Confirmatory adaptive group sequential designs for single-arm phase II studies with multiple time-to-event endpoints

Moritz Fabian Danzer 1, Tobias Terzer 2, Frank Berthold 3, Andreas Faldum 1, Rene Schmidt 1

1 Institute of Biostatistics and Clinical Research, University of Münster, Münster, Germany
2 Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany
3 Department of Pediatric Oncology and Hematology, Center for Integrated Oncology, University of Cologne, Cologne, Germany

Correspondence
Moritz Fabian Danzer, Institute of Biostatistics and Clinical Research, University of Münster, 48149 Münster, Germany.
Email: moritzfabian.danzer@ukmuenster.de

Funding information
Deutsche Krebshilfe, Grant/Award Number: 70107712; Deutsche Forschungsgemeinschaft, Grant/Award Number: 413730122

Abstract
Existing methods concerning the assessment of long-term survival outcomes in one-armed trials are commonly restricted to one primary endpoint. Corresponding adaptive designs suffer from limitations regarding the use of information from other endpoints in interim design changes. Here we provide adaptive group sequential one-sample tests for testing hypotheses on the multivariate survival distribution derived from multi-state models, while making provision for data-dependent design modifications based on all involved time-to-event endpoints. We explicitly elaborate application of the methodology to one-sample tests for the joint distribution of (i) progression-free survival (PFS) and overall survival (OS) in the context of an illness-death model, and (ii) time to toxicity and time to progression while accounting for death as a competing event. Large sample distributions are derived using a counting process approach. Small sample properties are studied by simulation. An already established multi-state model for non-small cell lung cancer is used to illustrate the adaptive procedure.

KEYWORDS
martingale theory, multi-state models, sample size recalculation, survival analysis

1 INTRODUCTION

Traditional designs for single arm phase II survival studies comprise hypothesis tests for prefixed survival rates or application of the one-sample log-rank test with respect to (w.r.t.) one primary time-to-event endpoint. The necessity to focus on
a single time-to-event endpoint constitutes a major restriction of those approaches. Especially in oncology, the multiple aspects of disease and treatment require a more comprehensive description, which often cannot be reflected adequately by a single time-to-event endpoint. In clinical practice, researchers therefore often tend to capture the different aspects of disease and treatment in a single composite endpoint, for example, time to progression, severe toxicity or death, whatever occurs first. The drawback of the latter approach is a quite coarse description of the course of disease. In oncology, it is instead of interest to describe progression, toxicity and death simultaneously as complementary and intertwined characteristics of disease and treatment. Of course, the relative importance of these events may depend on the characteristics of the disease and the stage of development of a new therapy. However, as soon as it is clarified which relationship shall be investigated in the respective situations, multi-state models are an appropriate tool to depict these aspects properly. While there are well-elaborated multi-state models for joint modelling of efficacy and toxicity or joint modelling of progression and death as time-to-event endpoints, there is limited methodology to assess the performance of an experimental treatment as compared to a reference w.r.t. all involved time-to-event endpoints. One-sample testing procedures for single transition probabilities have been suggested in Andersen et al. (1993). Existing methodology to evaluate whole models as in Tattar and Vaman (2008) is designed to find differences in the matrix of transition probabilities rather than the underlying multivariate survival distribution. Additionally, aforementioned approaches assume a single analysis at the end of the trial. To fill this gap, we instead provide methodology for group-sequential one-sample tests for multivariate survival distributions derived from a multi-state model, while making provision for design modifications based on interim data from all involved survival endpoints simultaneously.

Our research is essentially motivated by applications in pediatric oncology. As lined out in Le-Rademacher et al. (2018), multi-state models can be especially useful in clinical oncology trials where patients can experience multiple paths between study enrolment and final outcome. On the one hand, we take up the controversy on the choice between progression-free survival (PFS)/event-free survival (EFS) and overall survival (OS) as primary endpoint in phase II oncology trials. For this purpose, we provide group-sequential one-sample tests for the joint distribution of PFS and OS in the context of an illness-death model. We apply this using an example from Fleischer et al. (2009) which was derived from clinical data. On the other hand, the simultaneous assessment of long-term efficacy (PFS) and long-term safety (time to life-threatening toxicity or death) appears likewise crucial in pediatric oncology, because late effects of treatment may persist lifelong. For this purpose we also provide a group-sequential one-sample test for the joint distribution of time to toxicity and time to progression while accounting for death as a competing event as motivated by our second clinical example, the LOGGIC Europe trial (Eudra-CT number 2018-000636-10). LOGGIC Europe is an international multicentre trial for children and adolescents with low-grade glioma. During the planning phase, the trial committee asked for a study design with the two primary time-to-event endpoints neurotoxicity-free survival and progression-free survival. Two interim analyses with the option of data-dependent sample size recalculation based on interim data from both endpoints were also requested.

In concerns of this request, we additionally address and solve a common problem from the theory of adaptive designs for survival endpoints. The fundamental statistical challenge in adaptive survival trials is handling of data from patients who enter the trial before an interim analysis and remain event-free beyond it. This implies the difficulty to define stage-wise independent p-values, as required by common adaptive design methodology. Two approaches have so far been proposed in the literature to obtain stage-wise p-values: (i) The independent increments approach, and (ii) the strategy of patient-wise separation. Presently available methodology for the first approach (Schäfer and Müller, 2001; Schmidt et al., 2017; Wasmmer, 2006) allows design modifications to be based on a single time-to-event endpoint, only, because otherwise type I error rate inflation can occur (Bauer & Posch, 2004). The strategy of patient-wise separation allows design modifications to be based on all available interim data (see Jenkins et al., 2011 or Irle & Schäfer, 2012). But it is involved with the drawback that either some available primary outcome data will be discarded or some worst case adjustment has to be performed resulting in a conservative significance test (Magirr et al., 2016). Here, we are aiming for a multivariate extension of the independent increments approach. We provide survival tests where provision is made for data dependent design modifications based on all involved time-to-event endpoints while avoiding the problems arising from the strategy of patient-wise separation. Basically this is achieved by compensating the multivariate counting process w.r.t. the filtration generated by the multivariate process itself, rather than the filtration generated by a single time-to-event endpoint as it is done commonly. We therefore call our strategy the multivariate independent increments approach.

The paper is organized as follows. To motivate our proposed testing procedure, we start with a presentation of its application in the prominent example of PFS and OS in Section 2. In the following sections we give the required theoretical foundations. In Section 3 we settle basic notation which we will use in Section 4 to construct suitable counting process martingales and their asymptotics. These in turn will be used to derive statistical testing procedures. Foundations of
confirmatory sequential and adaptive designs are provided in Section 5. This theoretical part is followed by a brief treat-
mant of several issues which arise in practical implementations of the procedure in Section 6. In particular, we also elabor-
orate tests for the other setting mentioned above. Small sample properties are studied by simulations in Section 7 and we
conduct a case study in Section 8 in order to give an example for an application of the adaptive procedure. We conclude
with a discussion of our findings and prospects for future research.

Proofs of the mathematical statements and additional supporting information are shifted to the Appendices.

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. OS is
defined as time from start of a therapy to death and PFS is defined as time from start of a therapy to progression or death,
whatever occurs first. While there is no clear conclusion about which of these endpoints to choose as the primary endpoint
in general, our methodology yields the possibility to use both of them as primary endpoints under exploitation of their
dependence structure. The first step of outcome improvement in oncology is strictly associated with an extension of the
progression free survival time and an increase of the plateau of patients without tumor progression. The second step is
the hard criterion 'death yes or no' without any interpretation scope left. A second line treatment may or may not be in
between both steps; however, the link between PFS and OS is always in focus of the investigators and the simultaneous
presentation an important task.

As PFS and OS may occur simultaneously if a patient dies before any progress of the disease has been observed, we can
use a multi-state model as in Meller et al. (2019) to establish a corresponding probabilistic structure. The multi-state 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.

We denote the random time of transition to node 1 by $T^{(1)}$ and the time of transition to node 2 by $T^{(2)}$. Accordingly,
PFS can be defined as the minimum of $T^{(1)}$ and $T^{(2)}$ which will be denoted by $T^{(1)} \wedge T^{(2)}$ and OS is given by $T^{(2)}$. The
probabilistic structure of the multi-state model is given by hazards, also called intensities, for the three transitions. To be
consistent with our later notation, we will denote them by:

$\lambda^{(1,(1,2))}$ the hazard for a progression
$\lambda^{(2,(1,2))}$ the hazard for death without prior progression (e.g., by treatment related toxicity)
$\lambda^{(2,(2))}$ the hazard for death with prior progression.

Note that the first part of the upper index contains the state to which the corresponding transition refers while the second
part contains the set of states that have not yet been visited. These hazards can be defined by the limit

$$\lambda^{(j,(1,2))}(s,s^{(1)},s^{(2)}) ds := \frac{P[s < T^{(j)} \leq s + ds | T^{(1)} > s^{(1)}, T^{(2)} > s^{(2)}]}{P[s < T^{(j)} | T^{(1)} > s^{(1)}, T^{(2)} > s^{(2)}]} \quad \text{for } j \in \{1, 2\}$$

respectively (resp.)

$$\lambda^{(2,(2))}(s,s^{(1)},s^{(2)}) ds := \frac{P[s < T^{(2)} \leq s + ds | T^{(1)} = s^{(1)}, T^{(2)} > s^{(2)}]}{P[s < T^{(2)} | T^{(1)} = s^{(1)}, T^{(2)} > s^{(2)}]}.$$
From now on, we will assume that we can continuously monitor progression and death of the patients. Hence, we will only need to consider arguments of the form \((s, s, s)\) for \(\lambda^{(1,1,2)}\) and \(\lambda^{(2,1,2)}\) resp. arguments of the form \((s, s^{(1)}, s)\) with \(s^{(1)} \leq s\), which refers to the time of progression, for \(\lambda^{(2,2)}\). Depending on the assumptions of the model, the hazards can be simplified further. For the easiest case, a time-homogeneous Markov model, each hazard assumes a constant value and hence does not depend on the current time since start of the treatment \(s\) or the timing of observed progressions. An example of such a model will be investigated in Section 8.

In the context of a one-armed clinical trial, we assume that we are given a reference model. The hazards of this reference model may have been estimated using historical data of patients which have been treated under the current standard of care. Estimations of reference hazards will be the content of Section 6.2. The hazards of this reference model will be marked with a 0 in the lower index. We want to explore possible differences between the reference model and the survival history of patients under a new treatment which is examined in this trial. Then the null hypothesis can thus be formulated by

\[
H_0 : \begin{pmatrix} \lambda^{(1,1,2)}_0 & \lambda^{(2,1,2)}_0 & \lambda^{(2,2)}_0 \end{pmatrix} = \begin{pmatrix} \lambda^{(1,1,2)} & 0 & 0 \\ 0 & \lambda^{(2,1,2)} & 0 \\ 0 & 0 & \lambda^{(2,2)} \end{pmatrix}.
\]

(3)

We are now given \(n\) patients entering the trial at calendar times \(0 \leq y_1 < y_2 < \ldots < y_n < \infty\). For the \(i\)-th patient, let \(T^{[1]}_i\) the time of progression resp. \(T^{[2]}_i\) the time of death after entry. The patient is censored at the random time \(C_i\), whereby \(C_i\) is assumed to be independent of \((T^{[1]}_i, T^{[2]}_i)\). An analysis taking place at calendar time \(t_1\) can also induce censoring. Hence, we can only observe \(T^{[j]}_i \wedge C_i \wedge (t - y_i)\) at calendar time \(t_1\) for any patient \(i \in \{1, \ldots, n\}\) and \(j \in \{1, 2\}\).

For each of these counts, we need counterparts in which the expected number of these events under the null hypothesis is computed. These are given by

\[
A^{\text{PFS,OS}}_0(t) = \sum_{i=1}^n \left( \int_0^{T^{[1]}_i \wedge T^{[2]}_i \wedge C_i \wedge (t - y_i) \wedge} \lambda^{(1,1,2)}_0(s, s, s) ds \right),
\]

\[
A^{\text{PFS}}_0(t) = A^{\text{PFS,OS}}_0(t) + \sum_{i=1}^n \left( \int_0^{T^{[1]}_i \wedge T^{[2]}_i \wedge C_i \wedge (t - y_i) \wedge} \lambda^{(1,1,2)}_0(s, s, s) ds \right),
\]

\[
A^{\text{OS}}_0(t) = A^{\text{PFS,OS}}_0(t) + \sum_{i=1}^n \left( \int_0^{T^{[2]}_i \wedge C_i \wedge (t - y_i) \wedge} \lambda^{(2,2)}_0(s, T^{[1]}_i \wedge C_i \wedge T^{[2]}_i, s) ds \right).
\]

We can summarize these quantities in the matrices

\[
N(t) = \begin{pmatrix} N^{\text{PFS}}(t) & N^{\text{PFS,OS}}(t) \\ N^{\text{PFS,OS}}(t) & N^{\text{OS}}(t) \end{pmatrix} \quad \text{resp.} \quad A_0(t) = \begin{pmatrix} A^{\text{PFS}}_0(t) & A^{\text{PFS,OS}}_0(t) \\ A^{\text{PFS,OS}}_0(t) & A^{\text{OS}}_0(t) \end{pmatrix}.
\]

(4)

Together, we can form the multivariate process \(M_0(t) := (M^{\text{PFS}}_0(t), M^{\text{OS}}_0(t))\) with

\[
M^{\text{PFS}}_0(t) := n^{-1/2} \left( N^{\text{PFS}}(t) - A^{\text{PFS}}_0(t) \right) \quad \text{resp.} \quad M^{\text{OS}}_0(t) := n^{-1/2} \left( N^{\text{OS}}(t) - A^{\text{OS}}_0(t) \right),
\]

(5)

which is a martingale under the null hypothesis \(H_0\). In a group-sequential setting, we are given a sequence of analysis dates. For the sake of simplicity, we restrict ourselves to a design with one interim analysis at calendar time \(t_1\) and a final
analysis at calendar time $t_2$ here. The first stage test statistic will be based on $\mathbf{M}_0(t_1)$ and the second stage test statistic will be based on the increment $\mathbf{M}_0(t_2) - \mathbf{M}_0(t_1)$. For both, we need to estimate the covariance matrix under the null hypothesis for both stages. These are given by

$$
\hat{\Delta}_1 = \frac{\mathbf{N}(t_1) + \mathbf{A}_0(t_1)}{2n} \quad \text{and} \quad \hat{\Delta}_2 = \frac{(\mathbf{N}(t_2) - \mathbf{N}(t_1)) + (\mathbf{A}_0(t_2) - \mathbf{A}_0(t_1))}{2n}
$$

for the first resp. for the second stage. Let $\hat{\mathbf{C}}_j \hat{\mathbf{C}}_j^\top$ denote the Cholesky decompositions of $\hat{\Delta}_j$ for $j \in \{1,2\}$. The test statistic $\mathbf{Z}_{0,1}$ for the first resp. $\mathbf{Z}_{0,2}$ for the second stage can now be computed by

$$
\mathbf{Z}_{0,1} = \hat{\mathbf{C}}_1^{-1} \mathbf{M}(t_1) \quad \text{resp.} \quad \mathbf{Z}_{0,2} = \hat{\mathbf{C}}_2^{-1} (\mathbf{M}(t_2) - \mathbf{M}(t_1)).
$$

To construct statistical tests with a given level $\alpha$, it is crucial to know that under the null hypothesis $\mathbf{Z}_{0,1}$ and $\mathbf{Z}_{0,2}$ are bivariate standard normally distributed and that they are independent. There are several possibilities to obtain stage-wise $p$-values from these test statistics. Possible choices are the $L^2$- or the $L^\infty$ norm of the stage-wise test statistic. This would yield stage-wise $p$-values

$$
p_l = 1 - F_{\chi^2_2}(|\mathbf{Z}_{0,l}|^2_2) \quad \text{resp.} \quad p_l = 1 - F_{\chi^2_1}(|\mathbf{Z}_{0,l}|^2_\infty)^2
$$

for $l \in \{1,2\}$. Of course, this choice has to be made at the planning stage of the study. We can now apply any kind of combination test and sequential decision boundary to conduct an adaptive level $\alpha$ test. Such a test is given by a rejection region

$$
\mathcal{A} := \{ p_1 \leq \alpha_1 \} \cup \{ \alpha_1 < p_1 \leq \alpha_0, p_2 < \alpha_2(p_1) \} \quad \text{for a prefixed level } \alpha_1 \text{ and futility bound } \alpha_0 \text{ of the first stage and a strictly decreasing conditional error function } \alpha_2 : (\alpha_1, \alpha_0] \rightarrow [0,1] \text{ with } \int_{\alpha_0}^{\alpha_1} \alpha_2(x) dx = \alpha - \alpha_1.
$$

At the interim analysis, the design of the trial (e.g., sample size) may be adapted based on observations concerning PFS- and OS-events observed until the interim analysis. These topics are covered in more detail in Section 5.

### 3 \ NOTATION FOR THE GENERAL FRAMEWORK

In this section we will formalise the framework of our approach and introduce all the components we need to construct the processes and test statistics. While the previous section focused on a particular example concerning two time-to-event endpoints (PFS and OS), the general methodology allows comparisons based on an arbitrary number of composite events in the presence of an absorbing state.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space upon which all random variables are defined. For a trial with $n \in \mathbb{N}$ patients we are given a sequence of entry times $0 \leq y_{n,1} < y_{n,2} < \ldots < y_{n,n} < \infty$. As $n \to \infty$, the accrual process which is defined by

$$
a_n(t) := \sum_{i=1}^{n} \mathbb{1}_{[y_{n,i} \leq t]} 
$$

for all $t \geq 0$ and $n \in \mathbb{N}$ asymptotically behaves according to a cumulative accrual rate $r$ with $r(0) = 0$ and $\lim_{t \to \infty} r(t) = 1$, that is

$$
\lim_{n \to \infty} \frac{a_n(t)}{n} = r(t) \quad \forall t \geq 0.
$$

The asymptotic distribution of our test statistics will be derived in the limit $n \to \infty$. Hence, we will fix $n$ from now on and omit the double indexing for the sake of simplicity. The $i$-th patient in this trial, which enters the trial at calendar time $y_i$ is associated with a $(k+1)$-tuple $(C_i, T_{[1]}^i, \ldots, T_{[k]}^i)$ of $[0,\infty]$-valued random variables. The random variable $C_i$ is the time after entry of censoring and $(T_{[1]}^i, \ldots, T_{[k]}^i)$ is a vector of survival times where each entry refers to a distinct event $j \in \{1, \ldots, k\}$.

Please note that we explicitly include the possibility of these variables to be infinitely large. Whereas $T_{[1]}^i, \ldots, T_{[k]}^i$ are explicitly allowed to be dependent, $C_i$ is independent of $T_i := (T_{[1]}^i, \ldots, T_{[k]}^i)$, for any $i \in \{1, \ldots, n\}$. Tuples $(C_i, T_i)$ and $(C_j, T_j)$ for $i \neq j$ are independent and identically distributed (i.i.d.)

We additionally introduce the set-valued processes

$$
\mathbf{U}_i : [0,\infty) \times \Omega \to \mathcal{P}([1, \ldots, k])
$$

$$
\mathbf{U}_i(s, \omega) := \{ l \in \{1, \ldots, k\} : T_{[l]}^i(\omega) > s \land C_i(\omega) \}
$$

(10)
monitoring the set of yet unobserved events and define the vector \((s, T_i)_U\) for a set \(U \subset \{1, \ldots, k\}\) and any \(s \geq 0\) by

\[
(s, T_i)_U := \left(s \cdot 1_{j \in U} + T_i^{[j]} \cdot 1_{j \not\in U}\right)_{1 \leq j \leq k}.
\]

Notice that \(U_i(s) := U_i(s, \cdot)\) and \((s \wedge C_i, T_i)_{U_i(s)}\) contain all information about the time-to-event endpoints of patient \(i\) available at \(s\) time units after his entry.

We can imagine the situation as a multi-state model with \(k + 1\) different states. For now, we restrict ourselves with exactly one absorbing state which in practice is usually death. Of course, scenarios with several absorbing states or without any absorbing state can be handled similarly. The first state, which we will call state 0, is the initial state. We have \(k - 1\) transient states 1 to \(k - 1\) which are first entered at times \(T_i^{[1]}\) to \(T_i^{[k-1]}\) and the absorbing state \(k\) which is entered at time \(T_i^{[k]}\). Recurrent events can be represented by different states. As we do not want to restrict ourselves to a particular joint modeling approach for PFS and OS, we need to keep track of the processes introduced above.

Of course, this illustration is only meaningful if we suppose that no two events occur simultaneously in finite time \(P\)-almost surely. This limitation will later be eliminated, as it is not very reasonable for many constellations, especially when we deal with composite endpoints.

The described reasoning should be reflected by the representation of the hazard rates for the jump processes. For a fixed event \(j\) the hazard for this event to occur might be different depending on which events have happened before. Hence, for a set of events \(U \subset \{1, \ldots, k\}\) which have not yet happened, the hazard function for event \(j\) will be denoted by \(\lambda(U, j)\). Such conditional hazards, that is, hazards for one event under a certain knowledge about other events can be computed under \(P\) by

\[
\lambda(U, j)(s, s^{[1]}, \ldots, s^{[k]}) ds = \frac{P[s < T_i^{[j]} \leq s + ds | T_i^{[l]} > s^{[l]} \forall l \in U, T_i^{[l]} = s^{[l]} \forall l \not\in U]}{P[s < T_i^{[j]} | T_i^{[l]} > s^{[l]} \forall l \in U, T_i^{[l]} = s^{[l]} \forall l \not\in U]}.
\]

Please note that (11) also enables the computation of conditional hazards for cases in which the assessment of some time-to-event variables may not be executable continuously, but only in certain time-steps. In the following we assume that we can assess all events continuously.

As each state is only visited once, we have

\[ j \not\in U \Rightarrow \lambda(U, j) \equiv 0 \text{ on the set where } s \geq s^{[j]} \]

and for the absorbing event \(k\) additionally holds

\[ k \not\in U \Rightarrow \lambda(U, k) \equiv 0 \forall j \in \{1, \ldots, k\} \text{ on the set where } s \geq s^{[k]} \] (12)

As other existing methods for one-sample log-rank tests (see, e.g., Schmidt et al., 2017), the asymptotics of our test statistics rely on a martingale central limit theorem. In order to construct such a martingale it is necessary to describe a filtration that comprises exactly the information that is available for any fixed time \(t\) of patient \(i\)'s time in trial. This includes information about whether events under consideration could have been observed until time \(s\) after entry or not and if yes, when or if there has already been a censoring of the patient's data. Hence, let the \(\sigma\)-algebra \(F_{i,s}\) be generated by the random variables

\[
1_{[T_i^{[j]} \leq \max_j T_i^{[k]}]} \cdot T_i^{[j]} \cdot 1_{[T_i^{[j]} \leq \max_j T_i^{[k]}]},
\]

\[
1_{[C_i \leq \max_j T_i^{[k]}]} \cdot C_i \cdot 1_{[C_i \leq \max_j T_i^{[k]}]},
\]

for all \(j \in \{1, \ldots, k\}\). When it comes to measurability issues later on, one should make oneself clear that \(U_i\) is progressively measurable w.r.t. \((F_{i,s})_{s \geq 0}\) for all \(i \in \{1, \ldots, n\}\) and that \((s \wedge C_i, T_i)_{U_i(s)}\) is measurable w.r.t. \(F_{i,s}\) for all \(i \in \{1, \ldots, n\}\) and \(s \geq 0\).

Now, as we fixed the framework for a single patient, we also want to take into account, that we have several patients in our trial with different entry times \(y_i\). Hence we have to translate all the information which is so far given in a
patient’s time in trial, to calendar time. The events and the censoring of patient \( i \) take place at \( T_{ij} + y_i \) resp. \( C_i + y_i \) in calendar time.

The \( \sigma \)-algebra \( F_t \) containing all information available at calendar time \( t \) is hence given by

\[
F_t = \sigma\left( \bigcup_{i=1}^n F_{i,(t-y_i)^+} \right)
\]

where \( a^+ = \max(0, a) \) denotes the positive part of a real number \( a \in \mathbb{R} \). The filtration \( (F_i)_{i \geq 0} \) now describes the information available in calendar time w.r.t. which we want to construct a martingale leading to our test statistics.

In clinical practice (as in Sections 2 and 6.1), we often deal with composite events, for example, PFS, defined as time from start of therapy until death or progression, whatever occurs first. We can include composite events in our framework by taking a set \( E \subset \{1, \ldots, k\} \) and defining the event \( T_E := \min_{j \in E} T_{ij} \) for all \( i \in \{1, \ldots, n\} \). The original component events \( 1, \ldots, k \) are still included in this notation by choosing \( E \) as a singleton.

Finally, we want to introduce the processes with which we can monitor the information we collected for certain events. This includes expected and observed numbers of the composite events themselves as well as for their simultaneous occurrence. We will need them later to standardize our test statistics. Firstly, there are the counting processes for any \( E_1, E_2 \subset \{1, \ldots, k\} \) which are for all \( t \geq 0 \) defined by

\[
N_{E_1}(t) := \sum_{i=1}^n N_{i,E_1}(t) \quad \text{and} \quad N_{E_1,E_2}(t) := \sum_{i=1}^n N_{i,E_1,E_2}(t)
\]

\[
N_{i,E_1,E_2}(t) := 1_{\left\{ T_{i,E_1} \leq (t-y_i)^+ \land C_i \land T_{i,E_2} = T_{ij} \right\}}
\]

Secondly, we want to keep track of the hazard that has been accumulated for any possible component or composite event until any time \( t \geq 0 \). Therefore we define for any \( E_1, E_2 \subset \{1, \ldots, k\} \)

\[
A_{E_1}(t) := \sum_{i=1}^n A_{i,E_1}(t) \quad \text{and} \quad A_{E_1,E_2}(t) := \sum_{i=1}^n A_{i,E_1,E_2}(t)
\]

\[
A_{i,E_1,E_2}(t) := \int_{[0,(t-y_i)^+ \land C_i \land T_{i,E_1} \land T_{i,E_2}]} \lambda^{(j,U(s))}(s, (s \land C_i \land T_{i,E_1} \land T_{i,E_2})_{U_i(s)}) ds
\]

where the set-valued process \( U_i \) is defined as in (10). These can be interpreted as the number of events expected given the information about all of the component events in the model.

## 4 CONSTRUCTION OF THE MULTIVARIATE MARTINGALE AND ITS ASYMPTOTICS

After we set up the framework, we now just want to give a quick overview of the theoretical results before we proceed to their applications in relevant scenarios. Therefore, we need to state two conditions regarding the regularity of the joint distribution:

(R1) The distribution of \( T \) on \( [0, \infty)^k \) is absolutely continuous w.r.t. the \( k \)-fold product measure of \( \lambda_{[0,\infty)} + \delta_\infty \) where \( \lambda_{[0,\infty)} \) is the Lebesgue measure on \( [0, \infty) \) and \( \delta_\infty \) is the Dirac measure at \( \infty \).

(R2) For all \( j \in \{1, \ldots, k\}, U \in \mathcal{P}(\{1, \ldots, k\}) \) with \( j \in U \) and \( T \in [0, \infty)^k \) there exists a diverging sequence \( \max_{i \in U} T_i =: s_{0} < s_1 < s_2 < \ldots \) such that (s.t.) the mapping \( (s', s) \mapsto \lambda^{(j,U)}(s', (s, T)_{U}) \) is uniformly continuous on \( D_1 := \{(s', s) : s_{i-1} \leq s' < s_i, s' \leq s < s_j \} \) for all \( i \in \mathbb{N} \).

Proofs of the statements in this chapter and a more careful mathematical treatment can be found in the Appendix.

For any fixed component event \( j \) and any patient \( i \) we can now construct a martingale which is adapted to the filtration \( (F_{i,s})_{s \geq 0} \). The only difference to the construction of the martingale for the standard one-sample log-rank test is, that it has to take all the information given by the filtration into account.
Lemma 1. Let a vector of survival times $T_i := (T_1^i, \ldots, T_k^i)$ with a joint distribution on $[0, \infty)^k$ fulfilling (R1) and (R2) and an independent censoring variable $C_i$ be given. Then for all $j \in \{1, \ldots, k\}$ the process $(M_i^{[j]}(s))_{s \geq 0}$ defined by

$$M_i^{[j]}(s) := 1_{\{T_i^{[j]} \leq s \wedge C_i \wedge T_i^{[k]}\}} - \int_{[0,s \wedge C_i \wedge T_i^{[j]} \wedge T_i^{[k]}]} \lambda^{(j,U_i(u))}(u, (u, T_i \wedge C_i \wedge T_i^{[k]}))U_i(u) \, du$$

for all $s \geq 0$ is an $(F_{i,3})_{s \geq 0}$-martingale for any $i \in \{1, \ldots, n\}$ where the set-valued process $U_i$ is defined as in (10).

This formula also holds if we are in a situation with an absorbing state. The quite technical assumption of the Lemma only ensures that there are no simultaneous component events and that all the marginal distributions are absolutely continuous. Of course, a very similar statement, which is an easy consequence of the former, holds for composite events.

Corollary 2. Let a vector of survival times $T_i$ and an independent censoring variable $C_i$ as in Lemma 1 be given. Then for all $E \subset \{1, \ldots, k\}$ the stopped process $(M_E^i(s))_{s \geq 0}$ defined by

$$M_E^i(s) := \sum_{j \in E} M_i^{[j]}(s \wedge \min_{j \in E} T_i^{[j]} \wedge C_i)$$

for all $s \geq 0$ is an $(F_{i,3})_{s \geq 0}$-martingale for any $i \in \{1, \ldots, n\}$.

As in the previous chapter we can collect all the information and translate it to calendar time in order to get a martingale for any event comprising the information from all individuals in the trial.

Theorem 3. Let $(C_i, T_i)$ as in Lemma 1 be given for all $i \in \{1, \ldots, n\}$. Then for all $E \subset \{1, \ldots, k\}$ the process $(M_E(t))_{t \geq 0}$ defined by

$$M_E(t) := n^{-\frac{1}{2}} \sum_{i=1}^n M_i^E((t - y_i^+) +) = n^{-\frac{1}{2}} (N^E(t) - A^E(t))$$

for all $t \geq 0$ is an $(F_{i,3})_{t \geq 0}$-martingale.

Before we state the asymptotics we fix $d \in \mathbb{N}$ composite events defined by $d$ sets $E_1, \ldots, E_d \subset \{1, \ldots, k\}$. To state the asymptotics for our multivariate martingale $(M(t))_{t \geq 0} := (M_{E_1}(t), \ldots, M_{E_d}(t))_{t \geq 0}$, we have to establish two assumptions which ensure that the processes behave reasonably in the limit $n \to \infty$. These assumptions refer to the expected number of observable events and simultaneously observable events. Therefore we define

$$\sigma_{E_a}(t) := \lim_{n \to \infty} \sum_{i=1}^n \left( P\left[ T_i^{E_a} \leq C_i \wedge (t - y_i^+) \wedge T_i^{[k]} \right] / a_n(t) \right) \text{ resp.}$$

$$\sigma_{E_aE_b}(t) := \lim_{n \to \infty} \sum_{i=1}^n \left( P\left[ T_i^{E_a \cup E_b} \leq C_i \wedge (t - y_i^+) \wedge T_i^{[k]}, T_i^{E_a} = T_i^{E_b} \right] / a_n(t) \right)$$

for any $a, b \in \{1, \ldots, d\}$ and the matrix-valued function $\Sigma : [0, \infty) \to \mathbb{R}^{d \times d}$ with

$$\Sigma(t) := (\sigma_{E_a}(t) \delta_{ab} + \sigma_{E_aE_b}(t)(1 - \delta_{ab}))_{a,b \in \{1, \ldots, d\}}.$$  

As stated in Section II.5.1 of Andersen et al. (1993) we require $r \cdot \Sigma$, with $r$ from (9), to be a continuous $d \times d$ positive semidefinite matrix-valued function with positive semidefinite increments. This is generally fulfilled for any reasonable recruitment and censoring mechanism.
As we can see in the following theorem, increments of the matrix-valued function \( r \cdot \Sigma \) need to be estimated consistently. Possible choices of consistent estimators will be presented afterwards in Equations (19) and (20).

**Theorem 4.** Let \((C_i, T_i)\) as in Lemma 1 for all \( i \in \mathbb{N} \) and a set of composite events \( \{E_1, \ldots, E_d\} \subset P(\{1, \ldots, k\}) \) be given. Let \( 0 := t_0 < t_1 < \ldots < t_m < \infty \) be a sequence of \( m \) analysis times. Let \( \hat{\Delta}_l \) be a consistent estimator for \( \Delta_l := r(t_l)\Sigma(t_l) - r(t_{l-1})\Sigma(t_{l-1}) \) with Cholesky decomposition \( \hat{\Delta}_l = C_l C_l^\top \) for any \( l \in \{1, \ldots, m\} \). Then

\[
Z_l := (Z_{l,i})_{i \in \{1, \ldots, m\}} \xrightarrow{D} \mathcal{N}(0, \mathbf{1}_{d \cdot m}),
\]

where \( \mathbf{1}_{d \cdot m} \) is the identity matrix of size \( d \cdot m \) and where

\[
Z_{l,i} := \hat{C}_l^{-1}(M_{E_a}(t_l) - M_{E_a}(t_{l-1}))_{a \in \{1, \ldots, d\}}^\top \quad \forall l \in \{1, \ldots, m\}.
\]

Finally, the choice of the consistent estimators for the increments of \( r \cdot \Sigma \) mentioned in Theorem 4 needs further consideration before we proceed to the practical use cases for our methodology. Theorem II.5.1 from Andersen et al. (1993) directly yields two choices. We can either choose the quadratic variation or the predictable quadratic variation process to estimate the variance of \( M \). More exactly, we have

\[
[M] = \frac{1}{n} \mathbf{N} := \frac{1}{n} (N_{E_i,E_j})_{i,j \in \{1, \ldots, d\}} \xrightarrow{P} r \cdot \Sigma
\]

\[
\langle M \rangle = \frac{1}{n} \mathbf{A} := \frac{1}{n} (A_{E_i,E_j})_{i,j \in \{1, \ldots, d\}} \xrightarrow{P} r \cdot \Sigma
\]

uniformly on compact subsets of \([0, \infty)\). It has been studied in Tu and Gross (1996), Sun et al. (2011) and Wu (2014) that neither the former nor the latter one are perfect solutions in the case of standard one-sample log-rank tests. In particular for small sample sizes the tails of the distribution of the test statistics differs very much from tails of a normal distribution. Unfortunately, the behavior at the tails is even more important for us because we deal with sequential and multivariate testing. Hence we will stick to the suggestion from Wu (2014). There it is suggested to use the mean of the two choices mentioned above to estimate the variances. Obviously, it also holds that

\[
\frac{\mathbf{N} + \mathbf{A}}{2n} \xrightarrow{P} r \cdot \Sigma
\]

uniformly on compact subsets of \([0, \infty)\). As we will see in Section 7 this yields promising results regarding the empirical distributions of the test statistics and the empirical type I error probability which we obtained from simulations.

5 | TWO-STEP GROUP SEQUENTIAL AND ADAPTIVE DESIGNS

For the sake of simplicity we will restrict ourselves to two-step designs with one interim and one final analysis, but with an arbitrary number \( d \in \{2, 3, \ldots\} \) of (composite) endpoints here and in the following sections.

Let \( P_0 \) denote the null distribution, that is, the distribution under which the distribution of \( T = (T^{E_1}, \ldots, T^{E_d}) \) for the study cohort is equal to the reference distribution. As in Section 2 we will indicate hazards and processes computed under \( P_0 \) with an additional 0 in the lower index. If also \( F_0 \) denotes the joint null distribution function of the survival times \( T^{E_1}, \ldots, T^{E_d} \) and \( F \) is their true distribution function, the null hypothesis of our testing problem could be formulated as

\[
H_0 : F = F_0.
\]

Alternatively, we can also formulate the null hypothesis in terms of the transition hazards as we did it in (3) for our motivating example. For this, let \( \lambda \) be the collection of transition hazards in the multi-state model and let \( \lambda_0 \) be their exact specification in the null hypothesis. Then the null hypothesis can also be written as

\[
H_0 : \lambda = \lambda_0.
\]
This formulation also seems more appropriate as it may be more natural to formulate planning hypotheses on the level of the particular transitions.

Note that if the reference distribution is given in terms of the joint distribution function or the joint density, the hazards can easily be computed from these functions. This may be the case for models as in Yuan and Yin (2009) where copulas are used. If such a transformation is necessary, see Appendix A.3 for the corresponding formulas.

It is obvious that the space of alternative hypotheses for this simple null hypothesis is very big. Hence, the results of the test have to be interpreted with utmost care.

As already mentioned, the processes \( M_{Ei}^0 \) are defined by replacing the (true) unknown quantities \( \lambda \) with the known quantities \( \lambda_0 \) everywhere, which makes it possible to derive \( M_{Ei}^0 \) from the observed data. A two-step test of \( H_0 \) is defined by a sequence \( 0 = t_0 < t_1 < t_2 \) of calendar times. At any analysis \( t \in \{1, 2\} \) the test statistic \( Z_{0,t} \) can be calculated as in (18) with a proper choice for the estimator of the covariance matrix. By Theorem 4 we have

\[
(Z_{0,1}, Z_{0,2}) \overset{D}{\to} \mathcal{N}(0, \mathbf{I}_4) \quad \text{under } H_0.
\]

This asymptotic behavior needs to be exploited to obtain a stage-wise testing procedure.

## 5.1 The rejection region

At each analysis \( t \) one can measure the distance from the origin to \( Z_{0,t} \) and calculate a corresponding stage-wise \( p \)-value. Convenient choices for this distance could be the Euclidean norm \( |Z_{0,t}|_2 \) or the maximum norm \( |Z_{0,t}|_\infty \). The corresponding distributions to calculate the \( p \)-value would be the distribution of the square root of a \( \chi^2_2 \)-distributed random variable resp. the distribution of the square root of the maximum of two independent \( \chi^2_1 \)-distributed random variables. In these cases we get

\[
p_1 = 1 - F_{\chi^2_2}(|Z_{0,1}|_2^2) \quad \text{resp.} \quad p_1 = 1 - F_{\chi^2_1}(|Z_{0,1}|_\infty^2).
\]

Derivations of these \( p \)-values can be found in Appendix A.2.

The obtained \( p \)-values can be used with any kind of combination test and decision boundaries to get a sequential testing method. For a comprehensive overview on these topics see, for example, Wassmer and Brannath (2016).

We do not consider stopping for futility here for the sake of simplicity. Therefore, it suffices to specify the values \( 0 < \alpha_1 < \alpha < 1 \) where \( \alpha \) is the overall significance level and \( \alpha_1 \) is the level of the first stage and a conditional error function \( \alpha_2 : (\alpha_1, 1] \to [0, 1] \) which is monotonically decreasing such that

\[
\int_{\alpha_1}^{1} \alpha_2(\alpha) d\alpha + \alpha_1 = \alpha.
\]

This leads to the rejection region

\[
\{p_1 \leq \alpha_1\} \cup \{p_1 > \alpha_1, p_2 \leq \alpha_2(p_1)\}.
\]

If one fixes a norm for use, the rejection region in terms of stage-wise \( p \)-values, which is a subset of \([0, 1]^2\) in our two-stage case, can be translated to a rejection region for the vector of test statistics using (24) resp. (25). This region which we will call \( R \) in the following is in turn a subset of \( \mathbb{R}^4 \). An example of a translation to these principles of sequential testing to rejection regions for our method can be seen in Section 8.

The procedure can be extended to more than two stages. Futility stopping may also be included.

## 5.2 The adaptive test

If the times of analyses \( t_1 \) and \( t_2 \) are deterministic, we have \( P_0[R] = \alpha \). Since \( Z_{0,1} \) and \( Z_{0,2} \) are asymptotically independent, this still holds true in the adaptive case when the calendar time of the final analysis is chosen depending on the interim test statistic \( Z_{0,1} \). The exact requirements are given by the following theorem the proof of which can be found in Appendix A.4.
Theorem 5. Let \( \mathbf{Z}_0 = (Z_{0,1}, Z_{0,2}) \) be defined as in (18) using hazards from \( \mathcal{P}_0 \) and let \( R \) be a rejection region with level \( \alpha \). Furthermore, let a minimum follow-up time \( f > 0 \) be given, which is the time between the end of the recruitment period and the final analysis. Then the asymptotical property \( \mathcal{P}_0 \big[ R \big] = \alpha \) remains true if the final analysis time \( t_2 \) is replaced by any \( \mathcal{F}_{t_1} \)-measurable random variable \( \tau_2 \) with \( \tau_2 \geq t_1 + f \).

This enables a data-dependent sample size recalculation. On the one hand the number of events will change with modification of the final analysis time \( t_2 \). On the other hand the accrual period will be shortened or extended.

It is also possible to include unplanned interim analyses according to the CRP principle of Müller and Schäfer, see Schäfer and Müller (2001). In this case, the computation of conditional rejection probabilities requires a bit more attention. This is due to the fact that the exact knowledge of the conditional distribution of \( \mathbf{M}(t_1) \) given the data collected until an unplanned interim analysis \( t^* \) under the null hypothesis requires knowledge about the covariation matrix \( \mathbf{V}(t_1) \). This conditional distribution is given by \( \mathcal{N}(\mathbf{M}(t^*), \mathbf{V}(t_1) - \mathbf{V}(t^*)) \). One can now try to determine \( \mathbf{V}(t_1) - \mathbf{V}(t^*) \) via simulation or direct computation under assumptions about random censoring and recruitment mechanisms. If the accrual duration is extended or the final analysis is postponed a more accurate computation can be accomplished at the originally planned analysis date \( t_1 \). In this case the final rejection boundaries can only be computed at the time of the final analysis.

6 | PRACTICAL CONSIDERATIONS

After the presentation of our theoretical results and some brief explanations on how to use these results to construct valid adaptive group-sequential tests, we want to address some further problems which may arise upon application of the procedure.

6.1 | Application to further cases

Before going into another example, we quickly summarize how our motivating example from Section 2 fits into the general framework developed in Section 4. Obviously, there are \( k = 2 \) component events (progression and death) and \( d = 2 \) composite endpoints. One of them is PFS, which represents \( E_1 = \{1, 2\} \) while the other is OS which represents \( E_2 = \{2\} \). Thus, we have \( \mathbf{M} = (M^{\text{PFS}}, M^{\text{OS}}) = (M^{E_1}, M^{E_2}) \) and the rest follows as described in Section 4.

If another situation with multiple time-to-event endpoints shall be investigated with the presented procedure it is necessary to evaluate if there is a non-zero probability of these events happening simultaneously. If this is the case (as for PFS and OS), the cause of this coincidence has to be considered and a set of component events needs to be established to build up a latent multi-state model on this new set of events. After that, Theorem 4 can be applied to develop a testing procedure to execute a test for the joint distribution of all original endpoints. This approach will now be demonstrated with another example which corresponds to a typical setting of a phase IIa trial, in which we want to assess safety and efficacy simultaneously:

Similar to the setting of Section 2, efforts have been undertaken to investigate the interplay between efficacy and toxicity while accounting for death as a competing event. This is very important in the early clinical stages of drug development. Our methodology can be seen as a tool to detect in which way this interplay differs between the standard of care represented by the reference distribution and the treatment under examination.

As in Lee et al. (2020) we assume that we can observe toxicities (\( T^{[1]} \)), progressions (\( T^{[2]} \)) and death (\( T^{[3]} \)). In the spirit of a competing risk analysis (see, e.g., chapter 3 of Beyersmann et al., 2011) we define two events characterized by the sets \( E_1 = \{1, 3\} \) for toxicity and \( E_2 = \{2, 3\} \) for efficacy. The former corresponds to the time from the start of a therapy to toxicity or death, whatever occurs first, the latter corresponds to PFS. This setting is visualized by Figure 2. As this model is slightly more complex than the model in Section 2 we need more hazards to compute the martingale \( \mathbf{M}_0 \). But similar to above, the hazards \( \lambda_0^{(1,[1,2,3])}, \lambda_0^{(2,[1,2,3])}, \lambda_0^{(3,[1,2,3])} \) (i.e., the hazards for the component events without any prior event) do not need increased attention. But once again, the hazards \( \lambda_0^{(2,[2,3])}, \lambda_0^{(3,[2,3])}, \lambda_0^{(1,[1,3])} \) and \( \lambda_0^{(3,[1,3])} \) may depend on the time of an already observed toxicity resp. progression. It could be helpful to recognize that not all of these hazards need to be given explicitly. It is enough to know the sums \( \lambda^{(2,[2,3])} + \lambda^{(3,[2,3])} \) and \( \lambda^{(1,[1,3])} + \lambda^{(3,[1,3])} \) in order to compute \( \mathbf{M} \) and the test statistics \( \mathbf{Z} \). This becomes clearer if one takes a closer look at the test statistics for this special case, given in Appendix A.1, where the whole procedure for this application scenario is lined out, analogously to the presentation of the main example.
FIGURE 2  Representation of the simultaneous assessment of efficacy and toxicity as a multi-state model

in Section 2. Please note, that we do not need any hazard $\lambda^{(j, A)}$ where $A$ is a singleton in this case as both composite events will have happened by then.

6.2  Derivation of reference distributions

Of course, the creation of an adequate reference model is only possible if corresponding historical data is available. This may, for example, be the case in pediatric oncology thanks to the outstanding role of registries as the German Childhood Cancer Registry in which about 95% of cases in pediatric oncology in Germany are registered. Furthermore, about 90% of these registered patients are treated within a clinical trial (see Kaatsch, 2004 or Erdmann et al., 2020). Thus, the sensible use of such a database can prevent from selection bias and neglect of temporal trends.

If appropriate historical data is given, the choice of the modeling approach is also crucial. As already mentioned, multi-state models appear suitable. The most prominent multi-state modeling approaches for our main example, which will also be employed in our simulation study in Section 7, are summarized in Meller et al. (2019). A general treatise of multi-state modeling approaches can be found in chapter 6 of Hougaard (2000). In some cases it may be justifiable to assume that the resulting process is a (time-homogeneous) Markov Process as in Fleischer et al. (2009). In general, this might not be the case. In Titman and Sharples (2010) advice on how to choose the appropriate approach is provided.

Beyond that, the application of our procedure is not restricted to multi-state models. Also frailty- or copula-based approaches as presented, for example, in chapter 7 of Hougaard (2000) or Burzykowski et al. (2001) for the PFS/OS-example resp. in Yuan and Yin (2009) for the efficacy/safety-example. In these cases the joint distribution function of both endpoints is given. The quantities needed to execute our testing procedure can be calculated with the formulas provided in Appendix A.3.

Most of the publications cited here provide inference procedures for the transition hazards resp. the joint distribution function. As an additional support for users we provide functions for parameter estimations for the approaches applied in our simulation study in the supplementary R code.

The testing procedure itself relies on a correct specification of the reference distribution and does not account for sampling variability arising from estimates derived from historical data.

6.3  One-sided hypotheses

The methodology provided so far yields testing procedures for the null hypothesis as stated in (22) resp. (23). As any kind of deviation from this hypothesis can be detected here, it is commonly called a two-sided hypothesis. Nevertheless, the test underlying statistic $Z_0$ itself is multivariate and the rejection region for this statistic is therefore a subset of $\mathbb{R}^d$. If interest lies in showing superiority over the reference, a corresponding counterpart which qualifies as a one-sided hypothesis would then be given by

$$H_0 : \lambda \geq \lambda_0.$$  \hspace{1cm} (26)

where the inequality is to be understood componentwise. Obviously, a rejection of such a hypothesis goes along with the assumption that the value of at least one transition hazard is below the corresponding value of the reference model. Rejection regions should be given in such a way that the type I error levels monotonically decrease for any increase beyond
Given this monotonicity, it would be enough to show that the nominal type I error level is adhered to if equality holds in (26).

For this sake, we propose rejection regions as in Section 5.1 with the additional restriction that all entries of the multivariate martingale \( \mathbf{M} \) shall be negative. It is worth mentioning that this procedure is not a trivial extension of the preceding methodology. In particular it is necessary to compute the probabilities

\[
P_{\lambda_0}[M^E_i(t_l) < 0 \; \forall i \in \{1, \ldots, d\}]
\]

for any analysis date \( t_l \) which is only possible at the time of analysis itself, when the estimate of the covariance matrix of \( \mathbf{M} \) is identifiable.

For the sake of simplicity, we restrict ourselves to the two-sided hypothesis tests in the remainder of this article.

### 6.4 Power and sample size calculations

Clinically relevant differences of a new treatment compared with the reference distribution can be set separately for the different transition hazards. This way it is possible to depict different mechanisms of action. As in our simulation study in Section 7 or in the case study in Section 8, these differences can be quantified by separate hazard ratios for the transition hazards. This can be especially useful if the mechanism of action of a new treatment differs from previous ones which can be the case for new cancer immunotherapies as it is outlined in Hoos (2012).

If the planning alternative has been set, power for different sample sizes can be determined via simulation. The provided R code enables power simulations with several sample sizes, s.t. one can determine the sample size which is necessary to achieve a specified power under the planning alternative.

### 6.5 Interpretation and further analyses

As multiple time-to-event endpoints are evaluated simultaneously with the method presented here, we also obtain a multivariate test statistic (see Section 4). Thus, the interpretation of the results requires a bit more attention and shall be described here for the two examples presented so far.

In general, interpretation is easier on the level of the martingale \( \mathbf{M}_0 \), as defined in (15). The value of one entry of the vector \( \mathbf{M}_0 \), which is associated to an event \( E \), is the difference between the number of events of type \( E \) observed and expected given the history concerning component events outside of \( E \). As there are no component events outside PFS in the PFS/OS-example, the first component is just the difference between PFS-events observed and expected under the null hypothesis. The second component (OS) does not include the component event of progressions and is hence the difference of observed deaths and deaths expected given the progression history of all patients. Obviously, differences between the study cohort and the reference model concerning the hazard for transition to progression only affect the first entry, while differences in the transition from progression to death only affect the second component. This is also visualized in our case study in Section 8. Differences concerning the hazard for transition from the initial state to death affect both entries.

The same interpretation holds for the more complex example concerning efficacy and safety. The toxicity-component compares toxicity events of patients in the trial to their expected number under the null hypothesis given their progression history and the efficacy-component compares progressions of patients in the trial to their expected number given their history concerning toxicities. In both components, we additionally account for death as a competing event. Hence, any deviation from the null distribution concerning one of the hazards \( \lambda^{(2,1,2,3)} \), \( \lambda^{(1,1,3)} \) and \( \lambda^{(3,1,3)} \) affects the safety-component of \( \mathbf{M}_0 \), while any deviation from the null distribution concerning one of the hazards \( \lambda^{(2,1,2,3)} \), \( \lambda^{(2,2,3)} \) and \( \lambda^{(3,2,3)} \) affects the efficacy-component. Analogously to the PFS/OS-example, the transition \( \lambda^{(1,1,2,3)} \) affects both components.

Hence, interpretation requires a basic understanding of conditional probability. In both cases there is one transition emerging in both components, which is the direct transition to death. This is because we are accounting for death as a competing event. Unfortunately, this causes a correlation of the components of \( \mathbf{M}_0 \). The standardization procedure yields asymptotically independent components which we need to construct valid statistical tests.

While the examination of the entries of \( \mathbf{M}_0 \) reveals which of the endpoints under investigation is affected by differences between a novel therapy and the control group, one can also go further into detail to examine particular transitions.
Therefore, one can compare Nelson–Aalen estimates for the transitions hazards with the reference transitions, see Section IV.4 of Andersen et al. (1993) for more details.

7 | SIMULATION STUDY

The aim of this simulation study is twofold. On the one hand, we conduct it to demonstrate that the asymptotics from Theorem 4 already take effect for practically relevant sample sizes and assess empirical type I errors for our method in relevant scenarios. As already mentioned, the latter is an important issue as the missing concordance of the distribution of standard one-sample log-rank statistics and standard normal distribution at the tails as treated, for example, in Sun et al. (2011) may affect type I errors. On the other hand, we want to demonstrate how sample size and power calculations can be accomplished for a variety of alternative hypotheses.

7.1 | Design

We consider different scenarios for the PFS/OS main application example from Section 2 with \( m = 2 \) analyses (i.e., one interim analysis and a final analysis).

We use the scenarios from Meller et al. (2019) in which several commonly used joint modeling approaches are represented. The scenarios comprise processes for the latent progression-death model from Figure 1. More precisely, we have

\[
\begin{align*}
\lambda_{01}(s, s, s) &= \lambda_{01} \cdot s^{\gamma_{01}-1} \\
\lambda_{02}(s, s, s) &= \lambda_{02} \cdot s^{\gamma_{02}-1} \\
\lambda_{12}(s, s(1), s) &= \lambda_{12} \cdot s^{\gamma_{12}-1} \\
\lambda_{22}(s, s(1), s) &= \lambda_{12} \cdot (s - s(1))^{\gamma_{12}-1}
\end{align*}
\]

for scenarios 1-5,

for scenarios 1-5,

for scenarios 1, 4 and 5 resp.

for scenarios 2 and 3

with parameter values as in Table 1. Hence, we have a time-homogeneous Markov process in scenario 1, time-inhomogeneous Markov processes in scenarios 4 and 5 and a semi-Markov process in scenarios 2 and 3, where the hazard does not depend on the current time, but only the current duration of stay in the state (see, e.g., chapter 6 of Hougaard, 2000).

As the processes introduced and studied in Section 3 and Section 4 are given in calendar time, corresponding trials will be planned in calendar time, too. Please note, that unlike the classical one-sample log-rank test (see Breslow, 1975 or Schmidt et al., 2017), information time is not a univariate process, but a \( 2 \times 2 \) symmetric matrix-valued (resp. \( d \times d \) symmetric matrix-valued in the general case) process. The diagonal entries of these matrices refer to the expected PFS- resp. OS-events under the null hypothesis while the off-diagonal entry contains the expected number of joint PFS- and OS events, that is, direct deaths. The calendar time parameters were fixed as follows: accrual period \( a = 3 \) with uniform accrual rate, fixed follow-up time \( f = 2 \) and analyses at \( t_1 = 2.5 \) and \( t_2 = 5 \). Hence, the interim analysis takes place after half of the total time of the trial. The expected event rates of PFS and OS under the given parameters are also shown in Table 1 (denoted by \( \pi_{\text{PFS}} \) resp. \( \pi_{\text{OS}} \)). Given the sample size and assuming a uniform recruitment mechanism one can compute the diagonal entries of the expected information time matrix at the analysis dates with these values.

In the first part of the study, we investigated the behavior of \( Z_0 \) under the null hypothesis (22) to verify that the asymptotics from Theorem 4 already hold for small sample sizes. For each parameter constellation in combination with each of the sample sizes \( n \in \{50, 100, 250, 500\} \), we executed 100,000 simulation runs. In the second part, we
TABLE 2  Empirical type I errors for different sample sizes, sequential designs and scenarios

| n   | L² | Type | Scenarios  | 1   | 2   | 3   | 4   | 5   |
|-----|----|------|------------|-----|-----|-----|-----|-----|
| 50  | L² | P    | 0.0502     | 0.0493 | 0.0486 | 0.0491 | 0.0488 |
|     |    | OF   | 0.0515     | 0.0497 | 0.0496 | 0.0500 | 0.0499 |
| L∞  | P  | 0.0499 | 0.0489 | 0.0480 | 0.0486 | 0.0483 |
|     | OF | 0.0510 | 0.0494 | 0.0485 | 0.0498 | 0.0500 |
| 100 | L² | P    | 0.0483     | 0.0495 | 0.0489 | 0.0486 | 0.0490 |
|     |    | OF   | 0.0503     | 0.0499 | 0.0505 | 0.0499 | 0.0502 |
| L∞  | P  | 0.0491 | 0.0490 | 0.0494 | 0.0491 | 0.0487 |
|     | OF | 0.0496 | 0.0488 | 0.0502 | 0.0500 | 0.0508 |
| 250 | L² | P    | 0.0486     | 0.0485 | 0.0481 | 0.0483 | 0.0479 |
|     |    | OF   | 0.0498     | 0.0496 | 0.0490 | 0.0505 | 0.0491 |
| L∞  | P  | 0.0488 | 0.0481 | 0.0470 | 0.0486 | 0.0480 |
|     | OF | 0.0502 | 0.0496 | 0.0492 | 0.0508 | 0.0494 |
| 500 | L² | P    | 0.0493     | 0.0485 | 0.0484 | 0.0482 | 0.0492 |
|     |    | OF   | 0.0496     | 0.0506 | 0.0485 | 0.0497 | 0.0505 |
| L∞  | P  | 0.0495 | 0.0491 | 0.0488 | 0.0484 | 0.0483 |
|     | OF | 0.0501 | 0.0499 | 0.0489 | 0.0501 | 0.0499 |

determined sample sizes to reach a power of 80% for a variety of alternative hypotheses. We considered 9 different alternative hypotheses where we assumed that \( \lambda^{(2,\{1,2\})} \) is the same under the null and the alternative hypotheses and that we have proportional hazards for the transition hazards \( \lambda^{(1,\{1,2\})} \) and \( \lambda^{(2,\{2\})} \). With proportionality constants \( \delta_{01} \in \{1.3, 1.5, 1.7\} \) resp. \( \delta_{12} \in \{1.15, 1.25, 1.35\} \) we can compute the parameters under the alternative hypothesis by \( \delta_{01} \cdot \lambda_{1,01} = \lambda_{0,01} \) and \( \delta_{12} \cdot \lambda_{1,12} = \lambda_{0,12} \). The remaining parameters are the same under the null and alternative hypotheses. For each combination of hazard ratios, 10,000 simulation runs were executed.

The trials were designed for a type I error level of \( \alpha = 5\% \). We used inverse normal designs with equal weights and with Pocock and O’Brien-Fleming boundaries (abbreviated by P resp. OF in Table 2). We chose these designs as we consider them as well-known and -established in the wide range of group sequential and adaptive designs. The weights of the inverse normal combination could still be optimized for each scenario and alternative hypothesis separately. For the sake of simplicity, we restrict ourselves to the simple choice of equal weights here. The null hypothesis (22) was then tested using stage-wise \( p \)-values derived from the \( L^2 \)-norm or the \( L^\infty \)-norm of the stage-wise test statistics computed as in (18).

Hence, the tests can be interpreted as "two-sided". A brief discussion of testing procedures resembling one-sided tests can be found in Section 6.3.

The simulation study was performed with R 3.6.2 (see R Core Team, 2017).

7.2  Behavior under the null hypothesis

Please note that each set of four values in Table 2 concerning the same scenario and maximum sample size belong to the same set of simulated trials, and only the computation of \( p \)-values and decision boundaries was altered. For interpretation of the obtained values, notice that the 95%-confidence interval for an underlying true value of 0.05 has a breadth of about 0.0027 for 100,000 simulation runs.

The results stem from computations employing the standardization from (21). Using only the (predictable) quadratic (co)variation also yields asymptotically correct testing procedures. Anyhow, we can see here why we used the extension from the idea from Wu (2014) as the empirical type I errors for a sample size of 50 are in a range between 0.054 and 0.059 for the predictable quadratic (co)variation resp. 0.059 and 0.070 for the quadratic (co)variation. Additionally, we want to mention that the empirical distribution for the former standardization is right-skewed and the latter is left-skewed. Therefore, standardizations based on (19) and (20) were not further used here. Of course, the difference between empirical error and nominal error level shrink with increasing sample size. But as we can see in Table 2, the chosen approach based...
FIGURE 3 Empirical distribution of $Z_{0,1}^{E_1}$ (the first component of $Z_{0,1}$) and its joint empirical distribution with other entries of the vectors of test statistics for Scenario 1 with a sample size of 50.

on the standardization from (21) works quite well for sample sizes as small as 50. Furthermore it is recognisable that the empirical error levels differ slightly between the scenarios but are in an acceptable range.

Additionally, we take a closer look at scenario 1 with a sample size of 50. Exemplarily, we show the histogram of results for the first entry of the vector $Z_{0,1}$ and scatterplots illustrating its joint distribution with the other entries of the vectors of test statistics in Figure 3. Concordance with the standard normal distribution and asymptotical independence already seem to take effect at this sample size. This is confirmed by the empirical Pearson correlation coefficients between the different entries of the vector of test statistics. They range between $-0.0036$ and $0.0007$ and hence lie in the 95% confidence interval for a true correlation of zero. And also the tails of the marginal distributions agree to a great extent with the tails of the normal distribution, except from a very light left-skew.

7.3 Sample size and power

As mentioned in Section 6.4, the power of the procedure under alternatives can be determined via simulation. The supplementary R code provides simulation of trials with an arbitrary number of interim analyses for a set of sample sizes. The smallest sample size achieving the prespecified power can be determined this way to enable a sample size calculation.
### TABLE 3
Empirically determined sample sizes for different sequential designs, scenarios and alternative hypotheses necessary to reach a power of 80%

| \( \delta_{01} \) | \( \delta_{12} \) | Type | Scenarios |
|-----------------|-----------------|---------|-----------|
| \( L^2 \)       |                 |         | 1 | 2 | 3 | 4 | 5 |
| 1.3             |                 | P       | 246 | 245 | 220 | 215 | 196 |
| 1.15            |                 | OF      | 225 | 229 | 196 | 201 | 180 |
| 1.15            | \( L^\infty \)  | P       | 253 | 252 | 231 | 220 | 207 |
| 1.25            | \( L^\infty \)  | OF      | 231 | 236 | 206 | 206 | 189 |
| \( L^2 \)       |                 | P       | 202 | 212 | 173 | 187 | 160 |
| 1.3             |                 | OF      | 184 | 193 | 157 | 171 | 150 |
| 1.25            | \( L^\infty \)  | P       | 218 | 228 | 194 | 198 | 176 |
| 1.3             |                 | OF      | 197 | 204 | 173 | 181 | 157 |
| \( L^2 \)       |                 | P       | 161 | 176 | 139 | 153 | 129 |
| 1.3             |                 | OF      | 148 | 159 | 126 | 140 | 115 |
| 1.35            | \( L^\infty \)  | P       | 179 | 196 | 160 | 170 | 145 |
| 1.15            | \( L^\infty \)  | OF      | 162 | 175 | 141 | 152 | 129 |
| \( L^2 \)       |                 | P       | 111 | 106 | 98  | 94  | 89  |
| 1.5             |                 | OF      | 103 | 99  | 88  | 88  | 81  |
| 1.35            | \( L^\infty \)  | P       | 111 | 106 | 98  | 94  | 89  |
| 1.15            | \( L^\infty \)  | OF      | 104 | 100 | 88  | 89  | 81  |
| \( L^2 \)       |                 | P       | 105 | 101 | 89  | 87  | 80  |
| 1.5             |                 | OF      | 97  | 93  | 79  | 82  | 72  |
| 1.25            | \( L^\infty \)  | P       | 108 | 103 | 95  | 89  | 84  |
| 1.15            | \( L^\infty \)  | OF      | 98  | 95  | 84  | 83  | 76  |
| \( L^2 \)       |                 | P       | 94  | 97  | 81  | 84  | 74  |
| 1.5             |                 | OF      | 86  | 88  | 73  | 77  | 67  |
| 1.35            | \( L^\infty \)  | P       | 100 | 101 | 89  | 87  | 80  |
| 1.15            | \( L^\infty \)  | OF      | 91  | 92  | 79  | 81  | 71  |
| \( L^2 \)       |                 | P       | 68  | 63  | 59  | 56  | 53  |
| 1.7             |                 | OF      | 63  | 58  | 54  | 53  | 49  |
| 1.15            | \( L^\infty \)  | P       | 68  | 64  | 59  | 56  | 53  |
| 1.7             | \( L^\infty \)  | OF      | 63  | 59  | 54  | 53  | 49  |
| \( L^2 \)       |                 | P       | 66  | 62  | 56  | 54  | 50  |
| 1.7             |                 | OF      | 61  | 58  | 49  | 51  | 46  |
| 1.25            | \( L^\infty \)  | P       | 67  | 63  | 57  | 54  | 51  |
| 1.35            | \( L^\infty \)  | OF      | 62  | 58  | 51  | 51  | 47  |
| \( L^2 \)       |                 | P       | 62  | 60  | 52  | 53  | 47  |
| 1.7             |                 | OF      | 57  | 56  | 47  | 49  | 43  |
| 1.35            | \( L^\infty \)  | P       | 64  | 62  | 55  | 54  | 50  |
| 1.7             |                 | OF      | 59  | 58  | 49  | 50  | 45  |

The results can be found in Table 3. As in general, the maximum sample size is lower for the O’Brien-Fleming boundaries than for the Pocock boundaries. One can also observe that the maximum sample size for procedures employing the \( L^2 \)-norm of the sequential \( Z \)-statistics is always lower than it would be under application of the \( L^\infty \)-norm. Furthermore, one can see that this difference nearly vanishes if one of the hazard ratios is distinctly higher than the other one (see, e.g., results for \( \delta_{01} = 1.7 \) and \( \delta_{12} = 1.15 \)). While the sample size for all approaches are in a large part driven by the size of \( \lambda_{01} \), this effect is even more distinct for the \( L^\infty \)-norm, whose sample size is hence more stable under small changes in \( \lambda_{12} \). In contrast, the application of the \( L^2 \)-norm has clear advantages if there are effects of a similar magnitude for both components.
While the simulation study shows how our procedure enables the evaluation of PFS and OS as co-primary endpoints, we now also want to focus on the resulting advantage concerning interim analyses in such a study. With our new methodology, it is allowed to perform data-dependent sample size recalculation based on all involved time-to-event endpoints. This is not covered by previous methodology.

Here, we try to illustrate an adaptive testing strategy using a specific multi-state progression-death model from Fleischer et al. (2009) for patients with non-small cell lung cancer. The joint distribution of PFS and OS is given by a time-homogeneous Markov model, that is, the form of the transition hazards corresponds to those given for scenario 1 from the previous section (see (27)). The parameters were estimated using data from a two-armed study. The arms were summarized due to their similarity. The estimates are given by \( \lambda_{01} = 0.284, \lambda_{02} = 0.075 \) and \( \lambda_{12} = 0.128 \). As the model is time-homogeneous, we have \( \gamma_{01} = \gamma_{02} = \gamma_{12} = 1 \). Please note that the time scale for this model is given in months.

We assume that this model properly represents the course of disease under the standard of care. In this case, the model resp. the joint distribution of PFS and OS resulting from this model can be employed as a reference distribution in a fictional trial to investigate the effectiveness of a new treatment. The calendar time parameters of such a trial shall be fixed as follows: accrual period \( a = 24 \) with uniform accrual rate, fixed follow-up time \( f = 12 \) and analyses at \( t_1 = 18 \) and \( t_2 = 36 \). The expected PFS and OS rates under the null hypothesis with this design are given by 63.4% resp. 41.6% at \( t_1 \) and 99.8% resp. 91.8% at \( t_2 \). In the study, the testing procedure presented here shall be applied to detect differences between the distribution of PFS and OS under the new treatment and the standard of care. As in the previous section the trial is designed for a type I error level of \( \alpha = 5\% \) and stage-wise \( p \)-values will be combined via the inverse normal combination function with equal weights. We will apply "two-sided" tests in terms of what we lined out in Section 6.3. As it turned out to be the most powerful method among those which were investigated in the simulation study in Section 7 we will determine stage-wise \( p \)-values via the \( L^2 \)-norm of the stage-wise \( J \)-statistics (see (24)) and use the sequential O'Brien-Fleming boundaries.

As in the simulation study, the group-sequential trial is planned under proportional hazard assumptions for the three transitions with hazard ratios \( \delta_{01} = 1.5, \delta_{02} = 1 \) and \( \delta_{12} = 1.25 \). In order to reach a power of 80%, a sample size of 110 patients is needed. Obviously, if at least one of the hazard ratios is misspecified, the actual power may differ from 80%. Thus, we want to apply Theorem 5 and use the complete information about PFS- and OS-events available at calendar time \( t_1 \) to reschedule the remaining trial. The adaption rule at the interim analysis is inspired by the rule applied in the NB2004-HR trial (NCT number 03042429, see Berthold et al., 2020) and is as follows: At the interim analysis the length of the accrual period is recalculated. This period will always be followed by the follow-up period of 6 months. Recruitment will be continued for a minimum of 3 and a maximum of 30 months which corresponds to a doubling of the originally planned duration. We consider 10 different options in between these boundary scenarios by increasing the period in steps of 3 months. For each of these options, the conditional power based on the interim estimate for the three (constant) transition hazards is computed (as in Section 7.4.2 of Wassmer & Brannath, 2016). The interim estimates for the three transitions are computed according to the maximum likelihood procedure given in Fleischer et al. (2009) which is also implemented in the supplementary R code. As before, the conditional power is evaluated via simulation and takes into account the complete information about progressions of active patients available at the time of the interim analysis. If at least one of the 8 options reaches a conditional power of 80%, we choose the shortest continuation among these options. If this is not the case, but the continuation by 30 months leads to a conditional power exceeding 50%, we choose this option. If this is also not the case, we choose the shortest option, which implies a continuation of the accrual process by 3 months.

We simulated scenarios with any of the combinations of the following ”true” hazard ratios:

\[
\delta_{01} \in \{1.3, 1.4, 1.5, 1.6, 1.7\} \text{ and } \delta_{12} \in \{1.15, 1.2, 1.25, 1.3, 1.35\}.
\]

Afterwards, we compared the outcomes of group-sequential and adaptive trials. For each combination, 10,000 simulation runs of the group-sequential and the adaptive trial were executed. A simulation of the adaptive design is computationally intensive. For any simulated trial in which the null hypothesis is not rejected at the interim analysis, the conditional power of the 10 different rescheduling options needs to be determined via simulation as well. For these ”inner” simulations, 100 runs were performed.

In Figure 4 one can see the impact of the hazard ratios \( \delta_{01} \) and \( \delta_{12} \) on the stage-wise increments of the bivariate martingale \( M \). These increments are displayed for four different settings (combinations of \( \delta_{01} \in \{1, 1.7\} \) and \( \delta_{12} \in \{1, 1.35\} \) with
FIGURE 4  Empirical distribution of the bivariate martingale $M$ under different alternative hypotheses in a group-sequential test of the joint distribution determined by Fleischer et al. (2009) with sample size and design parameters as lined out in Section 8
TABLE 4  Comparison of empirically attained power of group-sequential and adaptive designs for several deviations from the planning hypothesis, originally planned sample size is set for hazard ratios \( \delta_{01} = 1.5 \) and \( \delta_{12} = 1.25 \). Upper value in each cell refers to power of group-sequential design, lower value refers to adaptive design. Numbers in brackets show average duration of accrual in the respective scenario (in months).

| \( \delta_{01} \) | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 |
|-------------------|-----|-----|-----|-----|-----|
| \( \delta_{12} \) |     |     |     |     |     |
| 1.15              | 0.443 (23.25) | 0.635 (22.61) | 0.781 (21.79) | 0.898 (20.89) | 0.948 (20.12) |
|                   | 0.476 (25.90) | 0.663 (24.42) | 0.809 (22.69) | 0.892 (21.27) | 0.942 (20.09) |
| 1.2               | 0.463 (23.27) | 0.642 (22.67) | 0.796 (21.80) | 0.901 (20.93) | 0.953 (20.13) |
|                   | 0.523 (25.80) | 0.686 (24.22) | 0.807 (22.81) | 0.892 (21.37) | 0.946 (20.10) |
| 1.25              | 0.511 (23.25) | 0.674 (22.60) | 0.816 (21.83) | 0.904 (20.98) | 0.956 (20.21) |
|                   | 0.558 (25.45) | 0.708 (24.31) | 0.831 (22.59) | 0.902 (21.13) | 0.945 (20.24) |
| 1.3               | 0.576 (23.19) | 0.705 (22.63) | 0.827 (21.76) | 0.911 (20.97) | 0.955 (20.18) |
|                   | 0.612 (25.17) | 0.743 (23.93) | 0.834 (22.50) | 0.911 (21.15) | 0.952 (20.11) |
| 1.35              | 0.636 (23.12) | 0.750 (22.52) | 0.849 (21.77) | 0.920 (20.94) | 0.967 (20.23) |
|                   | 0.666 (24.88) | 0.773 (23.52) | 0.864 (22.13) | 0.920 (21.14) | 0.954 (20.09) |

the first-stage values in the left and the second-stage values in the right column. As described in Section 6.5, an increase of \( \delta_{01} \) leads to a shift (to the left) of the first component and an increase of \( \delta_{12} \) leads to a (downward) shift of the second component. Furthermore, it can be observed that the effect of a decreased risk of progression becomes obvious rather in the first stage. On the contrary, a decreased risk of death after progression is apparent rather in the second stage. This becomes clear as many patients experience a PFS-event during the first stage while there are many patients which are in the progredient state during the second stage. Additionally, for the first stage results, we show how the \( L^2 \) - and \( L^\infty \) - rejection regions for the \( Z \) -statistics translate to the level of the martingale increments if the true covariance matrix (i.e., the matrix \( \mathbf{\Delta}_1 \) from Theorem 4) would have been known. In practice, this is not the case and this matrix has to be estimated as explained in Section 4. For the second stage, such a visualization is not possible as the rejection in the second stage depends from the outcomes in both stages.

Finally, we compared the empirical power of the two procedures. The results can be found in Table 4. The breadth of the 95%-confidence intervals for an underlying true value of 0.8 amounts to about 0.016. While the adaptive procedure assumes the preplanned power if the underlying truth has been met by the planning hypothesis, its power does not increase more than the power of the group-sequential design if one of the hazard ratios is underestimated. Additionally, the mean duration of the accrual period and hence also the sample size are similar. On the other hand, some of the power which is lost for the group-sequential design can be recovered by the adaptive design. Especially if \( \delta_{01} \) is overestimated, values of \( \delta_{12} \) which are close to the value used in the planning hypothesis can possibly help to show a difference in the joint distribution of PFS and OS. Furthermore, the adaptive procedure complies with the nominal type I error level (empirical type I error: 5.17%, 95%-confidence interval: [4.74%, 5.60%]).

9 | DISCUSSION

We proposed a procedure to evaluate the performance of an experimental cohort in comparison with a fixed reference distribution regarding several time-to-event endpoints. It allows testing hypotheses on the multivariate survival function of the patients while making provision for simultaneous use of interim data from all involved time-to-event endpoints for data-driven design modifications. These endpoints and their joint distribution have been embedded in a multi-state model. In comparison to other testing procedures for multi-state models, the presented method does not only focus on particular transitions, but evaluates the whole process with a focus on the relevant endpoints. It qualifies as a multivariate extension of the univariate independent increments approach for single-armed adaptive survival trials based on the one-sample log-rank test. As shown in Section 6.1, the methodology can be applied to relevant scenarios for confirmatory Phase IIa survival trials. If prognostic factors are available and incorporated in the reference distribution, these can of course be adjusted for by computing the martingales with hazards that are additionally conditioned on these factors. From a practical point of view, the bivariate testing problem reflects the requirements of early phase survival trials in oncology very suitably. In
such trials, recommendation of a new treatment for further assessment is often based on simultaneous consideration of treatment efficacy and treatment toxicity. This naturally leads to a bivariate testing problem.

If properly applied, our methodology yields several advantages. In the first place, it allows to choose multiple primary time-to-event endpoints. Especially our example from Section 2 addresses the broadly discussed topic on whether to choose PFS or OS as the primary endpoint in oncological survival trials. Subsequently, the result shows the performance of the cohort under consideration in several aspects and may help to detect structural differences compared to the reference model as it has been illustrated in Section 8. This is a valid point even if the procedure is applied in a non-sequential way, that is, with only one analysis. When used as a sequential adaptive procedure, design changes are allowed to be based on all included time-to-event endpoints. Hence, it also partially addresses the concern raised in Bauer and Posch (2004). Additionally, further refinements in view of the rejection region are possible. One could set different ways to determine the $p$-values in the different analyses in advance. In view of our case study (c.f. Section 8) this could mean that we spend the whole type I error level provided for the first stage on PFS as we may mostly observe progressions in this stage and the type I error level provided for the last stage on OS as we may mostly observe deaths at a later stage.

Choice of the reference distribution will typically be done in the light of historic data on standard of care. This entails a number of challenges. First, estimating a multivariate reference distribution from historic data is mathematically complex. However, several multivariate survival models and corresponding procedures for inference are available in the literature, especially for our main example from Section 2 concerning a joint model for PFS and OS. Apart from the estimation process, it is also essential to ensure that the chosen reference distribution is clinically valid and sensible. In this context, the common challenges in connection with historically controlled trials apply to our methodology as well: It is crucial that the composition of the historical control cohort (used for choosing the reference distribution) is consistent with the inclusion and exclusion criteria of the study in order to ensure that historic and study population are comparable (in particular concerning distribution of key prognostic factors). Furthermore, sufficient data has to be available for the entire length of follow-up in the planned study in both the historic control group and the experimental treatment group. These conditions are often fulfilled, for example, in pediatric oncology which is characterized by a great willingness of patients to participate in clinical trials (c.f. Kaatsch, 2004 or Erdmann et al., 2020). This implies the existence of population-based data sets on standard of care and thus the best possible prerequisites for choosing a sensible reference distribution. However, even under such ideal conditions, comparisons with a historical reference should be interpreted with the necessary caution, as any superiority over a historical reference may reflect not only the therapeutic effect but also any advances in diagnostics and concomitant therapy. Thus, our methodology appears very attractive for early phase survival trials in oncology. However, the results must be confirmed in a subsequent phase III and cannot replace a randomized comparison. For future research, we therefore aim for an extension of our methodology from a single-arm setting to a (randomized) multi-arm setting.

Apart from extensions toward a (randomized) multi-arm setting, our results could also be the basis to improve the univariate log-rank test by allowing interim decision making to be based on several (correlated) survival endpoints, simultaneously. This is even closer to the scenario envisioned in Bauer and Posch (2004) and requires an estimation of the dependence, which only seems possible under additional assumptions. That makes an accurate treatise of this particular problem necessary. Similar mathematical issues arise when study discontinuations occur while death might still be observable through administrative data. If the discontinuation occurs after progression, nothing needs to be changed. Otherwise, the OS-counting process for that particular patient needs to be compensated with the hazard conditional on the fact that progression has not occurred until the time of study discontinuation. The same conditional hazards need to be calculated in the aforementioned case.

Finally, we also addressed the problems arising for small sample sizes in classical one-sample log-rank tests. Deviations of the distribution of the classical one-sample log-rank test from the standard normal distribution for small sample sizes gains importance if applied to our sequential and multivariate procedure. We extended the suggestion from Wu (2014) to the multivariate case which yielded satisfying results for the scenarios considered in Section 7. However, it can still be deficient for either very short or very long follow-ups and could therefore be further improved.

For the future, we primarily strive to extend our methodology from a single-arm setting to a (randomized) multi-arm setting. We expect that the multivariate counting process martingale derived here can provide a basis for this.

**ACKNOWLEDGEMENTS**

The authors thank the two anonymous reviewers whose suggestions helped improve this manuscript, especially in drawing a clear connection to application and applicability. The work of the corresponding author was funded by the
German Science Foundation (Deutsche Forschungsgemeinschaft, DFG, grant number 413730122). The NB2004-HR trial was funded by a grant obtained by Frank Berthold from the Deutsche Krebshilfe (grant number 70107712).

CONFLICT OF INTEREST
The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT
The data that supports the findings of this study are available in the supplementary material of this article.

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 could fully be reproduced.

ORCID
Moritz Fabian Danzer https://orcid.org/0000-0002-0512-4227
Rene Schmidt https://orcid.org/0000-0003-4808-6707

REFERENCES
Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). Statistical Models Based on Counting Processes. New York: Springer.
Bauer, P., & Posch, M. (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, 1333–1334.
Berthold, F., Faldum, A., Ernst, A., Boos, J., Dilloo, D., Eggert, A., Fischer, M., Frühwald, M., Henze, G., Klingebiel, T., Kratz, C., Kremens, B., Krug, B., Leuschner, I., Schmidt, M., Schmidt, R., Schumacher-Kuckelkorn, R., von Schweinitz, D., Schilling, F. H., … Simon, T. (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, 422–429.
Beyersmann, J., Allignol, A., & Schumacher, M. (2011). Competing Risks and Multistate Models with R. New York: Springer.
Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. International Statistical Review / Revue Internationale de Statistique, 43, 45–57.
Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., & Renard, D. (2001). Validation of surrogate end points in multiple randomized clinical trials with failure time end points. Journal of the Royal Statistical Society: Series C (Applied Statistics), 4, 405–422.
Erdmann, F., Kaatsch, P., Grabow, D., & Spix, C. (2020). German Childhood Cancer Registry - Annual Report 2019(1980-2018). Mainz: Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) at the University Medical Center of the Johannes Gutenberg University.
Fleischer, F., Gaschler-Markefski, B., & Bluhmki, E. (2009). A statistical model for the dependence between progression-free survival and overall survival. Statistics in Medicine, 28, 2669–2686.
Hoos, A. (2012). Evolution of end points for cancer immunotherapy trials. Annals of Oncology, 23, VIII47–VIII52.
Hougaard, P. (2000). Analysis of Multivariate Survival Data. New York: Springer.
Irle, S., & Schäfer, H. H. (2012). Interim design modifications in time-to-event studies. Journal of the American Statistical Association, 107, 341–348.
Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharmaceutical Statistics, 10, 347–356.
Kaatsch, P. (2004). Das Deutsche Kinderkrebsregister im Umfeld günstiger Rahmenbedingungen. Bundesgesundheitsbl - Gesundheitsforsch - Gesundheitsschutz, 47, 437–443.
Karatzas, I., & Shreve, S. (1998). Brownian Motion and Stochastic Calculus. New York: Springer.
Lee, J., Thall, P. F., & Msaouel, F. (2020). A phase I-II design based on periodic and continuous monitoring of disease status and the times to toxicity and death. Statistics in Medicine, 39, 2045–2050.
Le-Rademacher, J., Peterson, R., Therneau, T., Sanford, B., Stone, R., & Mandrekar, S. (2018). Application of multi-state models in cancer clinical trials. Clinical Trials, 15, 489–498.
Magirr, D., Jaki, T., Koenig, F., & Posch, M. (2016). Sample size reassessment and hypothesis testing in adaptive survival trials. PLoS ONE, 11(2), e0146465.
Meller, M., Beyersmann, J., & Rufibach, K. (2019). Joint modeling of progression-free and overall survival and computation of correlation measures. Statistics in Medicine, 38, 4270–4289.
R Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
Supporting Information

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

How to cite this article: Danzer, M. F., Terzer, T., Berthold, F., Faldum, A., Schmidt, R. Confirmatory adaptive group sequential designs for single-arm phase II studies with multiple time-to-event endpoints. Biometrical Journal. 2022;64:312–342. https://doi.org/10.1002/bimj.202000205

Appendix

A.1 Computation of test statistics in Phase IIa setting

Analogously to the procedure in Section 2, we present the computation of test statistics for the scenario presented in Section 6.1 which corresponds to a typical Phase IIa trial which aims for a simultaneous assessment of efficacy and toxicity. The following conditional hazards need to be given:

\[
\lambda^{(1,1,2,3)} \quad \text{the hazard for toxicity without any prior event}
\]

\[
\lambda^{(2,1,2,3)} \quad \text{the hazard for progression without any prior event}
\]

\[
\lambda^{(3,1,2,3)} \quad \text{the hazard for death without any prior event}
\]

\[
\lambda^{(1,1,3)} \quad \text{the hazard for toxicity with prior progression}
\]

\[
\lambda^{(3,1,3)} \quad \text{the hazard for death with prior progression}
\]

\[
\lambda^{(2,2,3)} \quad \text{the hazard for progression with prior toxicity}
\]

\[
\lambda^{(3,2,3)} \quad \text{the hazard for death with prior toxicity}
\]

As already mentioned, \(\lambda^{(1,1,3)}, \lambda^{(3,1,3)}, \lambda^{(2,2,3)},\) and \(\lambda^{(3,2,3)}\) do not need to be given explicitly. It is sufficient to know \(\lambda^{(1,1,3)} + \lambda^{(3,1,3)}\) and \(\lambda^{(2,2,3)} + \lambda^{(3,2,3)}\), that is, the hazard for toxicity or death given a progression and the hazard for progression-free survival given a toxicity. The hazards can be computed using the joint distribution of \(T^{[1]}, T^{[2]},\) and \(T^{[3]},\) analogously to (1) resp. (2). If the hazards of a reference model are once again marked with a 0 in the lower index, the null hypothesis can, analogously to (3), be formulated by

\[
H_0 : \begin{pmatrix}
\lambda^{(1,1,2,3)}_0 \\
\lambda^{(2,1,2,3)}_0 \\
\lambda^{(3,1,2,3)}_0 \\
\lambda^{(1,1,3)}_0 + \lambda^{(3,1,3)}_0 \\
\lambda^{(2,2,3)}_0 + \lambda^{(3,2,3)}_0
\end{pmatrix} = \begin{pmatrix}
\lambda^{(1,1,2,3)}_0 \\
\lambda^{(2,1,2,3)}_0 \\
\lambda^{(3,1,2,3)}_0 \\
\lambda^{(1,1,3)}_0 + \lambda^{(3,1,3)}_0 \\
\lambda^{(2,2,3)}_0 + \lambda^{(3,2,3)}_0
\end{pmatrix}.
\]
For any $t \geq 0$ we count the following events:

- $N^E_1(t)$ is the number of observed PFS-events (i.e., progressions and deaths without prior progressions),
- $N^E_2(t)$ is the total number of toxicities and deaths without prior toxicities,
- $N^{E_1,E_2}(t)$ is the number of deaths without prior progressions and toxicities observed.

The number of events expected under the reference model given the filtration generated by the observable events in calendar time are

$$A_0^{E_1,E_2}(t) = \sum_{i=1}^{n} \left( \int_{0}^{t} \lambda_0^{(3,\{1,2,3\})}(s, s, s, s)ds \right),$$

$$A_0^{E_1}(t) = A_0^{E_1,E_2}(t) + \sum_{i=1}^{n} \left( \int_{0}^{t} \lambda_0^{(1,\{1,2,3\})}(s, s, s)ds \right),$$

$$+ \sum_{i=1}^{n} \left( \int_{0}^{t} \lambda_0^{(3,\{1,3\})}(s, s, T_i^{[3]} \land C_i \land t-Y_i) ds \right),$$

$$A_0^{E_2}(t) = A_0^{E_1,E_2}(t) + \sum_{i=1}^{n} \left( \int_{0}^{t} \lambda_0^{(2,\{1,2,3\})}(s, s, s)ds \right),$$

$$+ \sum_{i=1}^{n} \left( \int_{0}^{t} \lambda_0^{(3,\{2,3\})}(s, T_i^{[1]} \land C_i \land T_i^{[3]}, s) ds \right).$$

From here on, the remaining steps which lead to stage-wise $p$-values can be adapted from Section 2. While PFS need to be replaced by $E_1$ resp. OS needs to be replaced by $E_2$ in the computations (4) and (5), the following steps (6), (7) and (8) do not need any further adaptation. Once again, the standardization applied here is the extension of the approach from Wu (2014).

Please note that the effect of direct deaths without any prior event is only incorporated in the first component of the stage-wise test statistics. If it is desired otherwise, one might just interchange $E_1$ and $E_2$.

### A.2 Computation of stage-wise $p$-values

As indicated in Section 5, there are many options to derive $p$-values from the stage-wise test statistics. If $d$ different events are assessed, we obtain a standard normally distributed random vector of length $d$ after each stage. We fix the stage and call this vector $Z$. To derive a $p$-value from that vector we suggest to apply a function $g : \mathbb{R}^d \rightarrow [0, \infty)$, calculate the distribution of $g(Z)$ and compute the $p$-value for an observed value $Z$ by

$$p = P[g(Z) \geq g(z)].$$

In our opinion, $g$ should at least be a seminorm to get a meaningful test. Also, for the sake of applicability, the distribution of $g(Z)$ should be easy to calculate. From this point of view, choosing $g$ as the $L^2$-norm or the $L^{\infty}$-norm on $\mathbb{R}^d$ seem to be suitable.

For the former one we have

$$P[g(Z) \geq g(z)] = P\left[\sqrt{\sum_{i=1}^{d} Z_i^2} \geq g(z)\right]$$

$$= P\left[\sum_{i=1}^{d} Z_i^2 \geq g(z)^2\right]$$

$$= 1 - F_{\chi^2_d\left(g(z)^2\right)} = 1 - F_{\chi^2_d\left(|z|^2\right)}$$
where \( F_{\chi^2_d} \) denotes the distribution function of a \( \chi^2_d \)-distributed random variable. The last equation holds because 
\[
\sum_{i=1}^d Z_i^2 \sim \chi^2_d.
\]
For the latter one we have
\[
P\left[ g(Z) \geq g(z) \right] = P\left[ \max_{i=1,...,d} |Z_i| \geq g(z) \right] \\
= P\left[ \exists i \in \{1,...,d\} : |Z_i| \geq g(z) \right] \\
= 1 - P\left[ \forall i \in \{1,...,d\} : |Z_i| \geq g(z) \right] \\
= 1 - \prod_{i=1}^d P[|Z_i| < g(z)] \\
= 1 - P[|Z_1|^2 < g(z)^2]^d \\
= 1 - F_{\chi_1^2}(g(z)^2)^d = 1 - F_{\chi_1^2}(|z|_{\infty}^2)^d
\]
where we used that the components of \( Z \) are i.i.d. and standard normally distributed and that \( Z_i^2 \sim \chi^2_1 \).

### A.3 Computation of conditional hazards from joint densities and joint distribution functions

As indicated, we do not only want our procedure to be applicable for standard multi-state model approaches as time-homogeneous, time-inhomogeneous or Semi-Markov models, but also for other modeling approaches for multivariate survival times using copulas or frailties. In these cases, the conditional hazards which we need are not given directly. All we are given in the first place is the joint distribution function \( F \) or the joint density \( f \) of the random vector \( T \). So in the following we want to provide the formulas to compute the conditional hazards from joint distribution functions or densities. These formulas rely on the computation of regular versions of conditional distributions. A proof of the correctness of the following formulas can be given on request to the authors.

We fix \( j \in \{1,...,k\} \) and \( U \in P(\{1,...,k\}) \) with \( j \in U \) and a reference vector \( r = (r^{[1]},...,r^{[k]}) \in [0,\infty)^k \). For any \( l \in U \), \( r^{[l]} \) indicates the last known time for which the component event \( l \) has not yet happened. For any \( l \not\in U \), \( r^{[l]} \) indicated the time at which the component event \( l \) occurred. Then with \( u = (u_1,...,u_k) \), we have
\[
\lambda^{(U)}(s,r^{[1]},...,r^{[k]}) = \int \prod_{l \in U \setminus \{j\}} f((s,(u,r)_U)_{\{j\}}) \, du_{U \setminus \{j\}} \left( \int \prod_{l \in U \setminus \{j\}} f((u,r)_U) \, du_U \right)^{-1}
\]
where \((s,(u,r)_U)_{\{j\}}\) is just a nesting of the notation introduced above, which means that the \( j \)-th entry of this vector is \( s \), the other entries which index is contained in \( U \) are \( u_l \) and the entries which index is not contained in \( U \) are \( r^{[l]} \).

As
\[
f(s_1,...,s_k) = \frac{\partial}{\partial s_1} ... \frac{\partial}{\partial s_k} F(s_1,...,s_k)
\]
we can also give a similar formula in terms of the distribution function applying
\[
\int \prod_{l \in A_{\{j\}}} f((s,(u,r)_U)_{\{j\}}) \, du_{U \setminus \{j\}} = \frac{\partial}{\partial s} \prod_{l \in U \setminus \{j\}} \frac{\partial}{\partial r^{[l]}} \sum_{V \subset U \setminus \{j\}} (-1)^{|V|} F_{U \cup U \setminus V \cup \{j\}}((s,r)_{U \cup U \setminus V \cup \{j\}})\]
resp.

\[
\int_{\prod (r|U) = \infty} f((u,r) \mid U) \, du = \prod_{l \in U^c} \frac{\partial^{2}}{\partial r_l^2} \sum_{V \subset U} (-1)^{|V|} F_{U \cup V \mid (s,r)|U \cup V} \cdot f((u,r) \mid U) \, du \, U = \prod_{l \in U^c} C \frac{\partial}{\partial r_l} \sum_{V \subset U} \left( -1 \right)^{|V|} \, \left( \frac{s}{r_l} \right) \left( U \cup V \right) \cdot \left( u, r \right) \mid U \cup V.
\]

Here \( U^c \) denotes the complement of \( A \) in the set of events \( \{1, \ldots, k\} \) and \( F_U \) for a set \( U \subset \{1, \ldots, k\} \) indicates the joint distribution function of the components of \( T \) which index is in \( U \). Correspondingly \( r \mid U \) denotes the projection of the \( k \)-dimensional vector \( r \) to its components which index is in \( U \).

Although these formulas may appear a bit unwieldy, it is not very difficult to apply them to standard copula or frailty models.

### A.4 Proofs of main results

**Proof of Lemma 1.** At first, we consider the extended filtration \((\tilde{F}_{i,s})_{s \geq 0}\) where for any \( s \geq 0 \) the \( \sigma \)-algebra \( \tilde{F}_{i,s} \) is generated by the random variables \( C_i, \mathbb{1}_{[T_{i}^{[j]} \leq s]}, T_{i}^{[j]} \cdot \mathbb{1}_{[T_{i}^{[j]} \leq s]} \) for all \( j \in \{1, \ldots, k\} \). Note that \( F_{i,s} \subset \tilde{F}_{i,s} \) for any \( s \geq 0 \).

Clearly, \((\mathbb{1}_{[T_{i}^{[j]} \leq s]}(s))_{s \geq 0}\) is a \((\tilde{F}_{i,s})_{s \geq 0}\)-submartingale. As (R1) holds, all requirements of the Doob-Meyer decomposition theorem are fulfilled (see, e.g., Karatzas & Shreve, 1998). Hence, there exists an increasing \((\tilde{F}_{i,s})_{s \geq 0}\)-predictable process \((\tilde{A}_{i}^{[j]}(s))_{s \geq 0}\) s.t. \((\tilde{M}_{i}^{[j]}(s))_{s \geq 0}\) with

\[
\tilde{M}_{i}^{[j]}(s) = \mathbb{1}_{[T_{i}^{[j]} \leq s]}(s) - \tilde{A}_{i}^{[j]}(s) \quad \forall s \geq 0
\]

is an \((\tilde{F}_{i,s})_{s \geq 0}\)-martingale.

As in Karatzas and Shreve (1998), we will identify this process by a discrete approximation. For a fixed \( s \geq 0 \) we define the partitions \( \Pi_n = (u_{(n)}^0, \ldots, u_{(n)}^m) \) with \( u_{(n)}^m = (j/2^m)s \) for any \( n \in \mathbb{N} \) and \( m \in \mathbb{N} \cap [0,2^n] \) of the interval \([0,s]\). We want to express the probabilities

\[
P \left[ u_{m}^{(n)} < T_{i}^{[j]} \leq u_{m+1}^{(n)} \mid \tilde{F}_{i,u_{m}} \right] = \mathbb{E} \left[ \mathbb{1}_{[T_{i}^{[j]} \leq s]}(u_{m+1}^{(n)}) - \mathbb{1}_{[T_{i}^{[j]} \leq s]}(u_{m}^{(n)}) \mid \tilde{F}_{i,u_{m}} \right]
\]

in terms of the conditional hazard functions defined in Section 3. Therefor we additionally define for any \( j \in \{1, \ldots, k\} \) and \( U \subset \{1, \ldots, k\} \) the cumulative hazard functions

\[
\Lambda^{(j,U)}(s_{\text{end}},s^{[1]},\ldots,s^{[k]}) := \begin{cases} 
0 & \text{if } s_{\text{end}} \leq s^{[j]}, \\
\int_{s^{[j]}}^{s_{\text{end}}} \lambda^{(j,U)}(u,s^{[1]},\ldots,s^{[k]}) \, du & \text{else}.
\end{cases}
\]

Please note, that we only integrate w.r.t. the first argument of the corresponding hazard function while the reference times, that is, the information given concerning the events, remain fixed. We can now write

\[
P \left[ u_{m}^{(n)} < T_{i}^{[j]} \leq u_{m+1}^{(n)} \mid \tilde{F}_{i,u_{m}} \right] = \mathbb{1}_{[T_{i}^{[j]} > u_{m}^{(n)}]} \cdot \left( 1 - \exp \left( -\Lambda^{(j,U(u_{m}^{(n)}))}(u_{m+1}^{(n)},(u_{m}^{(n)},T_{i}^{[j]}))) \right) \right)
\]

to obtain

\[
A_{i}^{[j]}(s) := \sum_{m=1}^{2^n} P \left[ u_{m}^{(n)} < T_{i}^{[j]} \leq u_{m+1}^{(n)} \mid \tilde{F}_{i,u_{m}}^{(n)} \right] = \sum_{m=1}^{2^n} \mathbb{1}_{[T_{i}^{[j]} > u_{m}^{(n)}]} \cdot \left( 1 - \exp \left( -\Lambda^{(j,U(u_{m}^{(n)}))}(u_{m+1}^{(n)},(u_{m}^{(n)},T_{i}^{(n)}))) \right) \right).
\]
We split the sum into a group of summands for which
\[ (s', s) \mapsto \lambda(j, U(u^{(n)}_m))(s', (s, T)_{U(u^{(n)}_m)}) \]
is uniformly continuous on the triangle \[ \{(s', s) : u^{(n)}_m \leq s' < u^{(n)}_m, s' \leq s < u^{(n)}_m\} \] and another group of summands for which this is not the case. By condition (R2) there is only a bounded number of summands for which this is not the case and by (R1) their sum will converge to zero. Hence, we have
\[
\lim_{n \to \infty} A^{(j)}_{i,n}(s) = \lim_{n \to \infty} \sum_{m=1}^{2^n} 1_{\{|i| > u^{(n)}_m\}} \cdot \left(1 - \exp\left(-\lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})\right)\right)
\]
\[
= \lim_{n \to \infty} \sum_{m=1}^{2^n} 1_{\{|i| > u^{(n)}_m\}} \cdot \left(1 - \exp\left(-\lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})\right)\right)
\]
By the fundamental theorem of calculus and Taylor’s theorem, we can write
\[
1 - \exp\left(-\lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})\right) = \frac{1}{2^n} \cdot \lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) + \frac{1}{2^n} \cdot h(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})
\]
for these summands, where \( h(j, U)(s', s, t_U) = 0 \) for any \( U \) and \( t_U \). As the condition (R2) is met \( h \) can be written as
\[
h(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) = \int_0^1 \lambda(j, U(u^{(n)}_m))(u^{(n)}_m + v, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) - \lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) \, dv
\]
and hence the absolute values of these remainders can be bounded by
\[
|h(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})| \leq \sup_{u \in (u^{(n)}_m, u^{(n)}_m + 1)} \left|\lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) - \lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})\right|
\]
By (R2) we have that
\[
\lim_{n \to \infty} \sup_{m \in [1, \ldots, 2^n]} h(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)}) = 0
\]
and hence
\[
\lim_{n \to \infty} A^{(j)}_{i,n}(s) = \lim_{n \to \infty} \sum_{m=1}^{2^n} 1_{\{|i| > u^{(n)}_m\}} \cdot \frac{1}{2^n} \cdot \lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})
\]
As the sum of the missing summands converges to zero as their number is bounded and (R1) holds, we have
\[
\lim_{n \to \infty} A^{(j)}_{i,n}(s) = \lim_{n \to \infty} \sum_{m=1}^{2^n} 1_{\{|i| > u^{(n)}_m\}} \cdot \frac{1}{2^n} \cdot \lambda(j, U(u^{(n)}_m))(u^{(n)}_m, (u^{(n)}_m, T)_{U(u^{(n)}_m)})
\]
\[
= \int_{[0,s]} \lambda(j, U(u))(u, (u, T)_{U(u)}) \, du
\]
The convergence of the Riemann sum is once more guaranteed by (R2). Hence, given (R1) and (R2) we could uniquely identify the process \( (A_i^{(j)}(s))_{i \geq 0} \). Now, consider the process \( (M_i^{(j)}(s))_{i \geq 0} \) stopped at \( C_i \). This is still a \( (\bar{F}_t)_{t \geq 0} \)-martingale.
But this stopped process which is given by
\[
\tilde{M}^{(j)}_i(s) := \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i\}} - \int_{[0,s \land C_i \land T^{(j)}_i]} \lambda^{(j,U_i(u))(u,(u,T_i \land C_i)U_i(u))} du
\]
for any \(s \geq 0\), is also measurable w.r.t. the filtration \((\bar{F}_i,\mathcal{G})_{s \geq 0}\) where for any \(s \geq 0\) the \(\sigma\)-algebra \(\bar{F}_i,\mathcal{G}\) is generated by the random variables
\[
\mathbf{1}_{\{C_i \leq s \land T^{(j)}_i\}}, \mathbf{1}_{\{C_i \leq s \land T^{(j)}_i\}}, \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i\}}, \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i\}} \text{ for all } j \in \{1, \ldots, k\}.
\]
As \(\bar{F}_i,\mathcal{G}\) is a sub-\(\sigma\)-algebra of \(\mathcal{G}\), one can apply the tower property to see that it is now also a \((\bar{F}_i,\mathcal{G})_{s \geq 0}\)-martingale.

Now, \(\mathcal{F}_i,\mathcal{G}\) is in turn a sub-\(\sigma\)-algebra of \(\bar{F}_i,\mathcal{G}\) for any \(s \geq 0\). This can be seen from rewriting the indicator functions from the generators of the \(\sigma\)-algebras \(\bar{F}_i,\mathcal{G}\) by
\[
\mathbf{1}_{\{C_i \leq s \land \max_{j} T^{(j)}_i \land T^{(k)}_i\}} = \mathbf{1}_{\{C_i \leq s \land T^{(k)}_i\}} \cdot \left(1 - \prod_{j=1}^{n} \mathbf{1}_{\{C_i \leq s \land T^{(j)}_i\}}\right)
\]
resp.
\[
\mathbf{1}_{\{T^{(j)}_i \leq s \land C_i \land T^{(k)}_i\}} = \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i\}} \cdot \left(1 - \mathbf{1}_{\{T^{(k)}_i \leq s \land C_i\}} \cdot \mathbf{1}_{\{T^{(k)}_i \leq s \land C_i\}} < T^{(j)}_i \cdot \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i\}}\right).
\]
From (11) and (12), it follows that there is \(P\)-almost surely no event happening after \(T^{(k)}_i\). This can be seen from
\[
P[T^{(k)}_i < T^{(j)}_i < \infty] = \int_{(0,\infty)} P[T^{(k)}_i \land T^{(j)}_i < \infty | T^{(k)}_i = s^{(k)}] \cdot f_{T^{(k)}_i}(s^{(k)}) \, ds^{(k)}
\]
where the second factor of the integrand is the density of the distribution of \(T^{(k)}_i\) restricted to \((0,\infty)\) and the first factor can be rewritten as
\[
P[T^{(k)}_i \land T^{(j)}_i < \infty | T^{(k)}_i = s^{(k)}] = P[T^{(k)}_i \land T^{(j)}_i < \infty | T^{(k)}_i = s^{(k)}, T^{(j)}_i > 0 \forall l \in \{1, \ldots, k-1\}]
\]
\[
= 1 - \exp\left(-\int_{(s^{(k)},\infty)} \lambda^{(j,1\ldots,k-1)}(u,0,\ldots,0,s^{(k)}) \, du\right)
\]
\[
= 0.
\]
Hence, the process \((M^{(j)}_i(s))_{s \geq 0}\) defined by
\[
M^{(j)}_i(s) := \mathbf{1}_{\{T^{(j)}_i \leq s \land C_i \land T^{(k)}_i\}} - \int_{[0,s \land C_i \land T^{(j)}_i \land T^{(k)}_i]} \lambda^{(j,U_i(u))(u,(u,T_i \land C_i)U_i(u))} du
\]
for all \(s \geq 0\) is indistinguishable from the process \((\tilde{M}^{(j)}_i(s))_{s \geq 0}\) mentioned above. Together with \(\mathcal{F}_i,\mathcal{G}\) being a sub-\(\sigma\)-algebra of \(\bar{F}_i,\mathcal{G}\) for any \(s \geq 0\) this implies that \((M^{(j)}_i(s))_{s \geq 0}\) is a \((\mathcal{F}_i,\mathcal{G})_{s \geq 0}\)-martingale.

Proof of Corollary 2. For any \(j \in E\), \((M^{(j)}_i(s))_{s \geq 0}\) is an \((\mathcal{F}_i,\mathcal{G})_{s \geq 0}\)-martingale. Additionally \(\min_{j \in E} T^{(j)}_i \land C_i\) is a \((\mathcal{F}_i,\mathcal{G})_{s \geq 0}\)-stopping time. Hence, the result follows from the Optional Stopping Theorem.

Proof of Theorem 3. Let \(E \subset \{1, \ldots, k\}\) be fixed. We now need to incorporate the censoring induced by calendar time \(t\), that is, the censoring variable of patient \(i\) is now \(C_i \land (t - y_i)_+\).
Consider at first the filtration \((\mathcal{F}_s)_{s\geq 0}\) where \(\mathcal{F}_s\) is the smallest \(\sigma\)-algebra containing \(\sigma\)-algebras \(\mathcal{F}_{i,s}\) for any \(i \in \{1, \ldots, n\}\). We can apply the Optional Stopping Theorem to see that \((M^E_i(t,s))_{s \geq 0}\) with 
\[
M^E_i(t,s) := M^E_i(s \land (t - y_i)_+)
\]
is also a \(\mathcal{F}_s\)-martingale for any \(t \geq 0\). This procedure corresponds to stopping the patient's process at calendar time \(t\).

Now, for any \(s, t \geq 0\) we define a sub-\(\sigma\)-algebra \(\mathcal{F}_{t,s}\) of \(\mathcal{F}_s\), which is generated by the random variables
\[
1_{\{T_i^{(j)} \leq \sigma \land C_i \land \tau^{(k)} \land (t - y_i)_+\}} T_i^{(j)} \land 1_{\{T_i^{(j)} \leq \sigma \land C_i \land \tau^{(k)} \land (t - y_i)_+\}} C_i \land 1_{\{T_i^{(j)} \leq \sigma \land C_i \land \tau^{(k)} \land (t - y_i)_+\}}
\]
for all \(j \in \{1, \ldots, k\}\) and \(i \in \{1, \ldots, n\}\). This corresponds to the information which is actually available at calendar time \(t\) and time in trial \(s\) of the respective patients. Moreover, let \(\mathcal{F}_t := \mathcal{F}_{t,t}\) for any \(t \geq 0\).

As \(M^E_i(t,s)\) is measurable w.r.t. \(\mathcal{F}_{t,s}\) for any \(s, t \geq 0\), it follows that \((M^E_i(t,s))_{s \geq 0}\) is a \((\mathcal{F}_{t,s})_{s \geq 0}\)-martingale for any \(t \geq 0\). For any fixed \(s \geq 0\) also \((M^E_i(t,s))_{t \geq 0}\) is a \((\mathcal{F}_{t,s})_{s \geq 0}\)-martingale. This becomes clear as
\[
\mathbb{E}[M^E_i(t,s) | \mathcal{F}_{u,s}] = \mathbb{E}[M^E_i(s + y_i, (t - y_i)_+) | \mathcal{F}_{s+y_i, (u-y_i)_+}] = M^E_i(s + y_i, (u - y_i)_+) = M^E_i(u, s)
\]
for any \(s, t \geq 0\) and \(0 \leq u \leq t\) where we used that
\[
\mathcal{F}_{t,s} = \mathcal{F}_{s+y_i, (u-y_i)_+} \quad \text{and} \quad M^E_i(t,s) = M^E_i(s + y_i, (t - y_i)_+).
\]
As \(\mathcal{F}_{i,s} \subset \mathcal{F}_{t,s'}\) for any \(0 \leq s \leq s'\) and \(0 \leq t \leq t'\) we can see with
\[
\mathbb{E}[M^E_i((t' - y_i)_+) | \mathcal{F}_{t,s}] = \mathbb{E}[M^E_i((t' - y_i)_+) | \mathcal{F}_{t,s'}] = \mathbb{E}[\mathbb{E}[M^E_i((t' - y_i)_+) | \mathcal{F}_{t,s'}] | \mathcal{F}_{t,s}]
\]
that \((M^E_i((t - y_i)_+))_{i \geq 0}\) is a \((\mathcal{F}_{t,s})\)-martingale for any \(i \in \{1, \ldots, n\}\).

By summing over all patients and standardizing the sum with \(n^{-\frac{1}{2}}\) we obtain the mean zero \((\mathcal{F}_t)_{t \geq 0}\)-martingale \((M^E(t))_{t \geq 0}\) with
\[
M^E(t) := n^{-\frac{1}{2}} \sum_{i=1}^{n} M^E_i((t - y_i)_+)
\]
for all \(t \geq 0\). The second equality is obtained by inserting the formula for \(M^E((t - y_i)_+)\) and rearranging the terms to get the processes from (13) and (14).

**Proof of Theorem 4.** We aim for the approximate distribution of the multivariate mean zero \((\mathcal{F}_t)_{t \geq 0}\)-martingale \((M(t))_{t \geq 0} = (M^E_1(t), \ldots, M^E_d(t))_{t \geq 0}\) as sample size \(n\) increases. The proof strategy is to apply a multivariate version of Rebolledo’s martingale central limit theorem (see Theorem II.5.1 in Andersen et al., 1993). For this purpose, we need to study the asymptotic properties of the quadratic co(variation) process \([M^E_a, M^E_b]\) for \(a, b \in \{1, \ldots, d\}\) as \(n \to \infty\). We have
\[
[M^E_a](s) = \mathbb{1}_{\{T_i^{(j)} \leq \sigma \land C_i \land \tau^{(k)}\}} \quad \text{resp.}
\]
As (R1) guarantees that there are a.s. no two martingales \((M^E_a(t - y_i)_+), t \geq 0\) and \((M^E_b(t - y_{i'})_+), t \geq 0\) with \(i \neq i'\) jumping simultaneously in finite time, we have that

\[
[M^E_a](t) = n^{-1}N^E_a(t) \quad \text{and} \quad [M^E_a, M^E_b](t) = n^{-1}N^{E_a, E_b}(t)
\]

almost surely for all \(a, b \in \{1, \ldots, d\}\). Since

\[
M^E_a(t) = \frac{a_n(t)}{n^{1/2}} \cdot \sum_{i=1}^{a_n(t)} M^E_a(t),
\]

we conclude with the weak law of large numbers for independent random variables that

\[
[M](t) \xrightarrow{\mathbb{P}} V(t) : = r(t) \cdot \Sigma(t)
\]

where \([M](t)\) denotes the matrix \(\{[M^E_a, M^E_b](t)\}_{a, b \in \{1, \ldots, d\}}\) and \(\Sigma\) is the \(d \times d\)-matrix defined in (17). Since martingales have uncorrelated increments, it follows that \(V(t) - V(s) = \lim_{n \to \infty} \text{Cov}[M(t) - M(s)]\) for any \(0 \leq s < t\). Thus, as a limit of covariance matrices \((V_t)_{t \geq 0}\) has positive semidefinite increments. Moreover, \(V(0) = 0_{d \times d}\).

Likewise we need to check that \(\langle M^{E_a \varepsilon}(t) \rangle_+ \xrightarrow{\mathbb{P}} 0\) for any \(a \in \{1, \ldots, d\}\) and \(\varepsilon > 0\), where \(M^{E_a \varepsilon}(t)_+\) denotes the locally square integrable martingale containing all jumps of \((M^{E_a}(t))_{t \geq 0}\) larger in absolute value than \(\varepsilon\). This is guaranteed by (R1) once more, as it implies that any jump of \((M^{E_a}(t))_{t \geq 0}\) is of size \(n^{-1/2}\) almost surely. Hence, \(\langle M^{E_a \varepsilon}(t) \rangle_+ = 0\) for any \(a \in \{1, \ldots, d\}\) if \(n > \varepsilon^{-2}\).

Since the conditions of Theorem II.5.1 in Andersen et al. (1993) are thus met, we conclude that

\[
(M(t_1), \ldots, M(t_m)) \overset{D}{\to} (M_\infty(t_1), \ldots, M_\infty(t_m))
\]

as \(n \to \infty\), where \((M_\infty(t))_{t \geq 0}\) is a Gaussian process with independent increments and \(M_\infty(t) - M_\infty(s) \sim \mathcal{N}(0_d, V(t) - V(s))\) for all \(0 \leq s < t\). If \(\Delta_i = C_i C_i^\top\) is the Cholesky decomposition of \(\Delta_i\) for any \(i \in \{1, \ldots, m\}\), we have

\[
Z = (C_i^{-1}(M^{E_a}_\infty(t) - M^{E_a}_\infty(t_{i-1}))^\top)_{i \in \{1, \ldots, m\}} \sim \mathcal{N}(0, I_{d \times m}).
\]

Now, if a consistent estimator \(\hat{\Delta}_i\) for \(\Delta_i\) is given for any \(i \in \{1, \ldots, m\}\), the assertion follows from appeal to the Continuous Mapping Theorem and Slutsky’s Theorem.

\(\square\)

**Proof of Theorem 5.** Assume there is an infinite countable set of patients with entry times \(0 \leq y_1 < y_2 < \ldots\). The initially planned number of patients is hence be given by \(\sum_{i \in \mathbb{N}} 1_{|y_1 \leq s - f|_i}\). For the calculation of test statistics we need to redefine the summands in the definition of \(N\) and \(A\) from (13) resp. (14) by multiplying them with indicator functions \(1_{|y_i \leq s - f|_i}\) for any \(i \in \mathbb{N}\). Note that this formulation leads to the same asymptotical results as exhibited above. Let \(Z_0 = (Z_{0,1}, Z_{0,2})\) be the test statistic as if the trial was executed as originally planned. To indicate the formal dependence of \(Z\) on the analysis dates, we write \(Z_{0,1}(t_1)\) and \(Z_{0,1}(t_2)\) instead of \(Z_{0,1}\) resp. \(Z_{0,2}\). Let \(Z^*_0 = (Z^*_{0,1}, Z^*_{0,2}) = (Z_{0,1}(t_1), Z_{0,2}(t_2))\) be the test statistic after possible design adaptations at interim.

As the design adaptations are made at \(t_1\) and do not affect any design specifications in the interval \([0, t_1]\), we have \(Z^*_{0,1} \equiv Z_{0,1}\) as one can see from the results in Theorem 4.

We need to show that the rejection probability under the null hypothesis remains unchanged for the adapted trial, that is

\[
P_{H_0}[Z_0 \in \mathcal{R}] = P_{H_0}[Z^*_0 \in \mathcal{R}]
\]

(A.1)
We assume that \( \mathcal{R} \) is given via a conditional error function and we have a region \( \mathcal{R}_1 \) for early stopping and for each outcome \( \mathbf{Z}_{0,1} \) a region \( \mathcal{R}_2(\mathbf{Z}_{0,1}) \) for rejection at the final analysis date. We can split up the right hand side into

\[
P_{H_0}[\mathbf{Z}_0^* \in \mathcal{R}] = P_{H_0}[\mathbf{Z}_{0,1}^* \in \mathcal{R}_1] + P_{H_0}[\mathbf{Z}_{0,1}^* \notin \mathcal{R}_1, \mathbf{Z}_{0,2}^* \in \mathcal{R}_2(\mathbf{Z}_{0,1}^*)].
\]

The first summand amounts to \( P_{H_0}[\mathbf{Z}_{0,1} \in \mathcal{R}_1] \). For the second part we can compute

\[
P_{H_0}[\mathbf{Z}_{0,1}^* \notin \mathcal{R}_1, \mathbf{Z}_{0,2}^* \in \mathcal{R}_2(\mathbf{Z}_{0,1}^*)] = E_{H_0}[P_{H_0}[\mathbf{Z}_{0,1}^* \notin \mathcal{R}_1, \mathbf{Z}_{0,2}^* \in \mathcal{R}_2(\mathbf{Z}_{0,1}^*)|F_1]]
\]

= \[E_{H_0}[1_{\mathbf{Z}_{0,1}(t_1) \notin \mathcal{R}_1}P_{H_0}[\mathbf{Z}_{0,2}(\tau_2) \in \mathcal{R}_2(\mathbf{Z}_{0,1}(t_1))|F_1]]
\]

= \[E_{H_0}[1_{\mathbf{Z}_{0,1} \notin \mathcal{R}_1}P_{H_0}[\mathbf{Z}_2 \in \mathcal{R}_2(\mathbf{Z}_{0,1})|F_1]]
\]

= \[P_{H_0}[\mathbf{Z}_{0,1} \notin \mathcal{R}_1, \mathbf{Z}_{0,2} \in \mathcal{R}_2(\mathbf{Z}_{0,1})]
\]

where we used conditioning on \( F_1 \), the fact that \( \mathbf{Z}_{0,1}^* \) and \( \tau_2 \) are known given \( F_1 \) and \( \mathbf{Z}_{0,1}^* \equiv \mathbf{Z}_{0,1} \). Now, the equality (A.1) holds as the test statistics arising from both stages are independent under \( H_0 \) for any deterministic design. \( \square \)