AIMS Mathematics, 11(1): 66–84.
DOI: 10.3934/math.2026004
Received: 11 September 2025
Revised: 03 December 2025
Accepted: 16 December 2025
Published: 04 January 2026

https://www.aimspress.com/journal/Math

Research article

Testing probability of being in response

Ekkehard Glimm[1,2,*] **and Norbert Hollaender**[1]

[1] Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

[2] Otto-von-Guericke University, Institute of Biometry and Medical Informatics, Leipziger Str. 44, 39120 Magdeburg, Germany

* **Correspondence:** Email: ekkehard.glimm@novartis.com.

Abstract: The probability-of-being-in-response (PBR) curve is a graphical method that combines two time-to-event endpoints, namely time from study start to first response and time from first response to subsequent failure considering all patients of a study. We generalize the logrank-test to a test that compares PBR curves. We focus on the global null hypothesis of no difference between the multistate stochastic processes underlying the two curves. The test is designed in such a way that it has high power when the PBR is consistently higher for one of the two groups at all times. Simulations and the application to clinical trial data show that the proposed tests are useful additions to the visual comparison of PBR curves.

Keywords: time-to-event-analysis; multistate models; probability-of-being-in-response; duration of response
Mathematics Subject Classification: 62N01, 62P10

1. Introduction

Consider a randomized clinical trial comparing two treatments of a chronic disease such as cancer. At the time of randomization all patients are in the same disease state but require further therapy. If the treatment is successful, the patients' condition improves, hopefully to a point where they reach a state called "being in response". If, however, the treatment is not a cure, subsequently patients might enter a terminal state (e.g., they might die or their condition might deteriorate to an irreversible state of severe sickness).

Conceptually, this situation can be described by a discrete-state-continuous-time stochastic process with initial state 0, a transitory state 1 and a terminal state 2. Figure 1 shows a corresponding graph.

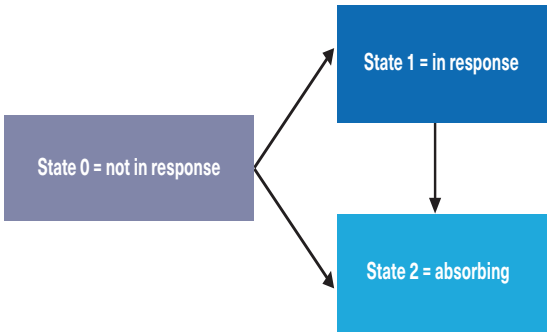

Figure 1. Three-state model for responder status.

A famous example of such a three-state-continuous-time model is the illness-death-model [3]. In this paper, we consider the probability of being in response (PBR). Several authors [5, 9, 18, 20] have investigated point and interval estimation of this quantity. A recent application focusing on the graphical presentation of PBR over time has been described by [10]. We will revisit this application in section 4.

In spite of this research, inference on the PBR has rarely been applied in practice. This is unfortunate in our view, because it offers several advantages over more commonly applied analyses strategies, such as calculating response rates in all patients and, thereafter, the time-to-event endpoint 'duration of response' in the subset of patients who responded [8]. Aggregating two time-to-event endpoints, namely time to first response (i.e., entering state 1) and time from first response to subsequent failure (i.e., entering state 2 from state 1), PBR combines several endpoints into one clinically meaningful measure for all patients. Comparison between treatment curves can be based on a visual inspection of the respective curves or its difference over time. An example is provided in figure 3 of section 4.

We suspect that one of the reasons for the poor uptake of the PBR in practice may be the lack of a corresponding statistical test. Statistical tests for the comparison of treatment and control have been suggested. For example, [11–13] have investigated tests for equal restricted mean duration of response (RMDOR). Additional tests for duration of response are discussed by [6]. Furthermore, tests for areas under the PBR curve (e.g., via a re-randomization approach), as well as logrank tests of the time from entering the response state until death or censoring are straightforward to derive. All of these tests, however, test null hypotheses of the equality of some derived quantity of two PBR curves, not directly the null hypothesis of equality of the entire curve. We are not aware that a statistical test for complete equality of two PBR curves, closely mimicking the logrank test, is readily available in the literature. We aim to fill this gap here.

The paper is structured as follows: In section 2, we briefly review the non-parametric estimation of PBR, introduce three statistical tests to compare two PBR curves and give some intuition regarding the derivation of their test statistics' distributions.

In section 3, the suggestions are investigated by simulations and in section 4, we apply the tests to data from a clinical phase 3 study in oncology. A discussion concludes the main text. A more formal derivation of the distributions in section 2 is given in the Appendix.

2. Proposed tests to compare PBR curves

Assume that $S_i(t) \in \{0, 1, 2\}$ denotes the state of patient i at time t of the trial where $S_i(0) = 0$ for all patients. Let $T_{s_1 s_2, i}$ be the time of transition of patient i from state s_1 into state s_2. We assume that patients are assumed to differ only by their treatment and are mutually stochastically independent, Hence we have that $\left(T^{(j)}_{01,i}, T^{(j)}_{02,i}, T^{(j)}_{12,i}\right) = \left(T^{(j)}_{01}, T^{(j)}_{02}, T^{(j)}_{12}\right)$ are i.i.d. and that $S_i(t) = S^{(j)}(t)\ \forall\ t$ for all patients in treatment $j \in \{0, 1\}$. We derive a statistical test for the null hypothesis

$$H_0 : F\left(T^{(0)}_{01}, T^{(0)}_{12}\right) = F\left(T^{(1)}_{01}, T^{(1)}_{12}\right) \tag{2.1}$$

of complete equality of the multivariate distribution $F(t_0, t_1)$ of the transition times $0 \to 1$ and $1 \to 2$. This hypothesis implies the slightly more general hypothesis

$$H^*_0 : P\left(S^{(0)}(t) = 1\right) = P\left(S^{(1)}(t) = 1\right) \text{ for all } t. \tag{2.2}$$

Hypothesis H^*_0 can in theory be fulfilled when H_0 is false. This would happen if the marginal distributions of transition times are the same, but the relations between transition times are different between treatments, e.g., if T_{01} and $T_{12} - T_{01}$ are independent in one treatment arm, but highly correlated in the other. In practice, it is difficult to conceive how such a case could arise, but for formal reasons, we must exclude it. Patients can be in one of three states: 0=stable disease, but not in response, 1=in response, 2=absorbing state (death or loss of response, e.g., disease progression). Regarding the alternative hypothesis, we will for the moment simply assume that this is "not H_0". We will return to this subsequently.

2.1. Notation

We will use the following notation which is similar to notation used for the logrank test in textbooks such as [17]:

- n_j, $j = 0, 1$ patients per treatment group j, $n = n_0 + n_1$.
- Observed event times $t_{s_1 s_2 k}$ where $s_1 = 0, 1$ denotes the state out of which patients transition to state $s_2 = 1, 2$. All patients start in state 0 at time $t_0 = 0$, i.e. for patient i T_{12i} is the total time elapsed from $0 \to 1 \to 2$, whereas the sojourn time in state 1 is $T_{12i} - T_{01i}$ (if not censored in state 1). For convenience, enumeration $k = 1, \ldots, K$ is consecutive, i.e. the number $d_{s_1 s_2 k}$ of transitions can be 0 for a given time t_k. Each $t_{s_1 s_2 k}$ corresponds to one or several observations of $T_{s_1 s_2 i}$ from one or several patients. Only transitions $0 \to 1$ and $1 \to 2$ are enumerated; transitions $0 \to 2$ do not need to be enumerated, as will become clear below. If different types of transitions occur at the same time, we will consider them in the order t_{12k}, t_{01k}. We do this so as to avoid a situation where a patient can have trajectory $0 \to 1 \to 2$ with sojourn time duration 0 in state 1.
- $d^{(j)}_{s_1 s_2 k}$ the number of patients from group j who transition from state s_1 into state s_2 at time $t_{s_1 s_2 k}$, $d_{s_1 s_2 k} = d^{(0)}_{s_1 s_2 k} + d^{(1)}_{s_1 s_2 k}$.
- $r^{(j)}_{sk}$ the number of patients at risk in state $s = 0, 1$ and group j with $r_{sk} = r^{(0)}_{sk} + r^{(1)}_{sk}$. For state 0, this number decreases as patients transition into state 1 or 2 over time. For state 1, the number increases when patients enter state 1 and decreases when they leave. If there are no censorings and no $0 \to 2$ transitions, updates are $r^{(j)}_{0(k+1)} = r^{(j)}_{0k} - d^{(j)}_{01k}$ and $r^{(j)}_{1(k+1)} = r^{(j)}_{1k} - d^{(j)}_{12k} + d^{(j)}_{01k}$.

Following the common convention in survival analysis, patients who transition out of a state s at time t or who are censored in state s at time t are assumed to be part of the corresponding risk set at time t. Likewise, $0 \to 2$ transitions at time t_k decrease $r_{0(k+1)}$, not r_{0k}.

2.2. Estimating the probability of being in response

Before introducing the test, we briefly review non-parametric estimation of the probability of being in response (PBR). The principles are very similar to those used for constructing the famous Kaplan-Meier-estimate [14].

Let $t_1 \leq ... \leq t_{k-1} \leq t_k \leq t_{k+1} ... \leq t_K$ denote all time points where at least one transition $0 \to 1$,$1 \to 2$ or $0 \to 2$ is observed (for simplicity we use the abridged notation t_k instead of $t_{s_1 s_2 k}$). Then the PBR $P\left(S^{(j)}(t) = 1\right)$ for $t_{k-1} \leq t < t_k$ is recursively estimated by

$$\hat{P}\left(S^{(j)}(t) = 1\right) = \hat{P}\left(S^{(j)}(t_{k-1}) = 1\right)\left(1 - \frac{d^{(j)}_{12k}}{r^{(j)}_{1k}}\right) + \hat{P}\left(S^{(j)}(t_{k-1}) = 0\right)\frac{d^{(j)}_{01k}}{r^{(j)}_{0k}} \tag{2.3}$$

with $d^{(j)}_{s_1 s_2 k}$ and $r^{(j)}_{sk}$ defined as above. Furthermore we define $\frac{0}{0} =: 0$, at the study start time $t_0 < t_1$ no patient is in state 1 or 2, i.e., $\hat{P}\left(S^{(j)}(t_0) = 1\right) = 0$, $\hat{P}\left(S^{(j)}(t_0) = 2\right) = 0$ and $\hat{P}\left(S^{(j)}(t_0) = 0\right) = 1$. The probability of being in the initial state $S = 0$ for $t_{k-1} \leq t < t_k$ is obtained by

$$\hat{P}\left(S^{(j)}(t) = 0\right) = \hat{P}\left(S^{(j)}(t_{k-1}) = 0\right)\left(1 - \frac{d^{(j)}_{01k} + d^{(j)}_{02k}}{r^{(j)}_{0k}}\right)$$

Thus, at each time point t, the PBR function $\hat{P}\left(S^{(j)}(t) = 1\right)$ is obtained by adding the probability that a patient was in state 0 at the previous time point and enters state 1 (=responder) at time t and the probability that a patient was in state 1 at the previous time point and is still in state 1 at time t (i.e., did not transition to state 2). Transitions from $0 \to 2$ before t_k impacts the PBR curve by reducing the risk set $r^{(j)}_{0k}$ only, the steps at which the PBR curves increases or decreases (i.e., the event times) occur at $0 \to 1$ or $1 \to 2$ transitions. An example of the resulting curve is given in Figure 3. The variance of $\hat{P}\left(S^{(j)}(t) = 1\right)$ can be approximated in a very similar way as is done for the simple Kaplan-Meier-curve using Greenwood's formula [14, 18].

We would like to emphasize here that the PBR could more precisely be called "probability of being in response conditional on not having reached the absorbing state already". Suppose we are at event time t where one $0 \to 1$ transition and no other transitions are observed. The probability of instantaneous risk of a $0 \to 1$ transition is then estimated as 1 divided by the number of patients in state 0 at time t. Assume the absorbing state is death. Then with this definition, the underlying true PBR curves of two treatments can be the same, even if the death rates in the two groups are not the same. This is similar to inference on the cause-specific hazard where events on competing risks are also indistinguishable from censored observations if we test for equality of the cause-specific hazard of two survival curves (see e.g., [4]).

In the subsequently derived tests, this behavior is reflected in the fact that $0 \to 2$-transitions and observations censored in state 0 are treated identically. This is consistent with the estimated PBR curves which also converge to the same underlying true PBR curve if the hazards of $0 \to 1$ and of $1 \to 2$ are the same in the two groups, respectively, even if the hazard of death is different.

2.3. Suggested tests

In analogy to the estimation of PBR with a Kaplan-Meier-style approach, we now derive tests for the null hypothesis (2.1). These tests can be viewed as generalized versions of the logrank test.

To derive the test statistic, we condition on $t_{s_1s_2k}, d_{s_1s_2k}, r^{(j)}_{s_1k}$ (time of event, total number of events and patients at risk per group). Consequently, $d^{(1)}_{s_1s_2k}$ is a random variable with $d^{(1)}_{s_1s_2k} \sim Hyp(r_{s_1k}, d_{s_1s_2k}, r^{(1)}_{s_1k})$ as its null distribution. Thus,

$$E(d^{(1)}_{s_1s_2k}) = d_{s_1s_2k} \cdot \frac{r^{(1)}_{s_1k}}{r_{s_1k}} \tag{2.4}$$

and

$$var(d^{(1)}_{s_1s_2k}) = d^{(1)}_{s_1s_2k} \cdot \frac{r^{(1)}_{s_1k}}{r_{s_1k}} \frac{r_{s_1k} - r^{(1)}_{s_1k}}{r_{s_1k}} \frac{r_{s_1k} - d_{s_1s_2k}}{r_{s_1k} - 1}. \tag{2.5}$$

$(d^{(1)}_{s_1s_2k} \,|d_{s_1s_2k}, r^{(0)}_{s_1k}, r^{(1)}_{s_1k})$ and $(d^{(1)}_{s_1s_2k'} \,|d_{s_1s_2k'}, r^{(0)}_{s_1k'}, r^{(1)}_{s_1k'})$ are stochastically independent if $k \neq k'$. For the transitions $0 \to 1$, this follows from the fact that $(d^{(1)}_{01k} \,|d^{(1)}_{01(k-1)}, d_{01k}, r^{(0)}_{0k}, r^{(1)}_{0k})$ has the same distribution as $(d^{(1)}_{01k} \,|d_{01k}, r^{(0)}_{0k}, r^{(1)}_{0k})$ since $d^{(1)}_{01k}$ depends on $d^{(1)}_{01(k-1)}$ only via $r^{(0)}_{0k}$ and $r^{(1)}_{0k}$ which we already conditioned on. For the transitions $1 \to 2$, the same logic can be applied: $(d^{(1)}_{12k} \,|d^{(1)}_{12(k-1)}, d^{(1)}_{01(k-1)}, d_{12k}, r^{(0)}_{1k}, r^{(1)}_{1k})$ has the same hypergeometric distribution $Hyp(r_{1k}, d_{12k}, r^{(1)}_{1k})$ as $(d^{(1)}_{12k} \,|d_{12k}, r^{(0)}_{1k}, r^{(1)}_{1k})$. It is thus stochastically independent of $d^{(1)}_{12(k-1)}$ since the influence of $d^{(1)}_{12(k-1)}$ is only via $r^{(1)}_{1k} = r^{(1)}_{1(k-1)} - d^{(1)}_{12(k-1)} + d^{(1)}_{01(k-1)}$ (apart from independent censoring). Note that this only holds under a global null hypothesis where treatment has no influence on patients trajectories at all, i.e., neither the sojourn time in state 1, nor the time to reaching state 1 nor the probability of transitioning into state 2 is at any point in time different between treatment 0 and treatment 1 under H_0.

The sojourn time in state 0 is the same for all remaining patients in this state at all event times (since all patients are in state 0 when recruited), but for patients in state 1, these times are not all the same. Superficially, this seems to call into question the hypergeometric distribution of $d^{(1)}_{12k}$. However, conditioning resolves this concern: The risk set of patients who are in state 1 contains patients who are in this state for different durations of time, but under H_0, the distribution of the durations in this state is identical in the two treatment groups. Hence, the distribution of $(d^{(1)}_{12k} \,|d_{12k}, r^{(0)}_{1k}, r^{(1)}_{1k})$ is not conditional on the individual sojourn times in state 1.* Assume that $\lambda_{12}(t\,|t_{01})$ denotes the conditional hazard rate for leaving state 1 to state 2 at time $t > t_{01}$ where t_{01} is the time of entry into state 1. It may well be that this hazard rate depends on t_{01}. The suggested test, however, does not use this information: $d^{(1)}_{12k}$ is a random variable which disregards individual risks that may vary with t_{01k}.

Special attention is required in the situation where both a $0 \to 1$-transition and a $1 \to 2$-transition are observed in the same patient. In this case, a correlation between $(d^{(1)}_{01k} \,|d_{01k}, r^{(0)}_{00k}, r^{(1)}_{0k})$ and $(d^{(1)}_{12k'} \,|d_{12k'}, r^{(0)}_{1k'}, r^{(1)}_{1k'})$ can be induced. The two quantities are independent if the Markov property

$$P\left(S_i(t) = s_1 \,|\{S_i(u^*)\}_{u^* \in [0,u)}, S_i(u) = s_0\right) = P\left(S_i(t) = s_1 \,|S_i(u) = s_0\right) \tag{2.6}$$

*Exactly the same phenomenon occurs when patients all have different individual probabilities p_i of responding, but we are drawing a simple random sample of size n. The number of responders is then $Bin(n, \bar{p})$-distributed where $\bar{p} = \sum_i p_i/n$. This distribution depends on p_i only via $\bar{p}$.

for all $0 \leq u < t$ and all s_0, s_1 holds. This property implies that knowing the state of time u, we do not need to know the history $[0, u)$ to predict the future. This in turn means that for a patient in state 1 at time u, it is irrelevant at what earlier time before u the patient moved into state 1.

The Markov assumption is commonly made in the literature [1, chapter A.2]. For completeness, we would like to mention here that (2.6) should not be confused with the stronger assumption

$$P(S_i(t+u) = s_1 | S_i(u) = s_0) = P(S_i(t) = s_1 | S_i(0) = s_0) \text{ for all } t, u.$$

This is sometimes called the time-homogeneous Markov assumption [1] and implies constant hazard rates for all transitions.

The (time-inhomogeneous) Markov assumption (2.6) is sufficient, but not necessary for independence of $(d_{01k}^{(1)} | d_{01k}, r_{0k}^{(0)}, r_{0k}^{(1)})$ and $(d_{12k'}^{(1)} | d_{12k'}, r_{1k'}^{(0)}, r_{1k'}^{(1)})$ Asymptotic independence of $(d_{01k}^{(1)} | d_{01k}, r_{0k}^{(0)}, r_{0k}^{(1)})$ and $(d_{12k'}^{(1)} | d_{12k'}, r_{1k'}^{(0)}, r_{1k'}^{(1)})$ may still hold as long as the number of events in every time interval of fixed length approaches infinity for both $0 \to 1$-transitions and $1 \to 2$-transitions. In that case, infinitely many event time observations fall between T_{01i} and T_{12i} from patient i, breaking the predictability of the $d_{12k'}^{(j)}$ at $t_{k'}$ which T_{12i} belongs to from the observed $d_{01k}^{(j)}$ at $t_k < t_{k'}$ to which T_{01i} contributed. This asymptotic independence would, for example, hold for inhomogeneous Poisson-processes where the hazard rates $\lambda_{s_1 s_2}(t) > 0$ for all t. However, this asymptotic may obviously be extremely slow and hence be a problematic assumption to rely on in practice. Alternatively, a robust sandwich estimate of covariance [19, chapters 7 and 8] can be used to approximate the variance of sums of the event counts.

In the following, we investigate three test statistics:

(1) Test relying on independence: If $d_{01k}^{(1)}$ and $d_{12k'}^{(1)}$ are assumed to be independent, then

$$LRT_{ext} = \frac{\sum_{k=1}^{K}(d_{01k}^{(1)} - E(d_{01k}^{(1)})) - \sum_{k=1}^{K}(d_{12k}^{(1)} - E(d_{12k}^{(1)}))}{\sqrt{\sum_{k=1}^{K} var(d_{01k}^{(1)}) + \sum_{k=1}^{K} var(d_{12k}^{(1)})}} \tag{2.7}$$

is approximately distributed according to $N(0,1)$ under H_0 where $E(d_{s_1 s_2 k}^{(1)})$ and $var(d_{s_1 s_2 k}^{(1)})$ are given in formulas (2.4) and (2.5), respectively. H_0 is rejected when $LRT < \Phi^{-1}(\alpha)$ where $\Phi^{-1}(\alpha)$ denotes the α-quantile from the standard normal distribution $N(0,1)$.

(2) A conservative test: The two separate test statistics

$$LRT_{01} = \frac{\sum_{k=1}^{K}(d_{01k}^{(1)} - E(d_{01k}^{(1)}))}{\sqrt{\sum_{k=1}^{K} var(d_{01k}^{(1)})}} \tag{2.8}$$

and

$$LRT_{12} = \frac{\sum_{k=1}^{K}(d_{12k}^{(1)} - E(d_{12k}^{(1)}))}{\sqrt{\sum_{k=1}^{K} var(d_{12k}^{(1)})}} \tag{2.9}$$

both are asymptotically $N(0,1)$-distributed under H_0 since they are sums of independent, centered and standardized random variables. As their correlation cannot be smaller than -1, a conservative test statistic for H_0 arises from

$$LRT_{cons} = \frac{(LRT_{01} - LRT_{12})}{2}, \tag{2.10}$$

rejecting H_0 when $LRT_{cons} < \Phi^{-1}(\alpha)$.
This is a conservative test, because $cov(LRT_{01}, LRT_{12}) \geq -1$ such that $var(LRT_{01} - LRT_{12}) = var(LRT_{01}) + var(LRT_{01}) - 2cov(LRT_{01}, LRT_{12}) \leq 4$. Since it uses an upper bound of $var(LRT_{01} - LRT_{12})$ in the denominator, it keeps the type I error even if the Markov assumption is violated. As shown in the simulations of section 3, however, its actual type I often remains substantially below the nominal α .

(3) A test using a robust variance estimate: As is well known, the logrank test is equivalent to the score test of a Cox-regression model with treatment as the only covariate. This fact can be used to derive an estimate of the variance of $\sum_{k=1}^{K}(d_{01k}^{(1)} - E(d_{01k}^{(1)})) - \sum_{k=1}^{K}(d_{12k}^{(1)} - E(d_{12k}^{(1)}))$ by the methods described in [19], chapter 8. These methods are implemented in statistical software tools such as the R function `coxph` or the SAS procedure PHREG. The variance of $\sum_{k=1}^{K}(d_{01k}^{(1)} - E(d_{01k}^{(1)})) - \sum_{k=1}^{K}(d_{12k}^{(1)} - E(d_{12k}^{(1)}))$ is approximated in this approach in the following way: Using the Cox-model $\lambda_{s_1 s_2}(\boldsymbol{x}, t) = \lambda_{0 s_1 s_2}(\boldsymbol{x}, t)\exp(\boldsymbol{\beta}'\boldsymbol{x})$ with $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_{12})'$ where $\boldsymbol{x} = (0,0,0)'$ for patients in the control treatment group and $\boldsymbol{x} = (1,1,1)'$ in the experimental treatment group, the variance of $\sum_{k=1}^{K}(d_{01k}^{(1)} - E(d_{01k}^{(1)})) - \sum_{k=1}^{K}(d_{12k}^{(1)} - E(d_{12k}^{(1)}))$ is

$$var(\hat{\boldsymbol{\beta}}) = (1,0,-1) \cdot \left(\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{D}\boldsymbol{D}'\mathcal{I}^{-1}(\boldsymbol{\beta})\right)^{-1} \cdot (1,0,-1)' \tag{2.11}$$

where $\mathcal{I}(\boldsymbol{\beta})$ is the information matrix (second derivative of the partial likelihood of the Cox-regression model with respect to $\boldsymbol{\beta}$ in the place of the true $\boldsymbol{\beta}$) and $\boldsymbol{D}$ is the matrix of score residuals. In the Appendix, some further comments are made regarding the derivation of this claim. The suggested test arises from replacing the denominator of LRT_{ext} in formula (2.7) with $\sqrt{\hat{var}(\hat{\boldsymbol{\beta}})}$ from formula (2.11) (with $var(\hat{\boldsymbol{\beta}})$ calculated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$), leading to LRT_{rob}. Again, H_0 is rejected if $LRT_{rob} < \Phi^{-1}(\alpha)$. Notice that since we do not know the true $\boldsymbol{\beta}$, we have replaced it by its estimate from the Cox-model. Alternatively, we could have calculated it under H_0 at $\boldsymbol{\beta} = \boldsymbol{0}$. Asymptotically, this leads to identical results if H_0 is true. We use this form for convenience: It can readily be obtained from a fit of a Cox-model using the R function `coxph`.

All three tests are designed to have power against the alternative

$$A : P\left(S^{(0)}(t) = 1\right) \leq P\left(S^{(1)}(t) = 1\right) \text{ for all } t \text{ with strict inequality somewhere.} \tag{2.12}$$

The tests presented here are in a sense "intrinsically two-sided" since they test a global null hypothesis. However, by leaving out the contributions from $0 \to 2$-transitions and taking the difference of "into response" ($0 \to 1$ transitions) and "out of response" ($1 \to 2$ transitions), they are set up to be sensitive against deviations from the global null pointing towards the alternative (2.12). In the Appendix, we elaborate a bit further on this notion.

3. Simulation study

3.1. Design of the simulation study

To investigate the operating characteristics of the suggested tests, we performed some simulations. Data was generated in the following way: Individual patient data is generated. For patient i in treatment group j,

- t_{01i} is randomly generated from the exponential distribution $Exp(\lambda_{01j})$,
- t_{02i} is randomly generated from the exponential distribution $Exp(\lambda_{02j})$,
- $t_{cens,i}$ is randomly generated from the exponential distribution $Exp(\lambda_{cens})$. This is an independent censoring time.
- $t_{0obs,i} = \min(t_{01i}, t_{02i}, t_{cens,i})$ is recorded as the observed time in state 0 and is flagged as $0 \to 1$, $0 \to 2$ or censored in state 0.
- If $t_{0obs,i} = t_{01i}$, i.e. if the transition to the response state 1 is observed, t_{12i} is generated from $Exp(\lambda_{12j} + \psi_{12j}t_{01i})$. This is the sojourn time in state 1. The rate of leaving state 1 linearily depends on ψ_{12j}. If ψ_{12j} is positive, then patients who reach state 1 earlier have a lower risk of losing state 1 again. The Markov property (2.6) requires $\psi_{12j} = 0$.
- For patients in whom transition to state 1 is observed, $t_{1obs,i} = \min(t_{01i} + t_{12i}, t_{cens,i})$ is recorded as the total time under observation and is flagged as observed $1 \to 2$ transition time or as censored in state 1 if $t_{cens,i} < t_{01i} + t_{12i}$. If a transition $0 \to 2$ is observed or the patient is censored in state 0, then $t_{1obs,i}$ is missing in the simulation dataset.

We implemented the three test statistics from section 2. We investigated different null and alternative scenarios, for each scenario 10000 simulations runs were performed. With N=600 (300 subjects per treatment and control arm, which we refer to as T ($j = 1$) and C ($j = 0$) below) the simulated data reflects the size of a typical study in Oncology. We also considered $N = 200$ (100 per arm) and $N = 400$ (200 per arm) to mimic smaller studies. For all simulations, the size of the tests was fixed as $\alpha = 0.025$ (= 2.5%). The simulations were performed on a high-performance computing platform using RStudio with R 4.1.0 [16]. The simulation code is available in the GitHub repository `https://github.com/glimmek2/PBRestimation`.

Nine scenarios (see Table 1) were considered to investigate the type-I error. Scenarios N1 to N3 uses the same response, i.e., $0 \to 1$ transition rate, but differ with respect to the response loss risk. Scenarios N4 and N5 have a low $0 \to 1$ transition rate and a high response loss risk. Scenarios N6 to N9 have a high $0 \to 1$ transition rate, where response loss risk is high for N6 and N7 and low for N7 and N8. With $\psi_{12} = 0$ only scenarios N2, N5, N7 and N9 fulfill the Markov property (2.6) whereas scenario N3 has the strongest violation of this property. For the sample size N=600 we also investigated the impact of censoring using $\lambda_{cens} = 0.3$ (low censoring) and $\lambda_{cens} = 1.2$ (high censoring), respectively.

Table 1. Null scenarios used for the assessment of the type-I error.

No.	Treatment	λ_{01}	λ_{02}	λ_{12}	ψ_{12}	λ_{cens}
N1	T=C	1	0.6	0.5	0.4	0.3* / 1.2**
N2	T=C	1	0.6	0.5	0	0.3* / 1.2**
N3	T=C	1	0.6	0.5	0.8	0.3* / 1.2**
N4	T=C	0.5	0.6	0.8	0.4	0.3* / 1.2**
N5	T=C	0.5	0.6	0.8	0	0.3* / 1.2**
N6	T=C	1.5	0.6	0.8	0.4	0.3* / 1.2**
N7	T=C	1.5	0.6	0.8	0	0.3* / 1.2**
N8	T=C	1.5	0.6	0.2	0.1	0.3* / 1.2**
N9	T=C	1.5	0.6	0.2	0	0.3* / 1.2**

*low censoring, **high censoring (N=600 only).

Power calculations were performed for 4 different alternative scenarios, the respective simulation parameters are given in Table 2, typical PBR curves are shown in Figure 2. These alternative scenarios were selected to reflect typical and partly challenging situations that may occur in practice. Scenarios A1 and A2 have both a higher response probability and a lower risk of response loss for T versus C, with a moderate difference between PBR curves for A1 and a large difference for A2. In Scenario A3 the $0 \rightarrow 1$ transition rate is lower in T but response, if achieved, is maintained for a long time (i.e., loss of response risk is low). In contrast, the response in C is lost very quickly. PBR curves for A3 are usually crossing between T and C, generating a challenging situation for any statistical test. Scenario A4 reflects the situation that T has a higher response but a quicker loss of response compared to C, such that the PBR curves increases quicker for T and then becomes similar to the PBR curve for C at later time points. Regarding the null scenarios, we investigated the impact of censoring for N=600 using the two different censoring rates.

Table 2. Alternative scenarios used for power calculation.

No.	Treatment	λ_{01}	λ_{02}	λ_{12}	ψ_{12}	λ_{cens}
A1	C	1	0.6	0.5	0.4	0.3* / 1.2**
	T	1.2	0.5	0.3	0.2	0.4* / 1.2**
A2	C	1	0.6	0.5	0	0.3* / 1.2**
	T	1.5	0.6	0.2	0.1	0.3* / 1.2**
A3	C	1.4	0.6	0.7	0.2	0.3* / 1.2**
	T	0.9	0.6	0.15	0.1	0.3* / 1.2**
A4	C	0.8	0.6	0.4	0.3	0.3* / 1.2**
	T	1.5	0.6	0.7	0.2	0.3* / 1.2**

*low censoring, **high censoring (N=600 only).

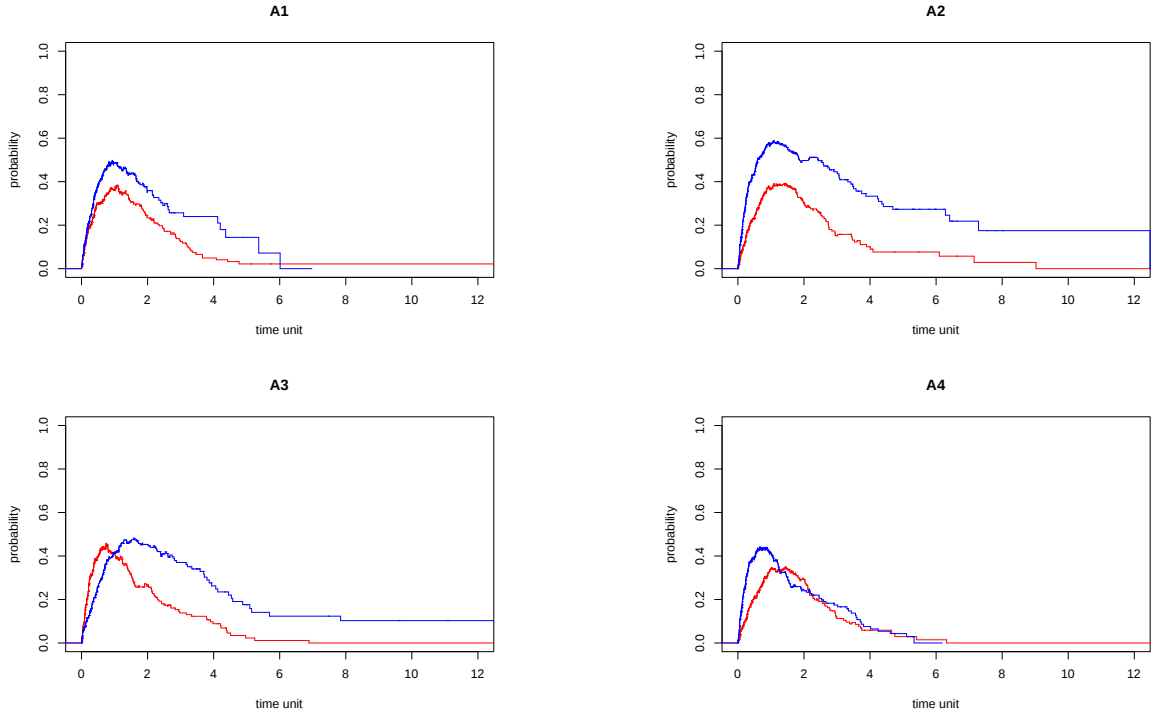

Figure 2. Typical PBR curves for scenarios A1 to A4 from one simulation run (T=blue, C=red)

3.2. Simulation results

For the null model the average number of $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions are equal between T and C (Table 3). The type-I error is maintained in all scenarios fulfilling the Markov property. A slight inflation for LRT_{ext} is seen for those scenarios for which the Markov property is violated (high value of ψ_{12}). The highest inflation is observed for scenario N3. The inflation can partly be corrected by applying the robust variance estimate used in LRT_{rob}. LRT_{cons} is very conservative in all scenarios. When a high censoring rate is assumed, the number of events (for both, the $0 \rightarrow 1$ transition and the $1 \rightarrow 2$ transition) is decreasing. Again, the type-I error is maintained if the Markov property is fulfilled and slightly inflated otherwise.

Table 3. Average number of transitions and type 1 error for the scenarios N1–N9 (nominal $\alpha = 2.5\%$).

Scenario	No. 0 → 1 transitions		No. 1 → 2 transitions		Type I error (%)		
	mean in T	mean in C	mean in T	mean in C	LRT_{ext}	LRT_{rob}	LRT_{cons}
N=600 with 300 subjects in each arm, low censoring							
N1	157.9	158.0	109.4	109.6	3.29	3.08	0.59
N2	157.9	158.0	98.6	98.9	2.59	2.52	0.34
N3	157.9	158.0	115.5	115.8	3.69	3.41	0.72
N4	107.2	107.2	86.3	86.2	3.39	3.55	0.60
N5	107.2	107.2	78.0	77.9	2.28	2.20	0.31
N6	187.5	187.6	140.6	140.9	2.95	2.77	0.47
N7	187.5	187.6	136.3	136.5	2.56	2.53	0.37
N8	187.5	187.6	80.7	80.8	2.51	2.26	0.37
N9	187.5	187.6	75.0	75.1	2.21	2.17	0.34
N=600 with 300 subjects in each arm, high censoring							
N1	107.2	107.2	36.9	37.0	2.57	2.50	0.38
N2	107.2	107.2	31.5	31.5	2.43	2.33	0.28
N3	107.2	107.2	41.2	41.3	2.93	2.69	0.43
N4	65.3	65.2	31.3	31.4	2.99	2.96	0.46
N5	65.3	65.2	26.1	26.2	2.59	2.51	0.31
N6	136.3	136.4	57.6	57.7	2.56	2.41	0.38
N7	136.3	136.4	54.5	54.6	2.49	2.44	0.43
N8	136.3	136.4	21.1	21.1	2.37	2.28	0.27
N9	136.3	136.4	19.5	19.5	2.38	2.29	0.29
N=400 with 200 subjects in each arm, low censoring							
N1	105.4	105.3	73.1	72.9	3.21	2.96	0.52
N2	105.4	105.3	65.9	65.8	2.46	2.35	0.26
N3	105.4	105.3	77.2	77.1	3.73	3.22	0.66
N4	71.3	71.4	57.3	57.5	3.48	3.68	0.51
N5	71.3	71.4	51.8	51.9	2.36	2.19	0.23
N6	125.1	125.0	93.9	93.8	2.92	2.76	0.40
N7	125.1	125.0	91.0	90.9	2.43	2.32	0.28
N8	125.1	125.0	53.8	53.8	2.77	2.44	0.34
N9	125.1	125.0	50.0	50.1	2.52	2.37	0.23
N=200 with 100 subjects in each arm, low censoring							
N1	52.6	52.6	36.5	36.5	3.58	3.16	0.73
N2	52.6	52.6	32.9	32.9	2.80	2.59	0.32
N3	52.6	52.6	38.5	38.5	4.20	3.52	0.83
N4	35.7	35.7	28.7	28.8	3.65	3.46	0.69
N5	35.7	35.7	25.9	26.0	2.71	2.32	0.31
N6	62.5	62.5	46.9	46.9	2.97	2.74	0.43
N5	62.5	62.5	45.5	45.4	2.68	2.50	0.35
N8	62.5	62.5	26.9	26.9	2.84	2.52	0.42
N9	62.5	62.5	25.0	25.1	2.74	2.60	0.31

Power simulations results are summarized in Table 4. For scenario A2 all 3 tests achieved a high power in all simulation scenarios, even with a high censoring rate or sample size of N=200. Scenario A1 had high simulated power for LRT_{ext} and LRT_{rob} if N=400 and N=600 (low censoring). As expected, the power is generally lower with smaller sample size or lower number of events.

Table 4. Average number of transitions and power for the scenarios A1–A4.

Scenario	No. $0 \to 1$ transitions		No. $1 \to 2$ transitions		Power (%)		
	mean in T	mean in C	mean in T	mean in C	LRT_{ext}	LRT_{rob}	LRT_{cons}
N=600 with 300 subjects in each arm, low censoring							
A1	171.4	158.0	82.6	109.6	96.3	95.6	90.1
A2	187.5	158.0	80.7	98.9	100	100	100
A3	150.0	182.7	61.0	131.1	39.3	39.4	64.8
A4	187.5	141.2	134.0	93.2	85.2	84.4	41.9
N=600 with 300 subjects in each arm, high censoring							
A1	124.1	107.2	28.4	37.0	63.3	62.6	44.5
A2	136.3	107.2	21.1	31.5	98.4	98.4	94.3
A3	100.0	131.3	13.7	50.3	0.3	0.4	3.5
A4	136.3	92.3	52.0	28.4	91.1	91.1	31.6
N=400 with 200 subjects in each arm, low censoring							
A1	114.4	105.3	55.1	72.9	86.2	85.1	69.8
A2	125.1	105.3	53.8	65.8	99.9	99.8	99.3
A3	100.1	121.8	40.7	87.3	27.6	27.5	42.0
A4	125.1	94.1	89.5	62.0	68.3	67.0	25.4
N=200 with 100 subjects in each arm, low censoring							
A1	57.1	52.6	27.5	36.5	58.7	56.0	34.9
A2	62.5	52.6	26.9	32.9	93.8	93.3	81.3
A3	50.0	60.8	20.3	43.6	15.4	15.2	17.3
A4	62.5	47.0	44.7	31.0	42.4	40.5	11.1

Overall, LRT_{ext} and LRT_{rob} should be preferred due to higher power versus LRT_{cons}. If PBR clearly differ between treatment arms over the entire study period LRT_{cons} might be useful to formally confirm this difference. All tests become difficult to interpret in case of crossing PBR curves with higher PBR for T versus C for one time period and C better than T for another time period (as in scenario A3).

4. Application to the REACH3 study

REACH3 was an open-label randomized controlled phase 3 study investigating the efficacy and safety of ruxolitinib (RUX) versus Best Available Therapy (BAT) in 329 patients 12 years or older with glucocorticoid-refractory or dependent cGvHD. Details on the study design and clinical results were reported in [22], see also ClinicalTrials.gov number NCT03112603. We illustrated the benefits of the probability of the being in response (PBR) function focusing mainly on the graphical presentation in a recent paper [10]. For treatment comparison, we calculated and displayed the difference of PBR

curves (ruxolitinib minus BAT) with pointwise 95% confidence intervals. In this paper we use the REACH3 data to illustrate the application of the proposed tests.

Figure 3 shows the PBR curves for RUX and BAT with pointwise 95% confidence intervals and Table 5 describes the calculation of the items required for the LRT tests for the first few time-points. For simplicity, we used only LRT_{ext} and LRT_{cons}. Using the sums displayed in the last row of Table 5 and the respective formulas from section 2, we obtain $LRT_{ext} = 4.5489$ ($p < 0.0001$) and $LRT_{cons} = 3.5005$ ($p = 0.0002$), which confirms the clear difference observed in the PBR curves for RUX vs BAT.

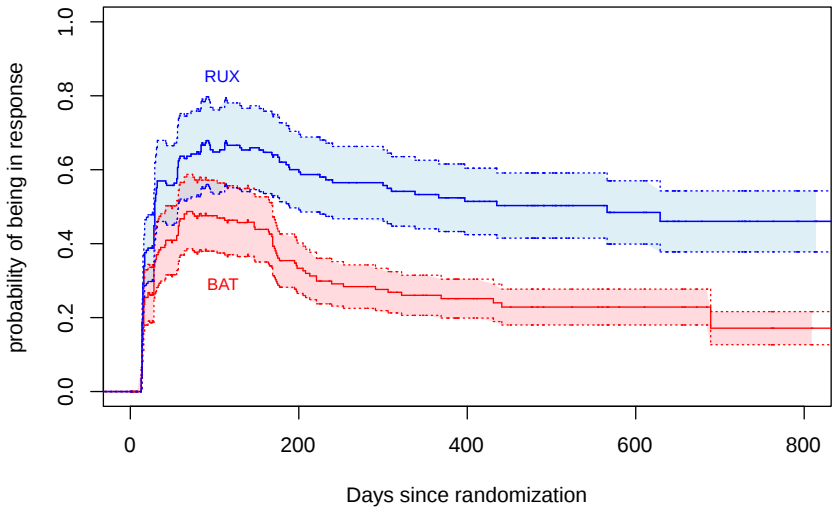

Figure 3. PBR curves with 95%CIs for REACH3.

Table 5. Steps to calculate the LRT test statistic for REACH3.

time	RUX arm ($j = 1$)				All patients				calculated for Rux arm ($j = 1$)					
k	$d^{(1)}_{01k}$	$d^{(1)}_{12k}$	$r^{(1)}_{0k}$	$r^{(1)}_{1k}$	d_{01k}	d_{12k}	r_{0k}	r_{1k}	$E(d^{(1)}_{01k})$	$E(d^{(1)}_{12k})$	Diff01	Diff12	$var(d^{(1)}_{01k})$	$var(d^{(1)}_{12k})$
0	0	0	165	0	0	0	329	0	.	.	.	.	.	.
13	6	0	156	0	12	0	304	0	6.1579	.	-.15789	.	2.88909	.
14	25	0	150	6	41	0	292	12	21.0616	.00000	3.93836	.00000	8.83443	.00000
15	17	0	124	31	23	0	249	53	11.4538	.00000	5.54618	.00000	5.23983	.00000
16	8	0	107	48	19	1	226	76	8.9956	.63158	-.99558	-.63158	4.35768	.23269
17	- 6	0	99	56	9	0	206	94	4.3252	.00000	1.67476	.00000	2.15893	.00000
19	1	0	93	62	1	0	197	103	.4721	.00000	.52792	.00000	.24922	.00000
21	1	0	92	63	3	0	195	104	1.4154	.00000	-.41538	.00000	.73991	.00000
22	0	0	91	64	1	1	192	107	.4740	.59813	-.47396	-.59813	.24932	.24037
									...continue					
									Sum over all time points k		17.2589	-22.2552	52.1637	23.2916

Diff01=$d^{(1)}_{01k} - E(d^{(1)}_{01k})$, Diff12=$d^{(1)}_{12k} - E(d^{(1)}_{12k})$.

5. Discussion

In this paper we propose three test statistics to compare PBR between two treatment arms. The PBR function aggregates two time to event-variables, time from study start to first response and time from first response to subsequent treatment failure. All three test statistics can be considered as an extension of the well known logrank test which is applied to compare right censored survival curves. Similar to the logrank test the derivation of the distribution of the test statistics is based on conditional probabilities of entering the response state given the risk sets in the treatment arms at each event time. In addition, the risks sets for leaving the response state at the event times are also considered. The type-I error and power are investigated in a simulation study and we illustrate the application using the data of REACH3, a clinical phase 3 study.

We apply the framework of multistate models to highlight that this application is a special case of a much more flexible general approach which accommodates more sophisticated model building. However, we also derive the properties of the suggested test statistics in close analogy with the logrank test to elucidate the connections of the general theory with simple and intuitively appealing ideas regarding the statistical inference on important scientific questions. Here, this is the question of differences between the time profiles of the probability of being in response. We believe that this can help to bridge a gap between the impressive theoretical work that has been done on stochastic state-space processes and the concrete application of these ideas in practice.

The multistate model considered in this paper refers to the classical illness-death model and is in alignment with the efficacy endpoints of the REACH3 study, allowing only one transition from the initial state 0 to the response state 1, not a return to the "neutral" state 0. The PBR can be easily extended by allowing both, $0 \to 1$ and $1 \to 0$ transitions with several switches over time. At least under the Markov assumption (2.6), it is straightforward to generalize the suggested tests to this situation. We did not do this here to keep notation simple and render the approach more accessible to readers.

The suggested tests ignore the probability of reaching the absorbing state (death, say) directly from the initial state (cf. section 2.2). For a test which is sensitive to differences in this "immediate death rate", we could keep the $0 \to 2$ patients in the risk set of state 0 forever. For an event time t with a single $0 \to 1$-transition and no other transitions, the estimate of instantaneous risk would then be 1 divided by the number of patients who either are still in state 0 at time t or who have died without ever reaching state 1 before time t. This would estimate an "unconditional probability of being in response". A PBR curve with more deaths would generally stay lower than one with identical conditional PBRs, but less deaths. The logic behind this would be that a patient in state 2 is not "in response" at time t and will never get into the response state 1, but nevertheless should impact the estimate of PBR.

We would like to emphasize that the tests we are suggesting here are certainly not the only useful tests in this situation. [11–13] present useful tests for the equality of restricted mean duration of response (RMDOR). These tests do not need the Markov assumption but require the determination of a truncation time. A direct comparison to our suggestions is difficult as the tests proposed by Huang and colleagues test a somewhat different hypothesis about the equality of the expected durations of response. Furthermore, as the setting of this truncation time ultimately is arbitrary, it remains unclear what specific version of an RMDOR test one should compare to in power simulations.

Furthermore, a logrank test of Kaplan-Meier-curves which start at time of entry into state 1 and end at entry into state 2 could be applied. Such a test could assign an event time of 0 to patients who

never reach state 1 (either because they directly enter state 2 from state 0 or because they do not reach response within a pre-defined period of time). Event time 0 would otherwise be treated as one of the event times at which contributions to the log-rank test statistic are calculated (also including patients who are censored in state 0). In this way, all patients affect the test statistic and any bias from selection of responders would be avoided. Finally, re-randomization tests offer an attractive, easy option for testing duration of response. They could be applied to the RMDOR as well as to other statistics such as the Wilcoxon test statistic for censored data. We have not included all these options in an exhaustive simulation study to keep the focus of this paper on a method which mimics the logrank test as closely as possible and on the derivation of the corresponding test statistic.

We suspect that the suggested tests also keep the type I error asymptotically in certain situations where the Markov condition (2.6) does not hold. A precise characterization of these situations requires clarification of both the kind of asymptotics used as well as the types of deviations from the Markov assumption. This is a topic for future research.

Our suggestions are motivated by a desire to have high power against the alternative of consistent inferiority of one treatment over the entire time axis. Again, there is an analogy with the logrank test:

- High power is achieved for event rate ratios (of the event types "going into state 1 = response" and "going out of state 1") between the treatments that are constant in time (and the test treatment is the better one).
- The type I error is preserved if the test treatment is consistently no better than the control treatment over the entire time axis.
- If event rate ratios vary in time, and are above 1 for some periods and below 1 for others, there are no statistical guarantees of the tests' properties. They will still tend towards "correct" decisions in the sense that the treatment with larger area under the PBR curve will in stochastic tendency be favored, but this does no longer come with formal guarantees of type I error control or high power.

Due to these limitations, we recommend to treat the suggested tests as a complement to a visual display of the PBR curves. This will protect against misinterpretation of significant results. In this respect, we are adopting the approach that most practitioners of clinical trials follow anyways regarding the interpretation of logrank test results and the results from fitting Cox regression models to time-to-event data.

In summary, we believe that the suggested tests support the proper interpretation of PBR graphs such as Figure 3 by indicating whether these are "truly" separated or not.

Author contributions

Both authors contributed equally to all aspects of this work.

Use of Generative-AI tools declaration

We have used ChatGPT in an assistive manner for spell checking and for debugging Latex and R codes.

Acknowledgements

We thank Jan Beyersmann for discussions about the topic. We are also grateful to three anonymous reviewers whose insightful comments led to several improvements of the manuscript.

Conflict of interest

Both authors are employees of Novartis Pharma AG and hold shares in the company.

References

1. O. Aalen, O. Borgan, H. Gjessing, *Survival and event hisftory analysis: A process point of view*, New York: Springer, 2008. https://doi.org/10.1007/978-0-387-68560-1
2. P. K. Andersen, O. Borgan, R. D. Gill, N. Keiding, *Statistical models based on counting processes*, New York: Springer, 1993. https://doi.org/10.1007/978-1-4612-4348-9
3. P. K. Andersen, S. Esbjerg, T. I. A. Sorensen, Multi-state models for bleeding episodes and mortality in liver cirrhosis, *Stat. Med.*, **19** (2000), 587–599. https://doi.org/10.1002/(sici)1097-0258(20000229)19:4¡587::aid-sim358¿3.0.co;2-0
4. P. C. Austin, D. S. Lee, J. P. Fine, Introduction to the analysis of survival data in the presence of competing risks, *Circulation*, **133** (2016), 601–609. https://doi.org/10.1161/CIRCULATIONAHA.115.017719
5. C. B. Begg, M. Larson, A study of the use of the probability-of-being-in-response function as a summary of tumor response data, *Biometrics*, **38** (1982), 59–66. https://doi.org/10.2307/2530288
6. Y. Cui, B. Huang, L. Mao, H. Uno, L. J. Wei, L. Tian, Inferences for the distribution of the duration of response in a comparative clinical study, *Clin. Trials*, **21** (2024), 541–552. https://doi.org/10.1177/17407745241264188
7. M. F. Danzer, A. Faldum, T. Simon, B. Hero, R. Schmidt, Confirmatory adaptive designs for clinical trials with multiple time-to-event outcomes in multi-state markov models, *Biometrical J.*, **66** (2024), e202300181. https://doi.org/10.1002/bimj.202300181
8. E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, et al., New response evaluation criteria in solid tumours: revised recist guideline (version 1.1), *Eur. J. Cancer*, **45** (2009), 228–247. https://doi.org/10.1016/j.ejca.2008.10.026
9. S. Ellis, K. J. Carroll, K. Pemberton, Analysis of duration of response in oncology trials, *Contemp. Clin. Trials*, **29** (2008), 456–465. https://doi.org/10.1016/j.cct.2007.10.008
10. N. Hollaender, E. Glimm, J. Gauvin, T. Stefanelli, R. Zeiser, A novel approach to visualize clinical benefit of therapies for chronic graft versus host disease (cGvHD): The probability of being in response (PBR) applied to the REACH3 study, *Bone Marrow Transpl.*, **59** (2024), 12–16. https://doi.org/10.1038/s41409-023-02128-8
11. B. Huang, L. Tian, Utilizing restricted mean duration of response for efficacy evaluation of cancer treatments, *Pharm. Stat.*, **21** (2022), 865–878. https://doi.org/10.1002/pst.2198

12. B. Huang, L. Tian, Z. McCaw, X. Luo, E. Talukder, M. Rothenberg, et al., Analysis of response data for assessing treatment effects in comparative clinical studies, *Ann. Intern. Med.*, **173** (2020), 368–374. https://doi.org/10.7326/M20-0104

13. B. Huang, L. Tian, E. Talukder, M. Rothenberg, D. Kim, L. Wei, Evaluating treatment effect based on duration of response for a comparative oncology study, *JAMA Oncol.*, **4** (2018), 874–876. https://doi.org/10.1001/jamaoncol.2018.0275

14. E. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481. https://doi.org/10.1080/01621459.1958.10501452

15. A. Nießl, A. Allignol, J. Beyersmann, C. Mueller, Statistical inference for state occupation and transition probabilities in non-markov multi-state models subject to both random left-truncation and right-censoring, *Economet. Stat.*, **25** (2023), 110–124. https://doi.org/10.1016/j.ecosta.2021.09.008

16. R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021.

17. P. J. Smith, *Analysis of failure and survival data*, New York: Chapman & Hall/CRC, 2002. https://doi.org/10.1201/9781315273150

18. N. Temkin, An analysis for transient states with application to tumor shrinkage, *Biometrics*, **34** (1978), 571–580. https://doi.org/10.2307/2530376

19. T. M. Therneau, P. M. Grambsch, *Modeling survival data: Extending the Cox model*, Statistics, 2000.

20. W. Y. Tsai, X. Luo, J. Crowley, The probability of being in response function and its applications, In: *Frontiers of biostatistical methods and applications in clinical oncology*, Singapore: Springer, 2017. https://doi.org/10.1007/978-981-10-0126-0_10

21. A. A. Tsiatis, The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time, *Biometrika*, **68** (1981), 311–315. https://doi.org/10.2307/2335832

22. R. Zeiser, N. Polverelli, R. Ram, S. Hashmi, R. Chakraverty, J. Middeke, et al., Ruxolitinib for glucocorticoid-refractory chronic graft-versus-host disease, *N. Engl. J. Med.*, **385** (2021), 228–238. https://doi.org/10.1056/NEJMoa2033122

Appendix

Derivation of the distibution of the test statistics

Several authors [2,7,21] give a rigorous derivation of tests for the global null of the equality of two multistate stochastic processes using martingale theory.

These publications use the Markov property (2.6). We will also assume it here. In the simulations, we have also investigated cases where this assumption is violated.

The above mentioned papers use formal notation with integrals. We simplify this here to a non-parametric notation replacing integrals by sums. We use the following notation:

- $Z_i \in \{0, 1\}$ is the treatment indicator of patient $i = 1, \ldots, n$.
- $\rho_{si}(t) = 1$ if patient i is in state s at time t, 0 otherwise.
- $r_s(t) = \sum_{i=1}^{n} \rho_{si}(t)$ is the number of patients in state s at time t.
- $r_s^{(1)}(t) = \sum_{i=1}^{n} Z_i \rho_{si}(t)$ is the number of patients in state s at time t in group 1.
- $\delta_{s_1 s_2 i}(t) = 1$ if patient i transitions from state s_1 to state s_2 at time t, 0 otherwise.
- $d_{s_1 s_2}(t) = \sum_{i=1}^{n} \delta_{s_1 s_2 i}(t)$ number of patients who transition $s_1 \to s_2$ at time t.
- $d_{s_1 s_2}^{(1)}(t) = \sum_{i=1}^{n} Z_i \delta_{s_1 s_2 i}(t)$ number of patients who transition $s_1 \to s_2$ at time t in group 1.
- $p_{s_1 s_2}^{(1)}(t) = \frac{r_{s_1}^{(1)}(t)}{r_{s_1}(t)} = p_{s_1}^{(1)}(t)$.

Note that $p_{s_1 s_2}^{(1)}(t)$ is the conditional probability that a randomly selected patient from the risk set corresponding to $r_{s_1}(t)$ of patients in state s_1 who transitions out of state s_1 and into state s_2 at time t is in the treatment group $Z_i = 1$, if we condition on $d_{s_1 s_2}(t)$, $r_{s_1}(t)$ and $r_{s_1}^{(1)}(t)$ and if a global null of identical stochastic processes in the two treatments holds. Under the global null, this probability does not depend on s_2.

This setup leads to the following building blocks for generating tests of the global null. We start by conditioning on the following:

- The event times t_l, $l = 1, \ldots, L \leq n$ where at least one transition occurs.
- $r_s(t_l)$, $r_s^{(1)}(t_l)$ and $d_{s_1 s_2}(t_l)$, the patients at risk (overall and in group 1) immediately before t_l and the total number of transitions $s_1 \to s_2$ at time t_l.

With this conditioning we get the following:

$$d_{s_1 s_2}^{(1)}(t_l) \sim Hyp\left(r_{s_1}(t_l), r_{s_1}^{(1)}(t_l), d_{s_1 s_2}(t_l)\right) \text{ under } H_0.$$

These are all mutually stochastically independent (due to the conditioning) for different t_l's. In addition, $\left(d_{02}^{(1)}(t)\right)_{t=t_{021},\ldots}$ are independent of $\left(d_{01}^{(1)}(t), d_{12}^{(1)}(t')\right)_{t=t_{011},t'=t_{121},\ldots}$. As discussed in section 2, $d_{01}^{(1)}(t)$ and $d_{12}^{(1)}(t')$ are independent under the Markov assumption (2.6) [2, theorem IV.1.2.], and are asymptotically independent in some more general settings too [1, 15]. From the hypergeometric distribution, we have

$$\begin{aligned} E\left(d_{s_1 s_2}^{(1)}(t_l)\right) &= d_{s_1 s_2}(t_l) \cdot p_{s_1}^{(1)}(t_l) \\ var\left(d_{s_1 s_2}^{(1)}(t_l)\right) &= d_{s_1 s_2}(t_l) p_{s_1}^{(1)}(t_l)\left(1 - p_{s_1}^{(1)}(t_l)\right) \cdot \frac{r_{s_1}(t_l) - d_{s_1 s_2}(t_l)}{r_{s_1}(t_l) - 1} \approx \\ & \quad d_{s_1 s_2}(t_l) p_{s_1}^{(1)}(t_l)\left(1 - p_{s_1}^{(1)}(t_l)\right). \end{aligned}$$

The approximation in the last formula is due to the convergence of the hypergeometric to the binomial distribution. If all event times are unique, this approximation becomes exact.

The joint null distribution of the centered, unstandardized event counts in group 1 at the event times t_l then becomes:

$$\mathbf{U}(t_l) := \begin{pmatrix} d_{01}^{(1)}(t_l) - E\left(d_{01}^{(1)}(t_l)\right) \\ d_{02}^{(1)}(t_l) - E\left(d_{02}^{(1)}(t_l)\right) \\ d_{12}^{(1)}(t_l) - E\left(d_{12}^{(1)}(t_l)\right) \end{pmatrix} \sim Hyp\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{V}(t_l)\right)$$

where

$$\mathbf{V}(t_l) = \begin{pmatrix} var\left(d_{01}^{(1)}(t_l)\right) & 0 & 0 \\ 0 & var\left(d_{02}^{(1)}(t_l)\right) & 0 \\ 0 & 0 & var\left(d_{12}^{(1)}(t_l)\right) \end{pmatrix}$$

and $Hyp(\mathbf{0}, \mathbf{V})$ denotes the three hypergeometric distributions, now described by their corresponding expected values, variances and covariances (with slight abuse of notation, since the hypergeometric distribution is not uniquely determined by its first two moments). Hence, for any fixed full rank matrix $\mathbf{A}$, we approximately have

$$\mathbf{AU}(t_l) \sim N\left(\mathbf{0}, \mathbf{AV}(t_l)\mathbf{A}'\right).$$

If $\sum_{l=1}^{L} d_{s_1 s_2}(t_l) \to \infty$ and if $\mathbf{AU}(t_l)$ are stochastically independent at different event times t_l,

$$\sum_{l=1}^{L} (\mathbf{AU}(t_l))' \left(\mathbf{AV}(t_l)\mathbf{A}'\right)^{-1} (\mathbf{AU}(t_l))$$

converges in distribution to a $\chi^2\left(rank(\mathbf{A})\right)$-distribution.

A "standard" approach would be to use $\mathbf{A} = \mathbf{I}_3$ for a global, two-sided test. [7] suggest to use $\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ which produces a bivariate test statistic where the first component can be associated with transitions $0 \to 1$ or $0 \to 2$ (like "PFS" if 1=progression and 2=death) and the second with transitions $0 \to 2$ or $1 \to 2$ ("death" in the example just mentioned).

Here, since our main interest is in PBR, we restrict attention to $0 \to 1$ and to $1 \to 2$ transitions. Large values of the test statistic should arise when patients in group 1 transition into state 1 earlier and leave it later than patients in group 0. This suggests $\mathbf{A} = \mathbf{a}' = (1 \;\; 0 \;\; -1)$. Since this produces a one-dimensional score $\mathbf{a}'\mathbf{U}(t_l)$, $\frac{\sum_l \mathbf{a}'\mathbf{U}(t_l)}{\sqrt{\sum_l \mathbf{a}'\mathbf{V}(t_l)\mathbf{a}}}$ is approximately standard normal $N(0,1)$. It is easy to see that this is in fact LRT_{ext} from formula (2.7). H_0 is rejected at level α if $LRT_{ext} < \Phi^{-1}(\alpha)$ where $\Phi^{-1}(\alpha)$ denotes the α-quantile from the standard normal distribution $N(0,1)$.

As mentioned in section 2, the assumption of independence of $\mathbf{U}(t_{l_1})$ and $\mathbf{U}(t_{l_2})$ may be questionable. To deal with this issue, we notice that $\sum_l \mathbf{U}(t_l)$ is the score statistic (first derivative of the partial likelihood) of a Cox-regression model in the point $\boldsymbol{\beta} = \mathbf{0}$ and $-\sum_l \mathbf{V}(t_l)$ is the information matrix (matrix of second derivatives) at $\boldsymbol{\beta} = \mathbf{0}$. This suggests that in cases where we do not know the correlation between $\mathbf{V}(t_l)$ at different time points, we might estimate the variance of $\sum_l \mathbf{U}(t_l)$ non-parametrically by the well-known sandwich estimation technique [19]. This leads to the form of the test statistic LRT_{rob} described in section 2.

AIMS Press