Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Confirmatory Adaptive Designs for Clinical Trials With Multiple Time-to-Event Outcomes in Multi-state Markov Models

Moritz Fabian Danzer[1] | Andreas Faldum[1] | Thorsten Simon[2] | Barbara Hero[2] | Rene Schmidt[1]

[1]Institute of Biostatistics and Clinical Research, University of Münster, Münster, Germany | [2]Department of Pediatric Oncology and Hematology, University Hospital Cologne, Cologne, Germany

Correspondence: Moritz Fabian Danzer (moritzfabian.danzer@ukmuenster.de)

Received: 30 June 2023 | **Revised:** 23 July 2024 | **Accepted:** 27 August 2024

Keywords: clinical trial | log-rank test | sample size recalculation | survival analysis

ABSTRACT

The analysis of multiple time-to-event outcomes in a randomized controlled clinical trial can be accomplished with existing methods. However, depending on the characteristics of the disease under investigation and the circumstances in which the study is planned, it may be of interest to conduct interim analyses and adapt the study design if necessary. Due to the expected dependency of the endpoints, the full available information on the involved endpoints may not be used for this purpose. We suggest a solution to this problem by embedding the endpoints in a multistate model. If this model is Markovian, it is possible to take the disease history of the patients into account and allow for data-dependent design adaptations. To this end, we introduce a flexible test procedure for a variety of applications, but are particularly concerned with the simultaneous consideration of progression-free survival (PFS) and overall survival (OS). This setting is of key interest in oncological trials. We conduct simulation studies to determine the properties for small sample sizes and demonstrate an application based on data from the NB2004-HR study.

1 | Introduction

Adaptive clinical trial designs for a single primary time-to-event endpoint are well established (see, e.g., Schäfer and Müller 2001; Wassmer 2006). These are based on the log-rank test by exploiting its independent increments structure as exhibited in Tsiatis (1981) and Sellke and Siegmund (1983) or in even broader generality by Scharfstein, Tsiatis, and Robins (1997). As long as only information on this single endpoint is used to inform an adaptation of the design in an interim analysis, the nominal type I error rate will be maintained. However, this no longer applies if information of further endpoints is used from patients, who have been recruited before this interim analysis and remain event-free beyond it (Bauer and Posch 2004). This is because these additional data can be used to predict the course of the disease in those same patients. For example, information on progression status can be used to predict individual mortality risk in a trial with primary endpoint overall survival (OS). Such misuse of surrogate interim data leads to inflation of the actual type I error level. Approaches to solving this problem make use of the strategy of patient-wise separation (Jenkins et al. 2011; Irle and Schäfer 2012; Jörgens et al. 2019). Although the initial approaches in Jenkins et al. (2011) and Irle and Schäfer (2012) have already been improved by Jörgens et al. (2019), some disadvantages cannot be resolved, such as partial discarding of primary endpoint data in the final analysis. Alternatively, worst-case adjustments can be made to avoid a type I error inflation (Magirr et al. 2016) that often result in a conservative procedure.

Similar issues arise as well for trials with multiple primary time-to-event endpoints. For one-sample studies, this situation has already been addressed by Danzer et al. (2022). Single-stage

procedures for the simultaneous assessment of multiple time-to-event endpoints in randomized trials have already been proposed in Wei and Lachin (1984). Roughly speaking, this method can be described as performing separate log-rank tests simultaneously for all endpoints involved. A final test decision is made based on an examination of the joint distribution of the individual test statistics. Corresponding group sequential procedures were introduced in Lin (1991). At first glance, an extension of Lin (1991) to adaptive designs seems obvious by following the strategy of Wassmer (2006), since the multivariate test statistic also has a property of independent increments. However, this property only applies to each component of the multivariate test statistic separately and not for the multivariate process as a whole. The reason for this is closely linked to the problem mentioned in Bauer and Posch (2004) because, again, information about some endpoints might be used to predict future outcome of other endpoints. At the same time, patients who are known to be in different disease states are compared to each other. We will solve this problem by taking into account the available information on all endpoints when calculating the test statistics, thus only comparing patients who have the same prognosis of disease course given the available information.

To this end, it is central to our approach that we can easily embed different time-to-event endpoints into a time continuous multistate model. Especially in oncology, which is of central importance to us as an area of application for our methods, such models can be very helpful in being able to depict different courses of disease (Le-Rademacher et al. 2018). Two of the most important endpoints in this field of clinical research are given by progression-free survival (PFS)/event-free survival (EFS) and OS. While the latter one is the most objectively defined endpoint, the former can often be regarded as its surrogate and has certain advantages in terms of time- and cost-effectiveness. The exact definition of the endpoint and its suitability as a primary endpoint strongly depends on the tumor entity and the patient collective to be considered (Bellera et al. 2013). Those two endpoints can be embedded in a simple illness-death model, which has been discussed extensively in Meller, Beyersmann, and Rufibach (2019). Provided that this model is a time continuous Markov chain, we can perform a two-group comparison that addresses the aforementioned issues. As in Lin (1991), this results in a consideration that refers to the clinical endpoints with the aid of a transition-wise consideration as in Tattar and Vaman (2014).

The paper is organized as follows. It starts with a presentation of the procedure for the prominent example of PFS and OS. Sections 3 and 4 introduce the general notation and generalize the procedure for broader applications. Building on that, planning and execution of a clinical trial are briefly sketched in Section 5. An application of the proposed method is demonstrated in Section 6 using the data from the NB2004-HR trial (NCT number NCT03042429). Properties of the method in practically relevant scenarios are studied by simulation in Section 7. We conclude with a discussion of our findings and prospects for future research.

Proofs of mathematical statements, further simulation results, and a further case study can be found in the Supporting Information.

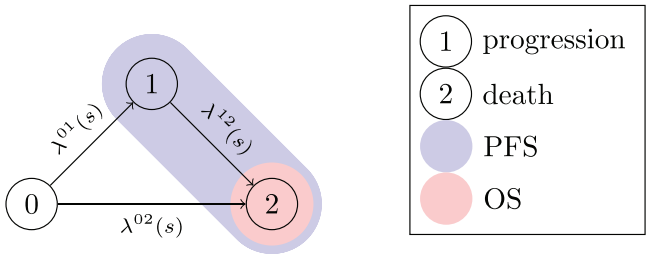

FIGURE 1 | Representation of the PFS/OS scenario as a multistate model.

2 | Main Application Example: PFS and OS

We illustrate our procedure using the example of a trial with the primary time-to-event endpoints PFS and OS. In a randomized clinical trial, PFS is defined as time from randomization to progression of the disease or death, whatever occurs first. OS denotes the time from randomization to death. While OS is obviously the most objectively defined time-to-event endpoint, the use of other endpoints such as PFS may also be justified in oncological Phase III clinical trials, depending on the nature of the disease and the mechanism of action of the experimental treatment. Outcome improvement can first be associated with longer PFS time or an increase of the rate of patients without tumor progression. In addition, there may be other advantages concerning death without prior progression or postprogression survival, which then additionally affect OS. The methods proposed here cover all of these aspects by allowing to use both of these endpoints as primary endpoints under exploitation of their dependence structure in a multistate model that is a time continuous Markov chain.

Such a model as presented in Meller, Beyersmann, and Rufibach (2019) establishes the corresponding probabilistic structure. The multistate model is visualized in Figure 1. A patient's history of disease from start of the therapy corresponds to a path along the arrows in this figure. At the beginning of the treatment, a patient starts in state 0. He may die directly without progression. This is represented by a jump to state 2. Otherwise, he may experience a progression of the disease, which is represented by a jump to state 1 and die afterwards which is represented to a subsequent jump to state 2.

In accordance with our general framework, we denote the random time of transition to node 1 of some patient i by $T_i^{\{1\}}$ and the time of transition to node 2 by $T_i^{\{2\}}$. Accordingly, the random time of PFS, which is the first hitting time of the set of nodes $\{1,2\}$, can be defined as $T_i^{\text{PFS}} := T_i^{\{1\}} \wedge T_i^{\{2\}}$, where $a \wedge b$ denotes the minimum of two real numbers a, b. The random time of OS, which is the first hitting time of node 2, is given by $T_i^{\text{OS}} := T_i^{\{2\}}$.

Such a model fulfills the Markov assumption if the conditional probability of future transitions does only depend on the present state. To introduce this more formally, let $X_i : \mathbb{R}_+ \to \{0,1,2\}$ denote the state occupation function for some patient i, that is, $X_i(s)$ yields the state of patient i at time s since randomization of that patient. The sample paths are assumed to be right continuous with left limits. These left limits $X_i(s-) := \lim_{h \searrow 0} X_i(s-h)$

denote the state of the patient just before s. Now, if

$$\mathbb{P}\left[X_i(s_2) \in S | (X_i(u))_{u\in[0,s_1]}\right] = \mathbb{P}[X_i(s_2) \in S | X(s_1)] \tag{1}$$

for any subspace of the state space $S \subset \{0,1,2\}$ and any $0 \le s_1 < s_2$, the stochastic process $(X_i(s))_{s\ge 0}$ is said to be a time continuous Markov chain. If this term is a function of $s_2 - s_1$, the process is called time-homogeneous. However, we do not require this as we also deal with the time-inhomogeneous case.

Given the current state of a patient, its instantaneous rate of transition to another state does only depend on the time elapsed since randomization. Hence, each of the transitions represented by the arrows is equipped with a univariate transition hazard or intensity function $\lambda^{jk} : \mathbb{R}_+ \to \mathbb{R}_+$ with $0 \le j < k \le 2$. These are defined by

$$\lambda^{jk}(s) := \lim_{h\searrow 0} \frac{\mathbb{P}[X_i(s+h) = k | X_i(s-) = j]}{h}. \tag{2}$$

Given those functions, the joint distribution of PFS and OS is fully specified. As we want to explore any kind of possible differences in the joint distribution of the two endpoints between the two groups, our testing procedure investigates the null hypothesis

$$H_0 : F^0_{\text{PFS,OS}} = F^1_{\text{PFS,OS}}, \tag{3}$$

where $F^g_{\text{PFS,OS}}$ denotes the joint distribution function of the time-to-event endpoints PFS and OS in group $g \in \{0,1\}$.

Following Tattar and Vaman (2014), this could also be reformulated in terms of the cumulative intensity matrices $(\mathbf{\Lambda}_g(s))_{s\ge 0} := (\Lambda^{jk}_g(s))_{j,k\in\{0,1,2\},s\ge 0}$, which is a 2×2 matrix of the transition-specific cumulative intensity functions. This means that at time s the (j,k)-th entry for $j \ne k$ of this matrix is the cumulative intensity for the transition from state j to k for the respective group g. This quantity is given by the integral

$$\Lambda^{jk}_g(s) := \int_0^s \lambda^{jk}_g(u)\, du, \tag{4}$$

where the additional index $g \in \{0,1\}$ describes a possible dependence from the treatment group. The diagonal entries are given so that each row and column sums to zero at any time $s \ge 0$. The corresponding hypothesis is then given by

$$H_0 : \mathbf{\Lambda}_0(s) = \mathbf{\Lambda}_1(s) \quad \forall s \ge 0. \tag{5}$$

Differing from Tattar and Vaman (2014), we do not compare the estimated transition intensity matrices, but pursue an approach that is motivated by the clinically relevant endpoints.

In univariate survival analysis, one-dimensional compensated counting processes form the basis for constructing adaptive designs. For the two endpoints considered here, these are given by $(\tilde{M}^{\text{PFS}}_i(s))_{s\ge 0}$ respectively (resp.) $(\tilde{M}^{\text{OS}}_i(s))_{s\ge 0}$ with

$$\tilde{M}^E_i(s) := \mathbb{1}_{\{T^E_i \le s\wedge C_i\}} - \int_0^{s\wedge T^E_i \wedge C_i} \lambda^E(u)\, du \tag{6}$$

for $E \in \{\text{PFS}, \text{OS}\}$ and any $s \ge 0$. The positive real-valued random variable C_i denotes the random censoring time, which is assumed to be independent from the process X_i. The endpoint-specific hazards λ^{PFS} and λ^{OS} do not take into account the current state of the patient. Since we have to do exactly this when constructing adaptive designs where all information on PFS and OS is allowed to be used at an interim analysis, we will instead consider the multivariate compensated counting processes $(\mathbf{M}_i(s))_{s\ge 0}$ with

$$\mathbf{M}_i(s) := \begin{pmatrix} M^{\text{PFS}}_i(s) \\ M^{\text{OS}}_i(s) \end{pmatrix} = \begin{pmatrix} \tilde{M}^{\text{PFS}}_i(s) \\ M^{\text{OS}}_i(s) \end{pmatrix}$$

and

$$\begin{aligned} M^{\text{OS}}_i(s) := \mathbb{1}_{\{T^{\text{OS}}_i \le s\wedge C_i\}} &- \int_0^{s\wedge C_i \wedge T^{\text{PFS}}_i} \lambda^{02}(u)\, du \\ &- \int_{s\wedge C_i\wedge T^{\text{PFS}}_i}^{s\wedge C_i \wedge T^{\text{OS}}_i} \lambda^{12}(u)\, du \end{aligned} \tag{7}$$

for any $s \ge 0$. The component for PFS can be adopted from the univariate setting (according to (6)) as there is no additional information to be taken into account for this endpoint. As soon as any transition occurs in our simple model, the process automatically stops.

In order to state the test statistics that arise in this way, we need to introduce some more notation. First, let $Z_i \in \{0,1\}$ denote the treatment indicator variable and $R_i \in \mathbb{R}_+$ the random time of trial entry of patient i. As we aim for adaptive sequential designs, we need to deal with two different time scales: We will always denote the calendar time by t and the individual time in trial by s. In this way, we can define the event counting processes

$$N^{\text{PFS}}_i(t,s) := \mathbb{1}_{\{T^{\text{PFS}}_i \le s\wedge C_i \wedge (t-R_i)_+\}} \quad \text{and}$$
$$N^{\text{OS}}_i(t,s) := \mathbb{1}_{\{T^{\text{OS}}_i \le s\wedge C_i \wedge (t-R_i)_+\}}$$

counting events that happen before calendar time t and trial time s. For any state $j \in \{0,1,2\}$ of our model from Figure 1, we can also define the corresponding at-risk processes

$$Y^j_i(t,s) := \mathbb{1}_{\{X_i(s-)=j\}} \cdot \mathbb{1}_{\{s\le C_i\wedge(t-R_i)_+\}} \quad \text{and}$$
$$Y^{j,Z=1}_i(t,s) := Z_i \cdot Y^j_i(t,s),$$

which indicate at some calendar time t whether patient i is known to be in state j just before trial time s and (for the latter one) whether the patient also is in treatment group 1. While these quantities are defined for each patient, the aggregates N^{PFS}, N^{OS}, Y^j, and $Y^{j,Z=1}$ over the whole study sample are given by summing the corresponding quantities over all patients i from 1 to n. In what follows, we will regularly obtain stochastic integrals of the form $\int_0^t H_i(t,s)\, N_i(t,ds)$. In the present cases, these equal $H_i(t,T_i)\cdot N_i(t,T_i)$ where T_i is the time at which N_i makes a jump.

At calendar time t, the component of our unstandardized multivariate test statistic concerning PFS is then given by

$$U^{\mathrm{PFS}}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \left(Z_i - \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \right) N_i^{\mathrm{PFS}}(t, ds),$$

which is just the common unstandardized log-rank statistic for PFS. For the second component, concerning the endpoint OS, we need to take the additional information of prior progressions into account. It is defined by

$$U^{\mathrm{OS}}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \left(Z_i - Y_i^0(t,s) \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} - Y_i^1(t,s) \frac{Y^{1,Z=1}(t,s)}{Y^1(t,s)} \right) N_i^{\mathrm{OS}}(t, ds).$$

These can also be expressed as sums. Let $K^{jk}(t)$ denote the number of transitions from state j to state k that could be observed until calendar time t. For each of the three transitions in this model, let $s^{jk}_{(1)} < \cdots < s^{jk}_{(K^{jk}(t))}$ denote the ordered event times of these transitions and let $Z^{jk}_{(1)}, \ldots, Z^{jk}_{(K^{jk}(t))}$ denote the treatment groups of the corresponding individuals. Then, the test statistics at calendar time t amount to

$$U^{\mathrm{PFS}}(t) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{K^{01}(t)} \left(Z^{01}_{(j)} - \frac{Y^{0,Z=1}(t, s^{01}_{(j)})}{Y^0(t, s^{01}_{(j)})} \right) + \sum_{j=1}^{K^{02}(t)} \left(Z^{02}_{(j)} - \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \right) \right)$$

resp.

$$U^{\mathrm{OS}}(t) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{K^{02}(t)} \left(Z^{02}_{(j)} - \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \right) + \sum_{j=1}^{K^{12}(t)} \left(Z^{12}_{(j)} - \frac{Y^{1,Z=1}(t, s^{12}_{(j)})}{Y^1(t, s^{12}_{(j)})} \right) \right).$$

Analogously to the adopted compensated counting process in (7), we need to distinguish between patients who did not experience a progression of the disease yet ($Y_i^0(t,s) = 1$) and those who did ($Y_i^1(t,s) = 1$). In contrast to Lin (1991), this distinction enables adaptive design changes based on all information from the illness-death model from Figure 1.

The variance of $U^{\mathrm{PFS}}(t)$ and $U^{\mathrm{OS}}(t)$ can be estimated by

$$\begin{aligned}
\hat{V}^{\mathrm{PFS}}(t) &= \frac{1}{n} \sum_{i=1}^{n} \int_{[0,t]} \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \left(1 - \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \right) N_i^{\mathrm{PFS}}(t, ds) \\
&= \frac{1}{n} \sum_{j=1}^{K^{01}(t)} \frac{Y^{0,Z=1}(t, s^{01}_{(j)})}{Y^0(t, s^{01}_{(j)})} \left(1 - \frac{Y^{0,Z=1}(t, s^{01}_{(j)})}{Y^0(t, s^{01}_{(j)})} \right) \\
&\quad + \frac{1}{n} \sum_{j=1}^{K^{02}(t)} \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \left(1 - \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \right)
\end{aligned}$$

resp.

$$\begin{aligned}
\hat{V}^{\mathrm{OS}}(t) &= \frac{1}{n} \sum_{i=1}^{n} \int_{[0,t]} Y_i^0(t,s) \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \left(1 - \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \right) \\
&\quad + Y_i^1(t,s) \frac{Y^{1,Z=1}(t,s)}{Y^1(t,s)} \left(1 - \frac{Y^{1,Z=1}(t,s)}{Y^1(t,s)} \right) N_i^{\mathrm{OS}}(t, ds) \\
&= \frac{1}{n} \sum_{j=1}^{K^{02}(t)} \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \left(1 - \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \right) \\
&\quad + \frac{1}{n} \sum_{j=1}^{K^{12}(t)} \frac{Y^{1,Z=1}(t, s^{12}_{(j)})}{Y^1(t, s^{12}_{(j)})} \left(1 - \frac{Y^{1,Z=1}(t, s^{12}_{(j)})}{Y^1(t, s^{12}_{(j)})} \right).
\end{aligned}$$

The covariance between the two random variables can be estimated by

$$\begin{aligned}
\hat{V}^{\mathrm{PFS,OS}}(t) &= \frac{1}{n} \sum_{i=1}^{n} \int_{[0,t]} Y_i^0(t,s) \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \\
&\quad \times \left(1 - \frac{Y^{0,Z=1}(t,s)}{Y^0(t,s)} \right) N_i^{\mathrm{OS}}(t, ds) \\
&= \frac{1}{n} \sum_{j=1}^{K^{02}(t)} \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \left(1 - \frac{Y^{0,Z=1}(t, s^{02}_{(j)})}{Y^0(t, s^{02}_{(j)})} \right).
\end{aligned}$$

As we integrate over the process $N_i^{\mathrm{OS}}(t, ds)$ and multiply each summand by $Y_i^0(t,s)$, we can see that the covariance is only driven by transitions from state 0 to node 2, that is, by patients for which PFS and OS happen simultaneously.

We will now consider the bivariate process $(\mathbf{U}(t))_{t \geq 0}$ with $\mathbf{U}(t) := (U^{\mathrm{PFS}}(t), U^{\mathrm{OS}}(t))$ for all $t \geq 0$ and the 2×2-matrix–valued process $(\hat{\mathbf{V}}(t))_{t \geq 0}$ with

$$\hat{\mathbf{V}}(t) = \begin{pmatrix} \hat{V}^{\mathrm{PFS}}(t) & \hat{V}^{\mathrm{PFS,OS}}(t) \\ \hat{V}^{\mathrm{PFS,OS}}(t) & \hat{V}^{\mathrm{OS}}(t) \end{pmatrix}$$

for all $t \geq 0$.

For the sake of simplicity, we only consider a design with one interim analysis at calendar time $t_1 > 0$ and a final analysis at calendar time $t_2 > t_1$ here. First stage test statistics will be based on $\mathbf{U}(t_1)$ and $\hat{\mathbf{V}}(t_1)$. Test statistics for the data from the second stage will be based on the increments since calendar time t_1, that is, $\mathbf{U}(t_2) - \mathbf{U}(t_1)$ and $\hat{\mathbf{V}}(t_2) - \hat{\mathbf{V}}(t_1)$. If the increments of the asymptotic covariance matrix $\mathbf{V}$, which is consistently estimated by $\hat{\mathbf{V}}$ has full rank, the quadratic form of the increments of $\mathbf{U}$ with the corresponding increments of $\hat{\mathbf{V}}$ is asymptotically χ^2 distributed with 2 degrees of freedom. The stagewise test statistics are thus given by

$$S_1 = \mathbf{U}(t_1)^T \hat{\mathbf{V}}(t_1)^{-1} \mathbf{U}(t_1) \quad \text{resp.}$$

$$S_2 = (\mathbf{U}(t_2) - \mathbf{U}(t_1))^T (\hat{\mathbf{V}}(t_2) - \hat{\mathbf{V}}(t_1))^{-1} (\mathbf{U}(t_2) - \mathbf{U}(t_1)).$$

In analogy to Wei and Lachin (1984), we obtain stagewise p-values by

$$p_r = 1 - F_{\chi^2_2}(S_r). \tag{8}$$

The stagewise p-values can then be further processed using the standard methods for adaptive designs of clinical trials.

3 | General Framework

In this section, we will introduce the framework and all its components we need to construct the multivariate process and resulting test statistics. This will allow us to expand upon the example from the previous section by considering an arbitrary number of composite events.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space upon which all random variables are defined. Any patient i enters the trial at the random time $R_i \geq 0$ and is assigned to treatment group $Z_i \in \{0, 1\}$. During the stay in the trial, the patients assume one state of the state space $\{0, 1, \dots, l\}$. The assumed state may change in course of time. For each i and any $s \geq 0$, let $X_i(s) \in \{0, 1, \dots, l\}$ denote the individual state of patient i at time s after its trial entry. We assume this process to be a time continuous time-(in)homogeneous Markov chain and $X_i(0) = 0$ for all $i \in \{1, \dots, n\}$. In particular it fulfills the Markov property from (1) and its transition intensities are defined as in (2). In both formulas, the state space of the illness-death model from Figure 1 needs to be replaced by the state space $\{0, \dots, l\}$. For a more detailed treatment of multistage models beyond what is required here, we refer to books on this topic (e.g., Hougaard 2000 or Beyersmann, Allignol, and Schumacher 2011).

On that basis, we can define hitting or first entry times for each node $j \in \{1, \dots, l\}$ by

$$T_i^{\{j\}} := \inf\{s \geq 0 | X_i(s) = j\}.$$

If the time of a composite event is of interest, this can be depicted by the hitting time of a set of nodes $E \subset \{1, \dots, l\}$ with

$$T_i^E := \inf\{s \geq 0 | X_i(s) \in E\}.$$

However, the observations for all our patients can be censored, either by administrative censoring at the time of an interim or the final analysis or by random dropout. In the former case, an analysis at calendar time t induces an administrative censoring at $(t - R_i)_+$. The latter case is depicted by the random variable $\tilde{C}_i$. Combining this information at calendar time t yields the censoring variable $C_i(t) := \tilde{C}_i \wedge (t - R_i)_+$. Note that censoring by some terminal event as, for example, death is not included here.

We assume the tuples $(R_i, Z_i, \tilde{C}_i, (X_i(s))_{s\geq 0})$ for $i \in \{1, \dots, n\}$ to be independent replicates of some tuple $(R, Z, \tilde{C}, (X(s))_{s\geq 0})$. Additionally, we assume independent censoring and recruitment mechanisms, that is, that the variables Z, R, and $\tilde{C}$ are mutually independent.

With the quantities given above, we can now define counting processes and at-risk indicators for the occurrence of certain events. First, for any event given via a set $E \subset \{1, \dots, l\}$, the multivariable process $(N_i^E(t, s))_{t\geq 0, s\geq 0}$ defined by

$$N_i^E(t, s) := \mathbb{1}_{\{T_i^E \leq s \wedge C_i(t)\}}$$

indicates whether a visit of patient i in the subset E of the state space (resp. the event associated with this set) has been observed before calendar time t and trial time s. We can also aggregate these individual counting processes to obtain the overall number of events $N^E(t, s) := \sum_{i=1}^n N_i^E(t, s)$ observed before calendar time t and trial time s.

As indicated by the Markov property in (1), the current state of a process at some trial time s determines the probability of future transitions. Hence, it will be of utmost importance for our procedure to keep track of the current state of each individual. The multivariable process $(Y_i^j(t, s))_{t\geq 0, s\geq 0}$ indicates whether it is known at calendar time t that individual i has been in state j just before its trial time s. It is defined by

$$Y_i^j(t, s) := \mathbb{1}_{\{X_i(s-)=j\}} \cdot \mathbb{1}_{\{s\leq C_i(t)\}}.$$

We can aggregate these indicators in the complete study sample or in the subsample of treatment group 1 to obtain the processes $(Y^j(t, s))_{t\geq 0, s\geq 0}$ resp. $(Y^{j,Z=1}(t, s))_{t\geq 0, s\geq 0}$ counting the number of patients in state j with

$$Y^j(t, s) := \sum_{i=1}^n Y_i^j(t, s) \quad \text{resp.} \quad Y^{j,Z=1}(t, s) := \sum_{i=1}^n Z_i \cdot Y_i^j(t, s). \tag{9}$$

As we only consider the first hitting time of a subset E of the state space, which corresponds to the event time of the corresponding (composite) event for now, we need to restrict these at-risk numbers to those patients, which did not already experience the event E. Those quantities are given by $(Y_i^{j\to E}(t, s))_{t\geq 0, s\geq 0}$ resp. $(Y_i^{j\to E, Z=1}(t, s))_{t\geq 0, s\geq 0}$ for any patient i. Those quantities are defined by

$$Y_i^{j\to E}(t, s) := Y_i^j(t, s) \cdot \mathbb{1}_{\{T_i^E \geq s\}} \quad \text{resp.}$$

$$Y_i^{j\to E, Z=1}(t, s) := Y_{i,Z=1}^j(t, s) \cdot \mathbb{1}_{\{T_i^E \geq s\}}$$

and the aggregates $(Y^{j\to E}(t, s))_{t\geq 0, s\geq 0}$ resp. $(Y^{j\to E, Z=1}(t, s))_{t\geq 0, s\geq 0}$ over the whole study sample are defined analogously to (9).

In the construction of our testing procedure, we will regularly obtain stochastic integrals of the form $\int_0^t H_i(t, s)\, N_i(t, ds)$. In the present cases, these equal $H_i(t, T_i) \cdot N_i(t, T_i)$ where T_i is the time at which N_i makes a jump.

4 | Construction of the Multivariate Process and Its Asymptotics

First, we consider only one composite event represented by a subspace of the state space except the initial state $E \subset \{1, \dots, l\}$. For this event, we define the stochastic process

$$U^E(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_{[0, t\wedge T_i^E]} \left(Z_i - \sum_{j\notin E} Y_i^{j\to E}(t, u) \frac{Y^{j\to E, Z=1}(t, u)}{Y^{j\to E}(t, u)} \right) \times N_i^E(t, du).$$

This and further stochastic integrals below are of the form $\int_0^t H_i(t,s)N_i(t,ds)$. In the present cases, these equal $H_i(t,T_i)\cdot N_i(t,T_i)$ where T_i is the time at which N_i makes a jump.

Different from a standard log-rank test for the composite endpoint E, we need to distinguish between the states from which a transition to one of the component events belonging to E occurs.

Now, let $d \in \mathbb{N}$ composite events of interest be given via some subsets of the state space $E_1, \dots, E_d$. Similar to the formulation of the null hypothesis for the example in Section 2, we can now formulate the null hypothesis in terms of the joint distribution by

$$H_0 : F^0_{T^{E_1},\dots,T^{E_d}} = F^1_{T^{E_1},\dots,T^{E_d}} \tag{10}$$

or in terms of the cumulative transition intensity matrix by

$$H_0 : \mathbf{\Lambda}_0(s) = \mathbf{\Lambda}_1(s) \quad \forall s \geq 0. \tag{11}$$

As in the case of Section 2, these are $(l+1)\times(l+1)$ matrices containing the cumulative intensity functions $(\Lambda_g^{jk}(s))_{s\geq 0}$ as in (4). To test these hypotheses, we will consider the multivariate process $\mathbf{U} : \mathbb{R}_+ \to \mathbb{R}^d$

$$\mathbf{U}(t) = (U^{E_1}(t), \dots, U^{E_d}(t)).$$

In Corollary 3 of the Supporting Information, it is shown that this process is asymptotically equivalent to a martingale with respect to the filtration incorporating any information about events in the multistate model. Please note that it is the same as the multivariate process introduced in Lin (1991) in a competing risks setting, where each of the states $1, \dots, l$ is a terminal state. In this special case, there is no difference between the two methods as there are no intermediate events that can be used to make predictions about later events of the same patient. The variance of $U^E(t)$ can then be estimated by

$$\hat{V}^E(t) := \frac{1}{n}\sum_{i=1}^n \int_{[0,t\wedge T_i^E]} \sum_{j\notin E} \left(Y_i^{j\to E}(t,u)\cdot \frac{Y^{j\to E,Z=1}(t,u)}{Y^{j\to E}(t,u)} \times \left(1 - \frac{Y^{j\to E,Z=1}(t,u)}{Y^{j\to E}(t,u)}\right)\right) N_i^E(t,du).$$

If at least two of the sets $E_1, \dots, E_d$ have a nonempty intersection, that is, if two of the composite events may occur at the same time, there is a nonzero covariance of the corresponding entries in $\mathbf{U}(t)$. Accordingly for $b,c \in \{1,\dots,d\}$, the covariance $\mathrm{Cov}(U^{E_b}(t), U^{E_c}(t))$ can be estimated by

$$\hat{V}^{E_bE_c}(t) := \frac{1}{n}\sum_{i=1}^n \int_{[0,t\wedge T_i^{E_b\cup E_c}]} \sum_{j\notin E} \left(Y_i^{j\to E_b\cup E_c}(t,u) \times \frac{Y^{j\to E_b\cup E_c,Z=1}(t,u)}{Y^{j\to E_b\cup E_c}(t,u)} \left(1 - \frac{Y^{j\to E_b\cup E_c,Z=1}(t,u)}{Y^{j\to E_b\cup E_c}(t,u)}\right)\right) \times N_i^{E_b\cap E_c}(t,du).$$

The covariance of the process $\mathbf{U}$ is thus estimated by the $d\times d$-matrix–valued function $\hat{\mathbf{V}} : \mathbb{R}_+ \to \mathbb{R}_+^{d\times d}$ with

$$\hat{\mathbf{V}}(t) = \left(\hat{V}^{E_bE_c}(t)\right)_{1\leq b,c\leq d}.$$

Invertibility of increments of the variance matrix $\mathbf{V}$ (which are estimated by the corresponding increments of $\hat{\mathbf{V}}$) should be checked first by applying the results in section B of the Supporting Information. Invertibility is of course not given if, for example, $E_b = E_c$ for some $b \neq c$ and $b,c \in \{1,\dots,d\}$. However, in the Supporting Information, we provide criteria for invertibility of $\mathbf{V}$. In most relevant cases as, for example, in those mentioned in the main paper, this can easily be ensured.

In a group sequential design, there is a sequence of calendar dates $t_1, \dots, t_m$ at which analyses shall be conducted. The asymptotic multivariate independent increments property of the process $\mathbf{U}$ is closely related to general results as presented, for example, in Scharfstein, Tsiatis, and Robins (1997), but also requires the Markov property. The covariance matrix of these increments can consistently be estimated by increments of $\hat{\mathbf{V}}$.

As in Section 2, we can consider the quadratic forms of the increments of $\mathbf{U}$ and $\hat{\mathbf{V}}$ as stagewise test statistics. Hence, with reference to Wei and Lachin (1984), we propose

$$S_r := (\mathbf{U}(t_r) - \mathbf{U}(t_{r-1}))^T(\hat{\mathbf{V}}(t_r) - \hat{\mathbf{V}}(t_{r-1}))^+(\mathbf{U}(t_r) - \mathbf{U}(t_{r-1})) \xrightarrow{\mathcal{D}} \chi^2_d \qquad \forall r \in \{1,\dots,m\}$$

as a natural test statistic for testing H_0 in stage $r \in \{1,\dots,m\}$. Here, $\mathbf{A}^+$ denotes the Moore–Penrose inverse of a quadratic matrix $\mathbf{A}$. Following Corollary 3 in the Supporting Information, $S_1, \dots, S_m$ are asymptotically independent and asymptotically follow a χ^2-distribution with d degrees of freedom. Stagewise p-values can accordingly be computed by

$$p_r := 1 - F_{\chi^2_d}(S_r) \tag{12}$$

for any $r \in \{1,\dots,m\}$.

Going beyond the joint assessment of PFS/EFS and OS, which has been explained in Section 2, the general framework can be used beyond this example. For example, it might also be of interest to assess long-term efficacy (PFS) and long-term safety (as time to life-threatening toxicity or death). This results in a slightly more complex illness-death model as depicted by Figure 2 with $k = 3$, $d = 2$, $E_1 = \{2,3\}$ (PFS) and $E_2 = \{1,3\}$ (safety). Additionally, it is notable, that this framework contains an adaptive design for a single-primary endpoint as a special case ($l = d = 1$) and coincides with the procedure of Lin (1991) in a competing risks setting ($l = d$ and $E_c = \{c\}\,\forall c \in \{1,\dots,d\}$).

5 | Group Sequential and Adaptive Designs

Based on the results obtained in the previous sections, we outline the procedure of a two-stage adaptive design for testing the null hypothesis (3) resp. (5) in the illness-death model from Figure 1. Procedures with more than two stages and/or different endpoints

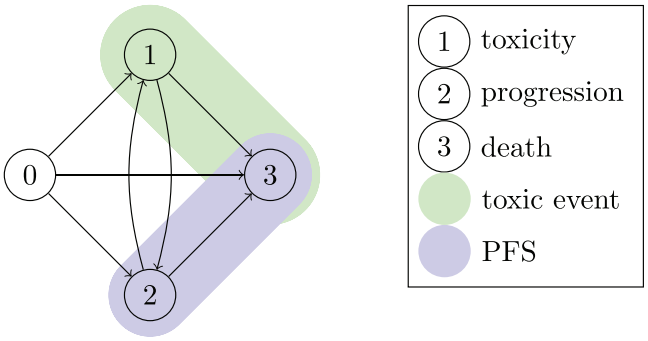

FIGURE 2 | Representation of the simultaneous assessment of efficacy and toxicity as a multistate model.

can be constructed analogously. Stagewise p-values computed as in (12) can be used with any kind of combination function and decision boundary to set up a group sequential testing procedure. An extensive overview on these topics is, for example, given in Wassmer and Brannath (2016).

We do not consider stopping for futility here. Hence, an adaptive level α test can be specified by values $0 < \alpha_1 < \alpha < 1$ where α is the overall significance level and α_1 the rejection level of the first stage, together with a conditional error function $\alpha_2 : (\alpha_1, 1] \to [0, 1]$, which is monotonically decreasing and fulfills the equality

$$\int_{\alpha_1}^{1} \alpha_2(x)\, dx = \alpha - \alpha_1.$$

In terms of the stagewise p-values, this leads to the rejection region $\{p_1 \leq \alpha_1\} \cup \{p_1 > \alpha_1, p_2 \leq \alpha_2(p_1)\}$. With regard to the stagewise increments of the multivariate process $\mathbf{U}$, α_1 and $\alpha_2(p_1)$ induce an ellipse as the decision boundary (see Figure 3). At the interim analysis, the design of the trial (e.g., its sample size) may be adapted based on observations concerning PFS- and OS-events observed until this analysis date.

5.1 | Theoretical Properties of the Multivariate Test Statistic

Decisive factors for the asymptotic behavior of the multivariate process $(\mathbf{U}(t))_{t\geq 0}$ introduced in Section 2 are the transition intensities in the illness-death model from Figure 1, the differences between the treatment groups concerning those intensities and the recruitment and censoring mechanism of the study to be planned. Transition-wise consideration in multistate models are, for example, presented in Le-Rademacher et al. (2018) and employed for planning purposes in Erdmann, Beyersmann, and Rufibach (2023).

In order to consider the theoretical properties of $\mathbf{U}$, we first take the transition intensity functions $\lambda^{0,01}, \lambda^{0,02}, \lambda^{0,12}$ in the control group ($Z = 0$) as given. Furthermore, we assume that the intensities in the control and treatment group ($Z = 1$) differ by fixed, time-independent factors $\delta^{01}, \delta^{02}, \delta^{12}$, that is,

$$\frac{\lambda_1^{jk}(s)}{\lambda_0^{jk}(s)} = \delta^{jk} \qquad \forall\, s \geq 0\,, (j,k) \in \{(0,1),(0,2),(1,2)\}.$$

Hence, we start from the assumption of transition-wise proportional hazards. Now, that transition intensities in both groups are determined by the above quantities, transition probabilities can be calculated. For our purposes, it is enough to calculate the probability of being in some state j at some time s when starting in state 0 at time 0, which is the case for all patients that will be recruited. In each treatment group $g \in \{0, 1\}$, these probabilities are denoted by

$$P_g^{0j}(0, s) := \mathbb{P}[X(s) = j | X(0) = 0, Z = g].$$

They can be calculated from the matrix exponential of $-\Lambda^g$. Explicit formulas for the illness-death model can also be found in the Appendix of Meller, Beyersmann, and Rufibach (2019). Concerning the recruitment and censoring mechanism, we assume that patients are recruited at a uniform rate r during the accrual period of length a and followed up for some additional time f after the end of the accrual period. They are assigned to the treatment group $Z = 1$ with probability $\pi \in (0, 1)$ in a randomized study. This information can be combined to compute the proportion of all patients that are randomized to group g and for which it is known at calendar time t that they are in state j at time s since their recruitment. This is given by

$$y^{j,Z=g}(t, u) := \mathbb{P}[Z = g, R \leq t - u, C \geq u, X(u) = j].$$

By omitting the index $Z = g$, we denote the sum over the two expressions for the treatment groups. Given that, we can now state the process $(\theta(t))_{t\geq 0}$ having two components, which describe the asymptotic mean of the process $\mathbf{U}$. These two components are

$$\theta^{\mathrm{PFS}}(t) := -\sqrt{n} \sum_{k=1}^{2} (1 - \delta^{0k}) \int_{[0,t]} \left(1 - \frac{y^{0,Z=1}(t,u)}{y^{0}(t,u)}\right) \times y^{0,Z=1}(t,u) \lambda_0^{0k}(u) du \tag{13}$$

and

$$\theta^{\mathrm{OS}}(t) := -\sqrt{n} \sum_{j=0}^{1} (1 - \delta^{j2}) \int_{[0,t]} \left(1 - \frac{y^{j,Z=1}(t,u)}{y^{j}(t,u)}\right) \times y^{j,Z=1}(t,u) \lambda_0^{j2}(u) du. \tag{14}$$

The elements of the 2×2-matrix–valued asymptotic variance function $\mathbf{V} : \mathbb{R}_+ \to \mathbb{R}_+^{2\times 2}$ with entries

$$\mathbf{V}(t) = \begin{pmatrix} V^{\mathrm{PFS}}(t) & V^{\mathrm{PFS,OS}}(t) \\ V^{\mathrm{PFS,OS}}(t) & V^{\mathrm{OS}}(t) \end{pmatrix}$$

for all $t \geq 0$ are given by

$$V^{\mathrm{PFS}}(t) = \sum_{k=1}^{2} \int_{[0,t]} \left(y^{0,Z=0}(t,u)\lambda_0^{0k}(u) + y^{0,Z=1}(t,u)\lambda_1^{0k}(u)\right) \times \left(\frac{y^{0,Z=1}(t,u)}{y^{0}(t,u)}\right)^2 du,$$

$$V^{\mathrm{OS}}(t) = \sum_{j=0}^{1} \int_{[0,t]} \left(y^{j,Z=0}(t,u)\lambda_0^{j2}(u) + y^{j,Z=1}(t,u)\lambda_1^{j2}(u)\right) \times \left(\frac{y^{j,Z=1}(t,u)}{y^{j}(t,u)}\right)^2 du,$$

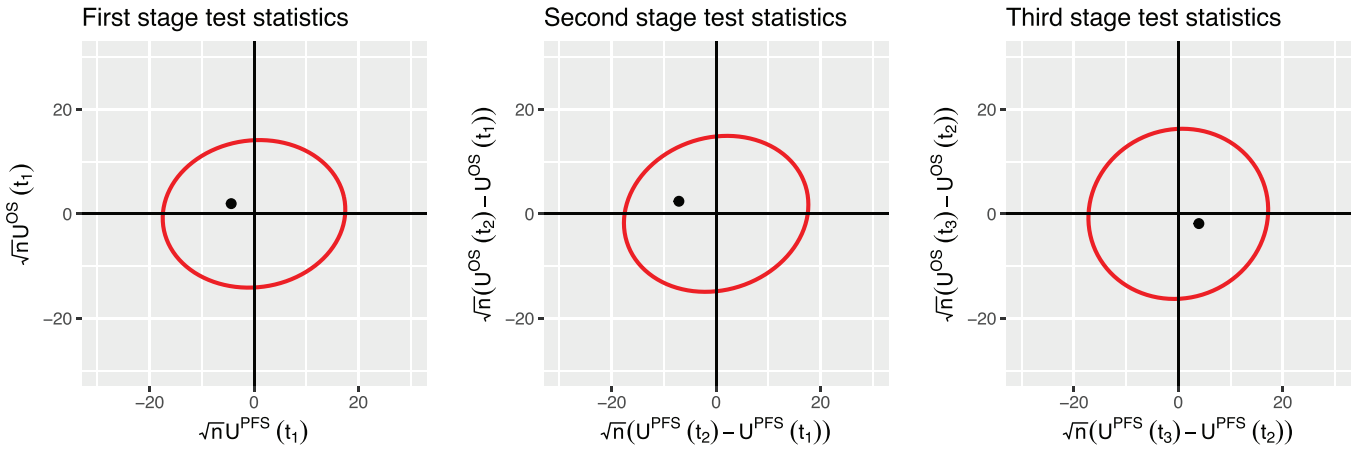

FIGURE 3 | Result of the evaluation of the NB2004-HR trial data without new method.

and

$$V^{\text{PFS, OS}}(t) = \sum_{j=0}^{1} \int_{[0,t]} \left(y^{j,Z=0}(t,u)\lambda_0^{j2}(u) + y^{j,Z=1}(t,u)\lambda_1^{j2}(u) \right) \times \left(\frac{y^{j,Z=1}(t,u)}{y^{j}(t,u)} \right)^2 du.$$

The incremental test statistics from (2) based on the quadratic form now follow a noncentral χ_2^2 distribution with noncentrality parameters

$$\eta_r := (\theta(t_r) - \theta(t_{r-1}))^T (\mathbf{V}(t_r) - \mathbf{V}(t_{r-1}))^{-1} (\theta(t_r) - \theta(t_{r-1}))$$

for $r \in \{0, 1\}$, $t_0 = 0$, a calendar date t_1 of the interim analysis, and the calendar date $t_2 = a + f$ of the final analysis. It should be noted that this can also be written as a multiple of the total sample size n.

For some fixed adaptive design as described above, this can directly be used to compute the distribution of stagewise p-values under the assumed alternative and to compute the resulting power based on the probability

$$\begin{aligned}&\mathbb{P}[p_1 \le \alpha_1] + \mathbb{P}[p_1 > \alpha_1, p_2 \le \alpha_2(p_1)] \\ &= \mathbb{P}[1 - F_{\chi_2^2}(\tilde{S}_1) \le \alpha_1] + \mathbb{P}[1 - F_{\chi_2^2}(\tilde{S}_1) > \alpha_1, 1 \\ &\quad - F_{\chi_2^2}(\tilde{S}_2) \le \alpha_2(1 - F_{\chi_2^2}(\tilde{S}_1))],\end{aligned}$$

where $\tilde{S}_r \sim \chi_2^2$ with noncentrality parameter η_r for $r \in \{1, 2\}$.

5.2 | Initial Planning

In this subsection, we try to shed light on some practical aspects of the power calculation sketched above. From historical data, it can be possible to estimate transition intensity functions $\lambda_0^{01}, \lambda_0^{02}, \lambda_0^{12}$ or integrated transition intensity functions $\Lambda_0^{01}, \Lambda_0^{02}, \Lambda_0^{12}$ of the control group ($Z = 0$). Such an estimation can be accomplished either nonparametrically (as in Aalen and Johansen 1978) or parametrically via standard maximum likelihood methods (as described in Meller, Beyersmann, and Rufibach 2019). The simplest case here may be a time-homogeneous model in which $\lambda_0^{01}, \lambda_0^{02}$, and λ_0^{12} are constants. These can be estimated unbiased and consistently by dividing the observed transitions between the respective states by the total time spent by all individuals in the first state. Hazard ratios that quantify the expected deviation of the treatment from the control group can be determined separately for each transition using prior data, external information, or a minimal clinically relevant effect.

When assuming a uniform recruitment mechanism over a period of length a and no loss to follow-up beyond administrative censoring, the computation of the functions $y^{j,Z=g}$ simplifies as follows under the independence assumption from Section 4:

$$\begin{aligned}y^{j,Z=g}(t,u) :=& \mathbb{P}[Z = g, R \le t - u, C \ge u, X(u) = j] \\ =& \mathbb{P}[R \le t - u] \cdot \mathbb{P}[X(u) = j | Z = g] \cdot \mathbb{P}[Z = g] \\ =& \frac{(t-u)_+ \wedge a}{a} \cdot P_g^{0j}(0,u) \cdot ((1-\pi) + g \cdot (2\pi - 1)).\end{aligned}$$

However, usually in a clinical trial, the sample size n cannot be chosen arbitrarily if the accrual duration a shall be fixed. The recruitment rate $r = n/a$ is rather given here, so that the accrual duration must be adjusted so that a target power is reached. In this case, the factor n appearing in the stagewise noncentrality parameters from the previous subsection shall be replaced by $r \cdot a$. As a also plays a role in the calculation of $y^{j,Z=g}$, a numerical root finding procedure has to be applied to find an appropriate accrual duration a to reach the target power.

5.3 | Sample Size Recalculation

At the time of interim analysis, the information received so far can be used to adapt the design of the trial. Commonly, the sample size is changed to meet a target of the conditional power. This is the probability of rejecting the null hypothesis given the p-value of the previous stages. Due to the construction of our test method and the associated circumvention of the problems mentioned in Bauer and Posch (2004), we can take into account all the information on deaths and progressive events that have taken place up to the interim analysis. In particular, it would be allowed to apply an adaptation rule as presented in

Bauer and Posch (2004). This is also confirmed by the results of the simulation study presented in Section 7.1. At the same time, no information needs to be discarded as it is the case for the designs presented in Jenkins et al. (2011) and Irle and Schäfer (2012).

Beyond that, this enables us to reassess all transition intensity functions λ_g^{jk} of the model. For the transition-wise hazard ratios, which constitute the decisive effect sizes in our design different approaches can be chosen. The most common ones are presented in section 7.4 of Wassmer and Brannath (2016). Of course, the interim estimate from the interim data can be used. However, this frequently made choice has been subject to some criticism (Bauer and Koenig 2006).

We now look at the most obvious way of sample size reassessment. If the recruitment rate r remains the same, only the recruitment period a is to be adjusted, whereby we assume that recruitment has already taken place until the time of the interim analysis t_1. Furthermore, the duration of the follow-up period f starting after the end of the recruitment period shall remain unchanged. One can now recalculate the noncentrality parameter η_2, inserting the newly estimated transition probabilities into formula (5.2), replacing the parameter a with $a_{\text{add}} + t_1$ and setting $t_2 = t_1 + a_{\text{add}} + f$. The duration of the accrual period beyond the interim analysis a_{add} is now the only free parameter when calculating the conditional power

$$\mathbb{P}\left[1 - F_{\chi_2^2}(\tilde{S}_2) \leq \alpha_2(p_1)\right] \quad \text{with } \tilde{S}_2 \sim \chi_2^2$$

$$\text{with non-centrality parameter } \eta_2$$

under the new assumptions. It can be chosen in such a way that this expression meets the targeted conditional power, which is often set to 80%. Please note that the first stage p-value p_1 is inserted in this expression.

Such a sample size recalculation procedure is also applied in our simulation studies in Section 7.

6 | Application Example

For further illustration of the methods introduced above, we now apply it to the data of the NB2004-HR trial (NCT number NCT03042429). This was an open-label, multicenter, prospective randomized controlled Phase III trial for treatment of children with high-risk neuroblastoma. The patients received six (control intervention) resp. eight (experimental intervention) cycles of induction chemotherapy. Afterwards, both groups received the same high-dose chemotherapy with autologous stem cell rescue and a consolidation therapy afterwards (see Berthold et al. 2020 for further details). The NB2004-HR trial had only one primary endpoint: EFS, defined as time from diagnosis to progression, recurrence, secondary malignant disease or death, whatever occurs first. Nevertheless, postprogression survival is of key interest both here and in many other studies with EFS as primary endpoint as well. In particular, the interaction of first- and second-line therapy given after progression is of special interest. The analysis did not reveal a relevant difference between the two interventions, neither in the primary endpoint EFS nor in the secondary endpoint OS. To illustrate our methodology, we will reanalyze the NB2004-HR trial using our testing method as in the context of Section 2 in order to compare the joint distribution of EFS and OS between the two interventions.

The NB2004-HR trial was originally designed group-sequentially according to Pampallona and Tsiatis (1994) including two interim analyses with futility stops and was later amended to an inverse normal adaptive design according to Wassmer (2006) using the same rejection region as the initial group-sequential design. On this basis, a data-dependent sample size recalculation was performed at the second interim analysis. We mimic this design by conducting interim analyses at the same time points. However, we do not make any binding futility stops. Stage-wise decision boundaries are determined by adopting the alpha-spending that resulted from the original procedure. Stagewise p-values are combined using the inverse normal combination function with equal weights for all stages.

The results are displayed in Figure 3. Each of the three plots in Figure 3 shows the value of the increment of $\sqrt{n}\mathbf{U}$ for the respective stage. As the test statistic $\sqrt{n}\mathbf{U}$ is bivariate, its observed value (as well as the corresponding rejection region) is located in the two-dimensional plane. The OS component is plotted in the vertical direction, the EFS component in the horizontal one. For both components of $\sqrt{n}\mathbf{U}$, negative values indicate an advantage of the experimental therapy in comparison with the control therapy. The red ellipses show the rejection bounds. If one of the test statistic increments would have been localized outside of the respective ellipse, the trial would have stopped with rejection of H_0. The exact shape of the ellipse that determines the rejection bound for the increments of $\sqrt{n}\mathbf{U}$ depends on the sequential decision boundaries in terms of p-values, the (estimated) variance of the increments of $\sqrt{n}\mathbf{U}$ given by the increments of $\hat{\mathbf{V}}$, as well as the results of previous analyses. The stagewise p-values resulting from our test turn out to be $p_1 = 0.536$, $p_2 = 0.227$, and $p_3 = 0.592$. Thus, the null hypothesis of no difference in the joint distribution of EFS and OS between the interventions cannot be rejected. This is qualitatively consistent with the results of the NB2004-HR trial as reported in Berthold et al. (2020).

In the original study, only primary outcome data on EFS were used for sample size recalculation as recommended by Wassmer (2006). The interim results from the first two phases suggested a slight benefit of the experimental treatment in terms of EFS, which is also evident from Figure 3 in the form of a slight shift to the left of the observed statistic. This led the researchers to increase the number of events after which the final analysis should take place. This increase resulted from the requirement to achieve a conditional power of 80% to reject the null hypothesis for EFS based on the original planning alternative. Information going beyond EFS-events has not been considered at the interim analyses. However, postprogression survival also plays a major role for a final assessment of a treatment for this disease. As one can see from the first two plots of Figure 3, a slightly unfavorable effect of the experimental treatment on postprogression survival has been observed at the interim analyses. This fact might have led the investigators to a different conclusion at the second interim analysis, if EFS and OS interim data had both been available in the context of the NB2004-HR trial.

7 | Simulation Study

In our simulation studies, we want to examine multiple aspects of our proposed procedure that were mentioned in our theoretical considerations. The first part demonstrates the extent to which the problem raised in Bauer and Posch (2004) also applies to the setting discussed here. Then, the compliance with the nominal type I error rate is checked in different settings and under different sample size recalculation rules. Finally, type II errors are assessed under correct specification and also under misspecification of the initial planning assumptions of the trial.

In all parts of our simulations study, we consider the illness-death model that has already been discussed in Section 2. Within this simple multistate model, we mainly consider transition intensities that have a Weibull form, that is, they will be given by

$$\lambda^{jk}(s) = \lambda^{jk} \cdot s^{\gamma^{jk}-1} \tag{15}$$

with shape parameter γ^{jk} and scale parameter λ^{jk} for any $(j,k) \in \{(0,1),(0,2),(1,2)\}$. In the special case in which $\gamma^{01} = \gamma^{02} = \gamma^{12} = 1$, the transition intensities are constant over time and the model is referred to as a time-homogeneous Markov model.

For sample size calculation and type II error considerations, we assume that the groups differ in each transition by a proportionality factor as in Section 5. This means that the transition intensities λ_1^{jk} in the experimental treatment group ($Z = 1$) are related to the intensities λ_0^{jk} in the control group ($Z = 0$) by

$$\lambda_1^{jk}(s) = \delta^{jk} \cdot \lambda_0^{jk}(s) \quad \forall s \geq 0$$

given hazard ratios δ^{jk}.

The tests based on our procedure were carried out at an overall significance level of $\alpha = 5\%$. Stagewise p-values were combined using the inverse normal combination function with equal weights for the two stages. We applied sequential decision boundaries according to Pocock as well as O'Brien-Fleming (abbreviated by P resp. OF). For any constellation in the following subsections, 10,000 simulation runs were executed. For underlying true values of 0.05 and 0.8, the half-width of the 95% confidence intervals amount to 0.0043 and 0.0078, respectively. The simulation study was performed with R 4.2.1 (see R Core Team 2014).

7.1 | Type I Error Rate Inflation Due to Informative Disease Progressions

As already mentioned in Section 5.3, we want to demonstrate that the type I error rate of the group-sequential procedure of Lin (1991) is inflated if information on disease progression is used that is informative for the further course of the disease (see Bauer and Posch 2004). We will show that this inflation goes beyond the inflation of the type I error rate that is already caused by a data-dependent redesign of a group-sequential testing procedure (cf. Proschan and Hunsberger 1995). The simulation results confirm that our approach simultaneously addresses both these aspects. In particular, our adaptive procedure adheres to the nominal type I error rate, even if information on the disease course is used to determine design changes.

For this purpose, we create two scenarios. In the first one, disease progression is not informative for the further course of disease, that is, $\lambda^{02} \equiv \lambda^{12}$. In particular, we set $\gamma^{02} = \gamma^{12} = 2$ and $\lambda^{02} = \lambda^{12} = 0.05$ in (15) for this scenario. In the second one, disease progression will deteriorate survival chances, that is, $\lambda^{02} < \lambda^{12}$. Therefore, we set $\lambda^{12} = 3$ and leave all other parameters unchanged. In both scenarios, the intensity for the transition from the initial state to the state of disease progression is chosen as

$$\lambda^{01}(s) = -\log(0.5) \cdot \mathbb{1}_{[0,1)}(s) \forall s \geq 0.$$

This means that in the first year in the trial, disease progression occurs at a constant rate of $-\log(0.5)$. After this first year, no further progressions will occur. In particular, this means that under the neglect of the other transitions, half of the patients would have experienced a disease progression after 1 year in the trial. In the case of informative disease progressions, this means that in an interim analysis there are groups of approximately the same size that have either good or poor chances of survival. In this regard, this scenario mimics one key property of the didactic scenario that has been sketched in Bauer and Posch (2004) to illustrate the statistical problem.

In the first stage, patients will be recruited uniformly over an interval of $t_1 = 3$ years. Recruitment rates are set in such a way that the number of recruited patients at the interim analysis amounts to $n \in \{50, 100, 200, 400, 1000\}$. The patients are allocated to the two treatment groups with equal probability.

At the interim analysis, the study may be terminated for an early success or continued. The adaptation rule used at this point shall mimic the one discussed in Bauer and Posch (2004). It is only based on the interim PFS test statistic $Z^{\text{PFS}}(t_1) := U^{\text{PFS}}(t_1)/\sqrt{\hat{V}^{\text{PFS}}(t_1)}$. This quantity is standard normally distributed under the null hypothesis. If it shows a certain deviation from its expected value of 0, recruitment will be stopped immediately. This option will be chosen if $Z^{\text{PFS}}(t_1) \notin [z_{\tilde{\alpha}}, z_{1-\tilde{\alpha}}]$ for some $\tilde{\alpha} \in [0, 0.5]$. The values $z_{\tilde{\alpha}}$ and $z_{1-\tilde{\alpha}}$ denote the $\tilde{\alpha}$ and $1-\tilde{\alpha}$ quantile of the standard normal distribution, respectively. We consider values of $\tilde{\alpha} \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.33\}$. The final analysis will be executed after the follow-up period of 5 years. Otherwise, the recruitment period will be extended by factor 10 under the same recruitment rate as in the initial stage. The final analysis will be executed after the subsequent follow-up period of 5 years.

In Table 1, empirical type I error rates for both scenarios (noninformative and informative disease progression), different choices of $\tilde{\alpha}$, and sequential decision bounds (Pocock and O'Brien-Fleming) from our simulations with 10,000 runs are shown. We display the results for the case where 100 patients per group are recruited up to the interim analysis. Results for different sample sizes look similar and can be found in the Supporting Information. First of all, it should be noted that the newly proposed design adheres to the nominal type I error rate in all constellations. Of course, even in the noninformative case, the type I error rate of Lin's design is inflated when the interim

TABLE 1 | Type I error rate inflation from informative disease progression for a sample size of 100 patients per group recruited up to the interim analysis.

Adaptation rule	Decision bounds	Noninformative progression		Informative progression	
		Lin's design	New design	Lin's design	New design
$\tilde{\alpha} = 0.05$	P	0.0559	0.0511	0.0588	0.0547
	OF	0.0600	0.0511	0.0678	0.0521
$\tilde{\alpha} = 0.10$	P	0.0602	0.0521	0.0623	0.0535
	OF	0.0657	0.0520	0.0737	0.0517
$\tilde{\alpha} = 0.15$	P	0.0614	0.0529	0.0632	0.0523
	OF	0.0673	0.0529	0.0759	0.0509
$\tilde{\alpha} = 0.20$	P	0.0615	0.0531	0.0635	0.0518
	OF	0.0664	0.0528	0.0763	0.0513
$\tilde{\alpha} = 0.25$	P	0.0619	0.0531	0.0637	0.0522
	OF	0.0656	0.0527	0.0755	0.0505
$\tilde{\alpha} = 0.33$	P	0.0610	0.0537	0.0625	0.0521
	OF	0.0633	0.0537	0.0711	0.0508

TABLE 2 | Parameter configurations and event rates for the three scenarios of the simulation study.

Scenario	γ^{01}	λ^{01}	γ^{02}	λ^{02}	γ^{12}	λ^{12}	$\pi^{\text{PFS}}(t_1)$	$\pi^{\text{OS}}(t_1)$	$\pi^{\text{PFS}}(t_2)$	$\pi^{\text{OS}}(t_2)$
1	1	0.6	1	0.075	1	0.9	0.431	0.241	0.889	0.745
2	1.3	0.85	1.3	0.1	1.3	0.3	0.522	0.189	0.980	0.694
3	1.5	0.57	0.5	0.065	0.85	1.1	0.441	0.235	0.957	0.772

PFS test statistic is used to make design adaptations. However, this inflation compared to our design is much bigger if disease progression is informative. This difference is most articulate for $\tilde{\alpha} = 0.2$ and the O'Brien–Fleming decision bounds. While there is an inflation of 1.64 percentage points for the noninformative case, the inflation amounts to 2.63 percentage points in the informative case. This underlines that the group sequential procedure is not suitable to be applied in an adaptive setting from several points of view.

7.2 | Type I Error Rate Compliance

We study the type I error rate compliance of our testing procedure in the illness-death model with parameter configurations as presented in Meller, Beyersmann, and Rufibach (2019). These can be found in Table 2. From our point of view, these configurations form a good basis, as a time-homogeneous model as well as a model with constant shape parameters across all transitions (as in Li and Zhang 2015) as well as a model with different shape parameters are considered. Initially, a sequential design with an interim analysis after $t_1 = 2.5$ and a final analysis after $t_2 = 5$ years is planned. The duration of the accrual period is set to $a = 3$. The recruitment date of each trial participant is simulated as uniformly distributed on the interval $[0, a]$. The trial participants are allocated to each of the two treatment groups with probability 0.5. Under these conditions, the expected proportion of all trial participants that will have experienced a PFS- or an OS-event by calendar time t_1 resp. t_2 is given in the last four columns of Table 2 for each scenario. Hereby, we consider three different adaptation strategies. First, a simple group-sequential procedure is applied in which no adaptation is made and the analyses take place at the initially planned dates. In addition, we inspect the decision rule introduced in Section 7.1 with $\tilde{\alpha} = 0.2$. Finally, we apply an adaptation strategy that is based on conditional power calculation as lined out in Section 5.3. The recruitment period is adjusted in such a way that the conditional power under the observed treatment effects amounts to 80%. For the sake of realism, the recruitment period is limited by twice the originally planned recruitment period. If a recruitment stop (i.e., $a_{\text{add}} = 0$ in terms of Section 5.3) already yields a power of 80%, recruitment is stopped and the final analysis is conducted at $t_2 = t_1 + f$. If no adjustment of the recruitment duration leads to the conditional power target, the maximum possible recruitment duration is selected and the final analysis is conducted at $t_2 = 2a + f$.

The empirical type I error rates obtained via 10,000 simulation runs are shown in Table 3. For an initially planned sample size of 200 and above our procedure adheres to the nominal type I error level of $\alpha = 0.05$ for any baseline scenario, sequential decision boundary, and adaptation rule. It can also be concluded that the presented method is slightly anticonservative for small sample sizes (below 100). However, the deviation from the nominal significance level is small. This tendency can also be observed for the standard log-rank test (Heller and Venkatraman 1996), which can be regarded as a special case of our framework (with

TABLE 3 | Empirical type I errors for different initially planned sample sizes, sequential decision bounds and scenarios in a group-sequential design (GS), an adaptive design as inspired by the adaptation rule in Bauer and Posch (2004) (BP) and an adaptive design based on conditional power calculations (CP).

		Scenarios								
		1			2			3		
n	Type	GS	BP	CP	GS	BP	CP	GS	BP	CP
50	P	0.0586	0.058	0.0544	0.0553	0.0557	0.0593	0.0546	0.0578	0.0598
	OF	0.0568	0.0564	0.0576	0.0552	0.0557	0.0603	0.0566	0.0568	0.0565
100	P	0.0554	0.0539	0.0531	0.053	0.0524	0.0543	0.0559	0.0539	0.0512
	OF	0.0535	0.0528	0.052	0.0512	0.052	0.0558	0.0544	0.0538	0.0518
200	P	0.0549	0.0525	0.053	0.0514	0.052	0.0498	0.0541	0.0519	0.0489
	OF	0.0558	0.0515	0.0537	0.0513	0.0514	0.055	0.0516	0.0511	0.05
400	P	0.0482	0.051	0.0518	0.0494	0.0517	0.0471	0.0512	0.0502	0.0485
	OF	0.0508	0.0515	0.0489	0.0493	0.0517	0.0549	0.0496	0.0505	0.0538
1000	P	0.0472	0.0508	0.0484	0.049	0.0508	0.0448	0.0526	0.0507	0.0493
	OF	0.0503	0.0505	0.0488	0.0494	0.0505	0.0528	0.0505	0.0503	0.0493

$k = 1$, $d = 1$, and $m = 1$ in terms of the framework introduced in Section 4). The choice of sequential decision boundaries does not seem to play a role for the actually achieved significance level.

7.3 | Type II Error Rates Under Correct Specification of Treatment Effects

For sample size calculation and type II error considerations, we assume that the groups differ in each transition by a proportionality factor as in Section 5. This means that the transition intensities in the experimental treatment group ($Z = 1$) are given by

$$\lambda_1^{jk}(s) = \delta^{jk} \cdot \lambda_0^{jk}(s) \quad \forall s \geq 0$$

given hazard ratios δ^{01}, δ^{02}, and δ^{12}. In our simulation settings, we assume values $\delta^{02} = 1$, $\delta^{01} \in \{0.8, 0.7, 0.6\}$, and $\delta^{12} \in \{0.85, 0.8, 0.75\}$. As in the previous subsection, a design with an interim analysis at $t_1 = 2.5$ and a final analysis at $t_2 = 5$ years is planned initially. The patients are planned to be recruited during an accrual period of length $a = 3$. For each combination of hazard ratios δ^{01} and δ^{12}, the sample size is calculated as lined out in Section 5 to achieve a power of 80%. We considered a purely group sequential plan without any adaptations as well as an adaptive plan where all transitions of the multistate model were assessed based on the interim data and the accrual duration was adjusted to meet a conditional power of 80% as lined out in Section 5.3. Sequential O'Brien–Fleming decision boundaries were applied throughout this part of the simulation study.

For the parameter constellation of scenario 1 (see Table 2), the results can be found in Table 4. First of all, they lead to the conclusion that the analytical determination of the sample size described above works reliably in terms of compliance with the targeted power. Adjustment of the accrual length at an interim analysis to meet a conditional power of 80% increases the overall power of the procedure by a bit more than 10 percentage points for the Pocock decision bounds and a bit less than 10 percentage points for the O'Brien–Fleming decision bounds. This goes along with an increase of the average sample size by 20%–25% and 15%–20%, respectively. Both increases can be attributed to the fact that a conditional power of 80% at an interim analysis leads to an overall power of more than 80% and the tendency of the conditional power recalculation rule toward extreme decisions when using interim effect estimates (see Bauer and Koenig 2006).

For the parameter constellations of the remaining scenarios 2 and 3 (see Table 2), the results do note deviate remarkably from the observations we made here for the first scenario. Corresponding tables can be found in the Supporting Information.

7.4 | Type II Error Rates Under Misspecification of Treatment Effects

Finally, we want to examine the behavior of the new procedure if the hazard ratios are misspecified in the planning stage of the trial. For the initial sample size calculation, we consider the same planning assumptions as in the previous subsection. In the simulation runs, the parameters for the control group remain the same, too. However, the differences between the two treatment groups in terms of the hazard ratios in the simulations differ from the planning assumptions. We consider settings in which the actual hazard ratios δ^{01} and δ^{12} used in the simulations differ from the planning assumptions by values in the set $\{-0.1, -0.05, 0, 0.05, 0.1\}$. We expect that some power is recovered from a sample size recalculation via a conditional power approach as in the previous subsection.

For the parameter constellation of scenario 1 (see Table 2) and initial planning assumptions of $\delta^{01} = 0.7$ and $\delta^{12} = 0.8$, the results can be found in Table 5. The power of the group sequential procedure without adaptations is determined analytically while the power of the adaptive procedure is determined by simulations

TABLE 4 | Results of simulations of type II error rates for scenario 1. Maximal sample size in group sequential design $n_{\max,GS}$, power of the group sequential design $1-\beta_{GS}$, average sample size in the group sequential design $\mathbb{E}[n_{GS}]$, power of the conditional power procedure $1-\beta_{CP}$, and average sample size of the conditional power procedure $\mathbb{E}[n_{CP}]$ are displayed.

δ^{01}, δ^{12}	Type	$n_{\max,GS}$	$1-\beta_{GS}$	$\mathbb{E}[n_{GS}]$	$1-\beta_{CP}$	$\mathbb{E}[n_{CP}]$
0.8, 0.85	P	620	0.7932	574.43	0.9169	711.57
0.8, 0.85	OF	577	0.7997	551.91	0.8957	652.99
0.7, 0.85	P	294	0.8012	271.08	0.9161	333.68
0.7, 0.85	OF	275	0.8087	262.38	0.8851	308.84
0.6	P	157	0.8055	144.35	0.9143	176.63
0.6, 0.85	OF	147	0.8003	140.26	0.8853	163.97
0.8, 0.8	P	512	0.7995	477.16	0.9182	587.76
0.8, 0.8	OF	473	0.8022	455.34	0.8963	542.86
0.7, 0	P	272	0.8036	251.09	0.9100	312.40
0.7, 0.8	OF	254	0.8033	243.01	0.8920	285.30
0.6, 0.8	P	153	0.8012	141.14	0.9151	175.19
0.6, 0.8	OF	143	0.8017	136.50	0.8872	161.03
0.8, 0.75	P	408	0.8025	381.70	0.9173	472.54
0.8, 0.75	OF	376	0.7992	363.47	0.8954	433.72
0.7, 0.75	P	244	0.7926	226.64	0.9132	281.34
0.7, 0.75	OF	227	0.8010	217.78	0.8930	258.67
0.6, 0.75	P	146	0.8026	134.85	0.9126	165.88
0.6, 0.75	OF	136	0.8030	130.10	0.8868	152.65

TABLE 5 | Comparison of the power of a group-sequential design and an adaptive design with sample size recalculation based on the conditional power for the observed effect sizes. The trial is initially planned with the parameter constellation of scenario 1 and hazard ratios $\delta^{01} = 0.7$ and $\delta^{12} = 0.8$. A sample size of 254 patients per group would be required in this case to reach a power of 80%. Upper value in each cell refers to power of group-sequential design, lower value refers to the adaptive design with the number in parentheses denoting the average sample size of the adaptive design.

		δ_{01}				
		0.6	**0.65**	**0.7**	**0.75**	**0.8**
δ_{12}	0.7	0.9864	0.9564	0.899	0.8183	0.7314
		0.9871 (231.37)	0.9695 (246.11)	0.9451 (266.99)	0.9045 (288.53)	0.8642 (309.53)
	0.75	0.9799	0.9348	0.8493	0.7319	0.6088
		0.9828 (233.69)	0.9586 (251.14)	0.9133 (277.63)	0.8614 (303.89)	0.7779 (328.93)
	0.8	0.9737	0.9141	0.8018	0.651	0.498
		0.9787 (236.77)	0.9471 (257.23)	0.8879 (287.51)	0.8011 (318.43)	0.6779 (348.75)
	0.85	0.9692	0.898	0.7641	0.5873	0.4133
		0.9763 (238.7)	0.9382 (261.77)	0.8672 (293.41)	0.744 (326.68)	0.5723 (359.96)
	0.9	0.9672	0.8892	0.7415	0.5473	0.3598
		0.974 (238.45)	0.9384 (262.99)	0.8541 (295.93)	0.7067 (332.35)	0.4948 (368.92)

with 10,000 simulation runs. Results for the other baseline scenarios and planning assumptions considered in the previous subsection can be found in the Supporting Information.

As already seen before, the sample size recalculation inflates the power by slightly less than 10 percentage points if the planning assumptions agree with the actual values. The power does not increase more than that of the group-sequential design if the difference between the two treatments is initially underestimated. In contrast, the power performance of the adaptive design is better than that of the group-sequential design in case differences between the two groups are overestimated. Especially when δ^{01} is overestimated, the adaptive design can recover some of the lost power if there is a relevant advantage in postprogression

survival. However, this may increase the average sample size by about 50% compared to the maximal sample size of the group-sequential trial.

8 | Discussion

An adaptive group-sequential testing procedure for multiple primary time-to-event endpoints has been introduced. It serves as a generalization to the adaptive log-rank test as presented, for example, by Wassmer (2006) and coincides with the group-sequential procedure of Lin (1991) in case of a competing risks setting. As a consequence of the concerns raised in Bauer and Posch (2004), an extension of Lin (1991) to an adaptive design is not straightforward. We do achieve that here by embedding these endpoints in a multistate model under the assumption of Markovianity of this model. Our approach enables data-dependent interim design modifications based on the information about all involved endpoints. Similar to the one-sample procedure presented in Danzer et al. (2022), this is based on conditioning on the prior history of each patient, which can be reduced to the current disease state under the Markov assumption.

As a particularly relevant application example from a practical point of view, we place a special focus on the joint consideration of PFS and OS in the framework of a simple illness-death model (see Figure 1). Both endpoints play a major role in oncology clinical trials. While OS is the most objectively defined endpoint, the choice of PFS as the primary endpoint is already established in many cases, depending on the specific indication. Often, as in our example in Section 6, both endpoints are of crucial importance, suggesting a joint consideration of both. For immunotherapies in particular, it is possible that therapy effects only become apparent or even after a progression event (Hoos 2012). This is another reason why a joint consideration of the endpoints OS and PFS appears helpful. Using the data from the NB2004-HR study, we also show how the different aspects of our multivariate test can be visualized and interpreted (see Figure 3). The benefits that can be gained from our adaptive design in terms of interim, data-driven design changes have also been demonstrated in Section 7.4.

Our simulation study has demonstrated that adherence to the nominal type I error rate is not only given asymptotically in the limit of large sample sizes, but is also acceptable at case numbers of practical relevance. However, adherence to the nominal type I error rate could be improved for small sample sizes by applying a resampling procedure, as was carried out, for example, in Ditzhaus and Friedrich (2020) for a test statistic that also results from a quadratic form. We also considered effects of several differences in the survival pattern between the two groups on power and sample size of a corresponding study. In this regard, it should be noted that our procedure appears particularly suitable and superior to an adaptive test of the single endpoint PFS in terms of power if there is a relevant effect for postprogression survival. If no or only a very small effect with respect to OS is expected, obviously, the restriction to a classical adaptive test of the single endpoint PFS in the sense of Wassmer (2006) appears more reasonable.

The methods presented here can be extended in several ways. To this end, it should be noted that the components of our general test statistic only take the first hitting time of some subset E of the state space into account. Hence, it is not only a test for the null hypothesis (11) formulated in terms of the cumulative intensity matrix, but also for the joint distribution of the d different endpoints as in (10). However, as an alternative to our approach, one could also think of test statistics that incorporate any hitting time of this set E and not only the first one. The derivation of such a procedure is analogous to the derivation of the procedure on which we are focussing here. It is also carried out in full detail in the Supporting Information. These two approaches are the same for our illness-death model from Section 2 but can already differ for slightly more complex cases as, for example, the setting in Figure 2. However, the latter approach can only be used as a test for the null hypothesis (11) as formulated in terms of the transition intensities.

Furthermore, we want to point out that we assumed transition-wise proportional hazards throughout our examples. Note that this generally does not imply proportional hazards for the endpoints (e.g., PFS and OS) considered within the multistate model. In addition, settings are possible where the transition-wise comparisons may also not be subject to the proportional hazards assumption. If this is known, it might be beneficial to apply weights as it is also common for the univariate log-rank test (see, e.g., section V.2 in Andersen et al. 1993). Such a weight can be selected separately for each individual transition. The theory lined out in the Supporting Information allows any weight fulfilling the standard assumptions.

An extension of our two-sample procedure to a k-sample procedure for some $k > 2$ follows analogously to the way that the multivariate testing procedure from Wei and Lachin (1984) is extended by Palesch and Lachin (1994), and is thus possible without further problems.

As stated in (10) resp. (11), we are testing a global null hypothesis. In Wei and Lachin (1984) and Lin (1991), tests against a more restricted alternative have also been suggested. These could be adopted to detect an advantage in one of the involved endpoints. However, these may be sensitive to undesirable alternatives as demonstrated in Bloch, Lai, and Tubert-Bitter (2001). To protect against this, methods such as those from Bloch, Lai, and Tubert-Bitter (2001) and Perlman and Wu (2004) could also be used. These are only sensitive to scenarios in which the new treatment is superior to the control group in one endpoint and noninferior in all other endpoints. However, a determination of transition-specific noninferiority bounds would be required here. In any case, rejection of the global null hypothesis should be followed by more in-depth analyses. This could be achieved by a closed testing procedure involving the various components of $\mathbf{M}$, similar to the suggestions made in Lehmacher et al. (1991). A separate analysis of the transition intensities as demonstrated in section IV.4.4 of Andersen et al. (1993) is also recommendable.

The correctness of the procedure requires the Markov assumption. This allows us to adequately incorporate the information gathered so far into the testing procedure. Before use, the appropriateness of this assumption should therefore be investigated. On the one hand, this can be based on the expertise of clinical investigators. On the other hand, it can be examined in historical data sets that reflect the population to be recruited for the present

trial. Corresponding testing procedures have been developed for the simple illness-death model of figure 1 in Rodríguez-Girondo and de Uña-Álvarez (2012) as well as for general multistate models in Titman and Putter (2020).

Considering the topics discussed here, we strive to further develop and improve our framework in future research to enhance applicability in clinical trials. In principle, analogous methods can be developed in non-Markov settings, for example, in the scenario of semi-Markov models (see, e.g., Meller, Beyersmann, and Rufibach 2019 for details on the semi-Markovian illness-death model). Furthermore, we aim to develop tests for marginal distributions of time-to-event endpoints in Markovian multistage models. Compared to the current methodology, these should not only consider the conditional distribution of an endpoint and still allow for the possibility of interim design adaptations based on the disease history data of all patients.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–413730122. The NB2004-HR trial was funded by a grant obtained by Frank Berthold from the Deutsche Krebshilfe (grant number 70107712). We thank Frank Berthold for the opportunity to scientifically use the anonymized EFS and OS data of the NB2004-HR trial (published elsewhere) as a real clinical example in the context of this work.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Due to legal restrictions, data of the NB2004-HR trial are not available.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

References

Aalen, O. O., and S. Johansen. 1978. "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations." *Scandinavian Journal of Statistics* 5, no. 3: 141–150.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding. 1993. *Statistical Models Based on Counting Processes*. New York: Springer.

Bauer, P., and F. Koenig. 2006. "The Reassessment of Trial Perspectives From Interim Data—A Critical View." *Statistics in Medicine* 25, no. 1: 23–36.

Bauer, P., and M. Posch. 2004. "Letter to the Editor: Modification of the Sample Size and the Schedule of Interim Analyses in Survival Trials Based on Data Inspections." *Statistics in Medicine* 23, no. 8: 1333–1334.

Bellera, C. A., M. Pulido, S. Gourgou, et al. 2013. "Protocol of the Definition for the Assessment of Time-to-Event Endpoints in CANcer Trials (DATECAN) Project: Formal Consensus Method for the Development of Guidelines for Standardised Time-to-Event Endpoints' Definitions in Cancer Clinical Trials." *European Journal of Cancer* 49, no. 4: 769–781.

Berthold, F., A. Faldum, A. Ernst, et al. 2020. "Extended Induction Chemotherapy Does Not Improve the Outcome for High-Risk Neuroblastoma Patients: Results of the Randomized Open-Label Gpoh Trial NB2004-HR." *Annals of Oncology* 31, no. 3: 422–429.

Beyersmann, J., A. Allignol, and M. Schumacher. 2011. *Competing Risks and Multistate Models With R*. Use R! New York: Springer.

Bloch, D. A., T. L. Lai, and P. Tubert-Bitter. 2001. "One-Sided Tests in Clinical Trials With Multiple Endpoints." *Biometrics* 57, no. 4: 1039–1047.

Danzer, M. F., T. Terzer, F. Berthold, A. Faldum, and R. Schmidt. 2022. "Confirmatory Adaptive Group Sequential Designs for Single-Arm Phase II Studies With Multiple Time-to-Event Endpoints." *Biometrical Journal* 64, no. 2: 312–342.

Ditzhaus, M., and S. Friedrich. 2020. "More Powerful Logrank Permutation Tests for Two-Sample Survival Data." *Journal of Statistical Computation and Simulation* 90, no. 12: 2209–2227.

Erdmann, A., J. Beyersmann, and K. Rufibach. 2023. "Oncology Clinical Trial Design Planning Based on a Multistate Model That Jointly Models Progression-Free and Overall Survival Endpoints." arXiv:2301.10059.

Heller, G., and E. S. Venkatraman. 1996. "Resampling Procedures to Compare Two Survival Distributions in the Presence of Right-Censored Data." *Biometrics* 52, no. 4: 1204–1213.

Hoos, A. 2012. "Evolution of End Points for Cancer Immunotherapy Trials." *Annals of Oncology* 23: viii47–viii52. Advances in Immuno-oncology.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data*. (1st ed.). Statistics for Biology and Health. New York: Springer-Verlag.

Irle, S., and H. Schäfer. 2012. "Interim Design Modifications in Time-to-Event Studies." *Journal of the American Statistical Association* 107, no. 497: 341–348.

Jenkins, M., A. Stone, and C. Jennison. 2011. "An Adaptive Seamless Phase ii/iii Design for Oncology Trials With Subpopulation Selection Using Correlated Survival Endpoints†." *Pharmaceutical Statistics* 10, no. 4: 347–356.

Jörgens, S., G. Wassmer, F. König, and M. Posch. 2019. "Nested Combination Tests With a Time-to-Event Endpoint Using a Short-Term Endpoint for Design Adaptations." *Pharmaceutical Statistics* 18, no. 3: 329–350.

Le-Rademacher, J., R. Peterson, T. Therneau, B. Sanford, R. Stone, and S. Mandrekar. 2018. "Application of Multi-State Models in Cancer Clinical Trials." *Clinical Trials* 15: 174077451878909.

Lehmacher, W., G. Wassmer, and P. Reitmeir. 1991. "Procedures for Two-Sample Comparisons With Multiple Endpoints Controlling the Experimentwise Error Rate." *Biometrics* 47, no. 2: 511–521.

Li, Y., and Q. Zhang. 2015. "A Weibull Multi-State Model for the Dependence of Progression-Free Survival and Overall Survival." *Statistics in Medicine* 34, no. 17: 2497–2513.

Lin, D. 1991. "Nonparametric Sequential Testing in Clinical Trials With Incomplete Multivariate Observations." *Biometrika* 78, no. 1: 123–131.

Magirr, D., T. Jaki, F. Koenig, and M. Posch. 2016. "Sample Size Reassessment and Hypothesis Testing in Adaptive Survival Trials." *PLOS ONE* 11, no. 2: 1–14.

Meller, M., J. Beyersmann, and K. Rufibach. 2019. "Joint Modeling of Progression-Free and Overall Survival and Computation of Correlation Measures." *Statistics in Medicine* 38, no. 22: 4270–4289.

Palesch, Y. Y., and J. M. Lachin. 1994. "Asymptotically Distribution-Free Multivariate Rank Tests for Multiple Samples With Partially Incomplete Observations." *Statistica Sinica* 4, no. 1: 373–387.

Pampallona, S., and A. A. Tsiatis. 1994. "Group Sequential Designs for One-Sided and Two-Sided Hypothesis Testing With Provision for Early

Stopping in Favor of the Null Hypothesis." *Journal of Statistical Planning and Inference* 42, no. 1: 19–35.

Perlman, M. D., and L. Wu. 2004. "A Note on One-Sided Tests With Multiple Endpoints." *Biometrics* 60, no. 1: 276–279.

Proschan, M. A., and S. A. Hunsberger. 1995. "Designed Extension of Studies Based on Conditional Power." *Biometrics* 51, no. 4: 1315–1324.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rodríguez-Girondo, M., and J. de Uña-Álvarez. 2012. "A Nonparametric Test for Markovianity in the Illness-Death Model." *Statistics in Medicine* 31, no. 30: 4416–4427.

Scharfstein, D. O., A. A. Tsiatis, and J. M. Robins. 1997. "Semiparametric Efficiency and Its Implication on the Design and Analysis of Group-Sequential Studies." *Journal of the American Statistical Association* 92, no. 440: 1342–1350.

Schäfer, H., and H.-H. Müller. 2001. "Modification of the Sample Size and the Schedule of Interim Analyses in Survival Trials Based on Data Inspections." *Statistics in Medicine* 20, no. 24: 3741–3751.

Sellke, T., and D. Siegmund. 1983. "Sequential Analysis of the Proportional Hazards Model." *Biometrika* 70, no. 2: 315–326.

Tattar, P. N., and H. J. Vaman. 2014. "The k-Sample Problem in a Multi-State Model and Testing Transition Probability Matrices." *Lifetime Data Analysis* 20, no. 3: 387–403.

Titman, A. C., and H. Putter. 2020. "General Tests of the Markov Property in Multi-State Models." *Biostatistics* 23, no. 2: 380–396.

Tsiatis, A. A. 1981. "The Asymptotic Joint Distribution of the Efficient Scores Test for the Proportional Hazards Model Calculated Over Time." *Biometrika* 68, no. 1: 311–315.

Wassmer, G. 2006. "Planning and Analyzing Adaptive Group Sequential Survival Trials." *Biometrical Journal* 48, no. 4: 714–729.

Wassmer, G., and W. Brannath. 2016. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Switzerland: Springer International Publishing.

Wei, L. J., and J. M. Lachin. 1984. "Two-Sample Asymptotically Distribution-Free Tests for Incomplete Multivariate Observations." *Journal of the American Statistical Association* 79, no. 387: 653–661.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.