JOINT MODELLING OF LONGITUDINAL AND TIME-TO-EVENT DATA APPLIED TO GROUP SEQUENTIAL CLINICAL TRIALS

Abigail J. Burdon1,∗, Lisa V. Hampson2 and Christopher Jennison3

1MRC Biostatistics Unit, University of Cambridge, Robinson Way, Cambridge, CB2 0SR, U.K
2Statistical Methodology, Novartis Pharma AG, Basel BA2 7AY, Switzerland
3Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, U.K
∗Email: abigail.burdon@mrc-bsu.cam.ac.uk

November 30, 2022

ABSTRACT

Often in Phase 3 clinical trials measuring a long-term time-to-event endpoint, such as overall survival or progression-free survival, investigators also collect repeated measures on biomarkers which may be predictive of the primary endpoint. Although these data may not be leveraged directly to support early stopping decisions, can we make greater use of these data to increase efficiency and improve interim decision making? We present a joint model for longitudinal and time-to-event data and a method which establishes the distribution of successive estimates of parameters in the joint model across interim analyses. With this in place, we can use the estimates to define both efficacy and futility stopping rules. Using simulation, we evaluate the benefits of incorporating biomarker information and the affects on interim decision making.

Keywords

Efficient designs, group sequential, joint modelling, longitudinal data, time-to-event data.

1 Introduction

Interest in joint modelling is motivated by clinical trials where the biomarker is predictive of a time-to-event (TTE) outcome. Overall survival and progression-free survival are examples of TTE endpoints. For example, Goldman et al. [1996] use CD4 lymphocyte cell counts as a surrogate endpoint for survival in a clinical trial comparing the efficacy and safety of two antiviral drugs for HIV-infected patients. Taylor et al. [2013] use joint models to predict survival times of patients with prostate cancer based on prostate specific antigen (PSA) levels measured by blood tests at multiple hospital visits.

To date, most group sequential designs for survival trials focus on a single primary endpoint. Designs which leverage repeated measurements on continuous or binary endpoints have been reported such as those by Galbraith and Marschner [2003] and this work extends this idea to survival trials with repeated measurements on a predictive biomarker. We shall focus on monitoring randomised control trials so that patients are randomised at baseline to receive either a novel treatment or control and the objective is to test for superiority of the treatment versus control. Hence, let η denote the treatment effect in a statistical model and let η0 be the true value of η, then we shall test the null hypothesis

\[ H_0 : \eta_0 \leq 0 \quad \text{vs} \quad H_A : \eta_0 > 0. \]  (1)

In the group sequential setting with a total of K analyses, we shall test \( H_0 \) at analysis \( k \) by defining treatment effect estimates \( \hat{\eta}^{(k)} \), information levels \( I_k \) and standardised Z−statistics given by \( -\hat{\eta}^{(k)} \sqrt{I_k} \) for \( k = 1, \ldots, K \). The aim will then be to determine the upper boundary constants \( b_1, \ldots, b_K \) and lower boundary constants \( a_1, \ldots, a_K \) for the group sequential test as in Figure 1. These boundaries are calculated to ensure such that the Type 1 error rate is \( \alpha \) and power is \( 1 - \beta \) when \( \eta = -0.5 \). In particular, we shall use the error spending test, also known as Gordon Lan and DeMets [1983], to calculate the boundary points.
We shall develop methods for designing and analysing a group sequential trial based on a joint model for longitudinal and TTE data. We may believe that a trend in the trajectory of the biomarker is predictive of the TTE outcome, and we would like to know whether these additional longitudinal data can be used to inform early stopping decisions. Suppose that biomarker observations are available but have not been used in the analysis. We shall assess the change in efficiency of the trial when these observations are included in the analysis. We shall focus on efficiency measured in terms of the number of patients that need be recruited to achieve a certain power, and we show that, in some scenarios, the trial using the longitudinal data is up to 1.7 times as efficient as the trial which discards the longitudinal data.

2 Joint modelling

2.1 Joint model

The joint model that we consider is given in Equation (2) by Tsiatis and Davidian [2001]. We shall refer to the authors as TD for short. There are two processes in this model which represent the survival and longitudinal parts separately, and these processes are linked through the hazard rate of the survival process. First we consider the longitudinal data. Suppose that $X_i(t)$ is the true value of the biomarker at time $t$ for subject $i$ and that $W_i(t)$ is the observed value of the biomarker at time $t$ for patient $i$. Then the longitudinal model takes the form

$$X_i(t) = b_{i0} + b_{i1}t$$

$$W_i(t) = X_i(t) + \epsilon_i(t)$$

where $b_i = (b_{i0}, b_{i1})$ is a vector of patient specific random effects and $\epsilon_i(t)$ is the measurement error. In general, the vector $b_i$ can have dimension $p$ and the function $X_i(t)$ need not be constrained to a linear function of $t$. We consider a random effects model where each $b_i$ is a random variable with density function $f(\cdot)$. The measurement errors are assumed to be independent and if the biomarker for patient $i$ is measured at times $t_{i1}, \ldots, t_{im}$, then
\(\epsilon_i(t_{ij})|b_i \sim N(0, \sigma^2)\) for \(j = 1, \ldots, m_i\) and \(\epsilon(t)\) and \(\epsilon(t')\) are independent for \(t \neq t'\). For now, we shall assume that \(\sigma^2\) is known and we later describe how \(\sigma^2\) can be estimated.

The model for the survival endpoint is a Cox proportional hazards model in which the underlying trajectory \(X_i(t)\) acts as a time-varying covariate with coefficient \(\gamma\). Let \(Z_i\) be an indicator function that patient \(i\) receives the experimental treatment and let \(\eta\) be the corresponding treatment effect. Then the hazard function is given by

\[
h_i(t) = h_0(t) \exp\{\gamma X_i(t) + \eta Z_i\},
\]

where \(h_0(\cdot)\) is the baseline hazard function. In general, \(Z_i\) may be a \(p \times 1\) column vector of coefficients and \(\eta\) is the corresponding coefficient vector of length \(p\). In this model, \(\gamma\) is the parameter that determines the correlation between the longitudinal data and the TTE endpoint. We show in Section 4 that if \(\gamma = 0\) then the investigator may pay a small penalty for having tried to leverage the biomarker data. However, if \(\gamma > 0\), then there is a large benefit from fitting this joint model to the data. Together, Equations (2), (3) and (4) define the joint model.

### 2.2 Conditional Score

In the fixed sample setting, TD present the “conditional score” method for fitting the joint model to the data. The method adapts the general theoretical work by Stefanski and Carroll [1987] who find unbiased score functions by conditioning on certain parameter-dependent sufficient statistics. This is a desirable method because the analysis is semi-parametric so that there are no distributional assumptions required for the random effects. Further, the authors show that the parameter estimates are normally distributed and unbiased. We shall now extend the fixed sample theory for the joint model to group sequential trials. To perform a group sequential trial with \(K\) analyses, we need to know the joint distribution of the sequence of treatment effect estimates that will be obtained at analyses \(k = 1, \ldots, K\).

To determine this distribution, we shall define group sequential versions of all objects included in the single-stage conditional score, which are calculated using data obtained at that analysis. Equivalent definitions of TD’s single-stage conditional score, and associated functions, can be found by setting \(K = k = 1\)

The conditional score methodology builds upon the theory of counting processes. In the general definition, a counting process is a step-function increasing in integer increments and the survival counting process is a step function jumping from 0 to 1 at the failure time for an uncensored observation.

The censoring mechanism is used to keep patients in the at-risk set who have yet to experience an event. For patient \(i\) with time-to-failure random variable \(F_i\), let \(C_i(k)\) be the time-to-censoring random variable at analysis \(k\). This censoring event includes “end of study” censoring for the total follow-up time of patient \(i\) at analysis \(k\), then at analysis \(k\) the event time random variable is \(T_i(k) = \min(F_i, C_i(k))\). The observed event time at analysis \(k\) is \(t_i(k)\) and the observed censoring indicator is \(\delta_i(k) = I\{F_i \leq C_i(k)\}\). In the conditional score approach, to be included in the at-risk set at time \(t\) the patient must have at least two longitudinal observations to fit the longitudinal regression model. The at-risk process at analysis \(k\) is an indicator for not yet observing the event, not yet censored, or not having enough longitudinal observations. Let \(v_{i2}\) be the time of the second longitudinal observation for patient \(i\), then at analysis \(k\), the at-risk process and counting process for the joint model are

\[
Y_i(k, t) = I\{t_i(k) \geq t, v_{i2} \leq t\}
\]

\[
N_i(k, t) = I\{t_i(k) \leq t, \delta_i(k) = 1, v_{i2} \leq t\}.
\]

The function

\[
dN_i(k, t) = N_i(k, t + dt) - N_i(k, t^-) = I\{t \leq t_i(k) < t + dt, \delta_i(k) = 1, v_{i2} \leq t\}
\]

presents us with useful notation: for any function or stochastic process \(f(\cdot)\), the stochastic integral

\[
\int_0^\infty f(u)dN_i(k, u) = f(t_i)
\]

is \(f\) evaluated at the place where \(N_i(k, t)\) jumps from 0 to 1 if \(\delta_i(k) = 1\) and \(v_{i2} \leq t\), and 0 otherwise.

By analogy to survival analysis, we seek a compensated counting process with expectation zero. This property leads us to define an estimating equation from which we can obtain treatment effect estimates that are asymptotically normally distributed by Section 5.3 by Van der Vaart [2000]. In the usual survival analysis setting, we can calculate the compensated counting process by subtracting the intensity process from the counting process itself. In the joint modelling setting, this is not as simple because the randomness of the nuisance parameters \(b\) mean that the intensity process is unpredictable. To overcome this, TD introduce a “conditional intensity process” which is conditional on
a certain “sufficient statistic”. The origins of the conditional intensity process and sufficient statistic are not crucial for our purpose. What we actually use are the definitions and properties that are derived from these. In what follows, we shall present TD’s definitions of the single-stage conditional intensity process, sufficient statistic and compensated counting process and we extend these to the group sequential versions. We shall then show that the group sequential compensated counting process has expectation zero.

For patient $i$, let $v_i(u)$ be set of all time points for measurements of the biomarker, up to and including time $u$. Let $X_i(u)$ be the ordinary least squares estimate of $X_i(u)$ for patient $i$ based on the set of measurements taken at times $v_i(u)$. That is, calculate $b_{0i}(u)$ and $\hat{b}_{1i}(u)$ based on measurements taken at times $v_i(u)$, then $X_i(u) = b_{0i}(u) + \hat{b}_{1i}(u)u$. As we pass time $v_{ij}$, a new observation $W_{ij}$ is generated and the formula for $X_i(u)$ is updated for larger values of $u$. This seems strange since at an early time point, $s$ where $s < u$, we use data $v_i(s)$ in the calculation of $X_i(s)$ even though there may be more data available at time points $v_i(u)$. However, this is necessary for the martingale property to hold in later results. Suppose that $\sigma^2\hat{\theta}_i(u)$ is the variance of the estimator $\hat{X}_i(u)$ at time $u$. TD define the sufficient statistic to be the function

$$S_i(k,t,\gamma,\sigma^2) = \hat{X}_i(t) + \gamma \sigma^2 \hat{\theta}_i(t)dN_i(k,t)$$

which is defined for all $t \in (v_{i2}, t_i)$ for patient $i$. The corresponding conditional intensity process and compensated counting process are defined by

$$\lambda_i^C(k,t) = \lim_{dt \downarrow 0} \frac{P(dN_i(k,t) = 1|S_i(k,t,\gamma,\sigma^2), t_i(t), Z_i, Y_i(k,t))}{dt}$$

$$= h_0(t) \exp\{\gamma S_i(k,t,\gamma,\sigma^2) - \gamma^2 \sigma^2 \hat{\theta}_i(t)/2 + \eta^T Z_i\}Y_i(k,t)$$

$$M_i(k,t) = N_i(k,t) - \int_0^t \lambda_i^C(k,t)du$$

$$dM_i(k,t) = dN_i(k,t) - \lambda_i^C(k,t).$$

We show below that this compensated counting process has expectation zero conditional on $S_i(k,t,\gamma,\sigma^2)$ and use this result to obtain the asymptotic distribution of some parameter estimates in the joint model. Specifically, we will determine the distribution of the estimates $\hat{\gamma}_i^{(k)}, \hat{\eta}_i^{(k)}$ and $\hat{\sigma}^{(k)}$ which are unbiased estimates, at analysis $k$, of $\gamma, \eta$ and $\sigma^2$ respectively.

**Lemma 1.** The function $dM_i(k,t)$ has expectation zero conditional on the sufficient statistic, that is

$$\mathbb{E}(dM_i(k,t)|S_i(k,t,\gamma,\sigma^2), t_i(t), Z_i, Y_i(k,t)) = 0.$$  

**Proof.** See Tsiatis and Davidian [2001].

TD present the fixed sample conditional score in their Equation (6). We present some functions that are needed to define the group sequential conditional score at analysis $k$ and the derivate of such an object with respect to parameters $\gamma$ and $\eta$. Let

$$J_i(k,t,\gamma,\sigma^2) = \{S_i(k,t,\gamma,\sigma^2) - \hat{X}_i(t) - \gamma \sigma^2 \hat{\theta}_i(t),0,\ldots,0\}^T, \quad \Gamma_i(k,t,\sigma^2) = \begin{bmatrix} \sigma^2 \hat{\theta}_i(t) & 0 \\ 0 & 0 \end{bmatrix}$$

be a $(p+1) \times 1$ vector and $(p+1) \times (p+1)$ matrix respectively. Now, for the following functions, we drop the dependency of all functions on the parameters $k$, $t$, $\gamma$, $\eta$ and $\sigma^2$. Then the functions of interest are
We shall now prove that the estimates $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\sigma}^2$ in the joint model

$$U_c(k, \gamma, \eta, \sigma^2) = \int_0^{\tau_k} \sum_{i=1}^n \left( \{ S_i(k, u, \gamma, \sigma^2), Z_i^T \}^T - E_c(k, u, \gamma, \eta, \sigma^2) \right) dN_i(k, u).$$

### 2.3 Differentiation of the conditional score

Another object of importance is the first derivative of the conditional score function, Equation (9), with respect to parameters $\gamma$ and $\eta$. This matrix plays a key role in the definition of the covariance matrix for the estimates $\hat{\gamma}$ and $\hat{\eta}$ and has a likeness to the Fisher information matrix which is the derivative of the score statistic for general statistical models. TD comment that the variance matrix can be found, however they do not present an equation for such an object. Details for the following calculation can be found in the supplementary materials. The first derivative of the conditional score of Equation (9) with respect to $\gamma$ and $\eta$ is

$$\frac{\partial}{\partial(\gamma, \eta)^T} U_c(k, \gamma, \eta, \sigma^2) = \sum_{i=1}^n \int_0^{\tau_k} \left[ \Gamma_i(k, u, \sigma^2) - V_c^{(1)}(k, u, \gamma, \eta, \sigma^2) + V_c^{(2)}(k, u, \gamma, \eta, \sigma^2) + V_c^{(3)}(k, u, \gamma, \eta, \sigma^2) \right] dN_i(k, u).$$

### 2.4 Asymptotic theory for parameter estimates $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\sigma}^2$ in the joint model

We shall now prove that the estimates $\hat{\gamma}^{(1)}$, $\hat{\eta}^{(1)}$, ..., $\hat{\gamma}^{(K)}$, $\hat{\eta}^{(K)}$ are asymptotically multivariate normally distributed and we shall derive an explicit form for the covariance matrix of this vector of parameters. In what follows, we shall determine the limiting distribution as $n \to \infty$. For clarity, we denote the objects $\hat{\gamma}_n$, $\hat{\eta}_n$ and $U_c(n)(k, \gamma, \eta, \sigma)$ as dependent on $n$.

**Theorem 1.** Suppose that $\gamma_0$, $\eta_0$ and $\sigma_0^2$ are the true values of the parameters $\gamma$, $\eta$ and $\sigma^2$ respectively. For each $k = 1, \ldots, K$, let $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$ be the values of $\gamma$ and $\eta$ which are the solution to the equation $(U_c(n)(k, \gamma, \eta, \sigma_0) = 0$ and suppose that Conditions 2 hold. Then the vector $(\hat{\gamma}_n^{(1)}, \hat{\eta}_n^{(1)}, \ldots, \hat{\gamma}_n^{(K)}, \hat{\eta}_n^{(K)})^T$ converges in distribution to a multivariate Gaussian random variable, specifically

$$n^{1/2} \begin{pmatrix} \hat{\gamma}_n^{(1)} - \gamma_0 \\ \hat{\eta}_n^{(1)} - \eta_0 \\ \vdots \\ \hat{\gamma}_n^{(K)} - \gamma_0 \\ \hat{\eta}_n^{(K)} - \eta_0 \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1K} \\ \Sigma_{12} & \Sigma_{22} & \cdots & \Sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{1K} & \Sigma_{2K} & \cdots & \Sigma_{KK} \end{bmatrix} \right)$$
where
\[ \Sigma_{k_1 k_2} = (A^{(k_1)})^{-1} B^{(k_1)} ((A^{(k_2)})^{-1})^T \] (11)
and the matrices \( A^{(k)} \) and \( B^{(k)} \) are defined by
\[ A^{(k)} = \int_0^{\tau_k} \left[ E(T_i(k, u, \sigma_0^2)) - v_c^{(1)}(k, u, \gamma_0, \eta_0, \sigma_0^2) - v_c^{(2)}(k, u, \gamma_0, \eta_0, \sigma_0^2) \right] s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) du \] (12)
\[ B^{(k)} = \int_0^{\tau_k} v_c^{(1)}(k, u, \gamma_0, \eta_0, \sigma_0^2) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) du. \] (13)

Proof. We begin this proof by considering the set of \( K(p+1) \) stacked equations
\[ \begin{bmatrix} U_c(1, \gamma, \eta, \sigma^2) \\ \vdots \\ U_c(K, \gamma, \eta, \sigma^2) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \] (14)
and we shall show that these equations define a multidimensional estimating equation. That is, for each \( k = 1, \ldots, K \), the function \( U_c(k, \gamma, \eta, \sigma^2) \) has expectation zero. The arguments given by TD about asymptotic normality in the fixed sample case apply for each \( k = 1, \ldots, K \). We present some of these arguments which are useful for the derivation of the covariance matrix in the group sequential setting. We shall write \( U_c(k, \gamma, \eta, \sigma^2) \) as
\[ \int_0^{\tau_k} \sum_{i=1}^n \left( \{ S_i(k, u, \gamma, \sigma^2), Z_i^T \}^T - c_c(k, u, \gamma, \eta, \sigma^2) \right) dM_i(k, u) \] (15)
\[ + \int_0^{\tau_k} \sum_{i=1}^n \left( c_c(k, u, \gamma, \eta, \sigma^2) - E_c(k, u, \gamma, \eta, \sigma^2) \right) dM_i(k, u). \] (16)
By the same argument as TD, \( n^{-1} \) times Expression (16) converges in probability to zero in a neighbourhood of \((\gamma_0, \eta_0)\) and we deduce that the behaviour of the estimates \( \hat{\gamma}(k) \) and \( \hat{\eta}(k) \) will be dictated by Expression (15). Then we have that the expectation of Expression (15) is
\[ E \left( \int_0^{\tau_k} \sum_{i=1}^n \left( \{ S_i(k, u, \gamma, \sigma^2), Z_i^T \}^T - c_c(k, u, \gamma, \eta, \sigma^2) \right) dM_i(k, u) \right) = 0 \] (17)
for each \( k = 1, \ldots, K \). Therefore combined with the fact that Expression (16) converges in probability to zero, we have that Equation (14) defines a multidimensional estimating equation. Therefore, the estimates \( (\hat{\gamma}(k), \hat{\eta}(k)) \) for each \( k = 1, \ldots, K \) are asymptotically multivariate normal and unbiased for parameters \( \gamma \) and \( \eta \) by Section 5.3 by Van der Vaart [2000]. It remains to determine the covariance matrix.

In the general setting for statistical models, the sandwich estimator can be used to robustly estimate the variance matrix for estimates that are the solutions to estimating equations as in Section 2.6 by Wakefield [2013]. The sandwich estimator requires deriving \( B = \text{Var}(U(\theta)) \) and \( A = \partial U(\theta) / \partial \theta \) where \( U(\cdot) \) is an estimating function for parameter \( \theta \). The similarity here is with matrices \( A^{(k)} \) and \( B^{(k)} \) in Equations (12) and (13) which are the sequential versions of such objects for the conditional score. Wakefield [2013] prove the asymptotic normality of estimates obtained using estimating equations and the asymptotic convergence of the sandwich estimator. The proof by Wakefield [2013] applies to scalar estimates and one-dimensional estimating functions so we shall extend this proof to derive the covariance matrix for the vector of parameters in the joint model as oppose to the scalar sandwich variance.

Following Wakefield [2013], we apply the standard Taylor expansion results to each row of Equation (14) and aggregate the results. Details of this calculation are found in the supplementary materials. Let \( A \) be the block diagonal matrix whose \( k^{th} \) diagonal matrix is the \( p \times p \) matrix \( A^{(k)} \). Then we obtain
\[ -n^{-\frac{1}{2}} \tilde{A}^{-1} \begin{bmatrix} U_c(1, \gamma, \eta, \sigma^2) \\ \vdots \\ U_c(K, \gamma, \eta, \sigma^2) \end{bmatrix} = \tilde{A}^{-1} \begin{bmatrix} n^{-1} \frac{\partial}{\partial (\gamma, \eta)} U_c(1, \gamma, \eta, \sigma^2) |_{(\gamma^{*(1)}, \eta^{*(1)})} \\ \vdots \\ n^{-1} \frac{\partial}{\partial (\gamma, \eta)} U_c(K, \gamma, \eta, \sigma^2) |_{(\gamma^{*(K)}, \eta^{*(K)})} \end{bmatrix} \left( \begin{bmatrix} \gamma^{(1)} - \gamma_0 \\ \eta^{(1)} - \eta_0 \\ \vdots \\ \gamma^{(K)} - \gamma_0 \\ \eta^{(K)} - \eta_0 \end{bmatrix} \right) \]
where \( (\gamma^{*(k)}, \eta^{*(k)}) \) lies on the line segment between \((\gamma_0, \eta_0)\) and \((\hat{\gamma}^{(k)}, \hat{\eta}^{(k)})\).
Suppose that we could show that
\[ n^{-1} \frac{\partial}{\partial(\gamma, \eta^T)} \mathbb{E}_c(U_{c(1)}(k, \gamma, \eta, \sigma^2)) \left|_{\gamma=\gamma^*, \eta=\eta^*} \right. \quad \text{and} \quad (18) \]
then by an application of Slutsky's Theorem, we have the desired result.

A simple application of the triangle inequality proves that condition (18) holds. The proof of convergence in probability closely follows the standard results for survival data seen in Theorem VII.2.2 by Andersen and Gill [1982] and the details of this step are found in the supplementary materials.

To prove condition (19) holds, first note that Expression (15) is a sum over \( n \) independent and identically distributed terms. Then the multivariate Central Limit Theorem can be applied to the vector \( \{U_c(1, \gamma, \eta, \sigma^2), \ldots, U_c(K, \gamma, \eta, \sigma^2)\} \) to establish asymptotic normality. It then remains to show that \( \text{Cov}(U_c(k_1, \gamma, \eta, \sigma^2), U_c(k_2, \gamma, \eta, \sigma^2)) \xrightarrow{p} \mathbf{B}(k_1, k_2) \) for \( k_1 \leq k_2 \).

We now follow a similar structure to the partial likelihood function for survival data given by Jennison and Turnbull [1997] and create a new counting process that allows the conditional score statistic to be written as the sum of corresponding compensated counting process is therefore given by
\[ DM_i(k, t) = D_{N_i}(k, t) - \int_0^t h_0(u)E_{0i}(t, \gamma, \eta, \sigma^2)(Y_i(k, u) - Y_i(k-1, u))du \]
for \( k = 1, \ldots, K \) and \( DM_i(0, t) = 0 \). The event for patient \( i \) can only occur in one interval, and therefore we have \( N_i(k, t) = \sum_{j=1}^k D_{N_i}(l, t) \). For consistency with Andersen and Gill [1982] and Jennison and Turnbull [1997], let Expression (15) be denoted by \( W_j^{(n)}(k, \tau, \gamma, \eta, \sigma^2) \). Then by the definition of the difference in counting process, we have
\[ W_j^{(n)}(k, \tau, \gamma, \eta, \sigma^2) = \sum_{i=1}^n \sum_{l=0}^{k-1} H_{ij}^{(n)}(l, u, \gamma, \eta, \sigma^2) dDM_i(l, u) \]
where
\[ H_{ij}^{(n)}(l, u, \gamma, \eta, \sigma^2) = n^{-1/2} \{ \{ S_i(l, u, \gamma, \sigma^2), Z_i \} \gamma_{ij} - e_{ij}(l, u, \gamma, \sigma^2) \} \]
and \( W_j^{(n)}(k, \tau, \gamma, \eta, \sigma^2) \) has expectation zero by a simple manipulation of Equation (17).

The main result, with details given in the supplementary materials, is as follows
\[ \text{Cov} \left( W_{j_1}^{(n)}(k_1, \tau_{j_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{j_2}, \gamma_0, \eta_0, \sigma_0^2) \right) \]
\[ = \mathbb{E} \left( \sum_{i=1}^n \int_0^{\tau_{j_1}} H_{ij_1}^{(n)}(k_1, u, \gamma_0, \eta_0, \sigma_0^2) H_{ij_2}^{(n)}(k_2, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u)E_{0i}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k, u)du \right) . \quad (20) \]

To complete the proof, it remains to show that Equation (20) converges in probability to \( \mathbf{B}^{(k_1)} \) and we shall do so by using the definitions of the objects \( s_i^{(j)} \) and \( c_i^{(j)} \) for \( j = 1, 0, 1, 2 \) and \( e_i^{(1)} \) and \( v^{(2)} \) given in Appendix 1. We are assuming that these limits exist by Conditions 2 and we shall exploit the relationship between these terms. This working is exactly the same as for standard survival data given by Andersen and Gill [1982], with the details given in the supplementary materials, and we see that
\[ \text{Cov} \left( W_{j_1}^{(n)}(k_1, \tau_{j_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{j_2}, \gamma_0, \eta_0, \sigma_0^2) \right) \]
\[ \xrightarrow{p} \left( \int_0^\infty v_i^{(1)}(k_1, u, \gamma_0, \eta_0, \sigma_0^2) s_i^{(0)}(k_1, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u)du \right) \xrightarrow{j_1, j_2} \mathbf{B}^{(k_1)} . \]
By the argument that the behaviour of the estimates \( \hat{\gamma}^{(k)} \) and \( \hat{\sigma}^{(k)} \) is dictated by Expression (15), we have
\[ \text{Cov} \left( U_c(k_1, \gamma_0, \eta_0, \sigma_0^2), U_c(k_2, \gamma_0, \eta_0, \sigma_0^2) \right) \xrightarrow{p} \mathbf{B}^{(k_1)} \text{ for } k_1 \leq k_2 \]
which is the result.
Similar to the fixed sample case, we have assumed that $\sigma^2_0$ is known in the derivation of the distribution of $\gamma^{(k)}_n$ and $\hat{\eta}^{(k)}_n$. This is not generally the case but by arguments in Carroll et al. [2006] Section A.3.3, we can find a consistent estimate to replace $\sigma^2_0$ with in the group sequential conditional score function. At analysis $k$ this estimate is given by

$$\hat{\sigma}^{(k)2} = \frac{\sum_{i=1}^n \mathbb{I}\{m_i(k) > 2\} R_i(k)}{\sum_{i=1}^n \mathbb{I}\{m_i(k) > 2\}(m_i(k) - 2)}$$

where $R_i(k)$ is the residual sum of squares for the least squares fit to all $m_i(k)$ observations for patient $i$ available at analysis $k$.

### 2.5 A group sequential trial design based on the conditional score analysis

We shall use Theorem 1 to create a group sequential test based on the joint model. Let $\gamma_0, \eta_0$ and $\sigma^2_0$ be the true values of the parameters $\gamma, \eta$ and $\sigma^2$ respectively. Using the group sequential conditional score method, let $\hat{\gamma}^{(k)}, \hat{\eta}^{(k)}$ be the values of the parameters $\gamma$ and $\eta$ such that $U_c(k, \gamma, \eta, \sigma^2) = 0$ where the conditional score function is calculated using Equation (9). Further, let $\hat{\sigma}^{(k)2}$ be the estimate for $\sigma^2$ given in Equation (21). By Theorem 1, for each $k = 1, \ldots, K$, the marginal distribution of the parameter $\hat{\eta}^{(k)}$ is

$$\sqrt{n}(\hat{\gamma}^{(k)} - \gamma_0) \xrightarrow{d} N(0, \Sigma_{22}^{(k)})$$

where

$$\Sigma^{(k)} = (A^{(k)})^{-1}B^{(k)}((A^{(k)})^{-1})^T$$

and the matrices $A^{(k)}$ and $B^{(k)}$ are defined by Equations (12)–(13). Note that the subscript notation in the covariance matrix represents that the parameter $\eta$ is the second parameter in the vector $(\gamma, \eta)$. In the more general case where $Z_i$ is a vector of dimension $p$ including other covariates apart from treatment indicator, then the information would be indexed by the corresponding dimension of the vector $(\gamma, \eta)^T$ which relates to the treatment effect. The matrices $A^{(k)}$ and $B^{(k)}$ are estimated using

$$\hat{A}^{(k)} = \frac{1}{n} \sum_{i=1}^n \int_0^\infty \left[ \Gamma_i(k, u, \hat{\sigma}^{(k)2}) - V_c^{(1)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) + V_c^{(2)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) \right] dN_i(k, u)$$

$$\hat{B}^{(k)} = \frac{1}{n} \sum_{i=1}^n \int_0^\infty \left[ V_c^{(1)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) \right] dN_i(k, u).$$

The information matrix at analysis $k$ of the group sequential trial is given by

$$I_k = \frac{1}{n} \left[ (\hat{A}^{(k)})^{-1} \hat{B}^{(k)}((\hat{A}^{(k)})^{-1})^T \right]_{22}^{-1}.$$ 

Further, a standardised test statistic at analysis $k$ is given by

$$Z_k = \hat{\eta}^{(k)} \sqrt{I_k}.$$ 

### 3 Designing group sequential trials when the canonical joint distribution does not hold

#### 3.1 Simulation evidence that the trial is conservative with respect to type 1 error rate

We discuss the implications for a group sequential test when the sequence of test statistics does not follow the usual canonical joint distribution and we shall show that in some scenarios, it is appropriate to make this assumption anyway and proceed with the trial as though the canonical joint distribution holds anyway. Let $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)}$ be the sequence of treatment effect estimates in a group sequential trial from data available at analyses $1, \ldots, K$ respectively and $I_1, \ldots, I_K$ are the associated observed information levels. The “canonical joint distribution” for the sequence of estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)}$ is such that

1. $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)})$ is multivariate normal
2. $\hat{\theta}^{(k)} \sim N(\theta, I_k^{-1}), \quad 1 \leq k \leq K$
3. $\text{Cov}(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) = I_k^{-1}, \quad 1 \leq k_1 \leq k_2 \leq K.$
Supposing that the canonical joint distribution holds for the sequence of treatment effect estimates in the joint model, then the library of standard group sequential designs, such as Pocock [1977] and O’Brien and Fleming [1979], could be directly applied in our setting.

In Section 2 we saw that asymptotically the sequence of treatment effect estimates in a group sequential trial obtained from the joint model is multivariate normally distributed. Further, each of these estimates is asymptotically unbiased. The first two conditions for the canonical distribution of the sequence of test statistics are satisfied. However, for the joint model, we have shown that

\[
\text{Var}(\hat{\eta}(k)) = \left[(A^{(k)})^{-1}B^{(k)}(A^{(k)})^{-1}\right]_{22} \quad \text{for } k = 1, \ldots, K
\]

and

\[
\text{Cov}(\hat{\eta}^{(k)}_1, \hat{\eta}^{(k)}_2) = \left[(A^{(k)})^{-1}B^{(k)}(A^{(k)})^{-1}\right]_{22} \quad \text{for } k_1 < k_2.
\]

This implies that the third condition for the canonical joint distribution, which is the Markov property, is not satisfied. In this section, we discuss the implications of performing a group sequential trial when the assumption of a canonical joint distribution fails. We first show that there are some small differences between the matrices \((A^{(k)})^{-1}B^{(k)}\) and the identity matrix and describe why this difference is important. Then, we discuss some alternative methods which aim to correct for this violation of the Markov property. In method 1, the trial is performed acting as though the canonical joint distribution holds, and we present some theory that this method controls the type 1 error rate conservatively.

The theory presented is for the case when a non-binding futility boundary is used which is where stopping for futility at an interim analysis is not mandatory as described in GUIDANCE [2018]. FDA guidance recommends using non-binding futility rules because if binding rules are employed and not followed, then type 1 error rates are inflated. The calculation of the type 1 error rate therefore only depends on the upper boundary \(b_1, b_2\) which is depicted in Figure 1.

The problem of not obtaining the canonical joint distribution is not unique to the conditional score method and the proof that this assumption holds is not always trivial. Slud and Wei [1982] design a group sequential test which uses the modified-Wilcoxon statistic for two-sample survival data. The authors show that the increments in test statistics are correlated and their alternative proposed method has a similar structure to our method 2. Then, Gu and Ying [1993] present a score process for the analysis of regression data under general right censorship and they implement the repeated significance method of Slud and Wei [1982] for the sequential analysis of such data. Further, Cook and Lawless [1996] discuss a modification of the error spending approach (also known as Gordon Lan and DeMets [1983]) for sequences of treatment effect estimates that do not have the independent increments structure.

If the relationship \(A^{(k)} = B^{(k)}\) holds for each \(k = 1, \ldots, K\), then

\[
\text{Cov}(\hat{\eta}^{(k)}_1, \hat{\eta}^{(k)}_2) = [\left(A^{(k)}\right)^{-1}]_{p+1,p+1} = \text{Var}(\hat{\eta}^{(k)}_2)
\]

and the third condition holds. Therefore, we shall measure the magnitude of the violation by considering the matrix \((A^{(k)})^{-1}B^{(k)}\) and to what extent it differs from the identity matrix. By definition, the variance of the estimate \(\hat{\eta}^{(k)}\) is found in the bottom right entry of the matrix \((A^{(k)})^{-1}B^{(k)}(A^{(k)})^{-1}\) and hence, the bottom row of \((A^{(k)})^{-1}B^{(k)}\) is of interest here.

We can find estimates \(\hat{A}^{(k)}\) and \(\hat{B}^{(k)}\) from simulated data. We have done this using a large sample size of \(n=4800\) patients to reduce noise in these estimates. This calculation is computationally expensive and takes roughly two hours to compute for this value of \(n\). This is appropriate because, although both matrices depend on the sample size \(n\), they can each be written in the form \(\hat{A}^{(k)} = (1/n) \sum_{i=1}^{n} X_i(k, \gamma, \eta)\) and \(\hat{B}^{(k)} = (1/n) \sum_{i=1}^{n} Y_i(k, \gamma, \eta)\) for some functions \(X_i(k, \gamma, \eta)\) and \(Y_i(k, \gamma, \eta)\). Therefore, in the formula \((A^{(k)})^{-1}B^{(k)}\), the value of \(n\) cancels out, and we are left with a function that converges in distribution to \((A^{(k)})^{-1}B^{(k)}\) as \(n \to \infty\). Further, to reduce simulation error, the true values of \(\gamma\) and \(\eta\) are used in this calculation, which is appropriate because of the consistency of the estimates \(\hat{\gamma}\) and \(\hat{\eta}\).

Table 1 shows the matrix \((A^{(1)})^{-1}B^{(1)}\) for \(\gamma = 0, 0.03, 0.06, 0.09, \sigma^2 = 0, 1, 10, 100\) and \(\eta = 0\). The matrices \(A^{(k)}, \hat{A}^{(k)}, B^{(k)}\) and \(\hat{B}^{(k)}\) are each of dimension \(2 \times 2\). The function \(V^{(2)}(k, u, \gamma, \eta)\) is calculated such that \(\left[V^{(2)}(k, u, \gamma, \eta)\right]_{12} = [V^{(2)}(k, u, \gamma, \eta)]_{22} = 0\), and hence by Equations (22) and (23) it can be shown that \(\left[A^{(k)}\right]_{12} = [\hat{B}^{(k)}]_{12}\) and \(\left[A^{(k)}\right]_{22} = [\hat{B}^{(k)}]_{22}\). Further simple algebraic manipulation gives \([\left(A^{(k)}\right)^{-1}B^{(k)}]_{12} = 0\) and \([\left(A^{(k)}\right)^{-1}B^{(k)}]_{22} = 1\) exactly, which is shown in Table 1. The fact that \([\left(A^{(1)}\right)^{-1}B^{(1)}]_{11}\) is a long way from 1 for \(\sigma^2 = 10\) and \(\sigma^2 = 100\) is therefore not a problem. As \(\sigma^2\) increases, the absolute value of \([\left(A^{(1)}\right)^{-1}B^{(1)}]_{21}\) increases.
but the value of $\gamma$ has a small impact on the value of $[(A^{(1)})^{-1}B^{(1)}]_{21}$. Thus, we may expect large values of $\sigma^2$ to affect the achieved type 1 error rate.

3.2 Method 1 - canonical joint distribution assumed

We consider alternative methods for creating a group sequential trial when the canonical joint distribution does not hold. In the first method, we construct the group sequential test by estimating $\text{Var}(\hat{\eta}^{(k)})$, $k = 1, \ldots, K$ from the data and supposing $\text{Cov}(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)})$ for $k_1 < k_2$ are as specified in the canonical joint distribution. We shall prove that we have type 1 error rate less than $\alpha$ and we also show, through simulation, that this method performs satisfactorily in practice with error rates diverging minimally from planned significance and power.

For this proof, we consider the case where $K = 2$ and the futility boundary is non-binding. We present a sketch proof in the supplementary materials for the case $K = 3$ and we believe that the results generalise for cases $K > 3$. To prove that the type 1 error rates are conservative, we shall compare the probabilities of crossing the boundaries of a group sequential trial for two sequences of treatment effect estimates; one where the canonical joint distribution does not hold and one where this assumption does hold. For a group sequential trial with $K = 2$, suppose that $\hat{\eta}_1, \hat{\eta}_2$ are the sequence of treatment effect estimates that are calculated using the conditional score method. Let the true variance-covariance matrices for these estimates be $\Sigma$. Proceeding using method 1, we let the information levels be calculated as $\mathcal{I}_k = (\Sigma_{k:k})^{-1}$ and the $Z-$statistic is given by $Z_k = \hat{\eta}_k \sqrt{\mathcal{I}_k}$ for $k = 1, 2$. Under $H_0$ we have

$$\begin{bmatrix} \hat{\eta}_1 \\ \hat{\eta}_2 \end{bmatrix} \sim N_K \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right]$$

$$Z_1 \sim N_K \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right] \quad (26)$$

where the correlation parameter $\rho$ is given by

$$\rho = \text{Cov}(Z_1, Z_2) = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}} \quad (27)$$

Suppose instead, that we have a different sequence of treatment effect estimates $\hat{\eta}_1^*, \hat{\eta}_2^*$ with distribution given below. The values of $\Sigma_{11}$ and $\Sigma_{22}$ are the same as in (26) and using the same information levels given by $\mathcal{I}_k = (\Sigma_{k:k})^{-1}$, we define the standardised statistics $Z_k^* = \hat{\eta}_k^* \sqrt{\mathcal{I}_k}$ for $k = 1, \ldots, K$. The joint distribution of the sequence $\hat{\eta}_1^*, \hat{\eta}_2^*$ and the distribution of the $Z-$statistics are given by

$$\begin{bmatrix} \hat{\eta}_1^* \\ \hat{\eta}_2^* \end{bmatrix} \sim N_K \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right]$$

$$Z_1^* \sim N_K \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^* \\ \rho^* & 1 \end{bmatrix} \right] \quad (28)$$

with correlation parameter

$$\rho_{12}^* = \text{Cov}(Z_1^*, Z_2^*) = \sqrt{\frac{\Sigma_{22}}{\Sigma_{11}}} \quad (29)$$

This sequence of treatment effect estimates $\hat{\eta}_1^*, \hat{\eta}_2^*$ therefore has the canonical joint distribution and $Z_1^*, Z_2^*$ has the canonical joint distribution for a sequence of $Z-$statistics, with information levels $\mathcal{I}_k = 1/\Sigma_{k:k}$ for $k = 1, \ldots, K$. 

| Parameter values | $\sigma^2$ | $\gamma$ | $\rho$ |
|------------------|-----------|---------|-------|
| $\gamma$          | 0, 0.03, 0.06, 0.09 and $\sigma^2$ = 0, 1, 10, 100 in the joint model simulated with 4800 patients. |

Table 1: Matrix $(A^{(1)})^{-1}B^{(1)}$ for the null hypothesis $\eta = 0$. 

| $\sigma^2$ | 0 | 0.03 | $\gamma$ | 0.06 | 0.09 |
|------------|---|------|---------|------|------|
| 0          | (1.00 0.00) | (1.00 0.00) | (1.00 0.00) | (1.00 0.00) |
| 1          | (1.00 0.00) | (1.01 0.00) | (1.01 0.00) | (1.02 0.00) |
| 10         | (1.03 0.00) | (1.06 0.00) | (1.12 0.00) | (1.22 0.00) |
| 100        | (1.28 0.00) | (1.63 0.00) | (2.32 0.00) | (3.49 0.00) |

Parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ in the joint model simulated with 4800 patients.
The upper boundary points $b_1, b_2$ are calculated under the assumption that the canonical joint distribution holds so as to give a group sequential test with the correct type 1 error rate $\alpha$. Hence, we have that

$$\Pr_{\eta=0}(Z_1^* > b_1 \cup Z_2^* > b_2) = \alpha.$$ 

We consider the probability of rejecting $H_0$ when we apply this boundary to the sequence $Z_1, Z_2$. We aim to prove that

$$\Pr_{\eta=0}(Z_1 > b_1 \cup Z_2 > b_2) \leq \alpha. \quad (30)$$

The following conditions are needed for the proof that type 1 error rate is conservative.

**Conditions 1.**

1. The upper boundary of a group sequential trial, on the $Z-$scale, is such that $b_1 \geq b_2 \geq 0$.

2. We have $\Sigma_{12} \geq \Sigma_{22}$.

Condition 1 of Conditions 1 holds under common error spending functions with increasing information sequences, which are most often used in practice. Condition 2 of Conditions 1 should be checked by simulation before proceeding with the analysis. To do so, the investigator would choose sensible values for all the parameters in the joint model, simulate a large dataset of 4800 patients using these parameter values, and calculate an estimate for the variance-covariance matrix $\Sigma$ for the sequence of estimates $\hat{\eta}^{(1)}, \ldots, \hat{\eta}^{(K)}$. We have found that calculations for various examples has always lead to condition 2 being satisfied. Further, the scenarios that we have checked span a 3-dimensional grid of $\eta, \gamma$ and $\sigma^2$ values each ranging from small to large and hence, we believe that the scenarios we have checked span a suitable range of the parameter values. In the rare event that this condition does not hold, a solution is to employ method 2, which will be described later. It can also be seen by simple algebraic manipulation that condition 2 implies $\rho \geq \rho^*$.

The following theorem shows that Equation (30) holds.

**Theorem 2.** Let $Z_1, Z_2$ be the standardised statistics of a group sequential trial with distribution given by (26) and let $Z_1^*, Z_2^*$ be the statistics with distribution given by (28). Let $\alpha$ be the planned type 1 error rate and suppose that $b_1, b_2$ are the upper boundary points on the $Z-$scale such that

$$\Pr_{\eta=0}(Z_1^* > b_1 \cup Z_2^* > b_2) = \alpha.$$ 

Suppose that Conditions 1 hold. Then the Type 1 error rate when applying the boundary for $Z_1^*, \ldots, Z_K^*$ to $Z_1, \ldots, Z_K$ is

$$\Pr_{\eta=0}(Z_1 > b_1 \cup Z_2 > b_2) \leq \alpha.$$ 

**Proof.** The problem is equivalent to proving that

$$\Pr_{\eta=0}(Z_1 > b_1 \cup Z_2 > b_2) \leq \Pr_{\eta=0}(Z_1^* > b_1 \cup Z_2^* > b_2).$$

and by another representation for the above probabilities, we aim to show that

$$\Pr_{\eta=0}(Z_1 > b_1 \cup Z_2 > b_2) = \Pr_{\eta=0}(Z_1^* > b_1 \cup Z_2^* > b_2) \leq \Pr_{\eta=0}(Z_1 > b_1 \cup Z_2 > b_2) \leq \Pr_{\eta=0}(Z_1^* > b_1 \cup Z_2^* > b_2).$$

Under $\eta = 0$, the marginal distributions of $Z_1, Z_2, Z_1^*$ and $Z_2^*$ are equivalent and are all $N(0, 1)$ random variables and hence the probabilities are such that $\Pr_{\eta=0}(Z_k > b_k) = \Pr_{\eta=0}(Z_k^* > b_k)$ for each $k = 1, 2$. Therefore, the problem is reduced to showing that

$$\Pr_{\eta=0}(Z_1^* > b_1 \cap Z_2^* > b_2) \leq \Pr_{\eta=0}(Z_1 > b_1 \cap Z_2 > b_2). \quad (31)$$

when $\rho^* \leq \rho$.

In the below calculations, we appeal to the fact that for two random variables which are bivariate normally distributed, the conditional distribution of one normal random variable on the other normal random variable is also normal. Specifically, we have that $Z_2|Z_1 = z_1 \sim N(\rho z_1, 1 - \rho^2)$. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative distribution function of a standard normal random variable respectively, then the probability on the left hand side in Equation (31) is

$$\Pr_{\eta=0}(Z_1 > b_1 \cap Z_2 > b_2) = \int_{b_1}^{\infty} \left[ 1 - \Phi \left( \frac{b_2 - \rho z_1}{\sqrt{1 - \rho^2}} \right) \right] \phi(z_1)dz_1.$$ 

The corresponding calculation where $Z_1^*$ and $Z_2^*$ replace $Z_1$ and $Z_2$ yields the following

$$\Pr_{\eta=0}(Z_1^* > b_1 \cap Z_2^* > b_2) = \int_{b_1}^{\infty} \left[ 1 - \Phi \left( \frac{b_2 - \rho^* z_1}{\sqrt{1 - \rho^{*2}}} \right) \right] \phi(z_1)dz_1.$$
Method φ σ^2 Type 1 error Percentage of time method fails
γ = 0 γ = 0.03 γ = 0.06 γ = 0.09 γ = 0 γ = 0.03 γ = 0.06 γ = 0.09
1 2.5 0 0.022 0.022 0.023 0.023 0.000 0.000 0.000 0.000
1 2.5 1 0.024 0.025 0.026 0.023 0.000 0.000 0.000 0.000
1 2.5 10 0.023 0.024 0.023 0.025 0.000 0.000 0.000 0.000
1 2.5 100 0.023 0.028 0.026 0.024 0.000 0.000 0.000 0.000
2 2.5 0 0.022 0.022 0.023 0.023 0.000 0.000 0.000 0.000
2 2.5 1 0.024 0.025 0.026 0.023 0.000 0.000 0.000 0.000
2 2.5 10 0.023 0.024 0.023 0.025 0.040 0.100 0.120 0.060
2 2.5 100 0.022 0.028 0.026 0.029 49.140 37.360 35.950 40.420
3 2.5 0 0.028 0.024 0.023 0.023 0.000 0.000 0.000 0.000
3 2.5 1 0.028 0.026 0.027 0.023 0.000 0.000 0.000 0.000
3 2.5 10 0.029 0.025 0.023 0.026 0.140 0.050 0.060 0.310
3 2.5 100 0.024 0.028 0.027 0.029 1.930 1.950 1.840 1.850

Calculated using a simulation study with 365 patients and 10^4 replicates.
Standard error is 0.0016 when the method does not fail.

Table 2: Type 1 error rates and percentage of times that each method fails.

and since \( \Phi(\cdot) \) is strictly increasing, it suffices to show that whenever \( z_1 > b_1 \), then

\[
\frac{b_2 - \rho z_1}{\sqrt{1 - \rho^2}} \leq \frac{b_2 - \rho^* z_1}{\sqrt{1 - \rho^*}}.
\]

(32)

We have by assumption that \( \rho^* \leq \rho \). Further note that by definition \( 0 \leq \rho \leq 1 \) and \( 0 \leq \rho^* \leq 1 \). The following shows a simple algebraic manipulation of this inequality, which gives

\[
\rho^* \leq \rho \iff \frac{\sqrt{1 - \rho^*^2} - \sqrt{1 - \rho^2}}{\rho \sqrt{1 - \rho^*^2} - \rho^* \sqrt{1 - \rho^2}} \leq 1.
\]

Finally, using the above inequality and Conditions 1 that \( 0 \leq b_2 \leq b_1 \), we have for \( z_1 \geq b_1 \) that

\[
\frac{b_2 - \sqrt{1 - \rho^*^2} - \sqrt{1 - \rho^2}}{\rho \sqrt{1 - \rho^*^2} - \rho^* \sqrt{1 - \rho^2}} \leq b_2 \leq b_1 \leq z_1
\]

and a simple rearrangement shows that Equation (32) is satisfied.

Table 2 show estimates of type 1 error rate. This was by simulating \( 10^4 \) data sets, each with a sample size \( n = 365 \). This sample size was chosen because it gives power 0.9 for \( \gamma = 0.06, \sigma^2 = 1, \phi = 2.5 \) and for other parameter values, the power was close to 0.9. For each data set, we calculate the estimates \( \hat{\eta}^{(1)}, \ldots, \hat{\eta}^{(K)} \) and estimates of the covariance matrices \( \Sigma^{(1)}, \ldots, \Sigma^{(K)} \) using estimates \( A^{(1)}, \ldots, A^{(K)}, B^{(1)}, \ldots, B^{(K)} \) given by Equations (22) and (23). The boundary points \( a_1, \ldots, a_K \) and \( b_1, \ldots, b_K \) are then calculated under the assumption that the canonical joint distribution holds and the type 1 error rate is calculated as the proportion of replicates that reject the null hypothesis. This simulation study is computationally expensive and with \( 10^4 \) replicates, there is noise in the simulation results. Taking this into account, the simulation results support the relevance of asymptotic theory since all empirical type 1 error rates are within 2 standard deviations of 0.025.

3.3 Method 2 - Use the complete structure of the covariance matrix

The second method for dealing with estimates from the joint model does not rely on the canonical joint distribution assumption. Instead, the group sequential boundaries are calculated using the complete structure of the variance-covariance matrix for the sequence of treatment effect estimates across analyses. This differs from when the canonical joint distribution is assumed because in such a case, only the variances are required and the covariances are ignored. To calculate the boundary points of the group sequential test, we are required to calculate a \( K - \)dimensional integral over the joint probability distribution of the sequence of test statistics. This integration calculation can be performed numerically using the R package mvtnorm by Genz et al. [2020].

12
However, this method poses some practical difficulties. For example, during the conditional score method, the variance-covariance matrix $\Sigma$ in Equation (11) is estimated with error which can sometimes result in a non positive-definite estimate $\hat{\Sigma}$. In such a case, the boundary calculations cannot be performed. Slud and Wei [1982] allude to a similar problem in their sequential analysis of modified-Wilcoxon scores for two-sample survival data. Table 2 shows the percentage of times that this problem occurs. We see that for extremely noisy longitudinal data with $\sigma^2 = 100$ this problem occurs roughly 40% of the time and for $\sigma^2 = 10$ we have the problem occurring infrequently. This is not a problem for small $\sigma^2$ and in such a case, the type 1 error rates are as expected. In summary, this method makes no assumption about the covariance structure of the sequence of treatment effect estimates and therefore the type 1 error rate is preserved.

### 3.4 Method 3 - Create an asymptotically efficient estimate

For the final method, we follow the approach of Van Lancker et al. [2022] where a new estimator is created which reaches asymptotic efficiency as it is designed in such a way to minimise the variance. The efficient estimate at analysis $k$ is a linear combination of the original estimates at analyses up to and including $k$. We choose the weights of the linear combination using a Lagrange multiplier method in such a way that the variance is minimised. Our motivation for this method follows the simple result by Jennison and Turnbull [1997], that all asymptotically efficient estimators have the canonical joint distribution. Van Lancker et al. [2022] show that this new estimator has the correct canonical distribution, and hence the group sequential methods can be used without hesitation.

Similarly to method 2, there are some limitations to this method as it relies too heavily on accurately estimating the covariance matrix of the sequence of treatment effect estimates. Van Lancker et al. [2022] do not observe these issues since their results focus on simulations with sample sizes much larger than our chosen $n = 365$. In some cases, because the variance-covariance matrix is estimated with error, it is possible to choose the weights of the Lagrange multiplier in such a way that the new estimate has negative variance. Table 2 shows a similar pattern to method 2, that this numerical problem occurs when the variance of the longitudinal data is extreme so that $\sigma^2 = 100$, occurs infrequently for $\sigma^2 = 10$ and no problems occur for small $\sigma^2$.

Given that $Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)})$ is very close to $Var(\hat{\eta}^{(k_2)})$ for $k_1 < k_2$, there is not a lot to be gained by the more complex methods 2 and 3. The calculations are sensitive to errors in estimates of $Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)})$ and $Var(\hat{\eta}^{(k_2)})$. In some cases, the errors in these calculations lead to these methods simply not working. This raises doubts about how well they work in less extreme cases. Despite these problems, methods 2 and 3 perform adequately in simulation studies. However, this does not change our view that method 1 is preferable.

### 4 Results

We aim to assess the efficiency gain when longitudinal data are included in the analysis compared to when this longitudinal data is available, yet ignored. In this case, we believe that our joint model is correctly specified and therefore, we shall simulate clinical trial data from the true joint model and analyse it in two separate ways. The first way is to fit the data to the joint model using the conditional score method to find a treatment effect estimate $\hat{\eta}_J$ in order to perform this hypothesis test.

We shall simulate using the joint model in Equation (4). To fit the joint model to the data, we do not need to assume a distribution for the random effects, however we must specify this for simulation purposes and we shall simulate using the following

$$
\begin{bmatrix}
  b_{00} \\
  b_{11}
\end{bmatrix} \sim N \left( \begin{bmatrix}
  \mu_0 \\
  \mu_1
\end{bmatrix}, \begin{bmatrix}
  \phi_0^2 & 0 \\
  0 & \phi_1^2
\end{bmatrix} \right).
$$

Unless otherwise stated, throughout the simulation studies, we shall use parameter values

$$(\mu_0, \mu_1) = (6, 3), \phi_0^2 = 12.25, \phi_1^2 = 6.25, \sigma^2 = 10, h_0(t) = 5.5, \gamma = 0.03, \lambda = 0.022 \text{ and } \eta = -0.5. \quad (33)$$

These parameter values are based on the aids dataset in the R package JM by Rizopoulos [2010].

We shall test the one-sided hypothesis

$$H_{0,J}^{(J)} : \eta_J \geq 0, \quad H_{A,J}^{(J)} : \eta_J < 0.$$  

Here the subscript notation $J$ represents that the parameter $\theta_J$ is from the “joint” model. We fit the joint model using the conditional score method to find a treatment effect estimate $\hat{\eta}_J$ in order to perform this hypothesis test.
Our aim is to find the sample size, \( n_J \), required using the conditional score method to achieve Type 1 error rate \( \alpha = 0.025 \) when the true treatment effect is \( \eta_J = 0 \) and power \( 1 - \beta = 0.9 \) when \( \eta_J = -0.5 \). An estimate of the sample size will be calculated by simulation. The trial is designed with 2 years recruitment and 3 years follow-up and the group sequential trial has analyses at 20, 30, 40, 50 and 60 months. When increasing the sample size, we do so by increasing the rate of recruitment so accrual and follow-up periods in the trial design stay fixed. This is to ensure that differences in power are purely due to the sample size and not changes in the trial design as sample size increases.

The trial uses an error spending design given by Gordon Lan and DeMets [1983]. That is, the boundary constants \( b_1, \ldots, b_K \) are chosen to satisfy
\[
\mathbb{P}_{\eta=0}(Z_1 > b_1) = \min\{\alpha(I_1/I_{\text{max}})^2, \alpha\}
\]
\[
\mathbb{P}_{\eta=0}(Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k) = \min\{\alpha(I_k/I_{\text{max}})^2, \alpha\} - \min\{\alpha(I_{k-1}/I_{\text{max}})^2, \alpha\} \text{ for } k = 2, \ldots, K
\]
where the value of \( I_{\text{max}} \) is calculated to ensure that the trial has power \( 1 - \beta \) when \( \eta = -0.5 \) as described by Jennison and Turnbull [2000].

Maximum sample sizes \( n_J \) are given in Table 3. The first analysis, at 20 months, occurs just before the end of the recruitment period which is 2 years. Trials that terminate at the first interim analysis may recruit less than \( n_J \) patients, however this occurs with very small probability. Hence, the expected sample size will be very close to the maximum sample size for each model and therefore the maximum sample size is a useful measure to compare methods. The sample size increases as \( \sigma^2 \) increases. This reflects that noisy longitudinal data is associated with high variance or small information levels. Sample sizes are particularly high in each case where \( \sigma^2 = 100 \), which has been chosen as an extreme value. Further, sample sizes appear to increase slightly with \( \gamma \) and decrease slightly with \( \phi^2 \).

We now consider the analysis when the longitudinal data is ignored. We believe the joint model to be true and correct, however we shall fit the data to a Cox model. To do so, we shall simulate data from the joint model and then fit this data to a misspecified Cox proportional hazards model. The Cox model is given by:
\[
\lambda_i(t) = \hat{\lambda}_0(t) \exp\{\eta_C Z_i\}.
\]
For this clinical trial, we test the hypothesis
\[
H_0^{(C)}: \eta_C \geq 0, \quad H_A^{(C)}: \eta_C < 0
\] (35)
and we find a treatment effect estimate \( \hat{\eta}_C \) using the maximum partial likelihood method as in Jennison and Turnbull [1997].

Although this model is misspecified, type 1 error is not affected. This is because, under \( H_0^{(J)} \) we have \( \eta_J = 0 \) and there is no difference between treatment groups in overall survival. When fitting the Cox model to the data, the longitudinal data trajectory is reflected in the function \( \hat{\lambda}_0(t) \) so that we also have that \( \eta_C = 0 \). Hence, \( H_0^{(C)} \) is also true.

Let \( n_C \) be the sample size such that we achieve type 1 error \( \alpha = 0.025 \) when \( \eta_J = 0 \) and power \( 1 - \beta = 0.9 \) when \( \eta_J = -0.5 \) when we perform the hypothesis test in (35). Values of \( n_C \) and \( n_J \) are given in Table 3. The values of \( \gamma \) and \( \phi^2 \) for simulation are varied. Notice that \( n_C \) does not change with \( \sigma^2 \) since this plays no role in simulating survival times, and the longitudinal data, which is affected by \( \sigma^2 \), is ignored. As the value of \( \gamma \) increases, the sample size \( n_C \) increases. This represents that as the longitudinal data has more weight in the survival hazard rate, ignoring the longitudinal data results in an increasingly inefficient clinical trial. When \( \gamma = 0 \), this represents the case where longitudinal data is available yet has no influence on the survival function. In this case, \( n_C < n_J \) and it is more efficient to fit the data to the simple Cox model. We see that the value of \( n_C \) increases with \( \phi^2 \). This is the variance amongst patients of the slopes of the longitudinal trajectories. This indicates that the simple Cox model is unable to account for large differences between individual patients.

To compare the sample sizes obtained using the joint model and the misspecified Cox model, we define “relative efficiency” to be \( RE = \frac{n_C}{n_J} \). Using this definition, when \( RE > 1 \) we interpret this as the joint model analysis being the more efficient model to use and similarly when \( RE < 1 \), the Cox model analysis is the more efficient analysis method.

Table 3 shows the relative efficiency results. We see that RE increases with \( \gamma \), increases with \( \phi^2 \) and remains constant with \( \sigma^2 \) apart from the case where \( \sigma^2 = 100 \) which reflects extremely noisy data. Also, we see that \( RE = 0.97 \) when \( \gamma = 0 \) and \( \sigma^2 = 100 \) which indicates that when the longitudinal data is not correlated with the survival endpoint and the longitudinal data is noisy, the simple Cox model is a slightly more efficient method for estimating the treatment effect. Apart from the case where \( \gamma = 0 \), it is always more efficient to analyse the data using the joint modelling approach. Even when \( \gamma = 0 \), fitting the data to the simple Cox model for survival data is only marginally more efficient than fitting the data to the joint model. In the extreme case, 2.63 times as many patients are required to analyse the data using the Cox model as when the joint modelling framework is used.
The conditional score method is used to find a treatment effect estimate in the joint model of longitudinal and time to event data and we have displayed new theoretical results for the distribution of the sequence of treatment effect estimates \( \eta_1, \ldots, \eta_K \) found using the conditional score method in a group sequential trial. Although the canonical joint distribution for the sequence \( \eta_1, \ldots, \eta_K \) does not hold, we show that it is sensible and practical to proceed assuming that the canonical joint distribution holds anyway. In particular, we have proven, by assuming the canonical joint distribution holds, and using a non-binding futility boundary, the trial is conservative with respect to type 1 error rates. Finally, using simulation studies we have seen that the deviations from planned type 1 error \( \alpha \) are minimal. Other benefits of using the conditional score method are that no distributional assumptions are required for the random effects distribution holds, and using a non-binding futility boundary, the trial is conservative with respect to type 1 error rates.

We have shown that by including the longitudinal data, compared to the case where the longitudinal data is observed but left out of the analysis, we can greatly improve the efficiency of the trial with respect to sample size. In some cases, 2.63 times as many patients are required to achieve the same power in the analysis where the longitudinal data is left out.

## Appendix 1

### Conditions for the asymptotic theory of parameter estimates

To ensure the existence of the asymptotic covariance matrix \( \Sigma \), we require the probabilistic limits of \( s_c^{(0)}(k), \ldots, s_c^{(2)}(k) \) and \( C^{(1)}(k), \ldots, C^{(3)}(k) \) given by (8) to exist. The limits are defined through the following conditions.

**Conditions 2.**

1. There exist neighbourhoods \( \Gamma \) of \( \gamma_0 \) and \( N \) of \( \eta_0 \) and for each \( k = 1, \ldots, K \) there are functions \( s_c^{(0)}(k, t, \gamma, \eta, \sigma^2), s_c^{(1)}(k, t, \gamma, \eta, \sigma^2), s_c^{(2)}(k, t, \gamma, \eta, \sigma^2), c^{(1)}(k, t, \gamma, \eta, \sigma^2) \) and \( c^{(2)}(k, t, \gamma, \eta, \sigma^2) \) defined on \( [0, \infty) \times \Gamma \times N \) such that

\[
\sup_{t \in [0, \infty), \gamma \in \Gamma, \eta \in N} \left\| S_c^{(j)}(k, t, \gamma, \eta, \sigma^2) - s_c^{(j)}(k, t, \gamma, \eta, \sigma^2) \right\| \xrightarrow{p} 0 \text{ for } j = 0, 1, 2
\]

\[
\sup_{t \in [0, \infty), \gamma \in \Gamma, \eta \in N} \left\| C^{(j)}(k, t, \gamma, \eta, \sigma^2) - c^{(j)}(k, t, \gamma, \eta, \sigma^2) \right\| \xrightarrow{p} 0 \text{ for } j = 1, 2.
\]

2. Each \( s_c^{(j)}(k, t, \gamma, \eta, \sigma^2) \) and \( c^{(j)}(k, t, \gamma, \eta, \sigma^2) \) is a continuous function of \( \gamma \in \Gamma \) and \( \eta \in N \) uniformly in \( t \in [0, \infty) \), and bounded on \( [0, \infty) \times \Gamma \times N \).

3. For each \( k = 1, \ldots, K \) \( s_c^{(0)}(k) \) and \( c^{(0)}(k) \) are bounded away from zero on \( [0, \infty) \times \Gamma \times N \).
It is clear that the probabilistic limits $e_c(k, t, \gamma, \eta, \sigma^2)$ of $E_{c}(k, t, \gamma, \eta, \sigma^2)$, $v_{c}^{(1)}(k, t, \gamma, \eta, \sigma^2)$ of $V_{c}^{(1)}(k, t, \gamma, \eta, \sigma^2)$ and $v_{c}^{(2)}(k, t, \gamma, \eta, \sigma^2)$ of $V_{c}^{(2)}(k, t, \gamma, \eta, \sigma^2)$ exist and can expressed in terms of $s_{c}^{(j)}(k, t, \gamma, \eta, \sigma^2)$ and $c^{(j)}(k, t, \gamma, \eta, \sigma^2)$ for $j = 0, 1, 2$ and these are

$$e_c = \frac{s_{c}^{(1)}}{s_{c}^{(0)}},$$

$$v_c^{(1)} = \frac{s_{c}^{(2)}}{s_{c}^{(0)}} - \frac{s_{c}^{(1)}s_{c}^{(1)^{T}}}{[s_{c}^{(0)}]^2},$$

$$v_c^{(2)} = \frac{c_{c}^{(2)}}{s_{c}^{(0)}} - \frac{s_{c}^{(1)c_{c}^{(1)^{T}}}}{[s_{c}^{(0)}]^2}.$$ 

**Software**

All statistical computing and analyses were performed using the software environment R version 4.0.2. Programming code for sample size calculations, is available at https://github.com/abigailburdon/Conditional-score-GST.

**Acknowledgements**

This research was funded by the Engineering and Physical Sciences Research Council.

**References**

Per Kragh Andersen and Richard David Gill. Cox’s regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982.

Raymond J Carroll, David Ruppert, Ciprian M Crainiceanu, and Leonard A Stefanski. *Measurement Error in Nonlinear Models: A Modern Perspective*. London: Chapman and Hall/CRC, 2006.

Richard J Cook and Jerald F Lawless. Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics*, pages 1311–1323, 1996.

Sally Galbraith and Ian C Marschner. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine*, 22(11):1787–1805, 2003.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Bjoern Bornkamp, Martin Maechler, Torsten Hothorn, and Maintainer Torsten Hothorn. Package ‘mvtnorm’. *Journal of Computational and Graphical Statistics*, 11:950–971, 2020.

Anne J Goldman, Bradley P Carlin, Lawrence R Crane, Cynthia Launer, Joyce A Korvick, Lawrence Deyton, and Donald I Abrams. Response of CD4 lymphocytes and clinical consequences of treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes*, 11(2):161–169, 1996.

KK Gordon Lan and David L DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

Minggao Gu and Zhiliang Ying. Sequential analysis for censored regression data. *Journal of the American Statistical Association*, 88(423):890–898, 1993.

**DRAFT GUIDANCE. Adaptive designs for clinical trials of drugs and biologics. Center for Biologics Evaluation and Research (CBER), 2018.**

Christopher Jennison and Bruce W Turnbull. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

Christopher Jennison and Bruce W Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman and Hall/CRC, 2000.

Peter C O’Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

Dimitris Rizopoulos. Jm: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

Eric Slud and LJ Wei. Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association*, 77(380):862–868, 1982.
Leonard A Stefanski and Raymond J Carroll. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716, 1987.

Jeremy MG Taylor, Yongseok Park, Donna P Ankerst, Cecile Proust-Lima, Scott Williams, Larry Kestin, Kyoungwha Bae, Tom Pickles, and Howard Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213, 2013.

Anastasios A Tsiatis and Marie Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458, 2001.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge: University Press, 2000.

Kelly Van Lancker, Joshua Betz, and Michael Rosenblum. Combining covariate adjustment with group sequential and information adaptive designs to improve randomized trial efficiency. *arXiv preprint arXiv:2201.12921*, 2022.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Berlin: Springer Science & Business Media, 2013.
Supplementary Materials for Joint Modelling of Longitudinal and Time-to-Event Data Applied to Group Sequential Clinical Trials

Abigail J. Burdon, Lisa V. Hampson and Christopher Jennison

November 30, 2022

A Differentiation of the conditional score

We shall differentiate the conditional score function at analysis $k$, $U_*(k, \gamma, \eta, \sigma^2)$, which is given in Equation (9) of the main paper, with respect to the vector $(\gamma, \eta)^T$. First we consider differentiating the function $E_{0i}(t, \gamma, \eta, \sigma^2)$ with respect to $\gamma$ and $\eta$ separately. These are

\[
\frac{\partial}{\partial \gamma} E_{0i}(t, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial \gamma} \exp\{\gamma S_i(t, \gamma, \sigma^2) - \gamma^2 \sigma^2 \theta_i(t)/2 + \eta Z_i\} \\
= \frac{\partial}{\partial \gamma} \exp\{\gamma^2 \sigma^2 \theta_i(t) dN_i(t) + \gamma \hat{X}_i(t) - \gamma^2 \sigma^2 \theta_i(t)/2 + \eta Z_i\} \\
= (2\gamma \sigma^2 \theta_i(t) dN_i(t) + \hat{X}_i(t) - \gamma \sigma^2 \theta_i(t)) E_{0i}(t, \gamma, \eta, \sigma^2) \\
= (2S_i(t, \gamma, \sigma^2) - \hat{X}_i(t) - \gamma \sigma^2 \theta_i(t)) E_{0i}(t, \gamma, \eta, \sigma^2)
\]

\[
\frac{\partial}{\partial \eta} E_{0i}(t, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial \eta} \exp\{\gamma S_i(t, \gamma, \sigma^2) - \gamma^2 \sigma^2 \theta_i(t)/2 + \eta Z_i\} \\
= Z_i E_{0i}(t, \gamma, \eta, \sigma^2).
\]

Derivatives of $E_{0i}(t, \gamma, \eta, \sigma^2)$ are needed for differentiating the function $S_c^{(0)}(t, \gamma, \eta, \sigma^2)$. Similarly for differentiation of the function $S_c^{(1)}(t, \gamma, \eta, \sigma^2)$ we now calculate derivatives for $S_i(t, \gamma, \eta, \sigma^2) E_{0i}(t, \gamma, \eta, \sigma^2)$ and $Z_i E_{0i}(t, \gamma, \eta, \sigma^2)$. These are

\[
\frac{\partial}{\partial \gamma} S_i(t, \gamma, \eta, \sigma^2) E_{0i}(t, \gamma, \eta, \sigma^2) = \left[ S_i(t, \gamma, \sigma^2)^2 + S_i(t, \gamma, \eta, \sigma^2)(S_i(t, \gamma, \sigma^2) - \hat{X}_i(t)) \\
- \gamma \sigma^2 \theta_i(t) \right] E_{0i}(t, \gamma, \eta, \sigma^2)
\]

\[
\frac{\partial}{\partial \eta} Z_i E_{0i}(t, \gamma, \eta, \sigma^2) = Z_i (2S_i(t, \gamma, \sigma^2) - \hat{X}_i(t) - \gamma \sigma^2 \theta_i(t)) E_{0i}(t, \gamma, \eta, \sigma^2)
\]

\[
\frac{\partial}{\partial \gamma} S_i(t, \gamma, \eta, \sigma^2) E_{0i}(t, \gamma, \eta, \sigma^2) = S_i(t, \gamma, \eta, \sigma^2) Z_i E_{0i}(t, \gamma, \eta, \sigma^2)
\]

\[
\frac{\partial}{\partial \eta} Z_i E_{0i}(t, \gamma, \eta, \sigma^2) = Z_i^T Z_i E_{0i}(t, \gamma, \eta, \sigma^2).
\]

The function $S_c^{(0)}$ is a scalar and its derivative will be a $(p+1) \times 1$ column vector. The matrices $S_c^{(1)}$ and $C^{(1)}$ are both $(p+1) \times 1$ column vectors. For differentiation, entry $(\partial S_c^{(1)}/\partial (\gamma, \eta)^T)_{i,j}$ represents the $i^{th}$ element of $S_c^{(1)}$ differentiated with respect to the $j^{th}$ element of the vector $(\gamma, \eta)^T$. The functions $S_c^{(2)}, C^{(2)}, C^{(3)}, V^{(1)}_c, V^{(2)}_c$ and $V^{(3)}_c$
are all \((p+1) \times (p+1)\) matrices. Using the above calculation, we find the following relationships hold
\[
\frac{\partial}{\partial (\gamma, \eta)^T} S_c^{(0)}(k, t, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial (\gamma, \eta)^T} \left[ \frac{1}{n} \sum_{i=1}^{n} E_{0i}(t, \gamma, \eta, \sigma^2)Y_i(k, t) \right]
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ 2S_i(t, \gamma, \sigma^2) - \dot{X}_i(t) - \gamma \sigma^2 \theta_i(t) \right\} E_{0i}(t, \gamma, \eta, \sigma^2)Y_i(k, t)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ S_i(Z_i - \dot{X}_i) \right\} E_{0i}(t, \gamma, \eta, \sigma^2)Y_i(k, t) + \frac{1}{n} \sum_{i=1}^{n} J_i(k, t, \gamma, \sigma^2) E_{0i}(t, \gamma, \eta, \sigma^2)Y_i(k, t)
\]
\[
= S^{(1)}_c(k, t, \gamma, \eta, \sigma^2) + C^{(1)}(k, t, \gamma, \eta, \sigma^2)
\]
A similar calculation is performed for the differentiation of \(S^{(1)}_c(k, t, \gamma, \eta, \sigma^2)\). Temporarily removing dependence of all functions on parameters \(k, t, \gamma, \eta\), we have
\[
\frac{\partial}{\partial (\gamma, \eta)^T} S^{(1)}_c(k, t, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial (\gamma, \eta)^T} \left[ \frac{1}{n} \sum_{i=1}^{n} \left\{ S_i(t, \gamma, \sigma^2) \right\} E_{0i}(t, \gamma, \eta, \sigma^2)Y_i(k, t) \right]
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ S_i^2 + S_i(S_i - \dot{X}_i - \gamma \sigma^2 \theta_i) + \gamma \sigma^2 \dot{\theta}_i dN_i \right\} E_{0i}Z_i Y_i
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ S_i Z_i \right\} E_{0i}Y_i
\]
\[
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ S_i \right\} E_{0i}Y_i + \frac{1}{n} \sum_{i=1}^{n} \left[ \sigma^2 \dot{\theta}_i(u) dN_i \right]
\]
\[
= S^{(2)}_c(k, t, \gamma, \eta, \sigma^2) + C^{(2)}(k, t, \gamma, \eta, \sigma^2) + C^{(3)}(k, t, \gamma, \eta, \sigma^2).
\]
We are now able to differentiate the function \(E_c(k, t, \gamma, \eta, \sigma^2)\). By the quotient rule and the relationships derived above, we have
\[
\frac{\partial}{\partial (\gamma, \eta)^T} E_c(k, t, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial (\gamma, \eta)^T} \left[ \frac{S^{(1)}_c(k, t, \gamma, \eta, \sigma^2)}{S^{(0)}_c(k, t, \gamma, \eta, \sigma^2)} \right]
\]
\[
= \frac{\partial}{\partial (\gamma, \eta)^T} S^{(1)}_c(k, t, \gamma, \eta, \sigma^2) \frac{\partial}{\partial (\gamma, \eta)^T} S^{(0)}_c(k, t, \gamma, \eta, \sigma^2) - \frac{S^{(1)}_c(k, t, \gamma, \eta, \sigma^2)}{S^{(0)}_c(k, t, \gamma, \eta, \sigma^2)} \frac{\partial}{\partial (\gamma, \eta)^T} S^{(0)}_c(k, t, \gamma, \eta, \sigma^2)
\]
\[
= S^{(2)}_c(k, t, \gamma, \eta, \sigma^2) + C^{(2)}(k, t, \gamma, \eta, \sigma^2) + C^{(3)}(k, t, \gamma, \eta, \sigma^2)
\]
\[
= S^{(1)}_c(k, t, \gamma, \eta, \sigma^2) + S^{(1)}_c(k, t, \gamma, \eta, \sigma^2) + C^{(1)}(k, t, \gamma, \eta, \sigma^2)
\]
\[
= V^{(1)}_c(k, t, \gamma, \eta, \sigma^2) + V^{(2)}_c(k, t, \gamma, \eta, \sigma^2) + V^{(3)}_c(k, t, \gamma, \eta, \sigma^2).
\]
Combining these results, the first derivative of the conditional score function with respect to the vector \((\gamma, \eta)^T\) is given by
\[
\frac{\partial}{\partial (\gamma, \eta)^T} U_c(k, \gamma, \eta, \sigma^2) = \frac{\partial}{\partial (\gamma, \eta)^T} \left[ \int_0^\infty \sum_{i=1}^{n} \left\{ S_i(t, \gamma, \sigma^2), Z_i^T \right\} E_{0i}(k, u, \gamma, \eta, \sigma^2) dN_i(k, u) \right]
\]
\[
= \sum_{i=1}^{n} \int_0^\infty \left[ \Gamma_i(k, u, \sigma^2) - V^{(1)}_c(k, u, \gamma, \eta, \sigma^2) + V^{(2)}_c(k, u, \gamma, \eta, \sigma^2) + V^{(3)}_c(k, u, \gamma, \eta, \sigma^2) \right] dN_i(k, u).
\]
B Details for the proof of Theorem 1

Lemma 1. Starting from

\[
\begin{bmatrix}
U_c(1, \gamma, \eta, \sigma^2) \\
\vdots \\
U_c(K, \gamma, \eta, \sigma^2)
\end{bmatrix} = \mathbf{0}
\]  

we have that

\[-n^{-\frac{1}{2}} \tilde{A}^{-1} \begin{bmatrix}
U_c(1, \gamma, \eta, \sigma^2) \\
\vdots \\
U_c(K, \gamma, \eta, \sigma^2)
\end{bmatrix} = \tilde{A}^{-1} \begin{bmatrix}
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c(1, \gamma, \eta, \sigma^2) |_{(\gamma^*(1), \eta^*(1))} \\
\vdots \\
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c(K, \gamma, \eta, \sigma^2) |_{(\gamma^*(K), \eta^*(K))}
\end{bmatrix} \cdot n^{\frac{1}{2}} \begin{bmatrix}
\hat{\gamma}^{(1)} - \gamma_0 \\
\hat{\eta}^{(1)} - \eta_0 \\
\vdots \\
\hat{\gamma}^{(K)} - \gamma_0 \\
\hat{\eta}^{(K)} - \eta_0
\end{bmatrix}
\]

where \((\gamma^*(k), \eta^*(k))\) lies on the line segment between \((\gamma_0, \eta_0)\) and \((\hat{\gamma}^{(k)}, \hat{\eta}^{(k)})\).

Proof. For a fixed, \(k = 1, \ldots, K\), we begin by applying a Taylor expansion to each element of \(U_c(k, \gamma, \eta, \sigma^2)\). Let \(U_c^j(k, \gamma, \eta, \sigma^2)\) be the \(j^{th}\) element of the \(p \times 1\) column vector \(U_c(k, \gamma, \eta, \sigma^2)\). Let the vector \((\gamma, \eta)^T\) be of dimension \(p \times 1\), then \(\frac{\partial}{\partial (\gamma, \eta)^T} U_c^j(k, \gamma, \eta, \sigma^2)\) is a \(1 \times p\) row vector and \((\gamma, \eta)^T - (\gamma_0, \eta_0)^T\) is a \(p \times 1\) column vector. Regularity conditions allow us to perform the following Taylor expansion. For each \(j = 1, \ldots, p\) the Taylor expansion of \(U_c^j(k, \gamma, \eta, \sigma^2)\) around \((\gamma_0, \eta_0)^T\) is

\[
U_c^j(k, \hat{\gamma}, \hat{\eta}, \sigma^2) = U_c^j(k, \gamma_0, \eta_0, \sigma^2) + \frac{\partial}{\partial (\gamma, \eta)^T} U_c^j(k, \gamma, \eta, \sigma^2) |_{(\gamma, \eta)^T = (\gamma^*(k), \eta^*(k))^T} \cdot ((\gamma, \eta)^T - (\gamma_0, \eta_0)^T)
\]

where \((\gamma^*(k), \eta^*(k))^T\) lies on the line segment between \((\gamma_0, \eta_0)^T\) and \((\hat{\gamma}, \hat{\eta})^T\). Each row of \(U_c(k, \gamma_0, \eta_0, \sigma^2)\) represents a dimension, so different rows are differing functions of \((\gamma, \eta)^T\) and hence each \((\gamma_j^*(k), \eta_j^*(k))^T\) is specific to the row \(j\).

Given the definition of the estimates \(\hat{\gamma}_n^{(k)}\) and \(\hat{\eta}_n^{(k)}\), we recognise that \(U_c^j(k, \hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)}, \sigma^2) = 0\). Therefore by substituting \((\hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)})^T\) for \((\gamma, \eta)^T\), and stacking the rows, we have

\[
0 = \begin{bmatrix}
U_c^1(k, \hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)}, \sigma^2) \\
\vdots \\
U_c^p(k, \hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)}, \sigma^2)
\end{bmatrix} = U_c(k, \gamma_0, \eta_0, \sigma^2) + \frac{\partial}{\partial (\gamma, \eta)^T} U_c^j(k, \gamma, \eta, \sigma^2) |_{(\gamma, \eta)^T = (\gamma^*(k), \eta^*(k))^T} \cdot ((\hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)})^T - (\gamma_0, \eta_0)^T)
\]

where each \((\gamma_j^*(k), \eta_j^*(k))^T\) lies on the line segment between \((\gamma_0, \eta_0)^T\) and \((\hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)})^T\).

The matrix \(A^{(k)}\) by definition is finite valued, symmetric and positive definite. Therefore \(A^{(k)}\) is invertible with inverse \((A^{(k)})^{-1}\) which is also finite valued, symmetric and positive definite. Multiplying the above equation by \((A^{(k)})^{-1}\) and a simple rearrangement gives

\[
-n^{-\frac{1}{2}} (A^{(k)})^{-1} U_c(k, \gamma_0, \eta_0, \sigma^2) = (A^{(k)})^{-1} \begin{bmatrix}
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c^1(k, \gamma, \eta, \sigma^2) |_{(\gamma, \eta)^T = (\gamma^*(1), \eta^*(1))^T} \\
\vdots \\
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c^p(k, \gamma, \eta, \sigma^2) |_{(\gamma, \eta)^T = (\gamma^*(p), \eta^*(p))^T}
\end{bmatrix} \cdot n^{\frac{1}{2}} ((\hat{\gamma}_n^{(k)}, \hat{\eta}_n^{(k)})^T - (\gamma_0, \eta_0)^T).
\]

Returning to the definition in the main paper, \(\tilde{A}\) is the block diagonal matrix whose \(k^{th}\) diagonal matrix is the \(p \times p\) matrix \(A^{(k)}\). Then, aggregating the above equation for \(k = 1, \ldots, K\), we have the result

\[-n^{-\frac{1}{2}} \tilde{A}^{-1} \begin{bmatrix}
U_c(1, \gamma, \eta, \sigma^2) \\
\vdots \\
U_c(K, \gamma, \eta, \sigma^2)
\end{bmatrix} = \tilde{A}^{-1} \begin{bmatrix}
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c(1, \gamma, \eta, \sigma^2) |_{(\gamma^*(1), \eta^*(1))} \\
\vdots \\
n^{-1} \frac{\partial}{\partial (\gamma, \eta)^T} U_c(K, \gamma, \eta, \sigma^2) |_{(\gamma^*(K), \eta^*(K))}
\end{bmatrix} \cdot n^{\frac{1}{2}} \begin{bmatrix}
\hat{\gamma}^{(1)} - \gamma_0 \\
\hat{\eta}^{(1)} - \eta_0 \\
\vdots \\
\hat{\gamma}^{(K)} - \gamma_0 \\
\hat{\eta}^{(K)} - \eta_0
\end{bmatrix}
\]
Lemma 2.

\[ n^{-1} \frac{\partial}{\partial (\gamma, \eta^T)^T} U_c^{(n)}(k, \gamma, \eta, \sigma^2) \bigg|_{\gamma=\gamma^*, \eta=\eta^*} \overset{p}{\longrightarrow} A^{(k)}. \]

Proof. We can write

\[ \frac{1}{n} \frac{\partial}{\partial (k, \gamma, \eta^T)^T} U_c^{(n)}(k, \gamma, \eta, \sigma^2) \bigg|_{\gamma=\gamma^*, \eta=\eta^*} - A^{(k)} = -D_0^{(k)} + D_1^{(k)} + D_2^{(k)} + D_3^{(k)} \]

where

\[ D_0^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \Gamma_i(k, u, \sigma^2_0) dN_i(k, u) - \int_0^\infty E(\Gamma_i(k, u, \sigma^2_0)) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma^2_0) h_0(u) du \]

\[ D_1^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty V^{(1)}(k, u, \gamma^*, \eta^*, \sigma^2_0) dN_i(k, u) - \int_0^\infty v^{(1)}(k, u, \gamma_0, \eta_0, \sigma^2_0) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma^2_0) h_0(u) du \]

\[ D_2^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty V^{(2)}(k, u, \gamma^*, \eta^*, \sigma^2_0) dN_i(k, u) - \int_0^\infty v^{(2)}(k, u, \gamma_0, \eta_0, \sigma^2_0) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma^2_0) h_0(u) du \]

\[ D_3^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty V^{(3)}(k, u, \gamma^*, \eta^*, \sigma^2_0) dN_i(k, u), \]

and it remains to show that each of the terms \( D_0^{(k)} \), \( D_1^{(k)} \), \( D_2^{(k)} \) and \( D_3^{(k)} \) converge in probability to zero. For the terms \( D_0^{(k)} \), \( D_1^{(k)} \) and \( D_2^{(k)} \), analogous results are presented by Andersen and Gill [1982] for survival data. For the term \( D_3^{(k)} \), we make use of the relationship \( dN_i(k, u) dN_j(k, u) = 0 \) for \( i \neq j \) and \( dN_i(k, u)^2 = 1 \) since \( dN_i(k, u) \) is an indicator function. The function \( V^{(3)} \) is a \((p+1) \times (p+1)\) matrix where all entries are zero apart from the top left. Therefore we shall only consider entry \((1,1)\). We have that

\[ \left[ D_3^{(k)} \right]_{11} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \left[ V^{(3)}(k, u, \gamma, \eta, \sigma^2) \right]_{11} dN_i(k, u) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \left[ C^{(3)}(k, u, \gamma, \eta, \sigma^2) \right]_{11} dN_i(k, u) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \sum_{j=1}^{n} \frac{E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)}{E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)} dN_i(k, u) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \frac{\sum_{j=1}^{n} \sigma^2 \theta_j(u) E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)}{\sum_{j=1}^{n} E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)} dN_i(k, u) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \frac{\sigma^2 \theta_j(u) E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)}{\sum_{j=1}^{n} E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)} dN_i(k, u), \]

A single element in the summand in Expression (5) can be written

\[ \int_0^\infty \frac{\frac{1}{n} \sigma^2 \theta_j(u) E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)}{\sum_{j=1}^{n} E_{0j}(k, u, \gamma, \eta, \sigma^2) Y_j(k, u)} \]

and it is clear that \( [D_3^{(k)}]_{11} \overset{p}{\longrightarrow} 0 \) as \( n \to \infty \). Therefore, we have the result

\[ \frac{1}{n} \frac{\partial}{\partial (\gamma, \eta^T)^T} U_c^{(n)}(k, \gamma, \eta, \sigma^2) \bigg|_{\gamma=\gamma^*, \eta=\eta^*} \overset{p}{\longrightarrow} A^{(k)}. \]

Lemma 3.

\[ \text{Cov} \left( W^{(n)}_{ij_1}(k_1, \tau_{k_1}, \gamma_0, \eta_0, \sigma^2_0), W^{(n)}_{ij_2}(k_2, \tau_{k_2}, \gamma_0, \eta_0, \sigma^2_0) \right) \]

\[ = \mathbb{E} \left( \sum_{i=1}^{n} \int_0^{\tau_{k_1}} H^{(n)}_{ij_1}(k_1, u, \gamma_0, \eta_0, \sigma^2_0) H^{(n)}_{ij_2}(k_1, u, \gamma_0, \eta_0, \sigma^2_0) h_0(u) E_{0i}(u, \gamma_0, \eta_0, \sigma^2_0) Y_i(k_1, u) du \right). \]
Proof. In the following, we shall drop the dependency of the function $H(n)$ on the parameters $\gamma$, $\eta$ and $\sigma^2$. In the following calculation, the second equality holds because patients are independent and the orders of summation and integration can be interchanged. Further, the event for each patient can only happen in one analysis so that if $l_1 \neq l_2$ then $dD_N_i(l_1, u) dD_N_i(l_2, u) = 0$ and because $dD_N_i(l, u)$ is an indicator function, we have that $dD_N_i(l, u)^2 = dD_N_i(l, u)$. Thus for $k_1 \leq k_2$, the covariance is given by

$$
Cov\left( W_{j_1}^{(n)}(k_1, \tau_{k_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{k_2}, \gamma_0, \eta_0, \sigma_0^2) \right)
= \mathbb{E}\left( \int_0^{\tau_{k_1}} \sum_{i=1}^{n} \sum_{l_1=1}^{k_1} H_{ij_1}^{(n)}(l_1, u) dD_N_i(l_1, u) \int_0^{\tau_{k_2}} \sum_{i=1}^{n} \sum_{l_2=1}^{k_2} H_{ij_2}^{(n)}(l_2, u) dD_N_i(l_2, u) \right)
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \sum_{l_1=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l_1, u) dD_N_i(l_1, u) \sum_{i=1}^{n} \sum_{l_2=1}^{k_2} \sum_{l_2=1}^{k_2} \int_0^{\tau_{k_2}} H_{ij_2}^{(n)}(l_2, u) dD_N_i(l_2, u) \right)
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l_1, u) H_{ij_2}^{(n)}(l_1, u) dD_N_i(l_1, u) \right).
$$

For the following, note that we have

$$
\mathbb{E}(dD_N_i(l, u)|S_i(t, \gamma, \sigma^2), t_i(t), Z_i, Y_i(k, u) = h_0(u)E_{0i}(u, \gamma, \eta, \sigma^2)Y_i(k, u)du
$$

which follows by definition of the conditional intensity process, $\lambda_i^{C}(k, t)$. The following calculations use the law of total expectation and the expectation of the counting process $dN_i(k, u)$ in a similar way to the fixed sample case.

$$
\mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l, u) H_{ij_2}^{(n)}(l, u) dD_N_i(l, u) \right)
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \sum_{l_1=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l, u) H_{ij_2}^{(n)}(l, u) dD_N_i(l, u)|S_i(u, \gamma_0, \sigma_0^2), Z_i, t_i(u), Y_i(l, u)|\right)
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \sum_{l_1=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l, u) H_{ij_2}^{(n)}(l, u) dD_N_i(l, u)|S_i(u, \gamma_0, \sigma_0^2), Z_i, t_i(u), Y_i(l, u)|\right)
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{l=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l, u) H_{ij_2}^{(n)}(l, u) h_0(u) E_{0i}(u, \gamma_0, \eta_0, \sigma_0^2)(Y_i(l, u) - Y_i(l - 1, u))du\right)
= \mathbb{E}\left( \sum_{i=1}^{n} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(k_1, u) H_{ij_2}^{(n)}(k_1, u) h_0(u) E_{0i}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k_1, u)du\right).
$$

Therefore we have the result

$$
Cov\left( W_{j_1}^{(n)}(k_1, \tau_{k_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{k_2}, \gamma_0, \eta_0, \sigma_0^2) \right)
= \mathbb{E}\left( \sum_{i=1}^{n} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(k_1, u, \gamma_0, \eta_0, \sigma_0^2) H_{ij_2}^{(n)}(k_1, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) E_{0i}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k_1, u)du\right).
$$

\[\square\]

Lemma 4.

$$
Cov\left( W_{j_1}^{(n)}(k_1, \tau_{k_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{k_2}, \gamma_0, \eta_0, \sigma_0^2) \right) \overset{p}{\rightarrow} B_{j_1, j_2}^{(k_1)}.
$$
Proof. The following calculations use the law of total expectation and the expectation of the counting process $dN_t(k, u)$ in a similar way to the fixed sample case.

\[
\mathbb{E}\left( \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{t_{i,j}} H^{(n)}_{ij}(l, u)H^{(n)}_{i,j}(l, u)dN_t(l, u) \right)
\]

\[
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{t_{i,j}} \mathbb{E}[H^{(n)}_{ij}(l, u)H^{(n)}_{i,j}(l, u)dN_t(l, u)|S_i(u, \gamma_0, \sigma_0^2), Z_i, t_i(u), Y_i(l, u)] \right)
\]

\[
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{t_{i,j}} H^{(n)}_{ij}(l, u)H^{(n)}_{i,j}(l, u)dN_t(l, u) - dN_t(l - 1, u)|S_i(u, \gamma_0, \sigma_0^2), Z_i, t_i(u), Y_i(l, u)] \right)
\]

\[
= \mathbb{E}\left( \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{t_{i,j}} H^{(n)}_{ij}(l, u)H^{(n)}_{i,j}(l, u)h_0(u)E_{o_1}(u, \gamma_0, \eta_0, \sigma_0^2)(Y_i(l, u) - Y_i(l - 1, u))du \right)
\]

\[
= \mathbb{E}\left( \sum_{i=1}^{n} \int_0^{t_{i,j}} H^{(n)}_{ij}(k, u)H^{(n)}_{i,j}(k, u)h_0(u)E_{o_1}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k, u)du \right).
\]

For the following calculation, dependency on parameters $\gamma_0, \eta_0$ and $\sigma_0^2$ is removed from the functions $S^{(0)}(k, t), S^{(1)}(k, t)$ and $S^{(2)}(k, t)$ for notational purposes. Further, the dependency on parameters $\gamma_0, \eta_0$ and $\sigma_0^2$ is also removed from the probabilistic limits $s^{(0)}(k, t), s^{(1)}(k, t), s^{(2)}(k, t)$ and $e_c(k, t)$. Then following calculation holds

\[
\sum_{i=1}^{n} H^{(n)}_{ij}(k, u)h_0(u)E_{o_1}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k, u)du
\]

\[
= \sum_{i=1}^{n} n^{-1}(\{S_i(u, \gamma_0, \sigma_0^2), Z_i^T\} - e_c(k, u)) \{S_i(u, \gamma_0, \sigma_0^2), Z_i^T\} - e_c(k, u)h_0(u)E_{o_1}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k, u)
\]

\[
= \left[ \left[S^{(2)}(k, u)\right]_{j_1,j_2} - \frac{\left[s^{(1)}(k, u)\right]_{j_2}}{s^{(0)}(k, u)} \left[S^{(1)}(k, u)\right]_{j_2} \right]
\]

\[
= \left[ \left[s^{(1)}(k, u)\right]_{j_2} + \frac{\left[s^{(1)}(k, u)\right]_{j_1}}{s^{(0)}(k, u)} \left[s^{(0)}(k, u)\right]_{j_2} \right]h_0(u)
\]

\[
= \left[ \left[s^{(1)}(k, u)\right]_{j_1,j_2} \frac{\left[s^{(0)}(k, u)\right]_{j_1} \left[s^{(0)}(k, u)\right]_{j_2}}{s^{(0)}(k, u)} \right]h_0(u)
\]

\[
\mathbb{P}\left( \int_0^{t_{i,j}} H^{(n)}_{ij}(k, u)H^{(n)}_{i,j}(k, u)h_0(u)E_{o_1}(u, \gamma_0, \eta_0, \sigma_0^2)Y_i(k, u)du \right)
\]

\[
= B^{(k)}_{j_1,j_2}
\]

\[\square\]

C Sketch proof that the trial is conservative with respect to type 1 error for a group sequential test with $K = 3$

We now extend the results of Section 3.2 from the main paper to show that for a group sequential test with $K = 3$ analyses, if the canonical joint distribution does not hold, yet we assume it does anyway, the resulting test will have type 1 error rate less than the planned significance level $\alpha$. 

6
We follow a similar approach to the proof of the case $K = 2$. Let $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3$ be the sequence of treatment effect estimates that are calculated using the conditional score method. Let the true variance-covariance matrix for this sequence of estimates be $\Sigma$. Proceeding using method 1, we let the information levels be calculated as $I_k = (\Sigma_{kk})^{-1}$ and the $Z$-statistic is given by $Z_k = \hat{\eta}_k \sqrt{I_k}$ for $k = 1, 2, 3$. Under $H_0$ we have

$$\begin{bmatrix} \hat{\eta}_1 \\ \hat{\eta}_2 \\ \hat{\eta}_3 \end{bmatrix} \sim N_K \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{pmatrix} \right], \quad \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} \sim N_K \left[ \begin{pmatrix} 0 \\ 1 \\ \rho_{12} \rho_{13} \rho_{23} \\ \rho_{12} \rho_{13} \rho_{23} \end{pmatrix} \right]$$

(6)

where the correlation parameters are given by

$$\rho_{k_1k_2} = Cov(Z_{k_1}, Z_{k_2}) = \frac{\Sigma_{k_1k_2}}{\sqrt{\Sigma_{k_1k_1}\Sigma_{k_2k_2}}} \text{ for } 1 \leq k_1 < k_2 \leq 3.$$ 

(7)

Suppose instead, that we have a different sequence of treatment effect estimates $\hat{\eta}_1^*, \hat{\eta}_2^*, \hat{\eta}_3^*$ with distribution given below. The values of $\Sigma_{11}, \ldots, \Sigma_{33}$ are the same as in (6) and using the same information levels given by $I_k = (\Sigma_{kk})^{-1}$, we define the standardised statistics $Z_k^* = \hat{\eta}_k^* \sqrt{I_k}$ for $k = 1, 2, 3$. The joint distribution of the sequence $\hat{\eta}_1^*, \hat{\eta}_2^*, \hat{\eta}_3^*$ and the distribution of the $Z$-statistics are given by

$$\begin{bmatrix} \hat{\eta}_1^* \\ \hat{\eta}_2^* \\ \hat{\eta}_3^* \end{bmatrix} \sim N_K \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{pmatrix} \right], \quad \begin{bmatrix} Z_1^* \\ Z_2^* \\ Z_3^* \end{bmatrix} \sim N_K \left[ \begin{pmatrix} 0 \\ 1 \\ \rho_{12} \rho_{13} \rho_{23} \\ \rho_{12} \rho_{13} \rho_{23} \end{pmatrix} \right]$$

(8)

with correlation parameter

$$\rho_{k_1k_2}^* = Cov(Z_{k_1}^*, Z_{k_1}^*) = \frac{\Sigma_{k_1k_2}}{\sqrt{\Sigma_{k_1k_1}}} \text{ for } 1 \leq k_1 < k_2 \leq 3.$$ 

(9)

The following conditions, which are an adaptation of the conditions for the case $K = 2$ are assumed to hold.

Conditions 1.

1. The upper boundary of a group sequential trial, on the $Z$-scale, given by $b_1, \ldots, b_3$ is such that $b_1 \geq b_2 \geq b_3 \geq 0$

2. For all $k_1 < k_2$, we have $\Sigma_{k_1k_2} \geq \Sigma_{k_2k_2}$

In a similar manner to the proof for the case $K = 2$ in the main paper, we aim to show that

$$P(Z_1 > b_1 \cup Z_2 > b_2 \cup Z_3 > b_3) \leq P(Z_k^* > b_1 \cup Z_k^* > b_2 \cup Z_k^* > b_3).$$

First, note another representation for the above probabilities. For example the probability on the left hand side can be written as

$$P(Z_1 > b_1) + P(Z_2 > b_2) + P(Z_3 > b_3) - P(Z_1 > b_1 \cap Z_2 > b_2) - P(Z_1 > b_1 \cap Z_3 > b_3) - P(Z_2 > b_2 \cap Z_3 > b_3) + P(Z_1 > b_1 \cap Z_2 > b_2 \cap Z_3 > b_3)$$

(10)

The marginal distributions are equivalent, that is we have that $P(Z_k > b_k) = P(Z_k^* > b_k)$ for each $k = 1, 2, 3$. Let the remaining probabilities of Equation (10) be summarised by the functions $f(\cdot)$ and $g(\cdot)$, which are given by

$$f(\rho_{12}, \rho_{13}, \rho_{23}) = f_{12}(\rho_{12}) + f_{13}(\rho_{13}) + f_{23}(\rho_{23}) = P(Z_1 > b_1 \cap Z_2 > b_2) + P(Z_1 > b_1 \cap Z_3 > b_3) + P(Z_2 > b_2 \cap Z_3 > b_3)$$

$$g(\rho_{12}, \rho_{13}, \rho_{23}) = P(Z_1 > b_1 \cap Z_2 > b_2 \cap Z_3 > b_3).$$

Then we need to show that

$$g(\rho_{12}, \rho_{13}, \rho_{23}) - g(\rho_{12}^*, \rho_{13}^*, \rho_{23}^*) \leq f(\rho_{12}, \rho_{13}, \rho_{23}) - f(\rho_{12}^*, \rho_{13}^*, \rho_{23}^*).$$

(11)

The right hand side of Equation (11) is greater than or equal to zero. This can be seen by a similar argument as for the case $K = 2$, it is clear that

$$P(Z_1^* > b_1 \cap Z_2^* > b_2) \leq P(Z_1 > b_1 \cap Z_2 > b_2)$$

$$P(Z_1^* > b_1 \cap Z_3^* > b_3) \leq P(Z_1 > b_1 \cap Z_3 > b_3)$$

$$P(Z_2^* > b_2 \cap Z_3^* > b_3) \leq P(Z_2 > b_2 \cap Z_3 > b_3)$$

7
Table 1: Parameter choices for three cases to compare rate of change of objects $\mathbb{P}(Z_{k_1} > b_{k_1} \cap Z_{k_2} > b_{k_2})$ and $\mathbb{P}(Z_{k_1} > b_{k_1} \cap Z_{k_2} > b_{k_2} \cap Z_{k_3} > b_{k_3})$ for a group sequential trial with $K = 3$ analyses.

| Case    | $\rho_{12}, \rho_{13}, \rho_{23}$ | $\rho_{12}, \rho_{13}^*, \rho_{23}^*$ |
|---------|-----------------------------------|--------------------------------------|
| Case 1  | 0.849, 0.693, 0.980               | 0.707, 0.577, 0.816                 |
| Case 2  | 0.850, 0.765, 0.990               | 0.772, 0.695, 0.900                 |
| Case 3  | 0.751, 0.531, 0.919               | 0.577, 0.408, 0.707                 |

and this implies that $f(\rho_{12}^*, \rho_{13}^*, \rho_{23}^*) \leq f(\rho_{12}, \rho_{13}, \rho_{23})$. Suppose that $g(\rho_{12}, \rho_{13}, \rho_{23}) > g(\rho_{12}, \rho_{13}, \rho_{23})$, then Equation (11) holds so it is sufficient to consider the case $g(\rho_{12}, \rho_{13}, \rho_{23})$. Both sides of Equation (11) are greater than or equal to zero, and we know by Conditions 1 that $\rho_{12}^* \leq \rho_{12}, \rho_{13}^* \leq \rho_{13}$ and $\rho_{23}^* \leq \rho_{23}$. For Equation (11) to hold, we need that the function $f(\cdot)$ is increasing at a greater rate than $g(\cdot)$ as each parameter $\rho_{12}, \rho_{13}$ and $\rho_{23}$ increases.

We shall check that Equation (11) holds graphically for each parameter $\rho_{12}, \rho_{13}$ and $\rho_{23}$ individually. First note that $f(\rho_{12}, \rho_{13}, \rho_{23})$ only depends on $\rho_{12}$ through the function $f_{12}(\rho_{12}) = \mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2)$ and so plotting this probability as a function of $\rho_{12}$ is sufficient for checking the rate of change of $f(\rho_{12}, \rho_{13}, \rho_{23})$ with respect to $\rho_{12}$. We cannot check this for every combination of the parameters and so we follow the steps below to check the condition in a systematic way.

1. Choose values of $\rho_{12}, \rho_{13}, \rho_{23}, \rho_{12}^*, \rho_{13}^*$ and $\rho_{23}^*$. 
2. Check that $f(\rho_{12}, \rho_{13}, \rho_{23})$ increases in $\rho_{12}$ at a greater rate than $g(\rho_{12}, \rho_{13}, \rho_{23})$. Do this by plotting $f_{12}(\rho_{12})$ and $g(\rho_{12}, \rho_{13}, \rho_{23})$ against $\rho_{12}$. 
3. Check that $f(\rho_{12}, \rho_{13}, \rho_{23})$ increases in $\rho_{13}$ at a greater rate than $g(\rho_{12}, \rho_{13}, \rho_{23})$. Do this by plotting $f_{13}(\rho_{13})$ and $g(\rho_{12}, \rho_{13}, \rho_{23})$ against $\rho_{13}$. 
4. Check that $f(\rho_{12}, \rho_{13}, \rho_{23})$ increases in $\rho_{23}$ at a greater rate than $g(\rho_{12}, \rho_{13}, \rho_{23})$. Do this by plotting $f_{23}(\rho_{23})$ and $g(\rho_{12}, \rho_{13}, \rho_{23})$ against $\rho_{23}$. 

We shall consider three combinations of the parameters $\rho_{12}, \rho_{13}, \rho_{23}, \rho_{12}^*, \rho_{13}^*$ and $\rho_{23}^*$. The first scenario is based on equally spaced information levels and the $\rho$ parameters are divided by a factor of 1.2 to get the $\rho^*$ parameters. The second scenario is where the two interim analyses have higher information levels (than if information was equally spaced) and the $\rho$ parameters are divided by 1.1. The final scenario is where the two interim analyses have lower information levels (than if information was equally spaced) and the $\rho$ parameters are divided by 1.3. The parameter values for the three cases are presented in Table 1. Figure 1 compares the pairwise probabilities with the probability of crossing all three boundaries for each of the three cases in Table 1. In cases 1 and 3, as the parameter $\rho_{13} \to 1$, it is not clear whether $f_{13}(\rho_{13})$ increases at a greater rate than $g(\rho_{12}, \rho_{13}, \rho_{23})$. We have checked this numerically and found it to be true. Further the case $\rho_{13} = 1$ is the very rare case where all three analyses have the same information levels. This is not a scenario of concern. Therefore, it is clear that the function $f(\rho_{12}, \rho_{13}, \rho_{23})$ increases at a greater rate than the function $g(\rho_{12}, \rho_{13}, \rho_{23})$ in each parameter $\rho_{12}, \rho_{13}$ and $\rho_{23}$ and hence we have shown convincing evidence to prove that Equation (11) holds.

References

Per Kragh Andersen and Richard David Gill. Cox’s regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982.
Figure 1: Comparison of rate of change of $f(\rho_{12}, \rho_{13}, \rho_{23})$ and $g(\rho_{12}, \rho_{13}, \rho_{23})$ for a group sequential trial with $K = 3$ analyses. Fixed values of $\rho_{12}, \rho_{13}, \rho_{23}, \rho^*_{12}, \rho^*_{13}$ and $\rho^*_{23}$ given by Table 1.