Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2024, Vol. 33(11–12) 2098–2114

DOI: 10.1177/09622802241287711
journals.sagepub.com/home/smm
Sage

Adaptive enrichment trial designs using joint modelling of longitudinal and time-to-event data

Abigail J Burdon[1], **Richard D Baird**[2] and **Thomas Jaki**[1,3]

Abstract

Adaptive enrichment allows for pre-defined patient subgroups of interest to be investigated throughout the course of a clinical trial. These designs have gained attention in recent years because of their potential to shorten the trial's duration and identify effective therapies tailored to specific patient groups. We describe enrichment trials which consider long-term time-to-event outcomes but also incorporate additional short-term information from routinely collected longitudinal biomarkers. These methods are suitable for use in the setting where the trajectory of the biomarker may differ between subgroups and it is believed that the long-term endpoint is influenced by treatment, subgroup and biomarker. Methods are most promising when the majority of patients have biomarker measurements for at least two time points. We implement joint modelling of longitudinal and time-to-event data to define subgroup selection and stopping criteria and we show that the familywise error rate is protected in the strong sense. To assess the results, we perform a simulation study and find that, compared to the study where longitudinal biomarker observations are ignored, incorporating biomarker information leads to increases in power and the (sub)population which truly benefits from the experimental treatment being enriched with higher probability at the interim analysis. The investigations are motivated by a trial for the treatment of metastatic breast cancer and the parameter values for the simulation study are informed using real-world data where repeated circulating tumour DNA measurements and HER2 statuses are available for each patient and are used as our longitudinal data and subgroup identifiers, respectively.

Keywords

Efficient designs, enrichment, joint modelling, longitudinal data, time-to-event data

1 Introduction

In current oncology practice and cancer clinical trials, it is crucial to focus testing of novel therapies on the patient subgroups most likely to benefit. Too many patients receive treatments that either do not work particularly well, are toxic, or sometimes both. Adaptive enrichment clinical trials enable the efficient testing of an experimental intervention on specific patient subgroups of interest.[1,2] At an interim analysis, if a particular subgroup of patients is identified as responding particularly well to treatment, then we can focus resources and inferences by recruiting additional patients from the subgroup which benefits.

Simon and Simon[3] showed the benefits of enrichment trials, in particular that patients who do not appear to benefit are removed from the experimental treatment with potentially harmful side effects. If the treatment is futile for all patients, we are able to terminate the trial at interim analyses.[4] Further, if patients respond overwhelmingly well to treatment, then

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2]Department of Oncology, Cancer Research UK, Cambridge Centre, University of Cambridge, Cambridge, UK
[3]Department of Machine Learning and Data Science, University of Regensburg, Regensburg, Bayern, Germany

Corresponding author:
Abigail J Burdon, MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Robinson Way, Cambridge, CB2 1TN, UK.
Email: abigail.burdon@mrc-bsu.cam.ac.uk

there is potential to stop the trial early for efficacy demonstrating that the experimental treatment is superior to control in this subgroup, and the usual benefits of group sequential tests apply.[5] To combine the data from multiple stages and ensure that type 1 error rates are controlled, either a combination function approach,[6] or conditional error rate approaches[7,8] were originally proposed. In recent years, the computation of such designs have been streamlined and optimised for different purposes.[9,10,4,11] Extending upon Simon and Simon,[3] more complex designs which allow for more generlised data structures and targeted selection rules have been proposed.[12–14] A further advance upon enrichment designs are adaptive signature trials[15] which simultaneously identify and validate subgroup structures within a single trial protocol. These designs are based on cross-validation techniques and suffer from inefficiencies in the way that data is analysed and are subject to bias. More recently, designs have been proposed[16] which consider subgroup identification using a continuous biomarker. Such designs are based on an assumption a priori of a nested structure among subgroups.

In recent years, there has been increased uptake in enrichment trials which consider a long-term time-to-event (TTE) endpoint, such as overall survival (OS), but this is still low compared to continuous endpoints.[17] In such trials, it is common for investigators to also collect repeated measures on biomarkers. Recent research proposes methods which use the short-term endpoint data for subgroup selection rules then focus on the primary endpoint data for hypothesis testing.[18,19] Our aim is to leverage the additional biomarker information to improve interim decision making, early stopping rules *and* hypothesis testing.

We present a joint model for longitudinal and TTE data and base an enrichment trial design on the treatment effect in the joint model. There has been significant interest in joint modelling of longitudinal and TTE data[20,21] with a focus on prediction and personalised medicine. However, the uses of joint modelling have yet to be established in clinical trial designs. We show that by incorporating the longitudinal data into the analysis via joint modelling, this results in the subgroup which benefits being selected more frequently and higher power (using the same number of patients) as the equivalent trial which ignores the biomarker observations. Our simulation results are based on data from a study which measured OS and plasma circulating tumour DNA (ctDNA) levels.[22] To define subgroups, we hypothesise that patients who are HER2 negative will benefit from the experimental treatment more than patients who are HER2 positive.

Similarly to Magnusson and Turnbull,[23] we use the 'threshold selection' rule combined with an error spending test to clearly predefine the subgroup selection and stopping rules before the trial commences. We present a method where, in the setting of TTE data and joint modelling, the relationship between number of observed events and information levels can be exploited to design an efficient clinical trial. The novel feature of this work is an enrichment trial which uses a modern joint model to make both interim decisions and perform hypothesis testing.

2 Motivating example

Fragments of ctDNA are detected in the blood of cancer patients and are routinely measured in many cancer clinical trials. These measurements, which we shall often refer to as 'biomarker measurements' or 'longitudinal data' are useful prognostic factors that can improve the precision of OS estimates. Throughout this article, we shall base our analyses on data from a study which compared different biomarkers and their accuracy in monitoring tumour burden among women with metastatic breast cancer.[22] The results of the study were conclusive that ctDNA was successfully detected and highly correlated with OS.

Another important factor in breast cancer studies is the presence or absence of the HER2 protein. Patients who are HER2 positive may be resistant to conventional therapies and treatments that specifically target the HER2 protein are very effective.[24] Not only is OS influenced by HER2 status, but it is expected that ctDNA measurements are similar across HER2 status upon arrival to the trial and HER2− patients' ctDNA trajectories will increase more rapidly than HER2+. Adaptive enrichment trials are therefore highly efficient in breast cancer settings because the eligibility criteria based on HER2 status can be updated during the trial, restricting entry to patients likely to benefit.

3 Joint modelling of ctDNA and OS in defined subgroups

3.1 Subgroup set-up and notation

For adaptive enrichment trials, a key assumption is that subgroup identification is known prior to commencement. For the metastatic breast cancer example of Section 2, let S_1 denote the HER2 negative subgroup and let S_2 denote the HER2 positive subgroup. Then, let $F = S_1 \cup S_2$ denote the full population. Extensions to more subgroups can be made following the same logic. Further, we denote K as the total number of analyses in the adaptive trial and for our metastatic breast cancer example, we shall use $K = 2$.

The aim of a clinical trial is to assess how effective a new experimental treatment performs compared to an existing standard-of-care drug or placebo. We make statistical inferences based on a treatment effect θ which is defined at the design stage. For a trial with multiple subgroups, let θ_j be the treatment effect in subgroup $j = 1, 2, F$. A mathematical consequence is that if the prevalence of S_1 in F is given by λ, then

$$\theta_F = \lambda\theta_1 + (1 - \lambda)\theta_2 \tag{1}$$

Throughout, it is assumed that λ is known and fixed, however methods are available that account for uncertainty and allow estimation of λ at each analysis.[25] We aim to test the hypotheses

$$H_{0,j} : \theta_j \leq 0 \text{ vs } H_{A,j} : \theta_j > 0 \quad \text{for } j = 1, 2, F \tag{2}$$

3.2 The joint model

The joint model that we consider is based on equation (2) of Tsiatis and Davidian[26] (referred to as 'TD' for short). There are two processes in this model which represent the survival and longitudinal parts, and these processes are linked using random effects. The difference between our joint model and that of TD is that we have chosen to model the longitudinal data trajectory as linear in time whereas in TD, the parametric form for the biomarker is not specified. This appears appropriate for the example dataset of Section 2 as we have seen ctDNA display this property. The methods can easily be extended to incorporate more complex trajectories for the longitudinal data.

Let the times of the measurements of the longitudinal data for patient i in subgroup $j = 1, 2$ be denoted by $v_{ji1}, \ldots, v_{jim_{ji}}$, then $X_{ji}(v_{jis})$ is the true value of the biomarker at time v_{jis} and $D_{ji}(v_{jis})$ is the observed value of the biomarker. Suppose that $\mathbf{b}_{ji} = (b_{0ji}, b_{1ji})$ is a vector of patient specific random effects and that $\epsilon_{ji}(v_{jis})$ is the measurement error. We make the assumptions that $\epsilon_{ji}(v_{jis})|\mathbf{b}_{ji} \sim N(0, \sigma_j^2)$ for $s = 1, \ldots, m_{ji}$ and $\epsilon_{ji}(v)$ and $\epsilon_{ji}(v')$ are independent for $v \neq v'$. For the survival endpoint, we shall assume a Cox proportional hazards model. Let ψ_{ji} be the indicator function that patient i in subgroup $j = 1, 2$ receives the experimental treatment and let θ_j and γ_j be a scalar coefficients. Then the hazard function for subgroup j is denoted $h_{ji}(t)$ and the joint model takes the form

$$\begin{aligned} X_{ji}(v_{jis}) &= b_{0ji} + b_{1ji}v_{jis} \\ D_{ji}(v_{jis}) &= X_{ji}(v_{jis}) + \epsilon_{ji}(v_{jis}) \quad \text{for } j = 1, 2 \\ h_{ji}(t) &= h_{0j}(t)\exp\{\gamma_j X_{ji}(t) + \theta_j\psi_{ji}\} \end{aligned} \tag{3}$$

Equation (3) defines the joint model and defines the working model from which we shall perform simulation studies in Section 6. Parameter estimates in the joint model can then be found by fitting both longitudinal and survival outcomes to the joint model simultaneously and we shall describe this process in Section 3.3.

We note here that there is no treatment effect included in the biomarker trajectory. The motivation for this follows the models that are presented in the literature given by TD. For a more general model including a treatment effect in the longitudinal data, we refer the reader to Section A of the Supplemental Material where we discuss the use of the restricted mean survival time (RMST) endpoint which can account for multiple treatment effect parameters. The RMST methodology requires additional modelling assumptions and performs poorly under model misspecification, and for this reason we do not consider it further. Another method which can account for a treatment effect in the long-term data is the p-value combination approach[19] where treatment selection is based solely on longitudinal data and confirmatory decisions assess survival outcomes. In Section A of the Supplemental Material, we make a comparison between the joint modelling method and the p-value combination approach. The joint modelling method makes full use of all the information at each analysis, whereas the p-value combination method neglects useful information at each stage; ignoring available survival outcomes at the interim and ignoring biomarker observations at the final analysis.

3.3 Conditional score

To perform the adaptive enrichment trial, we must find treatment effect estimates and their distributions at analyses $k = 1, \ldots, K$ and subgroups $j = 1, 2, F$. To do so, we shall use a modified version of the conditional score method by TD which is a method for fitting the joint model to the data. We present multi-stage adaptations of some functions presented in TD. Let $t_{ji}^{(k)}$ be the observed event time and let $\delta_{ji}^{(k)}$ be the observed censoring indicator for patient i in subgroup $j = 1, 2$ at analysis k. This censoring event includes censoring patients who remain in the study at analysis k but have not yet observed the event at the given analysis. We denote the maximum follow-up time at analysis k by τ_k. To be included in the at-risk set at time t, the patient must have at least two longitudinal observations to fit the regression model. At analysis k, we define the at-risk process, $Y_{ji}^{(k)}(t) = \mathbb{I}\{t_{ji}^{(k)} \geq t, v_{ji2} \leq t\}$, counting process, $N_{ji}^{(k)}(t) = \mathbb{I}\{t_{ji}^{(k)} \leq t, \delta_{ji}^{(k)} = 1, v_{ji2} \leq t\}$ and function $dN_{ji}^{(k)}(t) = \mathbb{I}\{t \leq t_{ji}^{(k)} < t + dt, \delta_{ji}^{(k)} = 1, v_{ji2} \leq t\}$ for the joint model.

The conditional score methodology is motivated by the work of Stefanski and Carroll[27] who find efficient score functions for nonlinear models by conditioning on sufficient statistics. The authors first present a functional likelihood for a given statistical model which is shown to reduce to the ratio of measurement-error variance to equation-error variance. In turn, the sufficient statistic is often a function of the variance of the nuisance parameters which are being conditioned out, in our case, the random effects of the longitudinal data model. For patient i in subgroup j, let $\hat{X}_{ji}(v)$ be the ordinary least squares estimate of $X_{ji}(v)$ based on the set of measurements taken at times $\{v_{ji1}, \ldots, v_{jis} | v_{jis} \leq v\}$. That is, let $\bar{D}_{ji} = 1/s \sum_{m=1}^{s} D_{ji}(v_{jim})$ be the mean biomarker observation and let $\bar{v}_{ji} = 1/s \sum_{m=1}^{s} v_{jim}$ be the mean measurement time. Then the OLS estimate is given by $\hat{X}_{ji}(v) = \hat{b}_{0ji} + \hat{b}_{1ji} v$ where

$$\begin{aligned} \hat{b}_{1ji} &= \frac{\sum_{m=1}^{s} (D_{ji}(v_{jim}) - \bar{D}_{ji})(v_{jim} - \bar{v}_{ji})}{\sum_{m=1}^{s} (D_{ji}(v_{jim}) - \bar{D}_{ji})^2} \\ \hat{b}_{0ji} &= \bar{D}_{ji} - \hat{b}_{1ji} \bar{v}_{ji} \end{aligned}$$

Suppose that $\sigma_j^2 \psi_{ji}(v)$ is the variance of $\hat{X}_{ji}(v)$. TD define the sufficient statistic to be the function

$$S_{ji}^{(k)}(t, \gamma_j, \sigma_j^2) = \hat{X}_{ji}(t) + \gamma_j \sigma_j^2 \psi_{ji}(t) dN_{ji}^{(k)}(t)$$

which is defined for all $t \in (v_{ji2}, t_{ji}^{(k)})$ for patient i in subgroup j. The multi-stage version of the scalar E_{0i} of TD, dependent on subgroup j, is given by

$$E_{0ji}^{(k)}(t, \gamma_j, \theta_j, \sigma_j^2) = \exp\{\gamma_j S_{ji}^{(k)}(t, \gamma_j, \sigma_j^2) - \gamma_j^2 \sigma_j^2 \psi_{ji}(t)/2 + \theta_j \psi_{ji}\}$$

and the multi-stage version of the quotient function E_1/E_0 in equation (6) by TD, dependent on subgroup j, is the 2 × 1 vector given by

$$E_j^{(k)}(t, \gamma_j, \theta_j, \sigma_j^2) = \frac{\sum_{i=1}^{n_j} \{S_{ji}^{(k)}(t, \gamma_j, \sigma_j^2), \psi_{ji}\}^T E_{0ji}^{(k)}(t, \gamma_j, \theta_j, \sigma_j^2) Y_{ji}^{(k)}(t)}{\sum_{i=1}^{n_j} E_{0ji}^{(k)}(t, \gamma_j, \theta_j, \sigma_j^2) Y_{ji}^{(k)}(t)}$$

Then, the conditional score function at analysis k for subgroup $j = 1, 2$, also a vector of dimension 2 × 1, is given by

$$\begin{aligned} & U_j^{(k)}(\gamma_j, \theta_j, \sigma_j^2) \\ & = \int_0^{\tau_k} \sum_{i=1}^{n_j} \left(\{S_{ji}^{(k)}(t, \gamma_j, \sigma_j^2), \psi_{ji}\}^T - E_j^{(k)}(t, \gamma_j, \theta_j, \sigma_j^2) \right) dN_{ji}^{(k)}(t) \end{aligned} \tag{4}$$

3.4 Estimates for the treatment effects θ_j and their distributions

The aim is now to find treatment effect estimates $\hat{\theta}_j^{(k)}$ for $j = 1, 2, F$ and analyses $k = 1, \ldots, K$. We define these estimates as the root of the conditional score. In doing so, it turns out that these estimates are asymptotically normally distributed and we derive the variance of the estimates.

Burdon et al.[28] showed that $\mathbb{E}(U_j^{(k)}(\gamma_j, \theta_j, \sigma_j^2)) = \mathbf{0}$ for each $k = 1, \ldots, K$, and $j = 1, 2$. Therefore, the conditional score function at analysis k and subgroup $j = 1, 2$ is an estimating function, and set equal to zero defines an estimating equation. Hence, asymptotically normal parameter estimates for γ_j and θ_j can be found as the root of the estimating equation. As in TD equation (13), define the pooled estimate $\hat{\sigma}_j^{(k)2} = \sum_{i=1}^{n_j} \mathbb{I}\{m_{ji}(k) > 2\} R_{ji}(k) / \sum_{i=1}^{n_j} \mathbb{I}\{m_{ji}(k) > 2\}(m_{ji}(k) - 2)$, where $R_{ji}(k)$ is the residual sum of squares for the least squares fit to all $m_{ji}(k)$ observations for patient i in subgroup j available at analysis k. Then, let $\hat{\gamma}_j^{(k)}, \hat{\theta}_j^{(k)}$ be the values of γ_j and θ_j, respectively, such that

$$U_j^{(k)}(\hat{\gamma}_j^{(k)}, \hat{\theta}_j^{(k)}, \hat{\sigma}_j^{(k)2}) = \mathbf{0}$$

We also need to know the distribution of these estimates and this requires knowledge of the variance of $\hat{\theta}_j^{(k)}$. We shall use the sandwich estimator, as in Section 2.6 by Wakefield,[29] to calculate a robust estimate for the variance of the parameter estimates. Firstly, define matrices

$$\begin{aligned} A_j^{(k)} &= \partial U_j^{(k)}(\gamma_j, \theta_j, \sigma_j^2)/\partial(\gamma_j, \theta_j)^T \\ B_j^{(k)} &= Var(U_j^{(k)}(\gamma_j, \theta_j, \sigma_j^2)) \end{aligned}$$

Burdon et al.[28] presented analytical forms for each of these 2×2 matrices including a detailed calculation for the derivative matrix $A_j^{(k)}$. In practice, $A_j^{(k)}$ can be calculated numerically and $B_j^{(k)}$ is found by considering the conditional score as a sum over n_j patients. Further, these matrices are estimated by substituting the estimates $\hat{\gamma}_j^{(k)}, \hat{\theta}_j^{(k)}$ and $\hat{\sigma}_j^{(k)2}$ for γ_j, θ_j and σ_j^2, respectively. Then the information for the treatment effect estimate is given by the following equation:

$$\mathcal{I}_j^{(k)} = 1/Var(\hat{\theta}_j^{(k)}) = n_j \left[(A_j^{(k)})^{-1} B_j^{(k)} ((A_j^{(k)})^{-1})^T \right]_{22}^{-1}$$

for $j = 1, 2$ and $k = 1, \dots, K$. The subscript represents that we are interested in the second parameter θ_j in the vector $(\gamma_j, \theta_j, \sigma_j^2)^T$.

In accordance with equation (1), the treatment effect estimate and corresonding information level in the full population at analysis $k = 1, \dots, K$ are given by the following equation:

$$\begin{aligned} \hat{\theta}_F^{(k)} &= \lambda \hat{\theta}_1^{(k)} + (1 - \lambda)\hat{\theta}_2^{(k)} \\ \mathcal{I}_F^{(k)} &= \left(\lambda^2/\mathcal{I}_1^{(k)} + (1 - \lambda)^2/\mathcal{I}_2^{(k)} \right)^{-1} \end{aligned}$$

Finally, standardised Z-statistic is given by the following equation:

$$Z_j^{(k)} = \hat{\theta}_j^{(k)} \sqrt{\mathcal{I}_j^{(k)}} \quad \text{for } j = 1, 2, F \text{ and } k = 1, \dots, K$$

For simplicity in notation and exposition, we now return to the example of Section 2 in which $K = 2$. In order for subsequent results to hold, we require $Z_j^{(1)}, Z_j^{(2)}$ to have the 'canonical joint distribution' (CJD) given in Section 3.1 of Jennison and Turnbull[5] for each $j = 1, 2, F$. The CJD of the standardised statistics across analyses is such that

$$\begin{bmatrix} Z_j^{(1)} \\ Z_j^{(2)} \end{bmatrix} \sim N \left(\begin{bmatrix} \theta_j^{(1)} \sqrt{\mathcal{I}_j^{(1)}} \\ \theta_j^{(2)} \sqrt{\mathcal{I}_j^{(2)}} \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\mathcal{I}_j^{(1)}/\mathcal{I}_j^{(2)}} \\ \sqrt{\mathcal{I}_j^{(1)}/\mathcal{I}_j^{(2)}} & 1 \end{bmatrix} \right) \tag{5}$$

Burdon et al.[28] showed that the Z-statistics calculated using the conditional score methodology have approximately the CJD, but not exactly. The authors show that by proceeding with a group sequential test assuming that this does hold is sensible since type 1 error rates are conservative and diverge minimally from planned significance level. We give simulation evidence that this is also true for an adaptive enrichment trial in Section 6.

The proposed methods make certain assumptions that are needed to validate the CJD in equation (5). In Section C of the Supplemental Material, sensitivity analyses are performed where some of these assumptions are verified. In particular, we find that the conditional score estimator is robust to the assumption that residual errors in the longitudinal data are independent and asymptotic properties hold under small sample sizes. The results of the sensitivity analyses suggest that a minimum of 20 events per subgroup are required at the interim analysis to ensure control of type 1 error rates.

4 Adaptive enrichment schemes for clinical trials with subgroup selection

4.1 The threshold selection rule

An adaptive enrichment scheme consists of two decisions; firstly a decision upon which subgroup, if any, to continue the trial with at the interim analysis and secondly, a decision upon whether or not to reject the null hypothesis at the final analysis. There are a collection of rules which can be used for subgroup selection, for example, the maximum test statistic[12] and a Bayes optimal rule.[4]

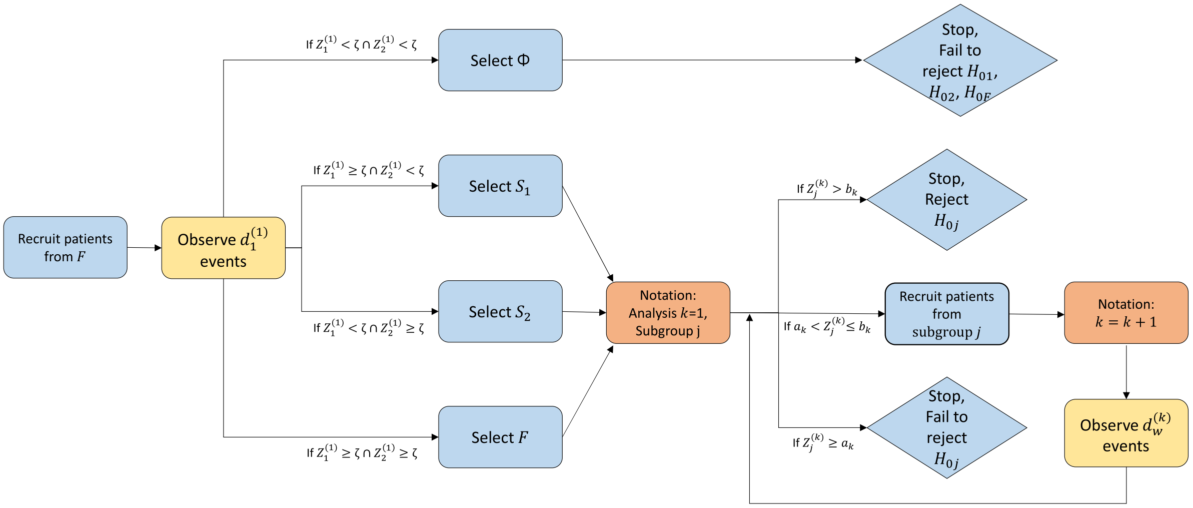
If $Z_1^{(1)} < \zeta \cap Z_2^{(1)} < \zeta$
Select Φ
Stop, Fail to reject H_{01}, H_{02}, H_{0F}
If $Z_1^{(1)} \geq \zeta \cap Z_2^{(1)} < \zeta$
Select S_1
Recruit patients from F
Observe $d_1^{(1)}$ events
If $Z_j^{(k)} > b_k$
Stop, Reject H_{0j}
Select S_2
If $Z_1^{(1)} < \zeta \cap Z_2^{(1)} \geq \zeta$
Notation: Analysis k=1, Subgroup j
If $a_k < Z_j^{(k)} \leq b_k$
Recruit patients from subgroup j
Notation: $k = k + 1$
Select F
If $Z_1^{(1)} \geq \zeta \cap Z_2^{(1)} \geq \zeta$
If $Z_j^{(k)} \geq a_k$
Stop, Fail to reject H_{0j}
Observe $d_w^{(k)}$ events

Figure 1. Flowchart for enrichment trial design which uses the threshold rule for subgroup selection at the interim analysis. Hypothesis testing is based on an error spending design with α-spending for the efficacy boundary and β-spending for the futility boundary including the opportunity for early stopping. The flowchart describes when the interim analysis should be performed based on the pre-planned number of events $d_1^{(1)}$ in subgroup S_1 at the interim and the total number of observed events $d^{(2)}$ in the selected subgroup at the final analysis.

Similarly to Magnusson and Turnbull,[23] we shall use the threshold selection rule. The definition is as follows; for some constant ζ, select all subgroups $j \in \{1, 2\}$ such that $Z_j^{(1)} > \zeta$ (Figure 1). If $Z_1^{(1)} > \zeta$ and $Z_2^{(1)} > \zeta$ then the trial continues in the full population and it should be noted that this is a stronger condition than $Z_F^{(1)} > \zeta$ as in the latter case, overwhelming benefit in one subgroup with poor effect in the other could still lead to selection of the full population. Finally, if $Z_1^{(1)} \leq \zeta$ and $Z_2^{(1)} \leq \zeta$ then the trial stops at the interim analysis declaring the treatment to be in-efficacious in all subgroups. This ensures that only subgroups which have a large enough treatment effect are followed to the second analysis. The threshold selection rule leads to an efficient enrichment trial design because we can find analytical forms for the type 1 and type 2 error rates and are, therefore, able to maximise power. As well as clearly providing the generic design framework for any test statistic, a novel aspect of this work will be applying this rule in the joint modelling setting.

To begin, we describe the probability distribution of the population index. At the interim analysis, let W be the random variable which represents the decision about which subgroup has been selected. Let w be the realisation of W and this can take any value in the set $\Omega = \{1, 2, F, \emptyset\}$. The notation $\emptyset$ indicates that it is possible to stop the trial for futility at the interim analysis without selecting a subgroup. Given the threshold selection rule and a configuration on parameters $\boldsymbol{\Theta} = (\theta_1, \theta_2)$, we have

$$\begin{aligned}
\mathbb{P}(W = 1; \boldsymbol{\Theta}) &= \mathbb{P}(Z_1^{(1)} > \zeta \cap Z_2^{(1)} \leq \zeta; \boldsymbol{\theta}) \\
\mathbb{P}(W = 2; \boldsymbol{\Theta}) &= \mathbb{P}(Z_1^{(1)} \leq \zeta \cap Z_2^{(1)} > \zeta; \boldsymbol{\theta}) \\
\mathbb{P}(W = F; \boldsymbol{\Theta}) &= \mathbb{P}(Z_1^{(1)} > \zeta \cap Z_2^{(1)} > \zeta; \boldsymbol{\theta}) \\
\mathbb{P}(W = \emptyset; \boldsymbol{\Theta}) &= \mathbb{P}(Z_1^{(1)} \leq \zeta \cap Z_2^{(1)} \leq \zeta; \boldsymbol{\theta})
\end{aligned} \tag{6}$$

In order for the proposed methods to apply and to ensure control of type 1 error rates, ζ must be specified in advance of the trial. To choose such a value, the desired operating characteristics are considered. First, we define the configuration of parameters under the global null as $\boldsymbol{\Theta}_G : \{\theta_1 = \theta_2 = \theta_F = 0\}$ and the alternative as $\boldsymbol{\Theta}_A : \{\theta_1 = \delta, \theta_2 = 0, \theta_F = \lambda\delta\}$. This represents that we believe there is an important effect of treatment in S_1. For the metastatic breast cancer example in Section 2, this reflects that the HER2 negative subgroup is expected to respond well to the treatment. Equation (6) can then be solved for ζ and $\mathcal{I}_1^{(1)}$. Since there are two unknowns, only two equations need be considered and we focus attention on those representing enrichment of the biomarker positive subgroup and continuing in the full population since these are the two most desirable outcomes in this order. As an example, with $\delta = -0.5$, $\mathbb{P}(W = 1; \boldsymbol{\Theta}_A) = 0.6$ and $\mathbb{P}(W = F; \boldsymbol{\Theta}_A) = 0.2$, we therefore need $\zeta = 0.674$ and $\mathcal{I}_1^{(1)} = 9.19$.

Sensitivity analyses for different threshold selection rules are included in Section B of the Supplemental Material. The choice of $\mathbb{P}(W = 1; \boldsymbol{\Theta}_A)$ is influential in the sample size calculation and should be at least 0.5 to ensure that asymptotic assumptions for the conditional score estimator are valid. The choice of $\mathbb{P}(W = F; \boldsymbol{\Theta}_A)$ has an effect on the number of required events at the final analysis.

We now present the joint distribution of the subgroup selection decision and the selected test statistic which will be needed for calculation of type 1 and type 2 error rates. Let $f_{Z_W^{(1)}|W}(z_w^{(1)}|W = w; \boldsymbol{\Theta})$ be the conditional distribution of the test statistic $Z_w^{(1)}$ given that w has been selected. Then the joint probability density function is

$$f_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}) = \mathbb{P}(W = w; \boldsymbol{\Theta}) f_{Z_W^{(1)}|W}(z_w^{(1)}|W = w; \boldsymbol{\Theta})$$

We note that the random variable $Z_\phi^{(1)}$ is not currently defined since if no subgroup is selected we cannot calculate a subgroup standardised statistic. However, it will be seen that the joint probability density function $f_{Z_W^{(1)},W}(z_\phi^{(1)}, \phi; \boldsymbol{\Theta})$ is independent of $z_\phi^{(1)}$ and this joint probability function still has meaning. By equation (5), the test statistics are such that $Z_w^{(1)} \sim N(\theta_w \sqrt{I_w^{(1)}}, 1)$ for $w = 1, 2$ and $Z_1^{(1)}$ and $Z_2^{(1)}$ are independent. The conditional distribution $f_{Z_W^{(1)}|W}(z_w^{(1)}|W = w; \boldsymbol{\Theta})$ is given by a truncated normal distribution bounded below by ζ. Hence, we have

$$\begin{aligned}
f_{Z_W^{(1)},W}(z_1^{(1)}, 1; \boldsymbol{\Theta}) &= \Phi\left(\zeta - \theta_2\sqrt{\mathcal{I}_2^{(1)}}\right) \phi\left(z_1^{(1)} - \theta_1\sqrt{\mathcal{I}_1^{(1)}}\right) \\
f_{Z_W^{(1)},W}(z_2^{(1)}, 2; \boldsymbol{\Theta}) &= \Phi\left(\zeta - \theta_1\sqrt{\mathcal{I}_1^{(1)}}\right) \phi\left(z_2^{(1)} - \theta_2\sqrt{\mathcal{I}_2^{(1)}}\right) \\
f_{Z_W^{(1)},W}(z_F^{(1)}, F; \boldsymbol{\Theta}) &= \frac{\sqrt{\mathcal{I}_1^{(1)}\mathcal{I}_2^{(1)}}}{\lambda(1-\lambda)\mathcal{I}_F^{(1)}} \\
&\quad \times \int_{-\infty}^{\infty} \phi\left(\frac{\sqrt{\mathcal{I}_1^{(1)}}(u - \lambda\sqrt{\mathcal{I}_F^{(1)}})}{\lambda\sqrt{\mathcal{I}_F^{(1)}}}\right) \phi\left(\frac{\sqrt{\mathcal{I}_2^{(1)}}(z_F^{(1)} - u - (1-\lambda)\sqrt{\mathcal{I}_F^{(1)}})}{(1-\lambda)\sqrt{\mathcal{I}_F^{(1)}}}\right) du \\
f_{Z_W^{(1)},W}(z_\phi^{(1)}, \phi; \boldsymbol{\Theta}) &= \Phi\left(\zeta - \theta_1\sqrt{\mathcal{I}_1^{(1)}}\right) \Phi\left(\zeta - \theta_2\sqrt{\mathcal{I}_2^{(1)}}\right)
\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density and cumulative distribution functions, respectively. We derive $f_{Z_W^{(1)},W}(z_F^{(1)}, F; \boldsymbol{\Theta})$ in Section B of the Supplemental Material.

The methods presented are unconventional in the fact that we allow enrichment of the biomarker-negative subgroup. We have chosen this structure to allow for maximum flexibility and a novel solution for the enrichment trial where the investigator really believes no hierarchy among subgroups. The proposed design can also be modified to adhere to conventional standards by making small adjustments. For example, the definition of the threshold selection rule becomes; select F if $Z_1^{(1)} > \zeta$ and $Z_2^{(1)} > \zeta$, select S_1 if $Z_1^{(1)} > \zeta$ and $Z_2^{(1)} \leq \zeta$, otherwise stop the trial at the interim analysis if $Z^{(1)} \leq \zeta$. The population index can now take values in the set $\Omega = \{1, F, \emptyset\}$. Then, the conditional distributions $f_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta})$ remain unchanged for $w = 1, F, \emptyset$ and all following equations hold under this new definition.

4.2 Calculation of type 1 error and power

We now consider the possible pathways of the enrichment trial. Then, given the definition of the Z-statistics, the threshold selection rule and the joint probability density function $f_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta})$, we are equipped to determine error rates for the study. We shall apply this method in Section 3.3 in order to create an enrichment trial using the joint model for longitudinal and TTE data. The family wise error rate (FWER), denoted by α, is defined as the probability of rejecting one or more true null hypotheses H_j and power is denoted by $1 - \beta$.

The testing procedure for this adaptive enrichment trial is described in Figure 1. At analysis k, let (a_k, b_k) be an interval that splits the real line into three sections. We stop for futility if the test statistic of the selected subgroup, $Z_w^{(k)}$, is below a_k, reject the corresponding null hypothesis and stop for efficacy if $Z_w^{(k)}$ is above b_k and otherwise continue to analysis $k + 1$.

Let H_G be the global null hypothesis, $\theta_1 = \theta_2 = \theta_F = 0$. There are many pathways which lead to rejecting H_G. Examples include select F and reject $H_{0,F}$ at the interim or select S_1 then reject $H_{0,1}$ at the final analysis. Considering all options, we have

$$\alpha = \sum_{w\in\Omega} \left\{ \int_{b_1}^{\infty} f_{Z_w^{(1)},W}\left(z_w^{(1)}, w; \boldsymbol{\Theta}_G\right) dz_w^{(1)} \right.$$
$$\left. + \int_{a_1}^{b_1} \int_{b_2}^{\infty} f_{Z_w^{(2)}|Z_w^{(1)}}\left(z_w^{(2)}|z_w^{(1)}; \boldsymbol{\Theta}_G\right) dz_w^{(2)} dz_w^{(1)} \right\} \tag{7}$$

Here, we have specified that we will only test the hypothesis corresponding to the selected subgroup, since it has the highest chance of being significant. For alternative configurations testing all hypotheses, fixed sequence testing[30] or other alpha propagation methods[31] can be applied.

As is common in the literature,[12,18,19] we define power as the conditional probability of rejecting $H_{0,1}$ given that subgroup S_1 is selected. Here, S_1 can be arbitrarily interchanged for S_2 or F. This reflects the belief that a 'successful' trial is one where the subgroup which benefits is selected and also reports a positive trial outcome. Following the same arguments as for type 1 error, type 2 error rates are calculated as

$$\beta = \int_{-\infty}^{a_1} f_{Z_W^{(1)},W}\left(z_1^{(1)}, 1; \boldsymbol{\Theta}_A\right) dz_1^{(1)}$$
$$+ \int_{a_1}^{b_1} \int_{-\infty}^{a_2} f_{Z_1^{(2)}|Z_1^{(1)}}\left(z_1^{(2)}|z_1^{(1)}; \boldsymbol{\Theta}_A\right) dz_1^{(2)} dz_1^{(1)} \tag{8}$$

It is now clear that the boundary points a_1, a_2, b_1 and b_2 can be calculated to satisfy pre-specified requirements of FWER α, under $\boldsymbol{\Theta}_G$, and power $1 - \beta$, under $\boldsymbol{\Theta}_A$. Further, to ensure that we have four equalities for the four boundary points, we make additional requirements that $\alpha^{(k)}$ is the type 1 error 'spent' and $\beta^{(k)}$ is the type 2 error spent at analysis k where $\alpha^{(1)} + \alpha^{(2)} = \alpha$ and $\beta^{(1)} + \beta^{(2)} = \beta$. Then solve

$$\alpha^{(1)} = \sum_{w\in\Omega} \int_{b_1}^{\infty} f_{Z_W^{(1)},W}\left(z_w^{(1)}, w; \boldsymbol{\Theta}_G\right) dz_w^{(1)}$$
$$\alpha^{(2)} = \sum_{w\in\Omega} \int_{a_1}^{b_1} \int_{b_2}^{\infty} f_{Z_w^{(2)}|Z_w^{(1)}}\left(z_w^{(2)}|z_w^{(1)}; \boldsymbol{\Theta}_G\right) dz_w^{(2)} dz_w^{(1)}$$
$$\beta^{(1)} = \int_{-\infty}^{a_1} f_{Z_W^{(1)},W}\left(z_1^{(1)}, 1; \boldsymbol{\Theta}_A\right) dz_1^{(1)}$$
$$\beta^{(2)} = \int_{a_1}^{b_1} \int_{-\infty}^{a_2} f_{Z_w^{(2)}|Z_w^{(1)}}\left(z_1^{(2)}|z_1^{(1)}; \boldsymbol{\Theta}_A\right) dz_1^{(2)} dz_1^{(1)}$$

The decomposition of the error rates also ensures that the boundary points a_1 and b_1 can be calculated at the first analysis before observing the information levels at the second analysis. Hence, there may be the opportunity to stop the trial early without needing to calculate the information levels at the second analysis. This is particularly helpful in trials which use TTE endpoints as information levels are estimated using the data.

There are many options for the break-down of the error rates. For the models considered, we shall use an error spending design.[32] In the group sequential setting (without subgroup selection), the error spending test requires specifying the maximum information $\mathcal{I}_{max}$ and then error is spent according to the proportion of information $\mathcal{I}^{(k)}/\mathcal{I}_{max}$ observed at analysis k. For the enrichment trial, we propose a similar structure considering $\mathcal{I}_{max}$ to be the maximum information in the full population. Specifically, we shall use the functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$ to determine the amount of error to spend. Then we set

$$\begin{aligned}
\alpha^{(1)} &= f\left(\mathcal{I}_F^{(1)}/\mathcal{I}_{max}\right)\\
\alpha^{(2)} &= f\left(\mathcal{I}_F^{(2)}/\mathcal{I}_{max}\right) - f\left(\mathcal{I}_F^{(1)}/\mathcal{I}_{max}\right)\\
\beta^{(1)} &= g\left(\mathcal{I}_F^{(1)}/\mathcal{I}_{max}\right)\\
\beta^{(2)} &= g\left(\mathcal{I}_F^{(2)}/\mathcal{I}_{max}\right) - g\left(\mathcal{I}_F^{(1)}/\mathcal{I}_{max}\right)
\end{aligned}$$

We shall discuss the calculation of $\mathcal{I}_{max}$ in the TTE (or joint modelling) setting in Section 4.4.

By construction, under $H_G : \theta_1 = \theta_2 = \theta_F = 0$, we have FWER α exactly by equations (7) and (8). Hence, the FWER is protected in the weak sense. To prove that we also have strong control of the FWER, we impose the condition that the treatment effect in the full population, is non-negative. This ensures that the subgroup selected does not differ under scenarios $\boldsymbol{\Theta} = (\theta_1, \theta_2)$ and $\boldsymbol{\Theta} = (0, 0)$ which is needed for the proof. The condition is not restrictive, since treatment effects other than θ_F are allowed to be negative and θ_F can equal zero.

Theorem 1. *For global null hypothesis H_G and any $\boldsymbol{\Theta} = (\theta_1, \theta_2)$ such that $\theta_F = \lambda\theta_1 + (1 - \lambda)\theta_2$ is non-negative, we have*

$$\mathbb{P}(\textit{Reject at least one true } H_j|\boldsymbol{\Theta}) \le \mathbb{P}(\textit{reject at least one } H_j|H_G)$$

Proof. See Section B of the Supplemental Material. □

In Section 6, we also show by simulation, that the FWER is protected at significance level $\alpha = 0.025$ and is not conservative.

4.3 Trials with unpredictable information increments: Events based analyses

To complete the calculation of the boundary points a_2 and b_2 in equations (7) and (8), it remains to find the information level at analysis 2 for the subgroups that have ceased to be observed. That is, suppose that $w \in \{1, 2, F\}$ is the subgroup that has been selected and the trial continues to analysis 2, then $\mathcal{I}_w^{(2)}$ is observed. However, we also need to know $\mathcal{I}_j^{(2)}$ for all $j = 1, 2, F$ such that $w \neq j$. Many enrichment trial designs focus on the simple example where the outcome measure is normally distributed with known variance. Hence, if the number of patients to be recruited is pre-specified, then information levels can be calculated in advance of the trial and this problem does not occur. However, in trials where the primary endpoint is a TTE variable, information is estimated using the data. We find that we can accurately forward predict the information levels at future analyses when we know the number of observed events. Hence, to mitigate the problem of not knowing $\mathcal{I}_j^{(2)}$, we shall pre-specify the number of observed events.

For subgroup $j = 1, 2$, let $d_j^{(k)}$ be the number of events observed in subgroup j by analysis k. We plan that if no early stopping occurs, then the total number of observed events in the selected subgroup is the same regardless of which subgroup has been selected so that $d_1^{(2)} = d_2^{(2)} = d_F^{(2)} = d^{(2)}$. Figure 1 identifies when the analyses are performed. Note that these values are set as design options and so will be known before commencement of the trial. We shall discuss how to choose these values in Section 4.4.

Further, we relate number of events and information so that we can predict the information level at the second analysis for the unobserved subgroups. Freedman[33] proves that, in the context of survival analysis, the variance of the log-rank statistic under H_G is such that $\mathcal{I}_j^{(k)} \approx d_j^{(k)}/4$. For analysis methods using test statistics other than the log-rank, we shall extend on this idea and assume that $\mathcal{I}_j^{(k)} = d_j^{(k)}/m_j$, where m_j is a constant. Figure 2 shows evidence that the assumed relationship between number of events and information holds.

For now, we need only the assumption of the structural form of this relationship. At the interim analysis, each $\mathcal{I}_j^{(1)}$ is observed for $j = 1, 2, F$. Hence, we can use the proportionality relationship to predict the information at the second analysis for the subgroup which is no longer observed. For $j \neq w$, we can predict $\mathcal{I}_j^{(2)}$ using

$$\mathcal{I}_j^{(2)} = d_j^{(2)}\frac{1}{m_j} = d^{(2)}\frac{\mathcal{I}_j^{(1)}}{d_j^{(1)}} \quad \text{for } j = 1, 2, F$$

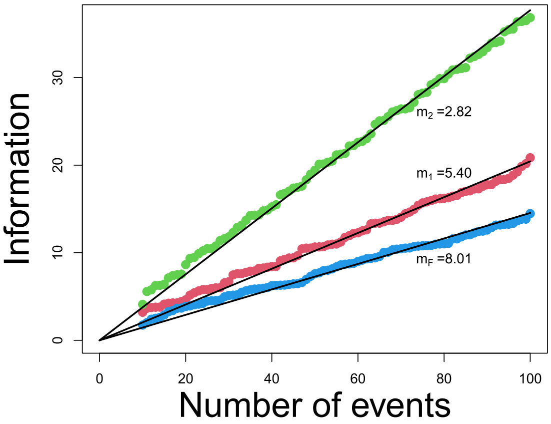

Figure 2. Calculation of constants m_1, m_2 and m_F. Result shows that information is proportional to number of events.

4.4 Trial design – number of events

We have so far presented the calculation of the boundary points for a trial where the number of events at the interim and final analyses are known prior to commencement. We now discuss the design of the trial, in particular, determining the constants m_j and information levels $\mathcal{I}_j^{(1)}$ for $j = 1, 2, F$ and maximum information level $\mathcal{I}_{max}$. These in turn mean that the required numbers of events $d_j^{(1)}$ for $j = 1, 2, F$ and $d^{(2)}$ can be planned. The driving design feature is that we will plan the trial to have power $1 - \beta$ under the parameterisation $\mathbf{\Theta}_A$. We now describe a simulation scheme to determine the constants m_j for $j = 1, 2, F$.

Algorithm 1.

```
*Under the parameterisation* $\mathbf{\Theta}_A$*, simulate a data set of* 5000 *patients*
*Let* $t_{j,1}, \ldots, t_{j,n_j}$ *be the event times in subgroup* $j$
***for*** $t_{j,s} = t_{j,1}, \ldots, t_{j,n_j}$
***do*** *Right-censor all patients at time* $t_{j,s}$
*Calculate* $\mathcal{I}_{j,s}^{(1)}$ *based on data up to time* $t_{j,s}$
***end for***
*Fit a linear model, without an intercept term, to the points* $(t_{j,1}, \mathcal{I}_{j,1}^{(1)}), \ldots, (t_{j,n_j}, \mathcal{I}_{j,n_j}^{(1)})$
*Use this linear model to estimate the value of* $m_j$.
```

Figure 2 gives a graphical representation of this scheme. It is now possible to calculate the required number of events at the first interim analysis. In the example in Section 4.1, we require $\mathcal{I}_1^{(1)} = 9.08$ which equates to $d_1^{(1)} = 9.08m_1$ events in subgroup S_1. Further, we find that $m_2 = (1-\lambda)m_1/\lambda$ and $m_F = m_1/\lambda$ which equates to $d_2^{(1)} = (1-\lambda)d_1^{(1)}/\lambda$ and $d_F^{(1)} = d_1^{(1)}/\lambda$ and this can be seen in Figure 2. The design of the trial does not require us to plan $d_2^{(1)}$ and $d_F^{(1)}$, but this provides us with estimates of the number of events that will be observed at the first analysis. We can also determine the timing of the final analysis at $K = 2$. Consider the sequence of information levels given by the following equation:

$$(\tilde{\mathcal{I}}_j^{(1)}, \tilde{\mathcal{I}}_j^{(2)}) = \left(d_j^{(1)}/m_j, m_F \mathcal{I}_{max}/m_j\right)$$

for $j \in= 1, 2, F$. The value of $\mathcal{I}_{max}$ is calculated such that boundary points satisfy $a_K = b_K$ when the information levels $\tilde{\mathcal{I}}_j^{(k)}$ replace $\mathcal{I}_j^{(k)}$ in equations (7) and (8) for $k = 1, 2$ and $j = 1, 2, F$. This is done using an iterative search method. Then, returning to the definition of $\mathcal{I}_{max}$, the total number of events can be found by solving $\mathcal{I}_F^{(2)} = \mathcal{I}_{max}$ for $d^{(2)}$. In Section 6, we present the sample sizes which have been calculated for a range of parameter choices.

5 Alternative models and their analysis methods

5.1 Cox proportional hazards model

Methods which leverage information from biomarkers in TTE data in enrichment trials are yet to be established. The current best practice for adaptive designs with a TTE endpoint is to base analyses on Cox proportional hazards models. We emulate this conventionality in order to assess the gain from including the longitudinal data in the analysis. To do so, we shall present a simple Cox proportional hazards model and define treatment effect estimates that can be used in accordance with the threshold selection rule to perform an enrichment trial.

Denote $h_{0j}(t)$ as the baseline hazard function, θ_j the treatment parameter and ψ_{ji} as the treatment indicator that patient i in subgroup $j = 1, 2$ receives the new treatment. Then the hazard function for the survival model is given by

$$h_{ji}(t) = h_{0j}(t)\exp\{\theta_j \psi_{ji}\} \tag{9}$$

We note the similarities and differences between this model and the joint model of Section 3.2. In the results that follow in Section 6.3, we shall assume that the joint model is true (and simulate data from the joint model). However, we fit the data to the Cox proportional hazards model which highlights that this will be a misspecified model.

When analysing data using this model, the null hypothesis in equation (2) can be tested at analysis $k = 1, \ldots, K$ by calculating treatment effect estimates $\hat{\theta}_j^{(k)}$, information levels $\mathcal{I}_j^{(k)}$ and Z-statistics for $j = 1, 2, F$. As described in Section D of the Supplemental Material, $\hat{\theta}_j^{(k)}$ is given as the root of the equation where the partial score statistic is set equal to zero[34] and the information $\mathcal{I}_j^{(k)}$ as the first derivative of the partial score statistic. Jennison and Turnbull[34] proved that the resulting Z-statistics have the CJD given in equation (5) and the methodology of Section 4 can be used to create an enrichment trial design.

5.2 Cox proportional hazards model with longitudinal data as a time-varying covariate

A final option for analysis is one where the longitudinal data is included but is assumed to be free of measurement error. This requires a more sophisticated model than the simple Cox proportional hazards model of Section 5.1 and represents a trial where the longitudinal data is regarded as important enough to be considered and included in the model. However, this is still a naive approach since the model will be misspecified in the presence of measurement error. For the purpose of assessing the necessity of correctly modelling the data, we shall fit a Cox proportional hazards model to the data where the longitudinal data is treated as a time-varying covariate.

In what follows, the definitions of the treatment indicator ψ_{ji} and longitudinal data measurements $D_{ji}(v_{ji1}), \ldots, D_{ji}(v_{jim_{ji}})$ remain the same as in Section 3.2. Let γ_j and θ_j be longitudinal data and treatment parameters, respectively, then the hazard function is given by

$$h_{ji}(t) = h_{j0}(t)\exp\{\gamma_j D_{ji}(t) + \theta_j \psi_{ji}\} \tag{10}$$

This model differs from the joint model because the assumption here is that $D_{ji}(t)$ is a function of time that is measured without error. In reality, we often have measurements $D_{ji}(v_{ji1}), \ldots, D_{ji}(v_{jim_{ji}})$ for patient i in subgroup j that include noise around a true underlying trajectory.

In a similar manner to Section 5.1, the hypothesis in equation (2) can be tested by finding Z-statistics, with the CJD of equation (5)[34] and following the enrichment trial design of Section 4.

6 Results

6.1 Simulation set-up

In what follows, we perform simulation studies to assess the type 1 error rates and observed power for the three analysis methods of Sections 3 and 5. These methods shall herto be referred to as 'Conditional score', 'Cox' and 'Cox with biomarker', respectively. The purpose of this comparison is to assess the gain by including the longitudinal data and to decide whether correctly modelling the measurement error is necessary.

For the presented analyses, we shall assume that the joint model is true. Hence, the working model for data generation is given by equation (3). Each of the analysis methods have the advantage that we need not specify the baseline hazard function since each method is semi-parametric and requires no assumptions regarding $h_{0j}(t)$. Even when the method includes the longitudinal data, there are no distributional assumptions about the random effects $\mathbf{b}_{j1}, \ldots, \mathbf{b}_{jn_j}$, ensuring it is robust to some model misspecifications. For the purpose of simulation however, we now describe the distributions used for data generation. We shall simulate data with baseline hazard function given by the following equation:

$$h_{0j}(t) = \begin{cases} c_{j1} & \text{if } t \leq 1 \\ c_{j2} & \text{if } t > 1 \end{cases} \tag{11}$$

We have chosen to simulate from a model where the baseline hazard function as piece-wise constant with a single knot-point at time $t = 1$ for simplicity. This is motivated by the metastatic breast cancer data where we see a sharp difference in the baseline hazard at one year. It is straight forward to extend this to a general piece-wise constant baseline hazard function with multiple knot-points. We consider a random effects model where $\mathbf{b}_{j1}, \ldots, \mathbf{b}_{jn}$ are independent and identically distributed with the following distribution:

$$\begin{bmatrix} b_{0ji} \\ b_{1ji} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{1j} \\ \mu_{2j} \end{bmatrix}, \begin{bmatrix} \phi_{1j} & \phi_{12j} \\ \phi_{12j} & \phi_{2j} \end{bmatrix} \right) \tag{12}$$

The parameter values for simulation studies are informed using the metastatic breast cancer dataset.[22] We removed patients whose ER status is negative and measurements of ctDNA which were ‘not detected’ were set to 1.5 (copies/mL).[35] The dataset contains multiple treatment arms and dosing schedules, hence, we use this dataset to represent standard of care (control group). The parameter values, which have been suitably rounded, shall remain fixed throughout the simulation studies are given by the following equation:

$$\begin{aligned} \lambda = 2/3, \quad \gamma = \gamma_1 = \gamma_2 = 0.8 \\ (\phi_1, \phi_{12}, \phi_2) = (\phi_{11}, \phi_{121}, \phi_{21}) = (\phi_{12}, \phi_{122}, \phi_{22}) = (2.5, 1.7, 5) \\ \sigma^2 = \sigma_1^2 = \sigma_2^2 = 0.25, \quad (\mu_{01}, \mu_{11}) = (\mu_{02}, \mu_{12}) = (4.23, 1.81) \\ c_{11} = c_{21} = 0.0085, \quad c_{12} = c_{22} = 0.0142 \end{aligned} \tag{13}$$

We shall perform simulation studies for a range of γ, σ^2 and ϕ_2 values. The interpretation of these parameter are now described. γ describes the association between the biomarker and TTE outcomes. Higher values of γ lead to higher correlation between the two endpoints. The parameter σ^2 controls the noise in the measurement error of the longitudinal data. Finally, ϕ_2 represents the variance of the slopes of the random effects terms and therefore the degree of similarity between patients’ longitudinal trajectories.

For our simulations, patients are recruited at a rate of 2 per week so that enrollment is slow and adaptive methods are appropriate. The recruitment ratio of control to experimental treatment is fixed as 1:1 for all subgroups and all simulations studies. ctDNA observations will be collected, via a blood test, at 2 weeks for the first 3 months following entry to study and then once per month. The final object of importance which is required for data generation is the mechanism which simulates censoring times, $y_1, \ldots, y_n$. We shall simulate these according to an exponential distribution with rate parameter 5×10^{-5} (years) and this is independent of the TTE outcome to reflect non-informative censoring. This results in roughly 10% of patients being lost to follow-up.

To complete the set-up, we now present the sample sizes used for each simulation study and these values have been calculated by employing the methods of Section 4.4. The trial is planned with FWER $\alpha = 0.025$ and planned power $1 - \beta = 0.9$. The number of events at the first analysis in subgroup S_1, denoted $d_1^{(1)}$, have been chosen to ensure that subgroup S_1 is selected roughly 60% of the time and the total number of events at the second analysis, $d^{(2)}$, have been chosen to attain power of 90% as described in Section 4.4. In all cases, the value of $d_1^{(1)}$ is large enough such that the survival data is mature at the interim analysis and decisions can be made with confidence. These number of events are displayed in Table 1 for a range of values of γ, σ^2 and ϕ_2. As γ increases, we see that required $d_1^{(1)}$ and $d^{(2)}$ increase. Similarly, the required number of events increase with σ^2. That is, more events and hence more information is needed to achieve power and selection probabilities when the longitudinal data is noisy. When $\sigma^2 = 2.25$ and with a small number of events at the first interim analysis, it is not always possible to find a root to equation (4). Consequently, the required $d_1^{(1)}$ and $d^{(2)}$ are high to ensure that large sample properties of the estimator hold. We have not seen this problem occur for $\sigma^2 \leq 2.25$. The values of $d_1^{(1)}$ and $d^{(2)}$ appear to be immune to changes in ϕ_2.

Table 1. Sample size calculations for the adaptive enrichment trial. $d_1^{(1)}$ is the required number of events in subgroup S_1 at the interim analysis and $d^{(2)}$ is the total number of events in the selected subgroup at the final analysis. Number of events calculated to satisfy family wise error rate (FWER) 0.025 and power 0.9.

γ	σ^2	ϕ_2	$d_1^{(1)}$	$d^{(2)}$
0	**0.25**	**5**	40	174
0.4	**0.25**	**5**	47	204
0.8	**0.25**	**5**	47	206
1.2	**0.25**	**5**	50	218
0.8	**0**	**5**	45	194
0.8	**0.25**	**5**	47	206
0.8	**1**	**5**	58	252
0.8	**2.25**	**5**	69	301
0.8	**0.25**	**0**	46	198
0.8	**0.25**	**2.5**	44	194
0.8	**0.25**	**5**	47	206
0.8	**0.25**	**7.5**	47	203

6.2 Type 1 error rate comparison

The first important comparison will be the type 1 error rate using each of the analysis methods conditional score, Cox and Cox with biomarker.

To represent no differences between control and treated groups under H_{0j}, let $\theta_j = 0$ for each $j = 1, 2, F$. Figure 3 shows the results of a simulation study assessing the FWER for each method and different parameter values. For each simulation, a dataset of patients is generated from the joint model, then subgroup selection and decisions about H_0 are performed after $d_1^{(1)}$ and $d^{(2)}$ events have been observed according to Table 1. All four methods are performed on the same dataset and after the same number of events so that differences can be attributed to the analysis methodology and not trial design features.

It is clear that for the majority of cases, the FWER is controlled when the conditional score method is used to estimate the treatment effect in the joint model. For a study with $N = 10^4$ simulations and planned significance value $\alpha = 0.025$, the simulation error bounds is $(0.0219, 0.0281)$. Hence, the observed FWER is within reasonable distance from $\alpha = 0.025$ in accordance with the number of simulations. The result of Theorem 1 together with the simulation result in Figure 3 give us confidence that FWER is controlled at the desired significance level using the joint modelling approach. The Cox model also appears to control the FWER but may be seen to be conservative for large values of γ. However, we see that the Cox with biomarker method has FWER considerably smaller than 0.025. This is particularly apparent for $\sigma^2 \geq 1$ and all values of ϕ_2.

6.3 Efficiency comparison

We shall focus on power as a measure of efficiency between the different methods and we compare some other outcome measures, such as number of hospital visits and expected stopping time, in Section C of the Supplemental Material. Under the alternative, only patients in subgroup S_1 will respond to treatment, represented by $H_{A1} : \theta_1 = -0.5$ and $H_{A2} : \theta_2 = 0$. Figure 4 shows the power comparison between the different methods. Power is calculated as the proportion of simulations which reject H_{01} out of those where subgroup S_1 is selected, as described in Section 4.2.

It is clear that the conditional score method is most efficient since power is highest across nearly all parameter combinations. When $\gamma = 0$, the conditional score method may suffer from a small loss in power in comparison to other methods. This is the case where longitudinal data has no impact on the survival outcome so including it in the analysis is futile. For $\gamma \neq 0$, however, a gain in power up to 0.46 is seen.

Fitting the data to the simple Cox model is very inefficient and in the extreme cases, power is below 0.5. The sample size that would be needed to increase power to 0.9 in such a scenario is excessive. This simple method has power lower than the conditional score method whenever $\gamma \neq 0$ and becomes increasingly inefficient as γ increases and as ϕ_2 increases. The efficiency of this method appears to increase slightly with σ. Hence, it is important to include the longitudinal data in the analysis when there is a suspected correlation between the longitudinal data and the survival endpoint.

The final method, where TTE outcomes are fit to a Cox proportional hazards model with the longitudinal data as a time-varying covariate, appears to be a simple yet effective way of including longitudinal data in the analysis. The achieved power

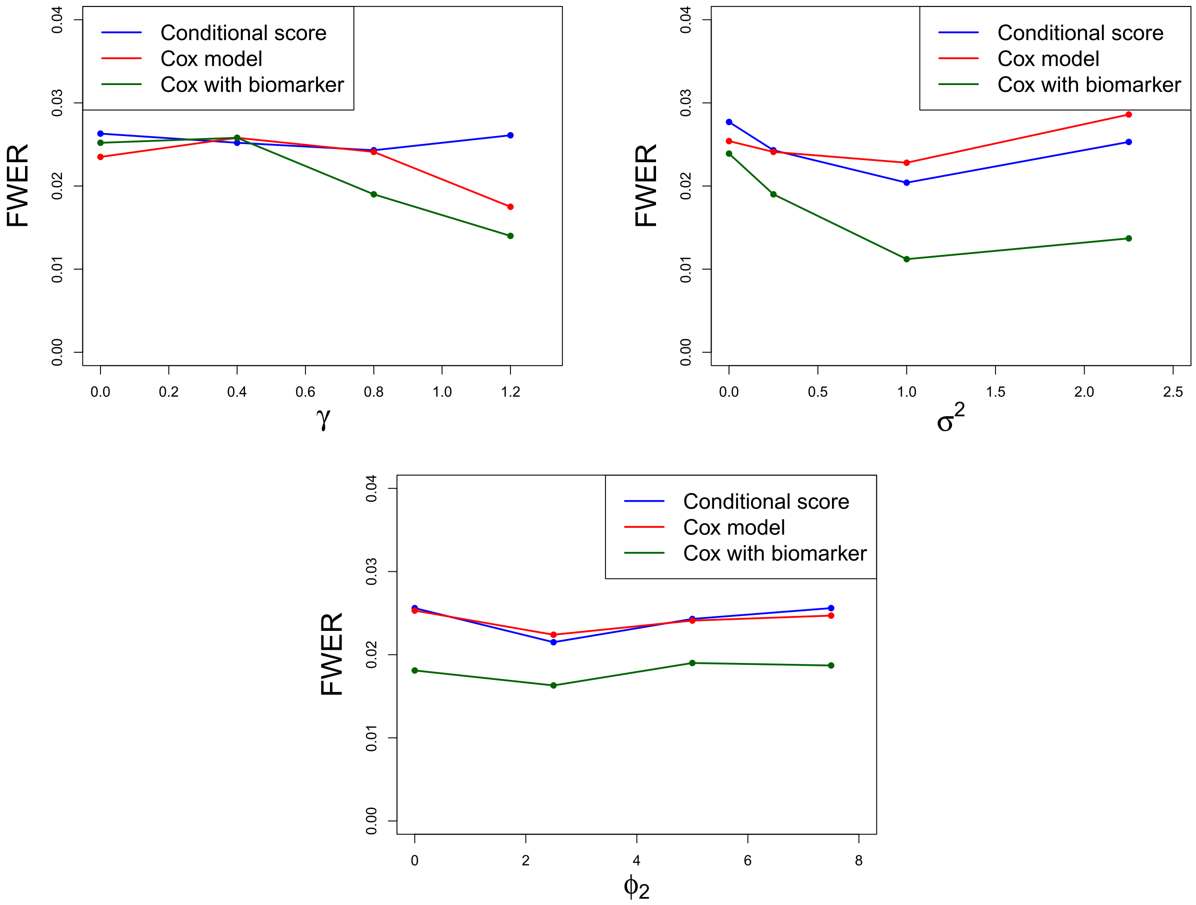

Figure 3. Type I error rates displaying changes in parameters γ, σ and ϕ_2. All other parameters are as in (13). Numeric values of the points are presented in Section C of the Supplemental Material. For a study with $N = 10^4$ simulations and family wise error rate (FWER) 0.025, simulation standard error is 0.00156.

is at least 0.78 but is usually lower than the conditional score method. However, scenarios where this method outperformes the conditional score are when $\sigma = 0$ or $\phi_2 = 0$ indicating that the longitudinal data is free of measurement error or there are no between-patient differences in the slopes of the longitudinal trajectories. The efficiency decreases as longitudinal data increase in noise or as patient differences become larger, that is as σ and ϕ_2 increase.

An advantage of the two alternative Cox models is that there is no criteria to have a minimum of two longitudinal observations to be included in the at-risk process. In fact, for these alternative models, we need not specify the functional form of the trajectory of the longitudinal data, for example that it is linear in time. Taking these considerations into account, we believe that the most efficient and practical method is the conditional score, which includes the longitudinal data and takes into account the measurement error.

7 Discussion

We have shown that the threshold selection rule can be combined with an error spending boundary to create an efficient enrichment trial. This is potentially suitable for any trial where the primary outcome is a TTE variable and we present a method to establish the required number of events at the design stage of the trial. A novel aspect of this work is that these methods can be applied to an endpoint which is the treatment effect in a joint model for longitudinal and TTE data. We have implemented the conditional score methodology to estimate the treatment effect and show that the estimator is robust to model assumptions provided that 20 events per treatment arm are observed at the interim analysis.

By including these routinely collected biomarker outcomes in the analysis to leverage this additional information, the enrichment trial has higher power compared to the enrichment trial where the longitudinal data is left out of the analysis. Bauer et al.[36] showed that bias is prevalent in designs with selection. In our case, selection bias occurs as the treatment effect estimate in the selected subgroup is inflated in later analyses which could affect the trial results. However, unlike

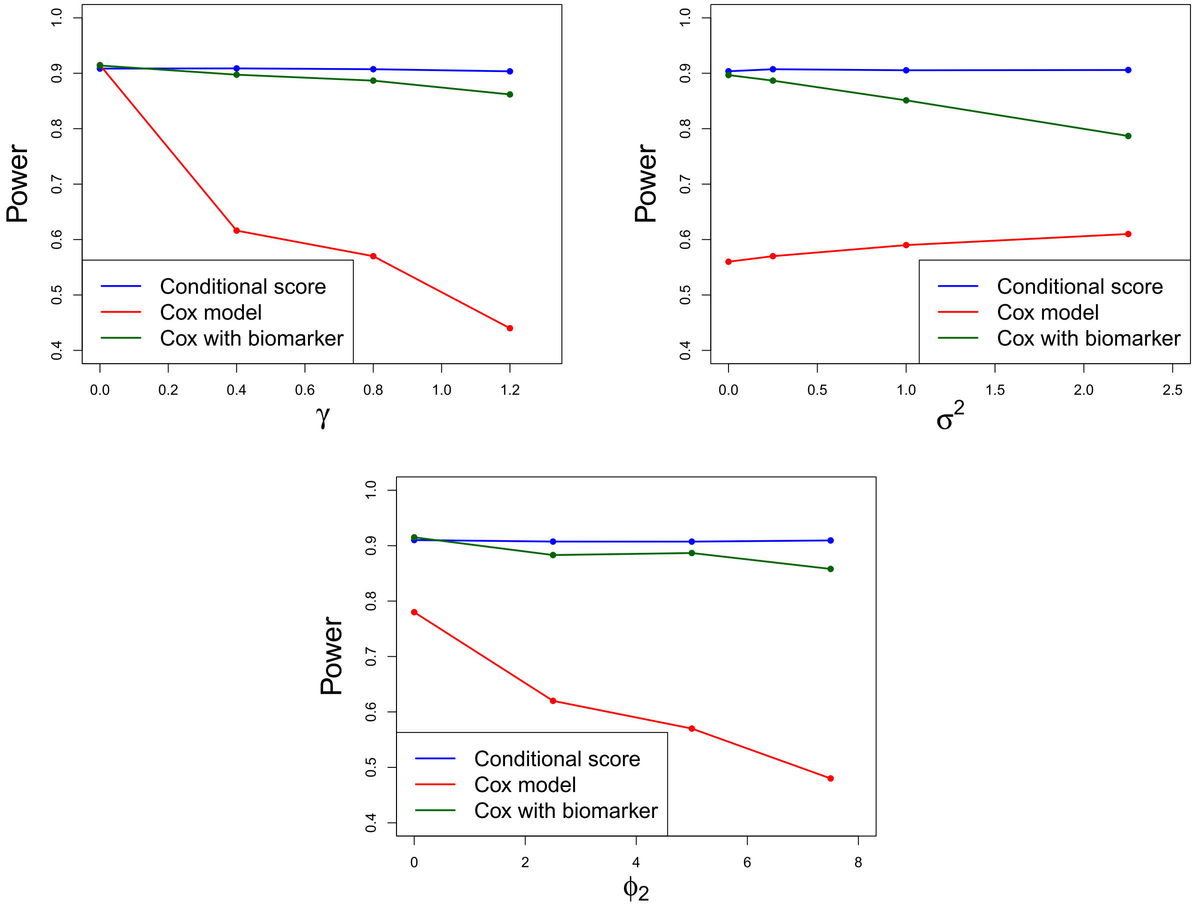

Figure 4. Obsered power displaying changes in parameters γ, σ and ϕ_2. All other parameters are as in (13). Numeric values of the points are presented in Section C of the Supplemental Material. For a study with $N = 10^4$ simulations and power 0.9, simulation standard error is 0.003.

most other selection schemes, the threshold selection rule adjusts for the magnitude of the treatment effect at the design stage so another advantage is that selection bias is incorporated into the decision making process.

We assessed the *p*-value combination approach as an alternative option for implementing enrichment designs using biomarker data for subgroup selection and survival outcomes alone for hypothesis testing, but we found the joint modeling approach to perform best due to more efficient use of available data. Further, we compared the joint modelling approach with a model which used the longitudinal data but naively assumed this was free of measurement error. Again, the joint model performed more effectively in most cases. This naive approach was more efficient when the longitudinal data was truly free from measurement error, there was no correlation between the two endpoints or there was no heterogeneity between patients' biomarker trajectories. However, we believe that these situations are rare in practice and the gain in power from joint modelling outweighs this downside.

Data availability statement

All data are simulated according to the specifications described.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 965397. TJ also received funding from the UK Medical Research Council (MC_UU_00002/14, MC_UU_00040/03). RB also acknowledges funding from Cancer Research UK and support for his early phase clinical trial work from the Cambridge NIHR Biomedical

Research Centre (BRC-1215-20014) and Experimental Cancer Medicine Centre. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any author accepted manuscript version arising.

ORCID iDs

Abigail J Burdon https://orcid.org/0000-0002-0883-4160
Thomas Jaki https://orcid.org/0000-0002-1096-188X

Supplemental material

Supplemental materials for this article are available online.
Software in the form of R code, is available at https://github.com/abigailburdon/Adaptiveenrichment-with-joint-models.

References

1. Burnett T, Mozgunov P, Pallmann P, et al. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC Med* 2020; **18**: 1–21.
2. Pallmann P, Bedding AW, Choodari-Oskooei B et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med* 2018; **16**: 1–15.
3. Simon N and Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics* 2013; **14**: 613–625.
4. Burnett T and Jennison C. Adaptive enrichment trials: what are the benefits? *Stat Med* 2021; **40**: 690–711.
5. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. London: Chapman and Hall/CRC, 2000.
6. Wang S-J, Hung HMJ and O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biom J: J Math Methods Biosci* 2009; **51**: 358–374.
7. Friede T, Parsons N and Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat Med* 2012; **31**: 4309–4320.
8. Mehta C, Schäfer H, Daniel H, et al. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Stat Med* 2014; **33**: 4515–4531.
9. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Stat Methods Med Res* 2019; **28**: 2096–2111.
10. Rosenblum M, Fang EX and Liu H. Optimal, two-stage, adaptive enrichment designs for randomized trials, using sparse linear programming. *J R Stat Soc Ser B: Stat Methodol* 2020; **82**: 749–772.
11. Lin Z, Flournoy N and Rosenberger WF. Inference for a two-stage enrichment design. *Ann Stat* 2021; **49**: 2697–2720.
12. Chiu YD, Koenig F, Posch M, et al. Design and estimation in clinical trials with subpopulation selection. *Stat Med* 2018; **37**: 4335–4352.
13. Lai TL, Lavori PW and Tsang KW. Adaptive enrichment designs for confirmatory trials. *Stat Med* 2019; **38**: 613–624.
14. Thall P. F.: Bayesian cancer clinical trial designs with subgroup-specific decisions. *Contemp Clin Trials* 2020; **90**: 105860.
15. Zhang Z, Li M, Lin M, et al. Subgroup selection in adaptive signature designs of confirmatory clinical trials. *J R Stat Soc Ser C: Appl Stat* 2017; **66**: 345–361.
16. Stallard N. Adaptive enrichment designs with a continuous biomarker. *Biometrics* 2023; **79**: 9–19.
17. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat* 2016; **26**: 99–119.
18. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Stat Med* 2010; **29**: 959–971.
19. Friede T, Parsons N, Stallard N, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Stat Med* 2011; **30**: 1528–1540.
20. Henderson R, Diggle P and Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**: 465–480.
21. Rizopoulos D. *Joint models for longitudinal and time-to-event data: with applications in R*. London: Chapman and Hall/CRC, 2012.
22. Dawson SJ, Tsui DWY, Murtaza M, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 2013; **368**: 1199–1209.
23. Magnusson BP and Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Stat Med* 2013; **32**: 2695–2714.
24. Slamon DJ, Clark GM, Wong SG, et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987; **235**: 177–182.
25. Wan F, Titman AC and Jaki TF. Subgroup analysis of treatment effects for misclassified biomarkers with time-to-event data. *J R Stat Soc Ser C: Appl Stat* 2019; **68**: 1447–1463.
26. Tsiatis AA and Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* 2001; **88**: 447–458.
27. Stefanski LA and Carroll RJ. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* 1987; **74**: 703–716.

28. Burdon AJ, Hampson LV and Jennison C. Joint modelling of longitudinal and time-to-event data applied to group sequential clinical trials. https://doi.org/10.48550/arxiv.2211.16138. arXiv. 2022. Creative Commons Attribution 4.0 International.
29. Wakefield J. *Bayesian and frequentist regression methods*. Berlin: Springer Science & Business Media, 2013.
30. Westfall PH and Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J Stat Plan Inference* 2001; **99**: 25–40.
31. Tamhane AC, Gou J, Jennison C, et al. A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics* 2018; **74**: 40–48.
32. Gordon Lan KK and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–663.
33. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982; **1**: 121–129.
34. Jennison C and Turnbull BW. Group-sequential analysis incorporating covariate information. *J Am Stat Assoc* 1997; **92**: 1330–1341.
35. Barnett HY, Geys H, Jacobs T, et al. Methods for non-compartmental pharmacokinetic analysis with observations below the limit of quantification. *Stat Biopharm Res* 2021; **13**: 59–70.
36. Bauer P, Koenig F, Brannath W, et al. Selection and bias – two hostile brothers. *Stat Med* 2010; **29**: 1–13.