Received: 24 May 2017 | Revised: 12 February 2018 | Accepted: 5 May 2018

DOI: 10.1002/sim.7834

RESEARCH ARTICLE

WILEY Statistics in Medicine

Incorporating baseline measurements into the analysis of crossover trials with time-to-event endpoints

Rengyi Xu[1] | Devan V. Mehrotra[2] | Pamela A. Shaw[1]

[1]Department of Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, USA

[2]Biostatistics and Research Decision Sciences, Merck & Co, Inc, Philadelphia, USA

Correspondence
Rengyi Xu, Department of Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: xurengyi@pennmedicine.upenn.edu

Two-period two-treatment (2×2) crossover designs are commonly used in clinical trials. For continuous endpoints, it has been shown that baseline (pretreatment) measurements collected before the start of each treatment period can be useful in improving the power of the analysis. Methods to achieve a corresponding gain for censored time-to-event endpoints have not been adequately studied. We propose a method in which censored values are treated as missing data and multiply imputed using prespecified parametric event time models. The event times in each imputed data set are then log-transformed and analyzed using a linear model suitable for a 2×2 crossover design with continuous endpoints, with the difference in period-specific baselines included as a covariate. Results obtained from the imputed data sets are synthesized for point and confidence interval estimation of the treatment ratio of geometric mean event times using model averaging in conjunction with Rubin's combination rule. We use simulations to illustrate the favorable operating characteristics of our method relative to two other methods for crossover trials with censored time-to-event data, ie, a hierarchical rank test that ignores the baselines and a stratified Cox model that uses each study subject as a stratum and includes period-specific baselines as a covariate. Application to a real data example is provided.

KEYWORDS

baseline measurement, crossover trials, model averaging, multiple imputation, time-to-event outcome

1 | INTRODUCTION

Crossover designs are commonly seen in clinical trials to compare the treatment effects on the same subject over different treatment periods. For trials with limited recruitment, crossover designs are ideal to use for higher efficiency than parallel designs. The ability of each person to serve as his or her own control also mitigates the influence of potential confounding factors. In commonly used two-period two-treatment (2×2) crossover designs, subjects are randomized to one of two sequences, AB or BA, where A and B are the treatment labels. A "washout" period is included between the two periods to ensure no carry-over effects. The use of a period-specific baseline measurement, which is taken before the subject is given the treatment in each period, is often considered. However, whether and how to use a baseline measurement is often challenging, given the extra cost and the need to determine which statistical methods can be used to fully utilize the information from the baselines. For a 2×2 crossover trial, each subject has four responses, ie, baseline (ie, pretreatment) in period 1, posttreatment in period 1, baseline in period 2, and posttreatment in period 2. There are many existing

methods for handling baseline information in the analysis of crossover trials with *continuous* endpoints, including ignoring baseline measurements, analyzing the change from baseline, using a function of the baselines as a covariate, and joint modeling of baseline and posttreatment responses.[1-6]

Mehrotra[7] evaluated and compared 13 different methods for analyzing 2×2 crossover trials to incorporate baseline measurements with continuous endpoints. Among all the competing methods, two methods were shown to have the highest efficiency, ie, analysis of covariance (ANCOVA) with the within-subject difference in baseline responses used as a covariate and joint modeling of the within-subject difference in treatment responses and difference in baseline responses. The commonly used method, analysis of the change from baseline, was shown to have poor efficiency, as also discussed by Kenward and Roger[1] and Metcalfe.[5]

All methods aforementioned are for continuous endpoints, but crossover trials with censored time-to-event endpoints are also commonly encountered in research. For example, blood thinners like Warfarin are important in preventing outcomes such as blood clots and stroke but can also induce undesirable increases in bleeding time from simple cuts or other injuries. In this setting, researchers are sometimes interested in studying the effect of an experimental anticoagulant drug on bleeding time using a crossover design with a baseline measurement at the beginning of each period. Kimchi et al[8] and Markman et al[9] both studied a drug's effect in a crossover trial with a time-to-event outcome and collected baseline measurements. However, neither incorporated the baseline information into their analysis. Our motivating data example is a crossover trial studying a drug's effect in preventing cardiac-related symptoms during a treadmill walking test. The outcome of interest for each subject is time to a specific cardiopulmonary event, with the outcome recorded as ">10 minutes" (ie, right censored) if the event has not yet occurred after 10 minutes of observation. Existing literature for examining treatment differences in crossover trials with censored time-to-event endpoints includes both regression-based and test-based approaches. A straightforward approach is to apply McNemar's test, which classifies patients by whether they respond better to one treatment or not. However, as noted by both France et al[10] and Brittain and Follmann,[11] this approach fails to incorporate event times into the analysis and ignores patients with events in both periods or with no events in either period. France et al[10] used a conventional stratified Cox regression approach to model the hazard ratio, where each subject was treated as a stratum and the same hazard ratio parameter assumed across strata.[12] The hazard ratio between the two treatment sequences can then be estimated. Feingold and Gillespie[13] proposed an approach based on the generalized Wilcoxon test. More recently, Brittain and Follmann[11] proposed a hierarchical rank (H-R) test, which they showed to have similar or greater power than both the Feingold and Gillespie's method and stratified Cox method under certain censoring patterns. The main idea behind the H-R test is that avoiding an event is more clinically meaningful than delaying an event. Therefore, each patient is assigned a rank that orders how much better an individual does on the novel treatment. The first order of ranking is based on whether patients have an event with the most extreme ranks going to those with an event only in one period. The second order of ranking is based on the times of the events for patients with events in both periods. Patients who do not have an event on either treatment receive the same rank, namely, the average of the remaining ranks. With assigned ranks for everyone, a two-group Wilcoxon test is then performed to test for a treatment effect. However, none of these Wilcoxon-type approaches utilizes baseline information. Moreover, the target parameter of interest for the two Wilcoxon approaches is more difficult to interpret, as it is dependent on both the underlying survival and censoring distributions.

In this paper, we propose a regression-based method using multiple imputation (MI) of censored values and ANCOVA to incorporate baseline measurements into the analysis of 2×2 crossover studies with censored time-to-event response outcomes. There is often uncertainty about the true underlying survival distribution in real data applications, and misspecification of the distribution can lead to a biased point estimator and/or inefficient analysis. To mitigate this risk, we propose to fit multiple survival models in the imputation step, and use frequentist model averaging to pool the final results from the ANCOVA step. Unlike Bayesian model averaging,[14,15] which requires setting a prior probability for each candidate model, frequentist model averaging does not require any priors.[16-18] To implement model averaging in the presence of MI, we need to account for both the uncertainty from model averaging and imputation.

We show that there is often a nontrivial efficiency gain in using baseline information for time-to-event endpoints in crossover trials compared with the H-R test and stratified Cox model. Furthermore, our proposed method is also able to provide a point and confidence interval estimate of a meaningful parameter of interest (treatment ratio of the geometric mean event times). Section 2 presents details of the proposed method. In Section 3, we contrast the numerical performance of our proposed method with that of the H-R test and stratified Cox model through simulation studies. Section 4 includes results from applying the different methods to our motivating real data example. Section 5 includes conclusions.

2 | METHODS

We consider a 2×2 crossover trial with two treatments, denoted by A and B. Subjects are randomized to either the AB or BA sequence, with a wash-out period between period 1 and 2. Let X_{ijk} and Y_{ijk} denote baseline and posttreatment event

times, respectively, for subject j from sequence k in period i, where $i = 1, 2; j = 1, 2, \ldots, n$; and $k = 1, 2$. It is sufficient to assume that, after a log transformation, $(X_{1j1}, Y_{1j1}, X_{2j1}, Y_{2j1})^T$ and $(X_{1j2}, Y_{1j2}, X_{2j2}, Y_{2j2})^T$ follow a multivariate distribution with different means and same variance-covariance structure Σ. We assume there is no censoring at baseline, and in each period, subjects without a posttreatment event are censored at the end of the period, denoted by time τ.

We propose a three-step procedure using MI and ANCOVA to estimate the ratio of geometric means of the event times for treatment A relative to B, denoted as θ, and test the null hypothesis $H_0 : \theta = 1$. For distributions that are symmetric on the log scale, the geometric mean is equivalent to the median. Thus, our parameter of interest can be used to approximate the ratio of median survival of the two treatments, which is commonly of interest in survival analysis. To implement our proposed method, we perform the following steps for each imputation iteration, details of which are given in the sections as follows.

Step 1: Fit two candidate parametric event models, ie, log normal and Weibull, to impute the posttreatment censored values sequentially, conditioning on the baseline event time in period 1 for period 1 imputation, and both baseline event times and posttreatment event time in period 1 for period 2 imputation.

Step 2: With the completed data set from each candidate model, perform ANCOVA on the log-transformed event times to estimate $\log\theta$, and obtained its standard error.

Step 3: Average across the $\log\theta$ estimates based on weights associated with Akaike information criterion (AIC) from each parametric model fit to get a model averaged estimate and standard error and synthesize for overall point and confidence interval estimation across the multiply imputed data sets using Rubin's rule.

It is important to note that, although we consider only two distributions in Step 1, our method can be easily generalized to include more prespecified candidate models in the imputation step. We chose log normal and Weibull because they are very flexible and in our experience provide reasonable fitting models for capturing commonly seen event time data. Through numerical studies in Section 3, we show that even averaging over a small number of models can deliver a good performance.

2.1 | Imputation

We generate M imputed data sets for each candidate model. Let $Z_{ijk} = 0, 1$ denote treatment A and B, respectively, for subject j in period i and sequence k. We impute the censored values in period 1 first, and then impute the censored values in period 2. In the mth imputed data set, we use the baseline value in period 1 and treatment indicator, ie, Z_{1jk}, as covariates and fit two candidate parametric survival models, ie, log normal and Weibull, respectively, to Y_{1jk}. Let

$$\log Y_{1jk} = \beta_{s,0} + \beta_{s,1}Z_{1jk} + \beta_{s,2}U_{s,1jk} + \sigma_{s,1}W_{s,1jk}, \tag{1}$$

where $s = 1, 2$ denotes the log normal and Weibull model, respectively; $W_{s,1jk}$ is the error distribution; and $U_{s,1jk}$ is the baseline covariate in the sth model. $W_{1,1jk}$ has the standard normal distribution for the log-normal distribution and $W_{2,1jk}$ has the standard extreme value distribution for the Weibull distribution. $U_{1,1jk} = \log X_{1jk}$ for the log-normal distribution, and $U_{2,1jk} = X_{1jk}$ for the Weibull distribution; sample R code for implementation is provided in the Supporting Materials. Equation (1) is a representation of the log normal and accelerated failure time model framework for the Weibull model that highlights the common linear regression model on the log scale. For fitting the parametric model, we analyze the log event times for the log-normal model and fit the traditional Weibull model for the event times on the original scale. We use robust sandwich standard errors in both candidate models to correct for potential model misspecification.

Let $\hat{\boldsymbol{\beta}}_s = (\hat{\beta}_{s,0}, \hat{\beta}_{s,1}, \hat{\beta}_{s,2}, \hat{\sigma}_{s,1})^T$ and $\hat{\Sigma}_s$ denote the estimated coefficients and variance-covariance matrix in the sth candidate model, respectively. At the mth imputation step, we draw $\hat{\boldsymbol{\beta}}_s^{(m)}$ from a multivariate normal distribution $N(\hat{\boldsymbol{\beta}}_s, \hat{\Sigma}_s)$. For subject with a censored posttreatment value, we then impute a right-censored value with an uncensored value by using $\hat{\beta}_s^{(m)}$, treatment indicator Z_{1jk} and subject-specific period 1 baseline values in Equation (1), and the distribution implied by model s truncated at the observed censoring time. The corresponding complete set of uncensored posttreatment values in period 1 are denoted by $Y_{s,1jk}^{(m)}$.

Now, with complete data in period 1, we can then use the observed/imputed posttreatment values in period 1, baseline values in both period 1 and period 2 as covariates, to impute post-treatment censored values in period 2 by fitting the sth model

$$\log Y_{2jk} = \alpha_{s,0} + \alpha_{s,1}Z_{2jk} + \alpha_{s,2}U_{s,1jk} + \alpha_{s,3}V_{s,2jk} + \alpha_{s,4}R_{s,1jk}^{(m)} + \sigma_{s,2}W_{s,2jk}, \tag{2}$$

where $U_{1,1jk} = \log X_{1jk}, V_{1,2jk} = \log X_{2jk}, R^{(m)}_{s,1jk} = \log Y^{(m)}_{s,1jk}$ for log-normal distribution, $U_{1,1jk} = X_{1jk}, V_{2,2jk} = X_{2jk}, R^{(m)}_{s,1jk} = Y^{(m)}_{s,1jk}$ for Weibull distribution, and Z_{2jk} is the treatment indicator in period 2. The imputation procedure described for period 1 is now implemented using random draws from the assumed multivariate normal distribution of the vector of estimated regression coefficients in Equation (2) for each of the two parametric models. The corresponding uncensored posttreatment values in period 2 are denoted by $Y^{(m)}_{s,2jk}$.

2.2 | Analysis of covariance

After each imputation, we have two sets of complete data on every subject from the two candidate models, ie, log normal and Weibull. Each imputed data set is analyzed using ANCOVA on the log-transformed event times. Specifically, we regress the difference between posttreatment event times $\Delta^{(m)}_{s,jk} = \log Y^{(m)}_{s,1jk} - \log Y^{(m)}_{s,2jk}$, on the difference between baseline measurements $D_{jk} = \log X_{1jk} - \log X_{2jk}$ and the sequence indicator Q_j

$$\Delta^{(m)}_{s,jk} = \gamma_{s,0} + \gamma_{s,1} D_{jk} + \gamma_{s,2} Q_j + \epsilon_{s,jk}, \tag{3}$$

where $\epsilon_{s,jk} \sim N(0, \eta^2)$.

The point estimator from the sth model in the mth imputed data set is $\log \hat{\theta}^{(m)}_s = \hat{\gamma}^{(m)}_{s,2}/2$, which is the logarithm of the ratio of geometric means for treatment A relative to B. The corresponding variance estimate for $\log \hat{\theta}^{(m)}_s$ from the sth model in the mth imputed data set is $\hat{v}^{(m)}_s$.

2.3 | Model averaging and Rubin's combination rule

For overall estimation and inference, we first combine the two estimators for $\log \theta$ from the candidate models in each imputed data set, and then pool the model-averaged estimators from all the imputed data sets and obtain the pooled variance estimate that accounts for both the uncertainty from model averaging and imputation.[19]

For model averaging, we need to assign a standardized weight. There are many different options for the choice of weights, including an information criterion,[17] Mallows' criterion,[20,21] and cross-validation criterion.[22] We propose the straightforward and commonly used AIC[23] to assign weights. Let $\boldsymbol{I}_s$ denote the AIC for the ANCOVA regression, Equation (3), from the sth candidate model, and then the weight is defined as[17]

$$w_s = \frac{\exp(-\boldsymbol{I}_s/2)}{\sum_{i=1}^{2} \exp(-\boldsymbol{I}_i/2)}.$$

The model-averaged estimator for the mth imputed data set is $\log \hat{\theta}^{(m)} = \sum_{s=1}^{2} w_s \log \hat{\theta}^{(m)}_s$, and the variance for the model averaging estimator is estimated by[17]

$$\hat{\mathrm{Var}}\left(\log \hat{\theta}^{(m)}\right) = \left[\sum_{s=1}^{2} w_s \sqrt{\hat{\mathrm{Var}}\left(\log \hat{\theta}^{(m)}_s\right) + \left(\log \hat{\theta}^{(m)}_s - \log \hat{\theta}^{(m)}\right)^2}\right]^2. \tag{4}$$

Now, we can pool the model-averaged estimators across the M imputed data sets, with the final estimator calculated as[19]

$$\log \bar{\hat{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \log \hat{\theta}^{(m)}. \tag{5}$$

When there is no model averaging, we can use Rubin's method[24] to combine the results from MI. As noted earlier, with the presence of model averaging, the uncertainty from both model averaging and imputation needs to considered. The between-imputation variance is

$$v_{\mathrm{btw}} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\log \hat{\theta}^{(m)} - \log \bar{\hat{\theta}}\right)^2.$$

The within-imputation variance is the average of the estimated variance from Equation (4) across M imputed data sets

$$v_{\mathrm{within}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\mathrm{Var}}\left(\log \hat{\theta}^{(m)}\right).$$

Therefore, the total variance of the estimator after MI is[19]

$$v_{\text{total}} = \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} \left(\log \hat{\theta}^{(m)} - \log \bar{\hat{\theta}}\right)^2 + \frac{1}{M} \sum_{m=1}^{M} \left[\sum_{s=1}^{2} w_s \sqrt{\hat{\text{Var}}\left(\log \hat{\theta}_s^{(m)}\right) + \left(\log \hat{\theta}_s^{(m)} - \log \hat{\theta}^{(m)}\right)^2}\right]^2. \quad (6)$$

To test the null hypothesis $H_0 : \theta = \theta_0$ (with $\theta_0 = 1$ in our application), we carry out a t-test with test statistic $(\log \bar{\hat{\theta}} - \log \theta_0)/\sqrt{v_{\text{total}}}$. To calculate the degrees of freedom d^* for the t-test, we follow the work of Barnard and Rubin[25] so that $d^* = (1/d + 1/\hat{d}_{\text{obs}})^{-1}$, where $d = (M-1)\left[1 + \frac{v_{\text{within}}}{(1+1/M)v_{\text{btw}}}\right]^2$ and $\hat{d}_{\text{obs}} = (1 - (1+1/M)v_{\text{btw}}/v_{\text{total}}) \left(\frac{d_{\text{com}}+1}{d_{\text{com}}+3}\right) d_{\text{com}}$, and d_{com} is the degrees of freedom for $\bar{\hat{\theta}}$ when there are no missing values.

3 | SIMULATION

3.1 | Simulation set-up

To compare the performance of our proposed approach to the H-R test and stratified Cox model, we carried out a simulation study to examine type I error and power among all three methods. Since our method utilized baseline information, we also included the period-specific baseline event times, in addition to the treatment indicator, as covariates in the stratified Cox model to make a fair comparison. The H-R test, however, does not incorporate baseline information, and thus, we used the method as is. We also examined the bias and 95% confidence interval (C.I.) coverage probability from our proposed estimator; of note, the other two methods cannot deliver an estimate of our parameter of interest (θ).

We simulated three underlying distributions for event times, namely, log normal, exponential, and gamma. Two of the distributions, log normal and exponential (a special case of the Weibull), are included in the candidate models in our method, while the gamma distribution is not. The density curves for each of the three distributions are shown in Supplementary Figure S.1 in the Supporting Materials. Under the log-normal distribution, for each of the N subjects in sequence AB and BA, we generated correlated log event times from a multivariate normal distribution with mean parameter $(0, \log\theta, 0, 0)^T$ for AB sequence and $(0, 0, 0, \log\theta)^T$ for BA sequence and common variance-covariance structure with common variance 1 and correlation coefficients $\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}$, and ρ_{34}. We considered three correlation structures, ie, compound symmetry (CS), first-order autoregressive (AR(1)), and equipredictability (EP), where $\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = \rho$ for CS, $\rho_{12} = \rho_{23} = \rho_{34} = \rho, \rho_{13} = \rho_{24} = \rho^2, \rho_{14} = \rho^3$ for AR(1), and $\rho_{23} = \rho_{14}, \rho_{24} = \rho_{13}, \rho_{34} = \rho_{12}$ for EP. The correlation structures are as follows:

$$\Sigma_{\text{CS}} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix} \qquad \Sigma_{\text{AR}} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \qquad \Sigma_{\text{EP}} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{14} & \rho_{13} \\ \rho_{13} & \rho_{14} & 1 & \rho_{12} \\ \rho_{14} & \rho_{13} & \rho_{12} & 1 \end{pmatrix}.$$

The EP matrix assumes that the correlation between baseline and posttreatment in period 1 and 2 is equivalent (ρ_{12}), the correlation between baseline response in period 1 and 2 is equal to the correlation between posttreatment response in period 1 and 2 (ρ_{13}), and the correlation between baseline in period 1 and posttreatment response in period 2 is equal to that between baseline period 2 and posttreatment response in period 1 (ρ_{14}).

We assumed no censoring in baseline event times in each period, and the posttreatment event times were right-censored at time τ. As discussed in the previous section, the parameter of interest θ is the ratio of the geometric means of the event times for treatment A and treatment B, and under the log-normal distribution, it is equivalent to the ratio of median event times.

For the exponential distribution, we used copulas[26] to generate correlated event times from a multivariate exponential with mean $(2, 2\theta, 2, 2)^T$ for AB sequence and $(2, 2, 2, 2\theta)^T$ for BA sequence and common variance-covariance structure and correlation coefficients as specified. Note that the ratio of arithmetic means is equivalent to the ratio of geometric means under exponential distribution. Since copulas only preserve the rank correlation coefficient but not the linear correlation coefficient,[27] the correlated exponential data follows approximately, but not exactly, the specified variance-covariance structure.

To further illustrate the performance of our proposed method, we also considered an underlying gamma distribution, which is not included in our two candidate models from the imputation step. Specifically, we used a gamma distribution with scale of 0.7 and shape of 2 for subjects in treatment B. Event times for subjects in treatment A was generated from a

gamma distribution with scale of 0.7θ and shape of 2. We again used copulas to generate the correlated event times. For AB sequence, the simulated event times followed a multivariate gamma distribution with mean $(1.4, 1.4\theta, 1.4, 1.4)^T$, and for BA sequence, the event times follows a multivariate gamma distribution with mean $(1.4, 1.4, 1.4, 1.4\theta)^T$. Note that it can be shown that the ratio of arithmetic means is equivalent to the ratio of geometric means in this setting. Details are provided in the Supporting Materials. Again, the event times in the two sequences followed a common variance-covariance structure and correlation coefficients as specified.

We varied the sample size, percentage of censoring, θ, correlation structure, and compared the performance of the different methods. Sample size per sequence was varied as $N = 12, 24, 48$, and percentage of censoring was controlled by changing the time τ to generate 10% and 50% censoring for the total sample.

The mean pairwise correlation coefficient $\bar{\rho}$, ie, the mean of all unique off-diagonal components, took values of 0.5 and 0.7. Under CS, $\rho = \bar{\rho}$. For AR(1), $\rho = 0.7$ for $\bar{\rho} = 0.5$ and $\rho = 0.83$ for $\bar{\rho} = 0.7$. For EP, we set $\rho_{12} = 0.6, \rho_{13} = 0.5$, and $\rho_{14} = 0.4$ when $\bar{\rho} = 0.5$ and $\rho_{12} = 0.8, \rho_{13} = 0.7$, and $\rho_{14} = 0.6$ when $\bar{\rho} = 0.7$. We generated $M = 50$ imputed data sets within each of the 5000 replications. Under the null hypothesis, $\theta = 1$. Under the alternative hypothesis, we chose a

TABLE 1 Type I error (target = 5%) for the hierarchical rank (H-R) test, stratified Cox model with baseline adjustment (SCB), and proposed multiple imputation with model averaging and analysis of covariance (MI^{MA}) for log-normal, exponential, and gamma distributions under the null hypothesis $H_0 : \theta = 1$ and bias in the estimate of $\log\theta$ using the proposed method (5000 simulations)

			$\bar{\rho} = 0.5$						$\bar{\rho} = 0.7$					
			10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Distribution	Σ	Measure \ N/seq	12	24	48	12	24	48	12	24	48	12	24	48
Log-normal	CS	H-R	4.6	4.3	4.9	4.4	4.3	4.8	4.5	5.0	4.3	4.6	4.8	4.8
		SCB	4.5	4.4	4.9	NC	4.4	4.6	4.5	5.0	4.8	NC	4.5	4.9
		MI^{MA}	4.3	4.8	4.9	2.4	3.9	4.9	4.8	4.9	4.8	1.9	3.8	4.2
		Bias	-0.002	-0.002	0.000	0.001	0.002	-0.002	0.005	-0.002	-0.001	0.001	0.003	0.002
	AR(1)	H-R	5.0	4.0	4.8	5.0	4.7	4.8	4.7	4.8	5.0	4.8	4.6	4.8
		SCB	4.2	4.4	4.5	NC	4.6	4.9	3.6	4.2	5.1	NC	4.6	4.7
		MI^{MA}	4.6	4.4	4.7	2.8	3.8	4.8	4.7	4.7	5.0	2.6	4.0	4.5
		Bias	0.002	0.002	0.001	-0.007	0.000	-0.002	-0.001	-0.002	0.001	0.001	-0.002	-0.001
	EP	H-R	4.4	5.1	4.7	4.8	4.4	4.4	4.8	4.3	4.4	4.4	5.1	4.9
		SCB	NC	4.9	4.5	NC	4.7	4.6	NC	3.7	4.3	NC	NC	4.8
		MI^{MA}	4.5	5.0	5.0	2.7	4.4	4.7	4.7	4.4	4.7	1.4	3.2	3.7
		Bias	-0.001	0.001	-0.002	0.001	0.001	0.001	0.001	-0.001	0.001	0.002	-0.000	0.002
Exponential	CS	H-R	4.8	4.8	5.0	4.9	4.5	4.8	4.7	4.6	5.0	4.1	4.3	4.3
		SCB	5.1	(5.6)	4.6	NC	4.7	5.0	4.0	4.7	5.3	NC	4.4	5.0
		MI^{MA}	4.7	5.0	4.4	2.2	3.8	5.1	4.4	4.8	4.5	1.8	2.9	4.0
		Bias	-0.003	-0.002	0.001	-0.071	-0.006	0.002	0.001	-0.001	0.002	0.063	0.002	-0.002
	AR(1)	H-R	4.6	4.6	4.6	4.6	5.0	4.4	4.3	5.0	5.1	4.5	4.5	4.8
		SCB	NC	4.5	5.2	NC	4.7	4.6	NC	4.8	5.2	NC	NC	4.7
		MI^{MA}	4.9	4.7	4.2	2.0	3.5	4.2	4.4	4.5	4.5	1.9	2.9	3.0
		Bias	-0.001	0.004	0.002	0.002	0.001	0.002	-0.001	-0.002	0.002	0.004	-0.001	-0.003
	EP	H-R	4.2	4.3	4.4	4.5	4.2	4.4	4.8	4.7	4.9	4.4	4.5	4.9
		SCB	NC	4.4	4.8	NC	4.3	4.5	NC	NC	4.2	NC	NC	4.2
		MI^{MA}	4.4	4.3	4.5	1.8	3.3	4.3	4.4	4.6	4.6	1.4	2.1	2.4
		Bias	-0.004	0.001	0.001	0.003	0.003	-0.001	0.002	-0.001	0.001	-0.015	0.001	0.001
Gamma	CS	H-R	4.2	4.4	4.7	4.4	4.4	4.6	4.2	4.7	4.4	4.1	4.8	5.1
		SCB	4.2	4.9	4.6	NC	4.6	4.8	NC	(5.9)	4.7	NC	4.9	(5.6)
		MI^{MA}	4.5	4.7	4.9	4.2	4.7	5.0	4.7	5.0	4.8	3.2	4.4	4.8
		Bias	0.001	0.002	-0.002	0.001	-0.000	0.001	0.002	0.002	-0.000	-0.002	-0.003	0.002
	AR(1)	H-R	4.3	4.6	4.4	4.7	5.1	4.5	4.4	4.2	4.3	4.3	4.7	4.8
		SCB	3.7	5.1	4.8	NC	4.1	4.5	NC	4.6	4.1	NC	4.1	5.1
		MI^{MA}	4.7	5.1	4.6	3.8	4.6	4.7	4.9	4.7	4.1	3.3	4.6	4.6
		Bias	-0.000	-0.000	-0.001	-0.006	0.001	0.000	-0.001	0.001	0.001	0.001	0.000	0.001
	EP	H-R	4.9	4.6	5.2	4.7	4.5	4.6	4.7	4.2	4.7	4.6	5.1	4.4
		SCB	4.3	5.0	5.2	NC	4.4	4.5	NC	3.8	5.0	NC	3.7	4.8
		MI^{MA}	4.7	4.3	4.9	3.5	4.7	5.0	4.8	4.7	4.5	3.0	3.9	4.0
		Bias	0.000	0.001	0.001	0.003	-0.000	-0.003	-0.000	-0.001	-0.001	0.001	-0.000	-0.001

Note: $\bar{\rho}$: mean pairwise correlation. Type I error more than $Z_{0.975}$ standard errors above 5% level is in parentheses. Abbreviations: AR(1), first-order autoregressive covariance structure; CS, compound symmetry covariance structure; EP, equipredicability covariance structure; NC, nonconvergence.

value of θ such that the power was about 80% for the H-R test, given the true underlying distribution, Σ, $\bar{\rho}$, and percentage censoring.

3.2 | Simulation results

Table 1 reports type I error for the three distributions for the H-R test, stratified Cox model with baseline adjustment, and our proposed MI and model averaging and ANCOVA method. As shown in Table 1, the stratified Cox model analysis had nonconvergence (NC) issues under several scenarios when the sample size was 12 and 24 subjects per sequence with 50% censoring and had an inflated type I error when there were 24 subjects per sequence with 10% censoring, $\bar{\rho} = 0.5$ and CS structure under exponential distribution. When the true distribution was gamma, the stratified Cox model analysis was associated with inflated type I error under CS structure with 24 subjects per sequence and $\bar{\rho} = 0.7$ and 10% censoring and with 48 subjects per sequence and $\bar{\rho} = 0.7$ and 50% censoring. The H-R test and our proposed model averaging method controlled type I error throughout all the scenarios considered. Table 1 also reports the bias in the estimate of $\log \theta$ using our proposed method under the null hypothesis. The bias was negligible under all simulated scenarios.

Figures 1–3 show the power for the three different methods for $N = 24$ subjects per sequence and different combinations of percentage censoring and variance-covariance structure under the log-normal, exponential, and gamma distributions, respectively; results for other sample sizes are provided in the Supporting Materials.

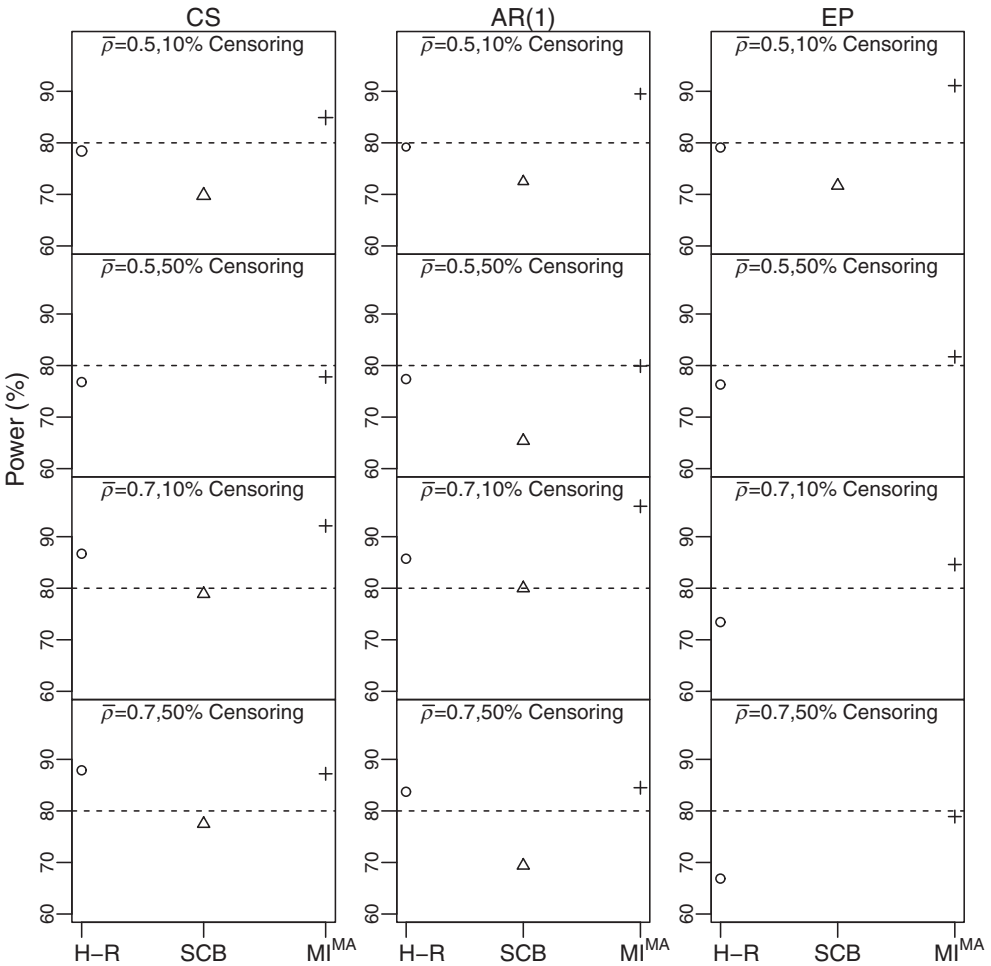

FIGURE 1 Power comparison for the hierarchical rank (H-R) test, stratified Cox model (SCB), and proposed multiple imputation with model averaging and analysis of covariance (MI^{MA}) under a log-normal distribution and varying assumptions for the true variance structure (compound symmetry (CS), first-order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline, and posttreatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%, 50%)), with 24 subjects per sequence. Stratified Cox model had nonconvergence issues under CS structure with $\bar{\rho} = 0.5$ and 50% censoring, and under EP structure with $\bar{\rho} = 0.5$ and 50% censoring, $\bar{\rho} = 0.7$ and 10% censoring, and $\bar{\rho} = 0.7$ and 50% censoring, and hence power is not reported

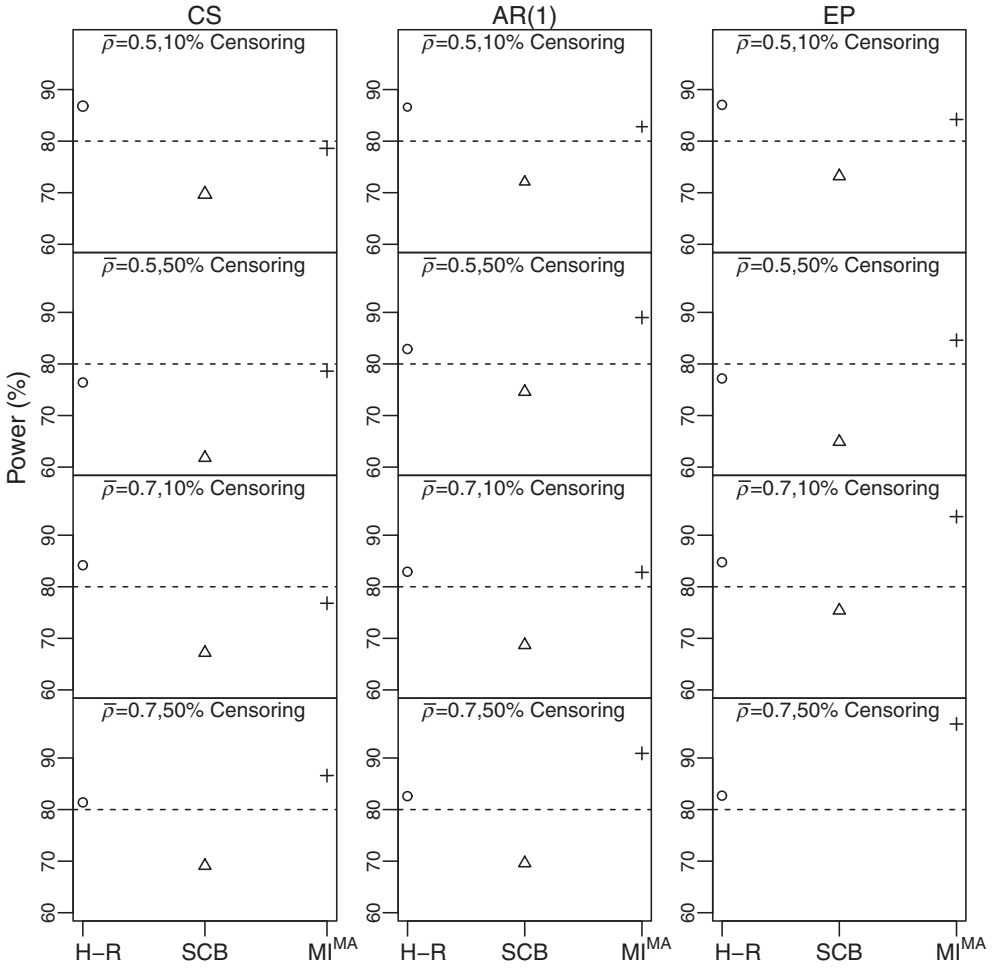

FIGURE 2 Power comparison for the hierarchical rank test (H-R), stratified Cox model (SCB), and proposed multiple imputation with model averaging and analysis of covariance (MI^{MA}) under an exponential distribution and varying assumptions for the true variance structure (compound symmetry (CS), first-order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline, and posttreatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%, 50%)), with 24 subjects per sequence. Stratified Cox model had nonconvergence issues under EP structure with $\bar{\rho} = 0.7$ and 50% censoring, and hence power is not reported

As shown in Figure 1, when the true distribution was log normal, our proposed method always provided a higher or similar power than the H-R test and stratified Cox model. For cases where the H-R test or stratified Cox failed to deliver 80% power, our method was able to achieve power close to or above 80%. The increase in power using our method was more significant under AR(1) and EP structures than under CS structure. The power gain compared with the H-R test likely comes from the fact that the H-R test fails to utilize baseline information. Likewise, our proposed method has a substantially higher power than the stratified Cox model that adjusts for baseline covariates in part because our method makes better use of the baseline information. In addition, the model averaging aspect provides the flexibility of assuming more than one distribution and further improves the efficiency of the analysis. Results from assuming only one distribution, either log normal or Weibull, in the imputation step is more prone to model misspecification in the imputation step.

Figure 2 displays the results when the true distribution was exponential. In this case, the true variance-covariance structure and percentage censoring affected the relative performance of the considered methods. When the true structure was CS, H-R test delivered higher power than the other considered methods. Of note, CS structure usually does not capture the true correlation pattern in most real data examples, since it assumes equal correlation among all pairs of with-subject event times, which has low plausibility. When the true structure was AR(1) or EP, which are a more realistic representation of the correlation structure in real data applications, our method again showed a substantial power gain compared with the H-R test and stratified Cox model under 50% censoring. When the percentage censoring was 10%, our

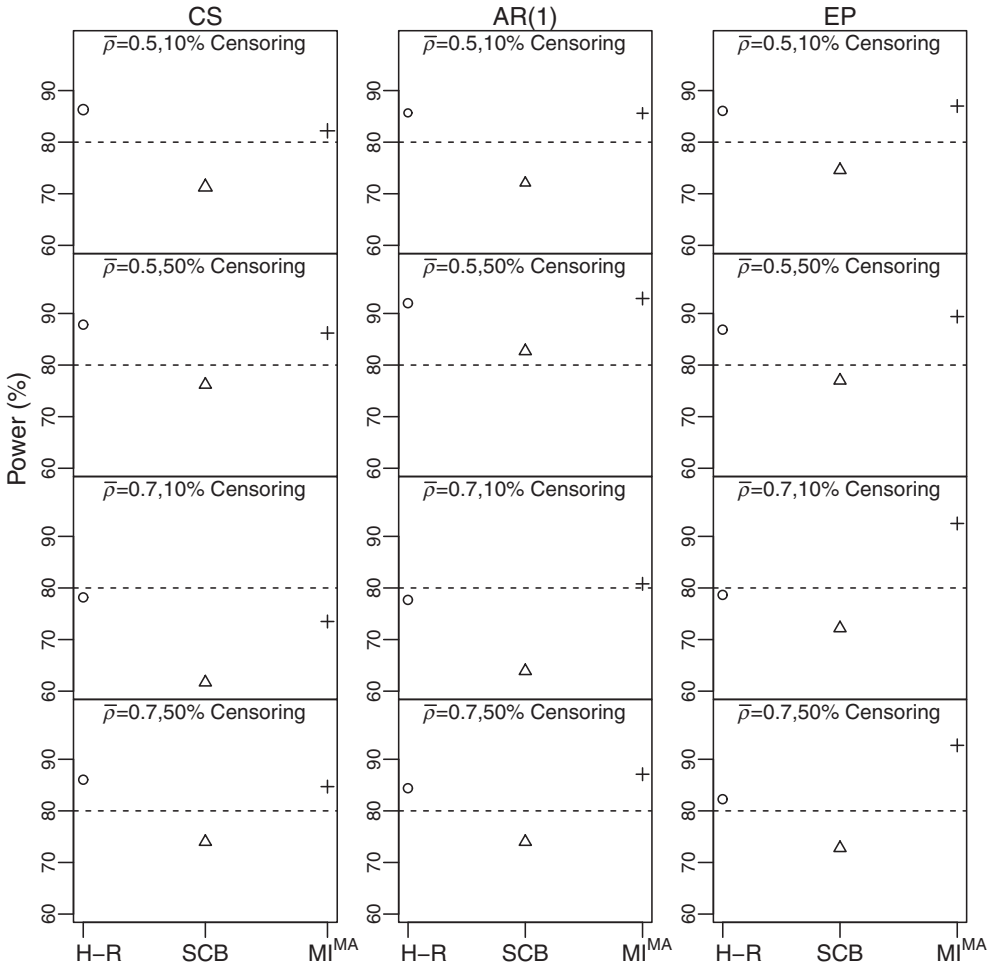

FIGURE 3 Power comparison for the hierarchical rank test (H-R), stratified Cox model (SCB), and proposed multiple imputation with model averaging and analysis of covariance (MI^{MA}) under a gamma distribution and varying assumptions for the true variance structure (compound symmetry (CS), first-order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline, and posttreatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%, 50%)), with 24 subjects per sequence

method delivered similar power as the H-R test. For all the other scenarios, where the stratified Cox model did not have NC issues, our proposed method was consistently more powerful than the stratified Cox model.

Finally, when the underlying distribution was gamma, our proposed method still provided higher power than the stratified Cox model throughout all scenarios, but slightly lower power than the H-R test under CS structures, as shown in Figure 3. Under AR and EP structures, using MI, model averaging and ANCOVA approaches delivered a more efficient analysis than both the H-R test and stratified Cox model. Recall that the true distribution, gamma, is not included as one of the candidate models in the imputation step; however, we are still able to provide a comparably efficient result. Additionally, our proposed method is able to provide a point and CI estimate of the treatment effect, while the other two methods do not.

Table 2 reports percentage bias and 95% C.I. coverage probability for $\log\theta$ using our proposed method under the alternative hypothesis. Our method was able to control bias within 10% under log-normal and exponential distribution. When the true distribution was gamma, it controlled bias within 10% under 10% censoring, and under 50% censoring, bias was no larger than 11%. Importantly, the 95% C.I. coverage probability was maintained at or above the nominal level under all the scenarios considered.

4 | DATA APPLICATION

We apply the three methods considered to a 2×2 crossover clinical trial of an investigation drug. The trial recruited 40 subjects in total and randomly assigned 20 to the placebo then drug sequence and 20 to the drug then placebo sequence.

TABLE 2 Percentage bias and 95% confidence interval coverage probability under the alternative hypothesis $H_1 : \theta \neq 1$ for the estimate of $\log\theta$ using the proposed method under log-normal, exponential, and gamma distributions (5000 simulations)

			$\bar{\rho} = 0.5$						$\bar{\rho} = 0.7$					
			10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Distribution	Σ	Measure \ N/seq	12	24	48	12	24	48	12	24	48	12	24	48
Log-normal	CS	%Bias	-4.9	-4.7	-3.2	-8.6	-7.9	-7.2	-4.1	-3.3	-4.0	-8.0	-8.3	-9.5
		Coverage	95.3	94.0	94.3	95.6	94.4	94.4	95.3	94.8	95.1	96.6	95.0	94.9
	AR(1)	%Bias	-4.5	-4.0	-4.1	-9.4	-8.2	-8.5	-2.2	-2.3	-2.5	-5.5	-6.1	-5.5
		Coverage	95.2	94.9	94.6	95.4	94.7	94.4	95.2	95.3	94.5	96.8	95.8	95.4
	EP	% Bias	-4.6	-4.8	-3.5	-9.3	-8.1	-7.3	-2.0	-5.0	-2.2	-5.5	-5.1	-5.7
		Coverage	95.1	94.8	94.7	96.1	94.4	94.4	95.8	96.5	95.3	97.8	96.8	96.0
Exponential	CS	%Bias	2.1	1.1	-0.1	1.4	0.8	0.8	1.7	1.3	0.9	2.4	2.7	3.3
		Coverage	95.0	95.2	94.7	97.0	95.3	95.1	95.0	95.3	94.2	96.6	96.2	95.6
	AR(1)	%Bias	0.6	1.8	0.4	2.2	3.1	2.9	1.4	1.2	1.3	3.9	5.9	5.7
		Coverage	94.8	94.9	95.2	96.3	95.7	95.4	94.9	95.3	95.1	97.1	96.3	95.7
	EP	%Bias	1.9	0.9	-0.1	1.5	1.8	1.4	1.8	1.7	2.4	3.6	4.1	4.1
		Coverage	95.3	95.2	94.8	96.6	95.1	95.0	95.1	95.2	95.0	97.5	97.0	97.1
Gamma	CS	%Bias	-5.4	-3.8	-3.6	-8.5	-3.3	-4.2	-3.9	-3.8	-3.0	-11.5	-5.1	-3.8
		Coverage	95.0	94.9	95.0	94.9	94.7	94.6	95.0	95.6	95.3	94.8	95.5	94.8
	AR(1)	%Bias	-5.2	-3.9	-3.3	-7.0	-4.8	-4.6	-3.7	-2.6	-1.7	-10.4	-10.6	-10.0
		Coverage	95.0	94.6	95.0	95.5	94.5	94.7	95.2	94.9	95.4	95.1	94.6	94.4
	EP	%Bias	-5.5	-3.9	-2.8	-7.4	-4.0	-4.4	-2.9	-2.7	-1.9	-9.9	-9.7	-7.9
		Coverage	94.8	94.9	95.3	95.1	94.7	94.6	95.3	94.7	94.7	95.3	94.5	95.4

Note: $\bar{\rho}$: mean pairwise correlation. True values of θ used for all the simulated scenarios are provided in Table S1 in the Supporting Materials. Abbreviations: AR(1), first-order autoregressive covariance structure; CS, compound symmetry covariance structure; EP, equipredicability covariance structure.

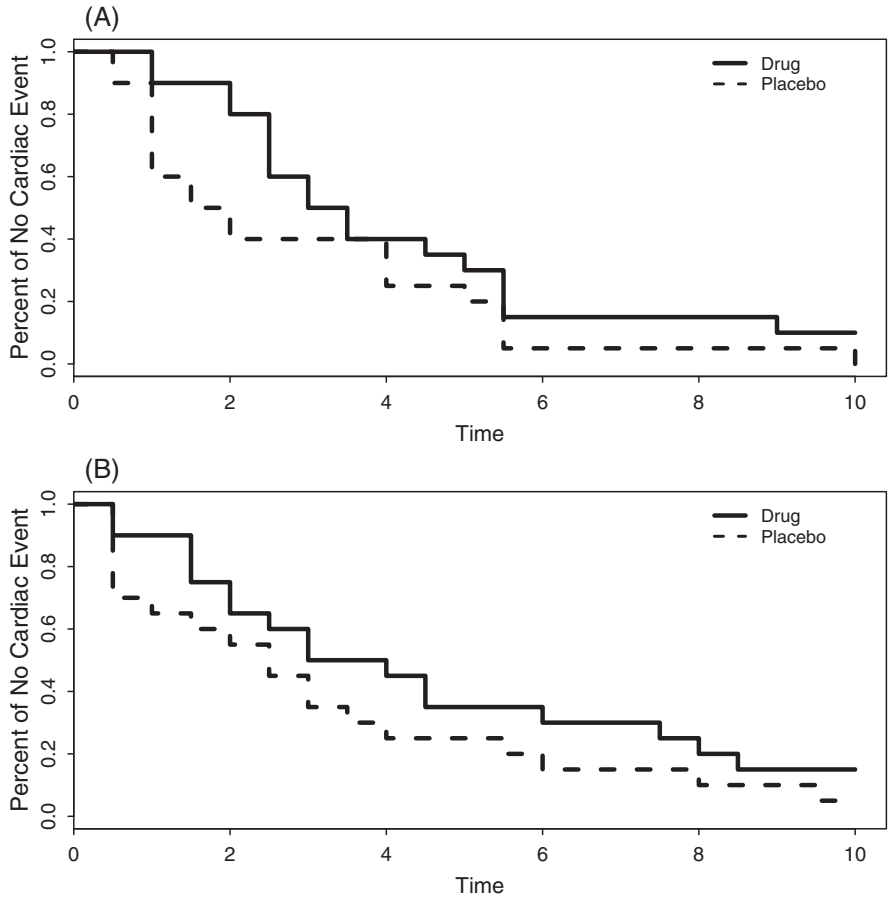

FIGURE 4 Kaplan-Meier curves for the time to a symptomatic cardiac-related event by treatment group from a 2×2 crossover trial; (A) is for period 1 and (B) is for period 2

TABLE 3 Event times (minutes) for a 10-minute treadmill test in a 2×2 crossover clinical trial

	Placebo-drug sequence					Drug-placebo sequence			
	Period 1 (placebo)		Period 2 (drug)			Period 1 (drug)		Period 2 (placebo)	
Subject	X_1	Y_1	X_2	Y_2	Subject	X_1	Y_1	X_2	Y_2
1	1.5	1	1	1.5	2	1	1	1	2.5
3	6	4	3.5	>10	4	6	>10	2.5	2.5
5	1	1	1.5	4.5	6	3	2	1	0.5
7	3.5	1.5	0.5	3	8	2.5	2.5	1.5	2
9	0.5	1	3.5	8	10	2	2.5	2.5	3
11	6	10	6	>10	12	1.5	4.5	2.5	1
13	0.5	0.5	1	>10	14	3.5	5.5	4.5	9.5
15	1	1	1	2.5	16	1	2	2	>10
17	1.5	1	0.5	0.5	18	6	>10	5	3.5
19	1	1.5	2	4	20	2	3	1.5	1.5
21	5	5.5	3	1.5	22	1.5	2.5	1.5	0.5
23	2.5	5	6	4.5	24	1.5	3.5	2.5	3
25	5	5.5	4.5	6	26	3.5	9	6	6
27	1	2	2.5	8.5	28	2	5.5	3.5	8
29	5	5.5	3.5	2	30	2.5	2.5	1	0.5
31	0.5	1	2	7.5	32	2.5	3.5	2.5	4
33	5	4	2	2	34	5.5	3	1	0.5
35	0.5	0.5	1	1.5	36	3	5.5	5	0.5
37	1.5	2	3	3	38	0.5	1	1	5.5
39	6	4	1.5	0.5	40	2.5	5	2.5	0.5
Median	1.5	1.75	2	3.5	Median	2.5	3.25	2.5	2.5

Note: X_1: baseline response in period 1. Y_1: post-treatment response in period 1. X_2: baseline response in period 2. Y_2: post-treatment response in period 2.

The outcome variable was time until a symptomatic cardiac-related event of interest during a 10-minute treadmill walking test. Each subject also had a measurement at baseline before taking the treatment. Figure 4 displays the Kaplan-Meier curves for posttreatment event times for placebo and drug in period 1 and period 2, separately.

The H-R test delivers a p-value of 0.052, indicating that there is not enough evidence at the two-tailed 5% level of significance to show a difference between the drug and placebo in delaying the event of interest. On the other hand, stratified Cox model adjusting for period-specific baseline and our proposed method deliver a p-value of 0.020 and 0.005, respectively. The ratio of geometric mean of time to the cardiac-related event for patients taking the drug to patients on placebo was estimated to be 1.67, with 95% C.I. of (1.18, 2.35). The raw data from this trial is provided in Table 3, and R code used to generate the analysis results for all the three methods is provided in the Supporting Materials. We also provided histograms of the imputed values for each of the 6 subjects with a censored time point from the log-normal and Weibull imputation models in the Supporting Materials (Figure S.2).

5 | DISCUSSION

While there are many methods for analyzing crossover trials with continuous endpoints, there are few studying crossover trials with censored time-to-event outcomes, which are often seen in practice. In this paper, we have proposed a method using MI, assuming two candidate parametric event time models, to impute censored posttreatment values. For each imputed data set, ANCOVA, with difference in period-specific baseline responses as a covariate, is applied to log-transformed event times to estimate the log treatment ratio of geometric means. Frequentist model averaging with AIC weighting in conjunction with Rubin's combination rule for MI is used for overall estimation and inference. We showed that, by utilizing baseline information, our method provided a more or as efficient result than some other existing methods, including H-R test and stratified Cox model, across different combinations of variance-covariance structures,

percentage censoring, and sample sizes. By using model averaging, we are able to provide a more flexible method than assuming only one distribution in the imputation step, which can be subject to misspecification of the true underlying distribution. Furthermore, the H-R approach does not provide a point estimator, and the underlying target parameter relies on the censoring distributions. Our regression-based method delivers an estimated ratio of geometric means of event times for one treatment relative to the other with small or no bias and adequate 95% C.I. coverage. The ratio of geometric means is a useful parameter in that it is equivalent to the ratio of median event times under a log-normal distribution and other distributions that are symmetric on the log scale.

The proposed method was motivated by clinical settings with repeatably observable time to events, where censored events were the minority. In other settings where censoring rates are high, the method might not be entirely appropriate. For our model averaging approach, we only used two candidate models, log normal and Weibull, to impute censored posttreatment values. More distributions can readily be used. The candidate distributions should include those that cover a spectrum of anticipated plausible shapes of the survival distribution for the outcome of interest. The relative success of our method, like other applications of MI, is not expected to be strong if the imputation model is grossly misspecified. We showed that using two candidate models provided efficient results with little bias for the settings considered, and thus, more candidate models could potentially improve these results. Lastly, as alluded to earlier, our research was motivated by applications where baseline values are not censored; for applications where that is not always the case, an extension of our method can be considered, but that requires further research and development.

ORCID

Rengyi Xu http://orcid.org/0000-0003-4135-668X
Devan V. Mehrotra http://orcid.org/0000-0002-0316-7362

REFERENCES

1. Kenward MG, Roger JH. The use of baseline covariates in crossover studies. *Biostatistics*. 2010;11(1):1-17.
2. Senn SJ. *Cross-over Trials in Clinical Research*. England, UK: John Wiley & Sons Chichester; 2002.
3. Chen X, Meng Z, Zhang J. Handling of baseline measurements in the analysis of crossover trials. *Stat Med*. 2012;31(17):1791-1803.
4. Yan Z. The impact of baseline covariates on the efficiency of statistical analyses of crossover designs. *Stat Med*. 2013;32(6):956-963.
5. Metcalfe C. The analysis of cross-over trials with baseline measurements. *Stat Med*. 2010;29(30):3211-3218.
6. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol*. 1979;8(1):7-20.
7. Mehrotra DV. A recommended analysis for 2 × 2 crossover trials with baseline measurements. *Pharm Stat*. 2014;13(6):376-387.
8. Kimchi A, Lee G, Amsterdam E, Fujii K, Krieg P, Mason DT. Increased exercise tolerance after nitroglycerin oral spray: a new and effective therapeutic modality in angina pectoris. *Circulation*. 1983;67(1):124-127.
9. Markman JD, Frazer ME, Rast SA, et al. Double-blind, randomized, controlled, crossover trial of pregabalin for neurogenic claudication. *Neurology*. 2015;84(3):265-272.
10. France LA, Lewis JA, Kay R. The analysis of failure time data in crossover studies. *Stat Med*. 1991;10(7):1099-1113.
11. Brittain E, Follmann D. A hierarchical rank test for crossover trials with censored data. *Stat Med*. 2011;30(30):3507-3519.
12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons; 2002.
13. Feingold M, Gillespie BW. Cross-over trials with censored data. *Stat Med*. 1996;15(10):953-967.
14. Bates JM, Granger CWJ. The combination of forecasts. *J Oper Res Soc*. 1969;20(4):451-468.
15. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc*. 1997;92(437):179-191.
16. Hjort NL, Claeskens G. Frequentist model average estimators. *J Am Stat Assoc*. 2003;98(464):879-899.
17. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics*. 1997;53(2):603-618.
18. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer-Verlag New York; 2003.
19. Schomaker M, Heumann C. Model selection and model averaging after multiple imputation. *Comput Stat Data Anal*. 2014;71:758-770.
20. Hansen BE. Least squares model averaging. *Econometrica*. 2007;75(4):1175-1189.
21. Mallows CL. Some comments on c p. *Technometrics*. 1973;15(4):661-675.
22. Hansen BE, Racine JS. Jackknife model averaging. *J Econ*. 2012;167(1):38-46.
23. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723.
24. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons; 1987.
25. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika*. 1999;86(4):948-955.

26. Sklar A. Random variables, joint distribution functions, and copulas. *Kybernetika*. 1973;9(6):449-460.
27. Genest C, MacKay J. The joy of copulas: bivariate distributions with uniform marginals. *Am Stat*. 1986;40(4):280-283.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Xu R, Mehrotra DV, Shaw PA. Incorporating baseline measurements into the analysis of crossover trials with time-to-event endpoints. *Statistics in Medicine*. 2018;37:3280–3292. https://doi.org/10.1002/sim.7834