Published in final edited form as:
Stat Med. 2011 August 15; 30(18): 2295–2309. doi:10.1002/sim.4263.

Sample size and power determination in joint modeling of longitudinal and survival data

Liddy M. Chen[a], **Joseph G. Ibrahim**[a,*,†], and **Haitao Chu**[b]

[a]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A

[b]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, U.S.A

Abstract

Owing to the rapid development of biomarkers in clinical trials, joint modeling of longitudinal and survival data has gained its popularity in the recent years because it reduces bias and provides improvements of efficiency in the assessment of treatment effects and other prognostic factors. Although much effort has been put into inferential methods in joint modeling, such as estimation and hypothesis testing, design aspects have not been formally considered. Statistical design, such as sample size and power calculations, is a crucial first step in clinical trials. In this paper, we derive a closed-form sample size formula for estimating the effect of the longitudinal process in joint modeling, and extend Schoenfeld's sample size formula to the joint modeling setting for estimating the overall treatment effect. The sample size formula we develop is quite general, allowing for *p*-degree polynomial trajectories. The robustness of our model is demonstrated in simulation studies with linear and quadratic trajectories. We discuss the impact of the within-subject variability on power and data collection strategies, such as spacing and frequency of repeated measurements, in order to maximize the power. When the within-subject variability is large, different data collection strategies can influence the power of the study in a significant way. Optimal frequency of repeated measurements also depends on the nature of the trajectory with higher polynomial trajectories and larger measurement error requiring more frequent measurements.

Keywords

sample size; power determination; joint modeling; survival analysis; longitudinal data; repeated measurements

1. Introduction

Censored time-to-event data, such as the time to failure or time to death, is a common primary endpoint in many clinical trials. Many studies also collect longitudinal data with repeated measurements at a number of time points prior to the event, along with other baseline covariates. One of the most original examples was an HIV trial that compared time to virologic failure or time to progression to AIDS [1, 2]. CD4 cell counts were considered as a strong indicator of a treatment effect and are usually measured at each visit as secondary efficacy endpoints. Although CD4 cell counts are no longer considered a valid surrogate for time to progression to AIDS in the current literature, the joint modeling

*Correspondence to: Joseph G. Ibrahim, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A.
†ibrahim@bios.unc.edu

strategies originally developed for these trials led to research on joint modeling in other research areas. As discoveries of biomarkers advance, more oncology studies collect repeated measurements of biomarker data, such as prostate-specific antigen (PSA) in prostate cancer trials, as secondary efficacy measurements [3]. Many studies also measure quality of life (QOL) or depression measures together with survival data, where joint models can also be applied [4–9]. Most clinical trials are designed to address the treatment effect on a time-to-event endpoint. Recently, there has been an increasing interest in focusing on two primary endpoints such as time-to-event and a longitudinal marker, and also to characterize the relationship between them. For example, if treatment has an effect on the longitudinal marker and the longitudinal marker has a strong association with the time-to-event, the longitudinal marker can potentially be used as a surrogate endpoint or as a marker for the time-to-event, which is usually lengthy to ascertain in practice. The issue of surrogacy of a disease marker for the survival endpoint by joint modeling was discussed by Taylor and Wang [10].

Characterizing the association between time-to-event and the longitudinal process is usually complicated due to incomplete or mis-measured longitudinal data [1, 2, 11]. Another issue is that the occurrence of the time-to-event may induce informative censoring of the longitudinal process [11, 12]. The recently developed joint modeling approaches are frameworks which acknowledge the intrinsic relationships between the event time and the longitudinal process by incorporating a trajectory for the longitudinal process into the hazard function of the event, or in a more general sense, introducing shared random effects in both the longitudinal model and the survival model [2, 7, 13–16]. Bayesian approaches that address joint modeling of longitudinal and survival data was introduced by Ibrahim *et al.* [4], Chen *et al.* [17], Brown and Ibrahim [18], Ibrahim *et al.* [19], and Chi and Ibrahim [8, 9]. It has been demonstrated through simulation studies that the use of joint modeling leads to correction of biases and improvement of efficiency when estimating the association between the event time and the longitudinal process [20]. A thorough review on joint modeling is given by Tsiatis and Davidian [11]. Further generalizations to multiple time-dependent covariates was introduced by Song *et al.* [21], and a full likelihood approach for joint modeling of a bivariate growth curve from two longitudinal measures and event time was introduced by Dang *et al.* [22].

Design is a crucial first step in clinical trials. Well-designed studies are essential for a successful research and drug development. Although much effort has been put into inferential and estimation methods in joint modeling of longitudinal and survival data, design issues have not been formally considered. Hence, developing statistical methods to address design issues in joint modeling is much needed. One of the fundamental issues is power and sample size calculations for joint models. In this paper, we will describe some basics of joint modeling in Section 2, and then provide a sample size formula associating the longitudinal process and the event time for study design based on a joint modeling in Section 3. In Section 4, we provide a detailed methodology to determine the sample size and power with an unknown variance–covariance matrix, discuss longitudinal data collection strategies, such as spacing and frequency of repeated measurements, to maximize the power. In Sections 5 and 6, we provide a sample size formula to investigate treatment effects in joint models, and discuss how ignoring the longitudinal process would lead to biased estimates of the treatment effect and a potential loss of power. In Section 7, we briefly compare the two-step inferential approach and the full joint modeling approach for the ECOG trial E1193, and show that the sample size formulas we develop are quite robust. We end this paper with some discussions in Section 8.

2. Preliminaries

For subject i, ($i = 1, \ldots, N$), let T_i and C_i denote the event and censoring times, respectively; $S_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \le C_i)$. Let Z_i be a treatment indicator, and let $X_i(u)$ be the longitudinal process (also referred to as the trajectory) at time $u \ge 0$. In a more general sense, Z_i can be a q-dimensional vector of baseline covariates including treatment. To simplify the notation, Z_i denotes the treatment indicator in this paper. Values of $X_i(u)$ are measured intermittently at times $u \le S_i$, $j = 1, \ldots, m_i$, for subject i. Let $Y(t_{ij})$ denote the observed value of $X_i(t_{ij})$ at time t_{ij}, which may be prone to measurement error.

The joint modeling approach links two sub-models, one for the longitudinal process $X_i(u)$ and one for the event time T_i, by including the trajectory in the hazard function of T_i. Thus,

$$\lambda_i(t)=\lambda_0(t)\exp\{\beta X_i(t)+\alpha Z_i\}. \tag{1}$$

Although other models for $X_i(u)$ have been proposed [7, 13, 14], we focus on a general polynomial model [17, 19]

$$X_i(u)=\theta_{0i}+\theta_{1i}u+\theta_{2i}u^2+\cdots+\theta_{pi}u^p+\gamma Z_i, \tag{2}$$

where $\boldsymbol{\theta_i} = \{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^{\mathrm{T}}$ is distributed as a multivariate normal distribution with mean $\boldsymbol{\mu_\theta}$ and variance–covariance matrix $\boldsymbol{\Sigma_\theta}$. The parameter γ is a fixed treatment effect. The observed longitudinal measures are modeled as $Y_i(t_{ij}) = X_i(t_{ij}) + e_{ij}$, where $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$, the $\boldsymbol{\theta_i}$'s are independent and $\mathrm{Cov}(e_{ij}, e_{ij'}) = 0$, for $j \ne j'$. The observed data likelihood for subject i is given by:

$$\int_{-\infty}^{\infty}\left[\prod_{j=1}^{m_j} f(Y_{ij}|\theta_i,\gamma,\sigma_e^2)\right] f(\theta_i|\mu_\theta,\textstyle\sum_\theta) f(S_i,\Delta_i|\boldsymbol{\theta_i},\beta,\gamma,\alpha)\mathrm{d}\boldsymbol{\theta_i}. \tag{3}$$

In expression (3), $f(Y_{ij}|\theta_i,\gamma,\sigma_e^2)$ is a univariate normal density function with mean $\theta_{0i}+\theta_{1i}t_{ij}+\theta_{2i}t_{ij}^2+\cdots+\theta_{pi}t_{ij}^p+\gamma Z_i$ and variance σ_e^2, and $f(\gamma_i|\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$ is the multivariate normal density with mean $\boldsymbol{\mu_\theta}$ and covariance matrix $\boldsymbol{\Sigma_\theta}$. The density function for the time-to-event, $f(S_i, \Delta_i|\boldsymbol{\theta_i}, \beta, \gamma, \alpha)$, can be based on any model. In this paper, we focus on the exponential model, where $f(S_i,\Delta_i|\boldsymbol{\theta_i},\beta,\gamma,\alpha)=\{\lambda_0\exp[\beta X(S_i)+\alpha Z_i]\}^{\Delta_i}\exp[-\int_0^{S_i}\lambda_0\exp[\beta X(t)+\alpha Z_i]\mathrm{d}t]$.

The primary objectives here are:

- **a.** To test the effect of the longitudinal process (H_0: $\beta = 0$) by the score statistic.
- **b.** To test the overall treatment effect (H_0: $\beta\gamma + \alpha = 0$) by the score statistic.

When the trajectory, $X_i(t)$ is known, the score statistic can be derived directly based on the partial likelihood given by Cox [23], namely

$$\prod_{i=1}^{N}\left\{\frac{\exp\{\beta X_i(S_i)+\alpha Z_i\}}{\sum_{k=1}^{N} I(S_k \le S_i)\exp\{\beta X_k(S_i)+\alpha Z_k\}}\right\}^{\Delta_i}. \tag{4}$$

When the trajectory is unknown, the observed hazard is $\lambda(t|\bar{Y}(t))$ instead of $\lambda(t|\bar{X}(t))$, where $\bar{Y}(t)$ denotes the observed history up to time t, and $\bar{X}(t)$ denotes the hypothetical true history up to time t. By the law of conditional probability, and assuming that neither the measurement error nor the timing of the visits prior to time t are prognostic, $\lambda(t|\bar{Y}(t)) = \lambda_0(t)E[f(X(t),\beta|\bar{Y}(t), S \geq t)]$ [1]. Then, an unbiased estimate of β can be obtained by maximizing the partial likelihood

$$\prod_{i=1}^{N}\left\{\frac{E[f(X(t),\beta|\bar{Y}(t),S \geq t)]}{\sum_{k=1}^{N} I(S_k \leq S_i)E[f(X(t),\beta|\bar{Y}(t),S \geq t)]}\right\}^{\Delta_i}$$

instead of modeling the observed history directly. The analytic expression of $E[f(X(t),\beta|\bar{Y}(t), S \geq t)]$ is difficult to obtain. Tsiatis *et al.* [1] developed a two-step inferential approach based on a first-order approximation, $E[f(X(t),\beta|\bar{Y}(t), S \geq t)] \approx f[E(X(t)|\bar{Y}, S \geq t, \beta)]$. Under this approximation, we can replace $\{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^{\mathrm{T}}$ in the Cox model with the empirical estimates $\{\hat{\theta}_{0i}, \hat{\theta}_{1i}, \ldots, \hat{\theta}_{pi}\}^{\mathrm{T}}$ described by Laird and Ware [24], so that $X_i(S_i)$ in (4) will be replaced by $\hat{X}_i(u) = \hat{\theta}_{0i} + \hat{\theta}_{1i}u + \hat{\theta}_{2i}u^2 + \cdots + \hat{\theta}_{pi}u^p + \gamma Z_i$. The partial likelihood (4) can then be used for inferences in obtaining parameter estimates without using the full joint likelihood (3).

3. Sample size determination for studying the relationship between event time and the longitudinal process

The sample size formula presented in this section is based on the assumption that the hazard function follows equation (1) and the trajectory follows a general polynomial model as specified in equation (2). No time-by-treatment interaction is assumed with the longitudinal process. Furthermore, we assume that if any Y_{ij}'s are missing, they are missing at random.

3.1. Known Σ_θ

We start by assuming a known trajectory, $X_i(t)$, so that the score statistic can be derived directly based on the partial likelihood. We show in the supplement that the score statistic converges to a function of $\mathrm{Var}\{X_i(t)\}$, and thus a function of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. When $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is known, and assuming that the trajectory follows a general polynomial function of time as in equation (2), we derive a formula for the number of events required for a one-sided significance level α test with power $\tilde{\beta}$ (see detailed derivation in the supplement‡). This formula is given by

$$D = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\sigma_s^2 \beta^2}, \tag{5}$$

where

$$\sigma_s^2 = \mathrm{Var}(\theta_{0k}) + \sum_{j=1}^{p}\mathrm{Var}(\theta_{jk})E\{I(T \leq \bar{t}_f)T^{2j}\}/\tau + 2\sum_{j=0}^{p}\sum_{l>j}^{p}\mathrm{Cov}(\theta_{jk},\theta_{lk})E\{I(T \leq \bar{t}_f)T^{j+l}\}/\tau. \tag{6}$$

p is the degree of polynomial in the trajectory, $\tau = D/N$ is the event rate, and $\bar{t}_f$ is the mean follow-up time for all subjects. $E\{(I \leq \bar{t}_f)T^q\}$ is a truncated moment of T^q, whose calculation

‡Supporting information may be found in the online version of this article.

is provided in Appendix A. It can be estimated by assuming a particular distribution of the event time T, and a mean follow-up time. Therefore, the power for estimating β depends on: (a) the expected log-hazard ratio associated with a unit change in the trajectory, or the size of β. As β increases, the required sample size decreases; (b) $\mathbf{\Sigma}_{\boldsymbol{\theta}}$. A larger variance and positive covariances lead to smaller sample sizes, while larger negative covariances imply less heterogeneity and require larger sample sizes; and (c) the truncated moments of the event time T, which depends on both the median survival and length of follow-up. Larger $E\{(I \leq \bar{t}_f)T^q\}$implies larger σ_s^2, and thus requires smaller sample size. Details for estimating $E\{(I \leq \bar{t}_f)T^q\}$are provided in Appendix A. Because τ, the event rate, also affects σ_s^2, censored observations do in fact contribute to the power when estimating the trajectory effect.

Specific assumptions regarding $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ are required in order to estimate σ_s^2, regardless of whether $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is assumed known or unknown (see Sections 3.2 and 4). It is usually difficult to find relevant information concerning each variance and covariance for the θ's, especially when the dimension of $\mathbf{\Sigma}_{\boldsymbol{\theta}}$, or the degree of the polynomial in the trajectory is high. A structured covariance matrix, such as an autoregressive or compound symmetry, can be used. One can simplify formula (6) with a structured covariance matrix. This also facilitates the selection of a covariance structure in the final analysis.

3.2. Unknown $\mathbf{\Sigma}_{\boldsymbol{\theta}}$

When $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is unknown, sample size determination can be based on the two-step inferential approach suggested by Tsiatis *et al.* [1]. Despite several drawbacks in this two-stage modeling approach [2], it has two major advantages: (a) the likelihood is simpler and standard statistical software for the Cox model can be used directly for inferences and estimation; (b) it can correct bias caused by missing data or mis-measured time-dependent covariates. Therefore, when $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is unknown, the trajectory is characterized by the empirical Bayes estimates of $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$. $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ in equation (6) can then be replaced with an overall estimate of $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}$, where $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}$ is the covariance matrix of $\{\hat{\theta}_{0i}, \hat{\theta}_{1i}, \ldots, \hat{\theta}_{pi}\}^{\mathrm{T}}$.

$\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}$ is clearly associated with the frequency and spacing of repeated measurements on the subjects, duration of the follow-up period, and the within-subject variability, σ_e^2 [25]. Since $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is never known in practice, sample size determination using $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ in equation (6) will likely over-estimate the power. Therefore, we need to understand how the longitudinal data (i.e. frequency of measurements, spacing of measurements, etc.) affect $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}$, and design a data collection strategy to maximize the power for the study. We defer the discussion of this issue to Section 4.

3.3. Simulation results

We first verified in simulation studies that when $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is known, formula (5) provides an accurate estimate of the power for estimating β. Table I shows a comparison of the calculated power based on equations (5) and (6), and empirical power in a linear trajectory with known $\mathbf{\Sigma}_{\boldsymbol{\theta}}$. In this simulation study, the event time was simulated from an exponential model with exponential parameter η and $\lambda_i(t) = \lambda_0(t)\exp\{\beta X_i(t) + \alpha Z_i\}$, where $X_i(t) = \theta_{0i} + \theta_{1i}\, t + \gamma Z_i$. To ensure a minimum follow-up time of 0.75 y (9 months), censoring was generated from a uniform [0.75, 2] distribution. $(\theta_{0i}\ \theta_{1i})$ was assumed to follow a bivariate normal distribution. We simulated 1000 trials and each trial has 200 subjects. Empirical power was defined as the % of trials with a p-value from the score test ≤0.05 for testing H_0: $\beta = 0$. The quantities D, η, and $\bar{t}_f$ were obtained based on the simulated data, η was obtained from the median survival of the simulated data, and $\bar{t}_f$ was the mean follow-up time of the simulated data using the product limit method. Thus, Table I shows that if the input parameters are correct, formula (5) returns an accurate estimate of power in various $\mathbf{\Sigma}_{\boldsymbol{\theta}}$.

4. Estimating $\Sigma_{\hat{\theta}_i}$ and maximization of power

4.1. Estimating $\Sigma_{\hat{\theta}_i}$

Following the notation in Section 2, Let

$$\mathbf{R}_i=\begin{pmatrix} 1 & t_{i1} & \dots & t_{i1}{}^p \\ 1 & t_{i2} & \dots & t_{i2}{}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \dots & t_{im_i}{}^p \end{pmatrix}$$

be an $m_i \times (1+p)$ matrix, and $\mathbf{Z}_i = \mathbf{1}_{m_i} Z_i$, $\mathrm{Var}(\boldsymbol{Y}_i)=\mathbf{V}_i=\mathbf{I}_{m_i}\sigma_e^2+\mathbf{R}_i\sum_\theta \mathbf{R}_i^{\mathrm{T}}$ and $\mathbf{W}_i=\mathbf{V}_i^{-1}$. Then $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\Sigma}_{\hat{\theta}_i}$ can be expressed as [24]

$$\widehat{\theta}_i-\mu_\theta=\sum_\theta \mathbf{R}_i^{\mathrm{T}}\mathbf{W}_i(\boldsymbol{Y}_i-\widehat{\gamma}\mathbf{Z}_i)$$

and

$$\mathrm{Var}(\widehat{\theta}_i)=\sum_{\widehat{\theta}_i}=\sum_\theta \mathbf{R}_i^{\mathrm{T}}\left\{\mathbf{W}_i-\mathbf{W}_i\mathbf{Z}_i\left(\sum_i^N \mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\right\}\mathbf{R}_i\sum_\theta. \quad (7)$$

4.2. Determinants of power

Based on equation (7), $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ is associated with the following: (a) the degree of the polynomial in (2); (b) $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, that is, the between-subject variability; (c) σ_e^2, the within-subject variability; (d) t_{ij}, the time of the repeated measurements of the longitudinal data. Larger t_{ij} imply a longer follow-up period, or more data collection points toward the end of the trial, and (e) m_i, the frequency of the repeated measurements. (a)–(c) above are likely to be determined by the intrinsic nature of the longitudinal data, and have little to do with the data collection strategy during the trial design. Based on (7), $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ is associated with the inverse of σ_e^2, meaning larger σ_e^2 will lead to smaller $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$, and thus a decrease in the power for estimating β. This is confirmed in the simulation studies (Table II).

Although σ_e^2, the within-subject variability, can be reduced by using a more reliable measurement instrument, this is not always possible. We therefore focus on investigating the impact of (d) and (e). Note that the hazard function can be written as $\lambda_i(t) = \lambda_0(t)\exp\{\beta(\theta_{0i} + \theta_{1i}t + \cdots + \theta_{pi}t^p) + \theta^* Z_i\}$, where $\beta^* = \beta\gamma + \alpha$. In the design stage, instead of considering a trajectory with $\gamma \neq 0$ and a direct treatment effect of α, we can consider a trajectory with $\gamma = 0$ and a direct treatment effect of $\alpha + \beta\gamma$. This will simplify the calculations for $\Sigma_{\hat{\theta}_i}$. Since formula (7) represents $\Sigma_{\hat{\theta}_i}$ when $Z_i = 0$, it should provide good approximation when $\Sigma_{\hat{\theta}_i}$ is similar between the two treatment groups. To see the relationship between m_i, t_{ij} and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$, let us consider the alternative trajectory with $\gamma = 0$. Equation (7) then simplifies to

$$\sum_{\widehat{\theta}_i} = \sum_{\theta} \mathbf{R}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{R}_i \sum_{\theta} \tag{8}$$

and

$$\sum_{\widehat{\theta}_i} = \sum_{\theta} \mathbf{Q} \sum_{\theta} = \sum_{\theta} \begin{pmatrix} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} W_{ijk} & \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} t_{ik} W_{ijk} \\ \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} t_{ij} W_{ijk} & \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} t_{ij} t_{ik} W_{ijk} \end{pmatrix} \sum_{\theta}. \tag{9}$$

When the trajectory is linear. $W_{i\,jk}$ is the element in the jth row and kth column of $\mathbf{W}_i$. Now we decompose $\mathbf{V}_i$ as $\mathbf{P}_i \mathbf{D}_{g_i} \mathbf{P}_i^{\mathrm{T}}$, where $\mathbf{P}_i$ is an $m_i \times m_i$ matrix with orthonormal columns, and $\mathbf{D}_{g_i}$ is a diagonal matrix with non-negative eigenvalues. Let $P_{i\,jk}$ denote the element of the jth row and kth column of $\mathbf{P}_i$, and $D_{g_{ij}}$ denotes the element in the jth row and jth column of $\mathbf{D}_{g_i}$. Then the diagonal elements of $\mathbf{Q}$ in (9) can be expressed as

$$\sum_{j=1}^{m_i} \sum_{k=1}^{m_i} W_{ijk} = \sum_{j=1}^{m_i} D_{g_{ij}}^{-1} \left(\sum_{k=1}^{m_i} P_{ijk} \right)^2 \tag{10}$$

and

$$\sum_{j=1}^{m_i} \sum_{k=1}^{m_i} t_{ij} t_{ik} W_{ijk} = \sum_{j=1}^{m_i} D_{g_{ij}}^{-1} \left(\sum_{k=1}^{m_i} t_{ik} P_{ijk} \right)^2. \tag{11}$$

We can see that both equations (10) and (11) are sums of m_i non-negative elements, and thus are non-decreasing functions of m_i. Equation (11) is also positively associated with $t_{i\,j}$, implying a larger variance with longer follow-up period or with longitudinal data collected at a later stage of the trial. However, we should keep in mind that some subjects may have failed or are censored due to early termination. If we schedule most data collection time point toward the end of the study, m_i could be reduced significantly in many subjects. An ideal data collection strategy should take into account drop-out and failure rates and balance $t_{i\,j}$ and m_i for a fixed maximum follow-up period.

The maximum follow-up period is usually prefixed due to timeline or budget constraints. We can observe more events with a longer follow-up and the increase in power is likely to be more significant due to an increased number of events. With a prefixed follow-up period, the most important decision is perhaps to describe an optimal number of data collection points. Here, we speculate that the power would reach a plateau as m_i increases. The number of data collection points required to reach the plateau is likely to be related to the degree of the polynomial in the trajectory function. A lower order polynomial may require smaller m_i.

4.3. Simulation studies and illustrations of using $\Sigma_{\hat{\theta}_i}$ in sample size calculation

We investigated the power assuming an unknown Σ_{θ} for different m_i 's in simulation studies. The results are summarized in Table II for a linear trajectory, and in Table III for a quadratic trajectory. We note that the longitudinal data, $Y_{i\,j}$, are missing after the event occurs or after the subject is censored and is assumed to be missing at random. Therefore, m_i varies among subjects. Let m_x denote the scheduled, or maximum number of data collection points if the subject has not had an event and is not censored at the end of the follow-up

period. In the simulation studies described in Tables II and III, m_x was assumed to be the same for all subjects, and t_{ij} was equally spaced. In the linear trajectory simulation studies, we further assumed that the longitudinal data was also collected when the subject exits the study due to an event or censoring, so that each subject would have at least two measurements (baseline and the end of the study). In the quadratic trajectory simulation studies, the longitudinal data was also collected when the subject exited the study before their first post-baseline scheduled measurement. Therefore, $\mathbf{R_i}$ in equation (8) was not the same for all subjects. Some had different numbers of measurements; and some had measurements at different t_{ij} 's. This results in a different $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ for each subject.

Note that $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ converges to $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ when $\sigma_e^2 \to 0$. However, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ does not converge to $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ when $\sigma_e^2 \neq 0$. It is not an estimator for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ as it is influenced by the magnitude of the residual (measurement error), σ_e^2. During the design stage, we need to find a single quantity that can represent an average effect of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$, which will take into account the impact of σ_e^2, to replace $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ in the sample size calculation. One choice of such a quantity is to assume that all subjects will have the same number of measurements at the same time points, and thus $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ will be the same for all subjects. As the measurement error will have a greater impact on the 'bias' when the number of measurements are small, assuming a maximum number of measurements for all subjects will result in an over-estimation of the power, while assuming a minimum number of measurements for all subjects will result in an under-estimation of the power. Assuming a median number of measurements may be adequate in assessing the average effect of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$. We recommend using the weighted average of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$'s because it takes into account the impact of σ_e^2 from the smallest to the largest number of measurements. For a fixed m_x, the weighted average can be calculated as

$$\sum_{m=1}^{m_x} \xi_m \sum\nolimits_{\theta} \mathbf{R}_{\cdot m}^{\mathrm{T}} \left(\mathbf{I}_m \sigma_e^2 + \mathbf{R}_{\cdot m} \sum\nolimits_{\theta} \mathbf{R}_{\cdot m}^{\mathrm{T}}\right)^{-1} \mathbf{R}_{\cdot m} \sum\nolimits_{\theta}, \tag{12}$$

where ξ_m is the % of non-censored subjects who have m measurements of the longitudinal data, $\mathbf{I}_m$ is the $m \times m$ identity matrix, and $\mathbf{R}_{\cdot m}$ is the $\mathbf{R}$ matrix with m measurements,

$$\mathbf{R}_{\cdot m} = \begin{pmatrix} 1 & t_{\cdot 1} & \dots & t_{\cdot 1}{}^p \\ 1 & t_{\cdot 2} & \dots & t_{\cdot 2}{}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{\cdot m} & \dots & t_{\cdot m}{}^p \end{pmatrix}.$$

$t_{\cdot k}$ in the $\mathbf{R}_{\cdot m}$ matrix represents the mean measurement time of the kth measurement in the subjects who had m measurements if not all measurements are taken at a fixed time point.

In the second to the last column of Tables II and III, we present the calculated power based on the maximum $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ instead of a weighted average of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$'s. The maximum $\sum_{\widehat{\theta_i}} = \sum_{\theta} \mathbf{R}_{\cdot m_{\mathbf{x}}}^{\mathbf{T}} \left(\mathbf{I}_{m_{\mathbf{x}}} \sigma_e^2 + \mathbf{R}_{\cdot m_{\mathbf{x}}} \sum_{\theta} \mathbf{R}_{\cdot m_{\mathbf{x}}}^{\mathbf{T}}\right)^{-1} \mathbf{R}_{\cdot m_{\mathbf{x}}} \sum_{\theta}$. The simulation setup in Tables II and III is the same as in Section 3.3. The longitudinal data Y_{ij} was simulated via a normal distribution with mean $\theta_{0i} + \theta_{1i} t_{ij} + \gamma Z_i$ (linear), or $\theta_{0i} + \theta_{1i} t_{ij} + \theta_{2i} t_{ij}^2 + \gamma Z_i$ (quadratic), and variance σ_e^2. Y_{ij} was set to be missing after an event or censoring occurred.

When the measurement error is relatively small and non-systematic, the two-step inferential approach yields nearly unbiased estimates of the longitudinal effect. The number of data collection points did not seem to be critical when the trajectory is linear as long as each subject had at least two measurements of the longitudinal data. There is a slight decrease in the power when $m_X < 5$ and σ_e^2 is large. When the trajectory is quadratic, m_X plays a more important role. The power for estimating β decreases as m_X decreases. Smaller numbers of measurements ($m_X < 4$) can also lead to a biased estimate of the longitudinal effect and result in a significant loss of power. The effect of m_X on estimates and power is more significant when σ_e^2 is large. Note that when $\sigma_e^2=0$, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ reduces to $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, and is unrelated to m_X. The effect of m_X comes from the magnitude of reducing the contribution of the within-subject variability, σ_e^2. If we have a very accurate and reliable measurement instrument, we can reduce the number of repeated measurements and can still obtain unbiased estimates and maximum power.

The power calculation under the assumption of known $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ or perfect data collection (maximum $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$) can result in a significant over-estimation of the power especially when σ_e^2 is large. We next demonstrate that if we use the weighted average of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$'s, we can obtain a good estimate of power based on formula (5).

Example 1 from Table II: For the scenario with $\sigma_e^2=0.64$ and $m_X = 2$, we observed that the mean measurement time for the subjects who had an event in the simulated data is about 0.5 y. We used $\mathbf{R}_{\cdot 2}=\begin{pmatrix} 1 & 0 \\ 1 & 0.5 \end{pmatrix}$ to calculate $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$ instead of setting $\mathbf{R}_{\cdot 2}=\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$ which assumes that the second measurement was taken at 2 y. As a result, the power based on formula (5) changed from 77.4 to 74.4 per cent, which is more closer to the empirical power of 74.0 per cent. We used the mean measurement time in the non-censored subjects, because the power calculation is mainly based on the number of events. In practice, we need to make certain assumptions about $t_{\cdot k}$ based on the median survival and length of the follow-up period.

Example 2 from Table III: For demonstration, we chose the scenario with $\sigma_e^2=0.81$ and $m_X = 4$. In this example, the second measurement was taken at 0.45 y (on average) in subjects who had only two measurements. For subjects who had more than two measurements, longitudinal data was collected at scheduled time points of 0, 0.5, 1, and 1.5. Therefore,

$$\mathbf{R}_{\cdot 2}=\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.45 & 0.20 \end{pmatrix}, \quad \mathbf{R}_{\cdot 3}=\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_{\cdot 4}=\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2.25 \end{pmatrix}.$$

A weighted average of the $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_i}$'s was calculated based on formula (12). The resulting power is 78.8 percent instead of 84.7 per cent, which is close to the empirical power of 79.1 per cent.

For trajectories that are quadratic or higher, it is important to schedule data collection to ensure m_i is large enough for a reasonable proportion of subjects. For example, when the trajectory is quadratic and only a small proportion of subjects had three measurements of the longitudinal data ($m_X = 3$ in Table III), we obtain a very biased estimate of β.

5. Sample size determination for the treatment effect

Using the same model as specified in Section 3, the overall treatment effect is $\beta\gamma + \alpha$. Thus, the null hypothesis is H_0: $\beta\gamma + \alpha = 0$. Following the framework of Schoenfeld [26], we show that Schoenfeld's formula can be extended to a joint modeling study design by taking into account the additional parameters β and γ. The number of events required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$, assuming the hazard and trajectory follow (1) and (2) in Section 2, is given by

$$D=\frac{(z_{\tilde{\beta}}+z_{1-\tilde{\alpha}})^2}{p_1(1-p_1)(\beta\gamma+\alpha)^2}, \tag{13}$$

where p_1 is the % of patients assigned to treatment 1 ($Z_i = 1$). The properties of the random effects in the trajectory do not play a significant role in the sample size and power determination for the overall treatment effect at the design stage. However, correct assumptions must be made with regard to the overall treatment effect ($\beta\gamma + \alpha$). If the longitudinal effect is a biomarker, $\alpha$and $\beta\gamma$ should have the same sign (aggregated treatment effect). We acknowledge that under the proposed longitudinal and survival model, the ratio of the hazard functions of the two treatment groups will be non-proportional, as the trajectory is time dependent. However, the method of using the partial likelihood can readily be generalized to allow for non-proportional hazards. It is unlikely that the proportional hazards assumption is ever exactly satisfied in practice. When the assumption is violated, the coefficient estimated from the model will be the 'average effect' over the range of time observed in the data [27]. Thus, the sample size formula developed using the partial likelihood method should provide good approximation of the power for estimating the overall treatment effect in a joint modeling setting.

The simulation studies presented in Table IV show that formula (13) works approximately well in the two-step inferential approach when the primary objective is to investigate the overall treatment effect. The power is not sensitive to $\mathbf{\Sigma}_{\boldsymbol{\theta}}$, and works well with different sizes of β and γ. We show in Sections 6 and the supplemental material that the two-step inferential approach and the full joint likelihood approach yield similar unbiased estimates of the overall treatment effect and have similar efficiency.

6. Biased estimates of the treatment effect when ignoring the longitudinal trajectory

When a treatment has an effect on the longitudinal process (i.e. $\gamma \neq 0$ in equation (2)) and the longitudinal process is associated with survival (i.e. $\beta \neq 0$ in equation (1)), the overall treatment effect on the time-to-event is $\beta\gamma + \alpha$. Thus, it is obvious that ignoring the longitudinal process in the proportional hazards model can result in a biased estimate of the treatment effect on survival. When the longitudinal process is not associated with the treatment (i.e. $\gamma = 0$ in equation (2)), it is not obvious that ignoring the longitudinal trajectory in the proportional hazards model would result in an attenuated estimate of the hazard ratio for the treatment effect on survival (i.e. bias toward the null). This attenuation is known in the econometrics literature as the attenuation due to unobserved heterogeneity [28, 29], and has been discussed in the work by Gail *et al.* [30].

We demonstrated in simulation studies (Table V) that the bias associated with ignoring the longitudinal effect is related to the size of β in the joint modeling setting.

7. Retrospective power analysis for the ECOG trial E1193

To illustrate parameter selection and the impact of incorporating $\boldsymbol{\Sigma}_{\hat{\theta}_i}$ in the power calculation, we apply the sample size calculation formula retrospectively based on the parameters obtained from the Eastern Cooperative Oncology Group (ECOG) E1193 trial [31, 32]. E1193 is a phase III cancer clinical trial of doxorubicin, paclitaxel, and the combination of doxorubicin and paclitaxel as front-line chemotherapy for metastatic breast cancer. Patients receiving single-agent doxorubicin or paclitaxel crossed over to the other agent at the time of progression. QOL was assessed using the FACT-B scale at two time points during induction therapy. The FACT-B includes five general subscales (physical, social, relationship with physician, emotional, and functional), as well as a breast cancer-specific subscale. The maximum possible score is 148 points. A higher score is indicative of a better QOL. In this subset analysis, we analyzed the overall survival after entry to the crossover phase (survival after disease progression), and its association with treatment and QOL. A total of 252 patients entered the crossover phase and have at least one QOL measurement, 124 patients crossed over from paclitaxel to doxorubicin (median survival is 13.0 months in this subgroup), 128 patients crossed over from doxorubicin to paclitaxel (median survival is 14.9 months in this subgroup). The data we used are quite mature, with only two subjects who crossed over to doxorubicin and six subjects who crossed over to paclitaxel being censored. We applied the Cox model with treatment effect only, the two step model incorporating the two QOL measurements, and the proposed joint model as specified in Section 2 of the paper, to analyze the treatment effect and effect of QOL. Since there are only two QOL measurements, we fit a linear mixed model. To satisfy the normality assumption for the longitudinal QOL, we transformed the observed QOL into $\mathrm{QOL}^{\frac{1}{2}}$. The results are report in Table VI. Similar results are also reported by the same authors [32].

Treatment effects are similar between the two-step model and the joint model. The difference in the QOL effect, $\hat{\beta}$, is similar to that of Wulfsohn and Tsiatis [2]. They reported a slightly larger $\hat{\beta}$ and standard error in the joint model as compared with the two-step model. In Section 6 of this paper, we used simulation studies to demonstrate that $\hat{\beta}$ is sensitive to whether the constant hazard assumption is satisfied in the joint model we used. We obtained the following parameter estimates for the retrospective power calculation: The median overall survival is 13.56 months $\sum_{\theta}^{\frac{1}{2}} = \begin{pmatrix} 0.8417 & 0 \\ 0 & 0.0025 \end{pmatrix}$, $\sigma_e = 0.7188$, the mean measurement time for the first QOL is 0.052 months, the mean measurement time for the second QOL is 2.255 months, and 35 per cent of the subjects had only one QOL measurement. If we assume a known $\boldsymbol{\Sigma}_{\theta}$, the power with 243 events and $\beta = 0.3$ is 98 per cent. When we assume an unknown $\boldsymbol{\Sigma}_{\theta}$ and use a weighted average of $\boldsymbol{\Sigma}_{\hat{\theta}_i}$, the power is reduced to 90 per cent. The relationship between sample size and power for both known and unknown $\boldsymbol{\Sigma}_{\theta}$ cases is illustrated in Figure 1.

8. Discussion

In this paper, we have provided a closed-form sample size formula for estimating the effect of the longitudinal data on time-to-event and discussed optimal data collection strategies. The number of events required to study the association between event time and the longitudinal process for a given follow-up period is related to the covariance matrix of the random effects (coefficients for the *p*-polynomial), within-subject variability, frequency of repeated measurements, and timing of the repeated measurements. Only a few parameters are required in the sample size formula. The median event time and mean follow-up time are needed to calculate the truncated moments. The mean follow-up time can be approximated by the average of the minimum and maximum follow-up times under the assumption of

uniform censoring. A structured covariance matrix can be used when we do not have prior data to determine each element of Σ_θ. More robust estimates can be achieved by assuming an unknown Σ_θ. An unknown Σ_θ requires further assumptions about the number and timing of repeated measurements, and the percentage of subjects who are still on-study at each scheduled measurement time. This is exactly what researchers should consider during the design stage. It is useful to consider a few scenarios and compare the calculated power. When the measurement error is small, estimates with known Σ_θ also provide good approximation of power.

We have also extended Schoenfeld's [26] sample size estimation formula to the joint modeling setting for estimating an overall treatment effect. When the longitudinal data was associated with treatment, the overall treatment effect is an aggregated effect on time-to-event directly and on the longitudinal process. When the longitudinal data is not associated with treatment, ignoring the longitudinal data will still lead to attenuated estimates of the treatment effect due to unobserved heterogeneity. The degree of attenuation depends on the degree of the association between the longitudinal data and time-to-event data. Use of a joint modeling analysis strategy leads to reduction of bias and increase in power in estimating the treatment effect. However, joint modeling is not yet commonly used in designing clinical trials. Most applications of joint modeling in the literature focus on estimating the effect of the longitudinal outcome on time-to-event.

The sample size formula we derived was based on a score test. Under the assumption that the fixed covariates are independent of the probability that the patient receives treatment assignment, the fixed covariates cancel from the final score statistic. Therefore, the number of patients required for a study does not depend on the effects of other fixed covariates. However, this does not mean that we should exclude all covariates from the analysis model, as they may be required to build an appropriate model.

The sample size formula we considered in this paper is based on the two-step inferential approach proposed by Tsiatis *et al.* [1], which is known to have several drawbacks [2]. In the supplementary material of this paper, we examined two joint modeling approaches: the two-step model and a model that is based on the full likelihood as specified in (3). The purpose of the supplemental section is to compare robustness and efficiency of the two joint models and evaluate whether the current sample size determination method can still provide an approximate estimate for the parametric joint model. We show that: (1) the two-step model may be more robust than the parametric joint model, especially when the parametric model is miss-specified; (2) if the parametric model is correctly specified, it is more efficient than the two-step model. Thus, a sample size based on the proposed method in this paper will be conservative for the parametric joint model for testing β; and (3) when testing the overall treatment effect, the two modeling approaches have similar efficiency and thus the method proposed can provide good estimate of sample size for both joint models.

Missing longitudinal data in practice is typically non-ignorably missing in the sense that the probability of missingness depends on the longitudinal variable that would have been observed. In order to examine the robustness of our sample size formulas to non-ignorable missingness, we conducted several simulation studies in which the empirical power was computed under a non-ignorable missing data mechanism using a selection model. Under several scenarios, our calculated powers based on the proposed sample size formulas were quite close to the empirical powers, therefore illustrating that our sample size formulas are quite robust to non-ignorable missing data. Developing closed-form sample size formulas in the presence of non-ignorable missing data is a very challenging problem that requires much further research.

One known drawback of the two-step method is that the random effects are assumed to be normally distributed in those at risk at every event time. This is unrealistic under informative dropout. It is also known that the empirical Bayes estimate of the random effect, $\hat{\boldsymbol{\theta}}_i$, from the Laird and Ware method [24] is biased under informative dropout. Therefore, non-informative censoring of the longitudinal process is an important assumption for the proposed method. Another limitation of this method is that we did not consider the treatment-by-time interaction in the model, which precludes the random slopes model. An extension of the classical Cox model introduces interaction between time and covariates, with the purpose of testing or estimating the interaction via a smoothing method. When testing a treatment effect alone, we are interested in showing that the treatment effect is constant over time (no interaction). Therefore, in the sample size calculation, we should not assume both a treatment effect and a treatment-by-time interaction. When the purpose is to test the effect of the longitudinal data on survival, if we need to assume a treatment-by-time interaction, we should fit a separate two-step model for each treatment.

Finally, we mention here that although simulations and distributional assumptions of the random effects in this paper were based on a Gaussian distribution, such distributional assumptions are not required for the formula. It may be applied to more general joint modeling design settings. To the best of our knowledge, this is the first paper that addresses trial design aspects using joint modeling.

[†]Calculated based on the mean number of deaths from simulations and fixed value of $p_1 = 0.5$, β, γ, and α.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. Journal of American Statistical Association. 1995; 90:27–37.
2. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997; 53:330–339. [PubMed: 9147598]
3. Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M, Vangeneugden T, Bijnens L. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. Journal of Applied Statistics. 2003; 30:235–247.
4. Ibrahim, JG.; Chen, MH.; Sinha, D. Bayesian Survival Analysis. Vol. Chapter 7. Springer; New York: 2001. Joint models for longitudinal and survival data.
5. Billingham LJ, Abrams KR. Simultaneous analysis of quality of life and survival data. Statistical Methods in Medical Research. 2002; 11:25–48. [PubMed: 11923992]
6. Bowman FD, Manatunga AK. A joint model for longitudinal data profiles and associated event risks with application to a depression study. Applied Statistics. 2005; 54:301–316.
7. Zeng D, Cai J. Simultaneous modeling of survival and longitudinal data with an application to repeated quality of life measures. Lifetime Data Analysis. 2005; 11:151–174. [PubMed: 15940822]
8. Chi Y-Y, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. Biometrics. 2006; 62:432–445. [PubMed: 16918907]
9. Chi Y-Y, Ibrahim JG. A new class of joint models for longitudinal and survival data accommodating zero and non-zero cure fractions: a case study of an international breast cancer study group trial. Statistica Sinica. 2007; 17:445–462.
10. Taylor JMG, Wang Y. Surrogate markers and joint models for longitudinal and survival data. Controlled Clinical Trials. 2002; 23:626–634. [PubMed: 12505241]

11. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: and overview. Statistica Sinica. 2004; 14:809–834.
12. Hogan JW, Laird NW. Model-based approaches to analyzing incomplete longitudinal and failure time data. Statistics in Medicine. 1997; 16:239–257. [PubMed: 9004395]
13. Henderson R, Diggle P, Dobson A. Joint modeling of longitudinal measurements and event time data. Biostatistics. 2000; 1:465–480. [PubMed: 12933568]
14. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodefinicency syndrome. Journal of American Statistical Association. 2001; 96:895–905.
15. Lin H, Turnbull BW, McCulloch EE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. Journal of American Statistical Association. 2002; 97:53–65.
16. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. Biometrics. 2002; 58:742–753. [PubMed: 12495128]
17. Chen M-H, Ibrahim JG, Sinha D. A new joint model for longitudinal and survival data with a cure fraction. Journal of Multivariate Analysis. 2004; 91:18–34.
18. Brown ER, Ibrahim JG. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. Biometrics. 2003; 59:221–228. [PubMed: 12926706]
19. Ibrahim JG, Chen M-H, Sinha D. Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. Statistica Sinica. 2004; 14:863–883.
20. Hsieh F, Tseng YK, Wand J-L. Joint modeling of survival and longitudinal data: likelihood approach revisited. Biometrics. 2006; 62:1037–1043. [PubMed: 17156277]
21. Song X, Davidian M, Tsiatis AA. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. Biostatistics. 2002; 3:511–528. [PubMed: 12933595]
22. Dang Q, Mazumdar S, Anderson SJ, Houck PR, Reynolds CF. Using trajectories from a bivariate growth curve as predictors in a Cox regression model. Statistics in Medicine. 2007; 26:800–811. [PubMed: 16612837]
23. Cox DR. Partial likelihood. Biometrika. 1972; 62:269–276.
24. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]
25. Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Vol. Chapter 15. Wiley; New York: 2004. Some aspects of the design of longitudinal studies.
26. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983; 39:499–503. [PubMed: 6354290]
27. Allison, PD. Survival Analysis Using SAS—A Practical Guide. Vol. Chapter 5. SAS Institute Inc; Cary. NC, U.S.A: 1995. Estimating Cox Regression Models with PROCPHREG.
28. Horowitz JL. Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. Econometrica. 1999; 67:1001–1028.
29. Abbring JH, Van den Berg GJ. The unobserved heterogeneity distribution in duration analysis. Biometrika. 2007; 94:87–99.
30. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984; 71:431–444.
31. Sledge GW, Neuberg D, Bernardo P, Ingle JN, Martino S, Rowinsky EK, Wood WC. Phase III trial of doxorubicin, paclitaxel, and the combination of doxorubicin and paclitaxel as front-line chemotherapy for metastatic breast cancer: an intergroup trial (E1193). Journal of Clinical Oncology. 2003; 21:588–592. [PubMed: 12586793]
32. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology. 2010; 28:2796–2801. [PubMed: 20439643]

Appendix A: Truncated moments of T

To obtain the truncated moments of T^q, $E\{(I \leq \tilde{t}_f)T^q\}$, in equation (6), we must assume a distribution for T. In practice, the exact distribution for T is unknown. However, the median

event time or event rate at a fixed time point for the study population can usually be obtained from the literature. It is a common practice to assume that T follows an exponential distribution with exponential parameter η in the study design stage. Thus, the truncated moment of T^q only depends on η and $\bar{t}_f$, and has the following form:

$$E\{I(T \le \bar{t}_f)T^q\} = \int_0^{\bar{t}_f} t^q \eta \exp(-\eta t) dt = \frac{1}{\eta^q}\Gamma(q+1, \bar{t}_f),$$

where $\Gamma(q + 1, \bar{t}_f)$ is a lower incomplete gamma function with $q = \{1, 2, 3, \ldots\ldots\}$. η can be estimated based on the median event time or event rate at a fixed time point. e.g. if the median event time, T_M, is known for the study population, $\eta = -\log(0.5)/T_M$. When the trajectory is a linear function of time,

$$\sigma_s^2 = \mathrm{var}(\widehat{\theta}_{0i}) + \frac{1}{\tau}E[I(T \le \bar{t}_f)T^2]\mathrm{var}(\widehat{\theta}_{1i}) + \frac{2}{\tau}E[I(T \le \bar{t}_f)T]\mathrm{cov}(\widehat{\theta}_{0i}, \widehat{\theta}_{1i}).$$

Both $E\{I(T \le \bar{t}_f)T^2\}$ and $E\{I(T \le \bar{t}_f)T\}$ have closed-form expressions, given by

$$E\{I(T \le \bar{t}_f)T^2\} = \int_0^{\bar{t}_f} t^2 \eta \exp(-\eta t)\mathrm{d}t = \frac{2}{\eta^2} - \exp(-\eta\bar{t}_f)\left(\bar{t}_f^2 + \frac{2\bar{t}_f}{\eta} + \frac{2}{\eta^2}\right),$$

and

$$E\{I(T \le \bar{t}_f)T\} = \int_0^{\bar{t}_f} t\eta \exp(-\eta t)\mathrm{d}t = \frac{1}{\eta} - \exp(-\eta\bar{t}_f)\left(\bar{t}_f + \frac{1}{\eta}\right).$$

There are certain limitations of this distributional assumption for T. It does not take into account covariates that are usually considered in the exponential or Cox model for S. A more complex distributional assumption can be used to estimate $E\{(I \le \bar{t}_f)T^q\}$ if more information is available. However, simple distributional assumptions for T, without the inclusion of covariates or using an average effect of all covariates, are easy to implement and it is usually adequate for sample size or power determination.

$E\{(I \le \bar{t}_f)T^q\}$ also depends on $\bar{t}_f$, the mean follow-up time for all subjects. It is truncated because we typically cannot observe all events in a study. Therefore, it is heavily driven by the censoring mechanism, and can be approximated by the mean follow-up time in censored subjects. One way to estimate $\bar{t}_f$ is to take the average of the minimum and maximum follow-up times if censoring is uniform between the minimum and maximum follow-up times. It can also be estimated based on more complex methods. If data from a similar study are available, $\bar{t}_f$ can be estimated with the product-limit method by switching the censoring indicator so that censored cases would be considered as events and events would be considered as censored.

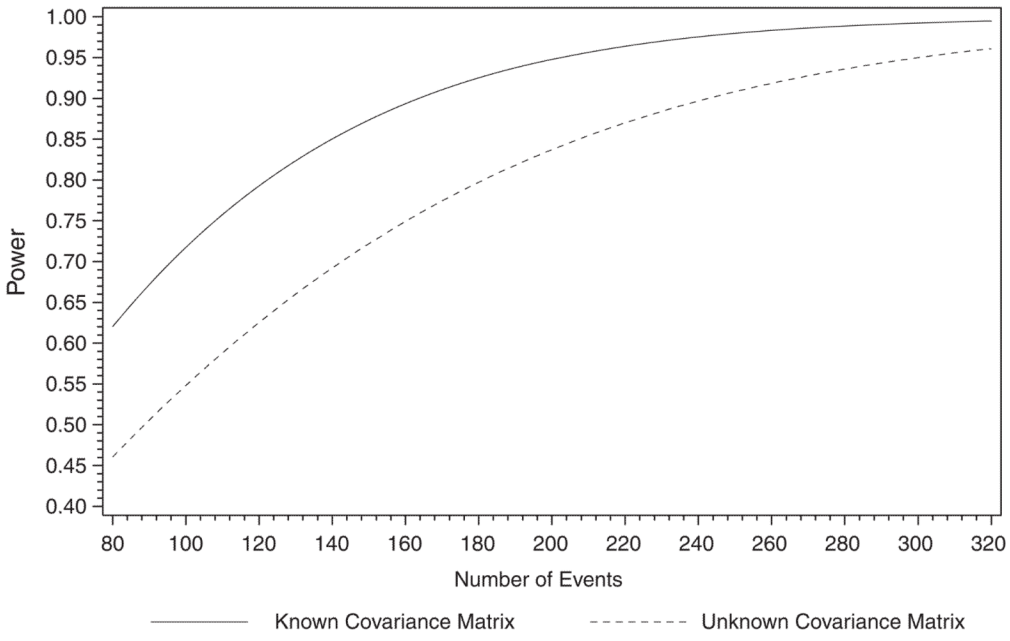

Figure 1.
Retrospective power analysis for the E1193 trial with known and unknown $\boldsymbol{\Sigma}_{\theta}$.

Table I

Validation of formula (5) for testing the trajectory effect β when $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is known.

				Power for estimating β*	
β	Var(θ_{0i})	Var(θ_{1i})	Cov(θ_{0i}, θ_{1i})	Empirical	Calculated
0.15	0.5	0.9	0	41.6	39.8
0.15	0.8	1	0	52.9	52.4
0.15	0.8	1	0.5	66.1	67.0
0.2	1.2	0.7	0	87.1	86.0
0.2	0.7	1.2	0	75.9	76.4
0.2	0.7	1.2	0.2	82.7	82.7
0.2	0.7	1.2	−0.2	69.8	68.4

*Covariance matrix of (θ_{0i}, θ_{1i}) is assumed known. Empirical power is based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 y (9 months), and maximum follow-up time is 2 y. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, and $\gamma = 0.1$. The $\boldsymbol{\theta}$'s are simulated from a normal distribution with $E(\theta_{0i}) = 0$, $E(\theta_{1i}) = 3$, and $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ as specified in columns 2–4.

Table II

Power for estimating β by maximum number of data collection points (m_x) and size of σ_e^2—linear trajectory.

σ_e^2	m_x	$\hat{\beta}$	Power for estimating β† Empirical	Calculated with maximum $\Sigma_{\hat{\theta}_i}$	Calculated with weighted average $\Sigma_{\hat{\theta}_i}$‡
True trajectory		0.2098	87.1	86.0*	
0.09	6	0.2080	86.6	85.4	82.7
0.09	5	0.2075	85.8	85.3	82.6
0.09	4	0.2071	86.3	85.1	82.7
0.09	3	0.2076	86.4	84.9	82.9
0.09	2	0.2065	85.3	84.6	83.3
0.64	6	0.1960	76.3	82.2	75.9
0.64	5	0.1978	76.9	81.6	75.5
0.64	4	0.1939	74.8	80.8	74.9
0.64	3	0.1972	75.0	79.6	74.4
0.64	2	0.1967	74.0	77.4	74.4
1	6	0.1919	71.9	80.5	72.2
1	5	0.1918	71.8	79.7	71.5
1	4	0.1917	69.7	78.5	70.7
1	3	0.1940	70.1	76.8	69.9

*Calculated with $\boldsymbol{\Sigma_{\theta_i}}$.

†β was estimated using the two-step inferential approach [1]. Empirical power was based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 y (9 months), and maximum follow-up time is 2 y. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, $\gamma = 0.1$, $\beta = 0.2$, $E(\theta_{0i}) = 0$, $E(\theta_{1i}) = 3$, $\mathrm{Var}(\theta_{0i}) = 1.2$, $\mathrm{Var}(\theta_{1i}) = 0.7$, and $\mathrm{Cov}(\theta_{0i}, \theta_{1i}) = 0$ (the same simulated data used in Row 4 of Table I).

‡Power based on the weighted average of $\boldsymbol{\Sigma_{\hat{\theta}_i}}$.

Table III

Power for estimating β by the maximum number of data collection points (m_x) and size of σ_e^2—quadratic trajectory.

			Power for estimating $\beta^{\dagger}$		
σ_e^2	m_x	$\hat{\beta}$	Empirical	Calculated with maximum $\Sigma_{\hat{\theta}_i}$	Calculated with weighted average $\Sigma_{\hat{\theta}_i}^{\ddagger}$
True trajectory		0.2212	91.6	90.6*	
0.09	10	0.2117	90.0	90.2	88.0
0.09	7	0.2102	89.5	90.0	88.0
0.09	5	0.2098	89.0	89.9	88.2
0.09	4	0.2014	89.3	89.8	88.4
0.09	3	0.1720	89.1	89.6	88.4
0.25	10	0.2135	89.7	89.5	86.1
0.25	7	0.2104	88.2	89.2	85.8
0.25	5	0.2089	86.7	88.8	85.8
0.25	4	0.2038	86.9	88.5	85.9
0.25	3	0.1621	86.6	88.0	85.8
0.81	10	0.2041	84.7	87.6	81.0
0.81	7	0.1984	81.5	86.6	79.7
0.81	5	0.2021	80.3	85.4	79.0
0.81	4	0.1818	79.1	84.7	78.8
0.81	3	0.1402	74.9	83.3	78.3

*Calculated with $\boldsymbol{\Sigma_{\theta_i}}$

†$\beta$was estimated with the two-step inferential approach [1]. Empirical power was based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 y (9 months), and maximum follow-up time is 2 y. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, $\gamma = 0.1$, $\beta = 0.22$, $\boldsymbol{\theta_i} = (0, 2.5, 3)^{\mathrm{T}}$, and $\boldsymbol{\Sigma_\theta} = \mathrm{diag}(1.2, 0.7, 0.8)$.

‡Power based on the weighted average of $\boldsymbol{\Sigma_{\hat{\theta}_i}}$.

Table IV

Validation of formula (13) for testing the overall treatment effect $\alpha + \beta\gamma$.

					Power for estimating overall treatment effect $\beta\gamma + \alpha$	
β	γ	Var(θ_{0i})	Var(θ_{1i})	Cov(θ_{0i}, θ_{1i})	Empirical*	Calculated†
0.3	−0.1	1.2	0.7	0.2	69.2	67.2
0.3	−0.4	1.2	0.7	0.2	85.8	85.9
0.3	−0.8	1.2	0.7	0.2	96.7	97.1
0.3	−1.2	1.2	0.7	0.2	99.4	99.6
0.1	−0.4	1.2	0.7	0.2	65.8	62.7
0.4	−0.4	1.2	0.7	0.2	92.3	92.6
0.8	−0.4	1.2	0.7	0.2	98.7	99.8
0.3	−0.4	1.2	1	0.2	86.2	85.9
0.3	−0.4	1.2	1.5	0.2	86.4	85.9
0.3	−0.4	1.2	2	0.2	86.5	85.9
0.3	−0.4	1.2	4	0.2	85.5	85.9
0.4	−0.4	1.2	0.7	−0.8	92.7	92.6
0.4	−0.4	1.2	0.7	−0.4	92.2	92.6
0.4	−0.4	1.2	0.7	0.4	91.4	92.6
0.4	−0.4	1.2	0.7	0.8	92.0	92.6

*Empirical power was based on the two-step inferential approach in 1000 simulations, each with 150 subjects per arm. Minimum follow-up time is 0.75 y (9 months), and maximum follow-up time is 2 y. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = -0.3$, $E(\theta_{0i}) = 0$, and $E(\theta_{1i}) = 3$. The longitudinal data are measured at years 0, 0.5, 1, 1.5 and at exit with a linear trajectory and $\sigma_e^2 = 0.16$.

Table V

Effect of β on the estimation of direct treatment effect on survival (α) based on different models.

	$\lambda_i(t) = \lambda_0(t)\exp(\alpha Z_i)$	$\lambda_i(t) = \lambda_0(t)\exp\{\beta(\theta_{0i} + \theta_{1i})t + \alpha Z_i\}$		
β	$\exp(\hat{\alpha})$ * based on Cox partial likelihood	$\exp(\hat{\alpha})$ based on known trajectory	$\exp(\hat{\alpha})$ based on two-step approach (partial likelihood)†	$\exp(\hat{\alpha})$ based on full joint likelihood as specified in (3)
0	0.668 (0.062)	0.667 (0.062)	0.667 (0.062)	0.667 (0.062)
0.4	0.697 (0.057)	0.668 (0.053)	0.667 (0.053)	0.667 (0.053)
0.8	0.755 (0.063)	0.670 (0.050)	0.673 (0.050)	0.668 (0.051)
1.2	0.800 (0.068)	0.670 (0.049)	0.684 (0.051)	0.668 (0.051)

*$\exp(\hat{\alpha})$ is the average value based on 1000 simulations, each with 200 subjects per arm. Minimum follow-up time is set to be 0.75 y (9 months), and maximum follow-up time is set to be 2 y. The baseline hazard is assumed constant with $\lambda_0 = 0.85$, and the true direct treatment effect on survival $\alpha = -0.4$ (i.e. HR = 0.670).

†Longitudinal data is measured at years 0, 0.5, 1, 1.5 and at exit with a linear trajectory and $\sigma_e^2 = 0.16$.

Table VI

Parameter estimates with standard errors for the E1193 data.

Parameters	Cox model with treatment only	Two-step model	Joint model
Overall treatment ($\hat{\alpha} + \hat{\beta}\hat{\gamma}$)	0.251 (0.1302)	0.261(0.1304)	0.271 (0.1413)
$\hat{\alpha}$			0.245 (0.1362)
$\hat{\gamma}$			−0.073 (0.1291)
β		−0.277 (0.0708)	−0.445 (0.1184)