and imputed times from study entry to Huntington's disease diagnosis in the PREDICT-HD study. The dashed line denotes the longest uncensored value observed in the data, $\widetilde{X} = 11.42$ years from study entry to diagnosis.

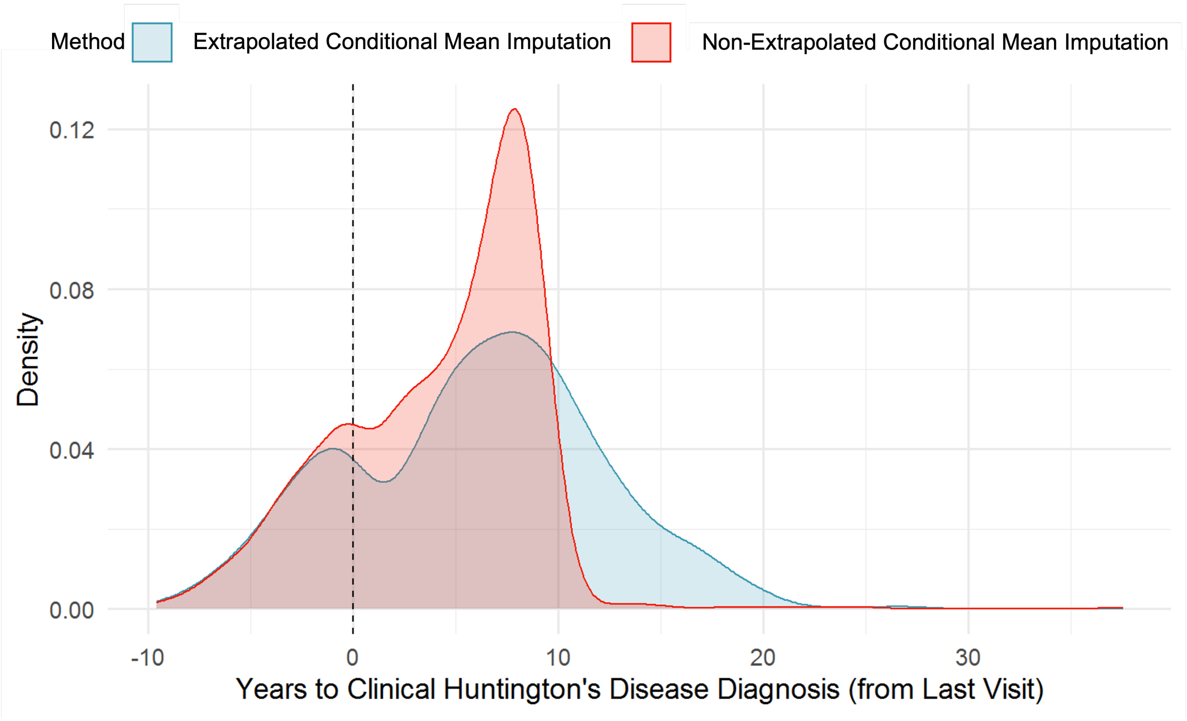

Figure S10: Histograms of observed and imputed times from last visit to Huntington's disease diagnosis in the PREDICT-HD study. The dashed line denotes the time of diagnosis.

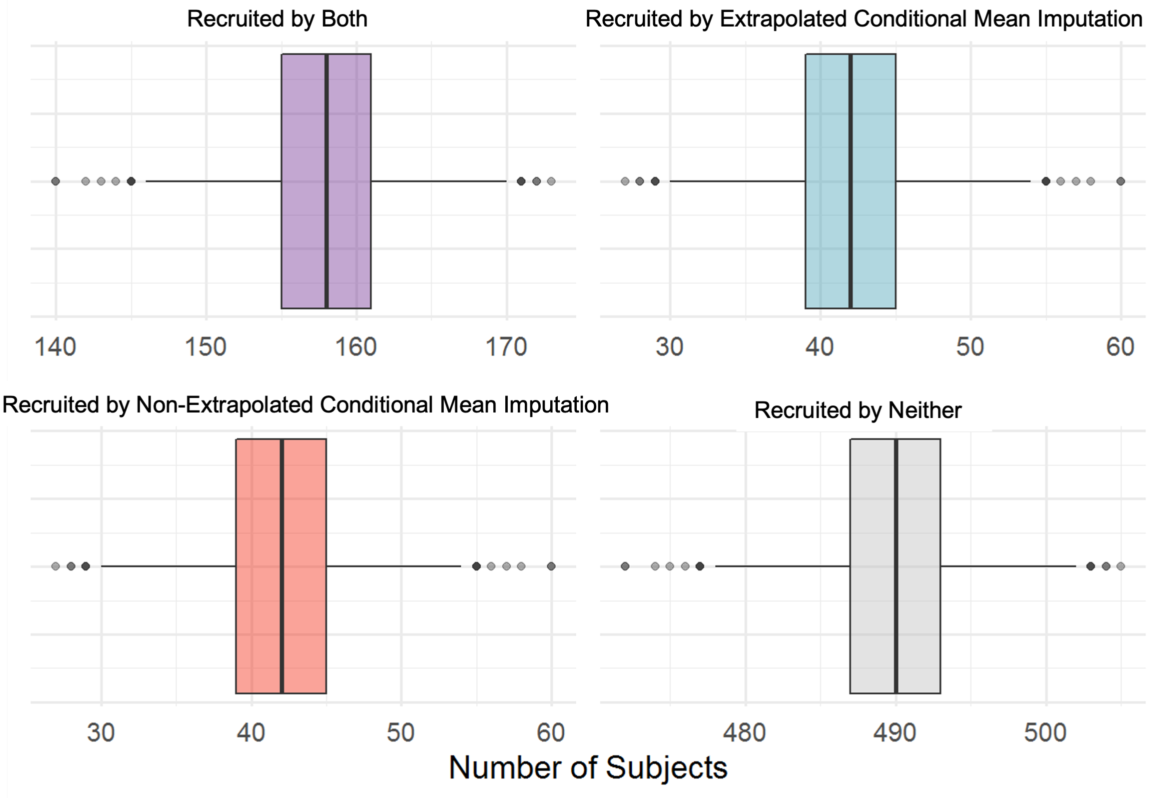

Figure S11: Statuses of $n = 732$ resampled subjects considered for recruitment into a hypothetical clinical trial based on Huntington's disease symptom progression models using the two imputation approaches in PREDICT-HD. New datasets of $n = 732$ subjects were created by resampling from censored subjects in PREDICT-HD with replacement 1000 times.

References

Brown, J. B. W., Hollander, M. & Korwar, R. M. (1974), 'Nonparametric tests of independence for censored data, with applications to heart transplant studies', In *Reliability and Biometry: Statistical Analysis of Lifelength*, Proschan, F. and Serfling, R.J., eds. Philadelphia: SIAM, pp. 327–354.

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/*

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Tierney, N., Cook, D., McBain, M. & Fay, C. (2021), *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.1.
URL: *https://CRAN.R-project.org/package=naniar*

U.S. Census Bureau (2017), '*National Population Projections*', Retrieved from https://www.census.gov/content/dam/Census/library/publications/2020/demo/p25-1145-supplemental-tables.pdf.

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., Paulsen, J. S., the PREDICT-HD Investigators & 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**(7), 751–763.