

# **On the use of surrogate endpoints in adaptive survival trials**

---

**W. Brannath, KKSB and IfS, FB 03**



**Colloquium in Honor of Hans Ulrich Burger on the Occasion  
of His Retirement at Roche, October 16, 2025**

## Why I am here today

---



**“Curiosity Never  
Retires”**

- ▶ We are colleagues and friends for more than 20 years
- ▶ We had always interesting and very fruitful discussions
- ▶ Joint involvement in activities of the IBS-DR and ROeS
- ▶ Involvement of Uli and Kaspar (Roche as industry partner) in the BMBF project BIMIT (with M. Kieser)
- ▶ Joint publications

# Some of Uli's work on surrogate endpoints

Received: 6 August 2017 | Revised: 14 July 2018 | Accepted: 16 July 2018  
DOI: 10.1002/sim.7936

## RESEARCH ARTICLE

WILEY *Statistics  
in Medicine*

## Nonparametric adaptive enrichment designs using categorical surrogate data

Matthias Brückner<sup>1</sup> | Hans U. Burger<sup>2</sup> | Werner Brannath<sup>1</sup>

<sup>1</sup>Competence Center for Clinical Trials and Institute for Statistics, University of Bremen, Bremen, Germany

<sup>2</sup>Hoffmann-La Roche AG, Basel, Switzerland

### Correspondence

Werner Brannath, Competence Center for Clinical Trials and Institute for Statistics, University of Bremen, Faculty 03, Bremen, Germany.

Email: werner.brannath@uni-bremen.de

### Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: 05M13VHC

Adaptive survival trials are particularly important for enrichment designs in oncology and other life-threatening diseases. Current statistical methodology for adaptive survival trials provide type I error rate control only under restrictions. For instance, if we use stage-wise  $P$  values based on increments of the log-rank test, then the information used for the interim decisions need to be restricted to the primary survival endpoint. However, it is often desirable to base interim decisions also on correlated short-term endpoints like tumor response. Alternative statistical approaches based on a patient-wise splitting of the data require unnatural restrictions on the follow-up times and do not permit to efficiently account for an early rejection of the primary null hypothesis. We therefore suggest new approaches that enable us to use discrete surrogate endpoints (like tumor response status) and also to incorporate interim rejection boundaries. The new approaches are based on weighted Kaplan-Meier estimates and thereby have additional advantages. They permit us to account for nonproportional hazards and are robust against informative censoring based on the surrogate endpoint. We will show that nonproportionality is an intrinsic and relevant issue in enrichment designs. Moreover, informative censoring based on the surrogate endpoint is likely because of withdrawals and treatment switches after insufficient treatment response. It is shown and illustrated how nonparametric tests based on weighted Kaplan-Meier estimates can be used in closed combination tests for adaptive enrichment designs, such that type I error rate control is achieved and justified asymptotically.

### KEYWORDS

combination test, flexible design, surrogate endpoint, time-to-event data, type I error rate control

Received: 21 August 2018 | Revised: 15 May 2019 | Accepted: 15 May 2019  
DOI: 10.1002/sim.800250

## RESEARCH PAPER

Biometrical Journal

## A multistate model for early decision-making in oncology

Ulrich Beyer<sup>1</sup> | David Dejardin | Matthias Meller | Kaspar Rufibach |  
Hans Ulrich Burger

Department of Biostatistics, MDDB 663, F. Hoffmann-La Roche Ltd., Basel, Switzerland

### Correspondence

Ulrich Beyer, Department of Biostatistics, MDDB 663, F. Hoffmann-La Roche Ltd., 4070 Basel, Switzerland.  
Email: ulrich.beyer@roche.com

### Abstract

The development of oncology drugs progresses through multiple phases, where after each phase, a decision is made about whether to move a molecule forward. Early phase efficacy decisions are often made on the basis of single-arm studies based on a set of rules to define whether the tumor improves ("responds"), remains stable, or progresses (response evaluation criteria in solid tumors [RECIST]). These decision rules are implicitly assuming some form of surrogacy between tumor response and long-term endpoints like progression-free survival (PFS) or overall survival (OS). With the emergence of new therapies, for which the link between RECIST tumor response and long-term endpoints is either not accessible yet, or the link is weaker than with classical chemotherapies, tumor response-based rules may not be optimal. In this paper, we explore the use of a multistate model for decision-making based on single-arm early phase trials. The multistate model allows to account for more information than the simple RECIST response status, namely, the time to get to response, the duration of response, the PFS time, and time to death. We propose to base the decision on efficacy on the OS hazard ratio (HR) comparing historical control to data from the experimental treatment, with the latter predicted from a multistate model based on early phase data with limited survival follow-up. Using two case studies, we illustrate feasibility of the estimation of such an OS HR. We argue that, in the presence of limited follow-up and small sample size, and making realistic assumptions within the multistate model, the OS prediction is acceptable and may lead to better early decisions within the development of a drug.

### KEYWORDS

clinical trial, decision-making, multistate model

# The work I will focus on



Received: 6 August 2017 | Revised: 14 July 2018 | Accepted: 16 July 2018

DOI: 10.1002/sim.7936

## RESEARCH ARTICLE

WILEY **Statistics**  
in Medicine

# Nonparametric adaptive enrichment designs using categorical surrogate data

Matthias Brückner<sup>1</sup>  | Hans U. Burger<sup>2</sup> | Werner Brannath<sup>1</sup> 

<sup>1</sup>Competence Center for Clinical Trials  
and Institute for Statistics, University of  
Bremen, Bremen, Germany

<sup>2</sup>Hoffmann-La Roche AG, Basel,  
Switzerland

Adaptive survival trials are particularly important for enrichment designs in oncology and other life-threatening diseases. Current statistical methodology for adaptive survival trials provide type I error rate control only under restrictions. For instance, if we use stage-wise  $P$  values based on increments of the

## Content of my talk

---

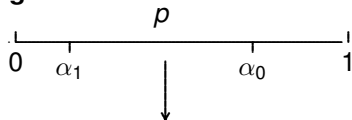
- ▶ Brief introduction to adaptive survival trials (AST)
- ▶ Difficulties with surrogate endpoints in AST
- ▶ Information unrestricted and information restricted adaptive designs
- ▶ Summary and discussion

# **Adaptive Survival Trials with Surrogate Endpoints**

# Confirmatory Adaptive Designs

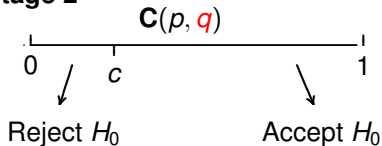
## Combination Tests

### Stage 1



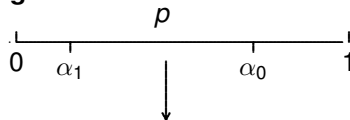
**Adaptations for Stage 2**

### Stage 2



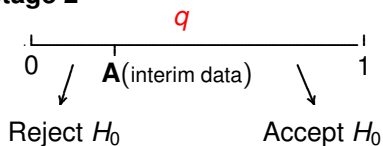
## Conditional Error Function Approach

### Stage 1



**Adaptations for Stage 2**

### Stage 2



## The p-clud condition (BRANNATH ET AL., 2002, LIU & PLEDGER, 2006)

---

Adaptive designs control the type I error rate  $\alpha$  under the following "p-clud" condition:

$$\mathbf{P}_0(q \leq u | p) \leq u \quad \text{for all } 0 \leq u \leq 1 \text{ and all } 0 \leq p \leq 1$$

This condition holds when

- ▶  $q$  is computed from an *independent* second stage cohort with a *conservative test* for the selected second stage design,
- ▶ or more general:  $q$  is *conditionally conservative*, i.e.

$$\mathbf{P}_0(q \leq u | \text{used interim data}) \leq u \quad \text{for all } 0 \leq u \leq 1 \text{ and any interim data.}$$

This condition is useful for overlapping (dependent) first and second stage data!

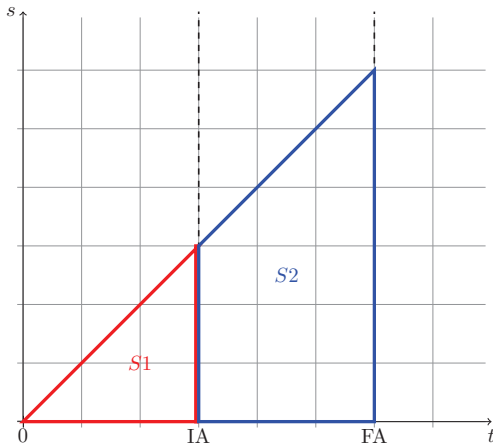


# Adaptive Survival Trials (AST)

---

- ▶ Adaptive designs with primary and possible also secondary time to event endpoints, like overall and/or progression free survival
- ▶ Usually, statistical inference based on logrank test or Cox's proportional hazard model by utilizing right and left truncation (or independent increments)  
→ “follow-up-wise splitting” with stoch. independent stage-wise p-values
- ▶ *Alternative:* Inference based on *restricted mean survival* or *average hazard ratio* (requires finite time horizon; see Brückner, Burger & Brannath, 2018)

# Follow-up-wise separation of stages



- ▶ Stage 1 p-value:  
follow-up **till** IA, i.e. **right** censoring at IA
- ▶ Stage 2 p-value:  
follow-up **from** IA, i.e. **left-truncation** at IA or using increments.
- ▶ P-clud condition via independent increments or left-truncation
- ▶ From the patients censored at the IA, no (other) information can be used for the adaptations.

## Use of surrogate endpoints in AST

---

- ▶ Usually short-term surrogate endpoints (SEP) like response rate (categorical) or progression free survival with OS a primary or co-primary endpoint are available.
- ▶ Surrogate endpoints may be used for ...
  - interim treatment and/or subgroup selection;
  - interim sample size and/or event number reassessment;
  - an interim futility decision in an AST or GSST (= Group Sequential Survival Trial);
  - an interim efficacy testing in an AST or GSST:
    - with the intention for an accelerated approval, or
    - to enhance an early inference for the primary endpoint with a statistical model.

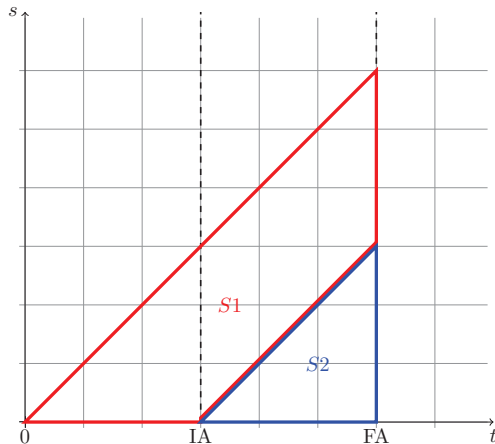
## Difficulties with surrogate endpoints (BAUER & POSCH, 2004)

---

- ▶ Surrogate endpoints (SEP) are usually correlated with the primary endpoint (PEP) (e.g. tumor response with progression free survival)
- ▶ A randomly promising interim result in SEP is indicative for a promising result in the primary endpoint (PE) of the interim patients after the interim analysis  
⇒ a **reduction** in second stage sample size or follow-up time transfers the randomly promising result to the second stage ⇒ **positive** bias and type I error rate **inflation**
- ▶ A randomly unpromising interim result can be **diluted** by **increasing** the second stage sample size or follow-up time ⇒ **positive** bias and type I error rate **inflation**
- ▶ Independent increments no longer guaranteed ⇒ p-clad property difficult to achieve!

# **Information Unrestricted AST with Surrogate Endpoints**

# Patient-wise separation of stages



- ▶ Stage 1 p-value:  
use patients recruited **before** IA
- ▶ Stage 2 p-value:  
use patients recruited **after** IA
- ▶ P-clud property follows from separation into independent cohorts
- ▶ All interim data can be used for the adaptations

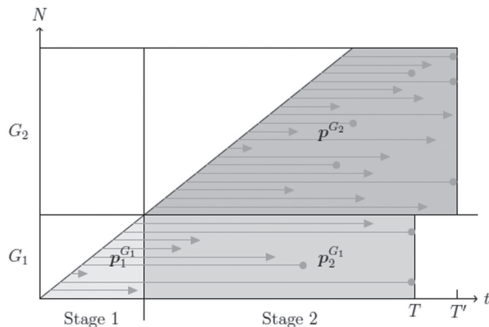
# Difficulties with patients-wise separation

---

1. The first stage p-value can only be computed at end of the trial  
⇒ no early rejection possible
2. For strict type I error rate control, patients from stage 1 must be followed-up as pre-planned and their later events be ignored (Jenkins et al., '11; Magirr et al., '14)
  - ▶ To overcome the issue of not using all observed events:  
Magirr et al. consider a conservative adjustment of critical boundaries, but also show that this approach is too conservative for applications.
  - ▶ **An always valid and efficient solution is still lacking!!**
  - ▶ *Possible practical solution:* Additional descriptive analysis with all events to (hopefully) confirm the adaptive test result.

## Three-fold separation approach (JÖRGNES ET AL., 2017)

Permits early rejections while using all interim data for the adaptations:



- ▶ Separation of data in three parts;
- ▶ early inference with  $p_1^{G_1}$ ;
- ▶ combination of stoch. independent p-values

$$p_1^{G_1}, p_2^{G_1}, p^{G_2}$$

with inverse normal method.

*Important remark:* The design for  $p_1^{G_1}$  and  $p_2^{G_1}$  must be as pre-specified (like in a GSST); (only) the design for  $p^{G_2}$  can be adapted at the IA using all interim data!



# **Information Restricted AST with Surrogate Endpoints**

## Conditional p-value approach

---

- ▶ The information used for the adaptations is restricted to specific interim data / estimators / test statistics  $\mathbf{D}_1$ .
- ▶ The second stage p-value is based on a second stage test statistics  $Z_2$  whose conditional null distribution function  $\mathcal{F}_2(z|\mathbf{D}_1) = \mathbf{P}_0(Z_2 \leq z|\mathbf{D}_1)$  is known.
- ▶ In this case the second stage p-value

$$q = 1 - \mathcal{F}_2(Z_2|\mathbf{D}_1)$$

satisfies the (generalized) p-clud condition:

$$\mathbf{P}_0(q \leq u | \mathbf{D}_1) \leq u \quad \text{for all } 0 \leq u \leq 1 \text{ and any interim data } \mathbf{D}_1$$

- ▶ *Disadvantage:* **Interim surrogate information** that is not included in the first stage p-value  $p$  or in the conditional error function  $A(\text{interim data})$  **remains unused**.

## Conditional p-value approach with normal statistics

- ▶ Often  $(\mathbf{D}_1, Z_2) = (X_1, \dots, X_m, Z_2)$  is under the null hypothesis (at least asymptotically) multivariate normal with  $\mathbf{E}_0(X_1) = \dots = \mathbf{E}_0(X_m) = \mathbf{E}_0(Z_2) = 0$  and known or estimable

$$\mathbf{V}_1 = \text{Cov}(\mathbf{D}_1) = (\text{Cov}(X_i, X_j))_{1 \leq i, j \leq m} \quad \text{and} \quad \mathbf{v}_2 = (\text{Cov}(X_i, Z_2))_{1 \leq i \leq m}$$

- ▶ Then

$$Z_2 | \mathbf{D}_1 \sim N(\mathbf{v}_2^T \mathbf{V}_1 \mathbf{D}_1, \mathbf{v}_2^T \mathbf{V}_1 \mathbf{v}_2)$$

and the second stage p-value

$$q = 1 - \Phi\left((Z_2 - \mathbf{v}_2^T \mathbf{V}_1 \mathbf{D}_1) / \sqrt{\mathbf{v}_2^T \mathbf{V}_1 \mathbf{v}_2}\right)$$

fulfils the generalized p-clud property (at least asymptotically).

- ▶ Liu & Pledger (2006) applied this to *partial tumor response*, *PFS* and *overall survival*, with score statistics from logistic regression and Cox models.

# Joint Modeling approach

---

- ▶ Methods for using the surrogate endpoints to predict the primary endpoint (e.g. Beyer et al. 2020)
- ▶ Requires a joint model for the used surrogate and primary endpoint (e.g. multistate models)
- ▶ With a categorical surrogate endpoint (e.g. tumor response) a non-parametric modeling is possible (Brückner, Burger, Brannath, 2018)
- ▶ In more complex situations, like with a continuous or time to event SEP, the statistical inference may rely model assumptions and also a joint null hypothesis!
- ▶ *A still open research question:*  
When and how is it possible to utilize surrogate information in a model robust way (e.g. via double robust estimation techniques from causal inference)?

## Utilizing categorical SEP (BRÜCKNER, BURGER AND BRANNATH, 2018)

---

- ▶ Assume that adaptations depend (only) on a categorical SEP and the PE.  
⇒ the second stage follow-up time is determined for each

$$\text{tcs-stratum} = (\text{treatment} \times \text{SEP-category} \times \text{stage}) - \text{stratum}$$

**before** the respective stage starts.

- ▶ With a *patient-wise separation* of stages, we can unbiasedly estimate the primary endpoint's survival function **within each tcs-stratum** using **all events**.
- ▶ With a *follow-up-wise separation* of stages, we obtain unbiased **tcs-strata-wise estimates** (e.g.) by using right censoring at stage 1, and left-truncation (plus right censoring) at stage 2, in each stratum and also **using all events**.  
⇒ First and second stage estimators are stoch. independent (asymptotically)

## Non-parametric adaptive survival trials (BBB, 2018)

---

- ▶ The trt-specific **overall survival functions** are **weighted means of the strata-wise survival functions** with weights = trt-specific probabilities of the SEP-categories.
- ▶ With **stage-wise estimates of the SEP-category probabilities**, we obtain stoch. independent first and second stage estimates of the overall survival functions.
- ▶ This method can be applied with **patient-wise** and **follow-up-wise separation** of the primary time-to-event data to **each treatment group**.
- ▶ The stage-wise and trt-specific survival function estimates can be used to obtain **stoch. independent stage-wise p-values** for **adaptive non-parametric tests** on the *restricted mean survival time* or *average hazard rate*.
- ▶ This provides valid **adaptive survival trials** where adaptations can be **based on the categorical SEP and time-to-event PE** using **all observed events**.

## Some simulation results (BBB, 2018)

- ▶ Adaptive enrichment design with treatments  $E$  and  $C$ , full population  $F$ , biomarker sub-population  $B$  and complement  $B^c = F \setminus B$
- ▶ SEP = binary response with response rates  $\pi_{E,B}$ ,  $\pi_{E,B^c}$  and  $\pi_C$ .
- ▶ *Simulation*: Multiplicative hazard model for PE (OS) with hazard rate  $r$  for responder vs. non-response in  $C$ , and hazard rates  $c_B$  and  $c_{B^c}$  for  $E$  vs.  $C$ .

Hazard Ratios			Response Rates			Power Patient-wise Splitting				Power Follow-up-wise Splitting			
$r$	$c_B$	$c_{B^c}$	$\pi_{E,B}$	$\pi_{E,B^c}$	$\pi_C$	SLR	AHR	RMS	DIFF	SLR	AHR	RMS	DIFF
0.7	1	1	0.4	0.4	0.4	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.7	0.74	1	0.4	0.2	0.2	0.73	0.88	0.88	0.71	0.79	0.91	0.90	0.74
0.7	0.7	0.7	0.4	0.4	0.4	0.89	0.87	0.87	0.68	0.95	0.91	0.91	0.74
0.7	0.7	0.8	0.5	0.3	0.2	0.88	0.98	0.98	0.91	0.92	0.99	0.99	0.92
0.5	1	1	0.8	0.65	0.2	0.02	0.96	0.96	0.84	0.02	0.97	0.97	0.86

# Comments and Discussion



## Some practical and general comments

---

- ▶ A pre-specified adaptation rule ensures that only the permitted interim data is used for the adaptations.
- ▶ Using (and staying) with a pre-defined adaptation rule may also lead to more efficient designs but may not always be desirable.
- ▶ A stepwise disclosure of interim data can enforce the use of the only permitted interim data.
- ▶ A complete futility/safety stop can always be based on all observed interim data.
- ▶ With multiple treatments and/or populations, usually the FWER need to be controlled. This can be achieved via the closed testing principle.

## Summary and discussion

---

- ▶ The use of (short term) surrogate endpoints in AST complicates matters, since the p-clad condition is not so easy to achieve.
- ▶ A patient-wise separation provides a simple solution, that permits to use all interim data for the adaptations.
- ▶ However, it does not permit any explicit or implicit change of the interim patient's follow-up time which require to ignore some of the observed event data.
- ▶ Restricting the information used for the adaptations provides additional solutions.
- ▶ However, this requires statistical modeling and the validity of the design may rely on specific model assumptions.
- ▶ Further methodological research is required to obtain valid and robust methods that make efficient use of common types of surrogate information.

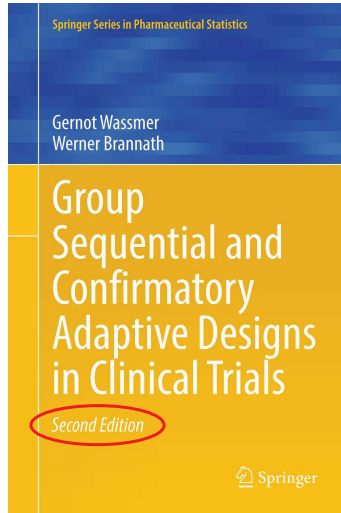
## Selected References - Thank You! (→)

---

- ▶ Brückner M, Burger HU, Brannath W (2018). Nonparametric adaptive enrichment designs using categorical surrogate data. *Statistics in Medicine* 37:4507–4524.
- ▶ Beyer U, Dejardin D, Meller M, Rufibach K, Burger HU (2020). A multistate model for early decision-making in oncology. *Biometrical Journal* 62:550–567.
- ▶ Brannath W, Posch M, Bauer P (2002). Recursive combination tests. *Journal of the American Statistical Association* 97:236–244.
- ▶ Liu Q, Pledger G (2006). On design and inference for two-stage adaptive clinical trials with dependent data *Journal of Statistical Planning and Inference* 136:1962–1984.
- ▶ 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* 23:1333–1335.
- ▶ 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.
- ▶ Magirr D, Jaki T, Koenig F, Posch M (2016). Sample Size Reassessment and Hypothesis Testing in Adaptive Survival Trials. *PLoS ONE* 11(2):e0146465.
- ▶ Jörgens S, Wassmer G, König F, Posch M (2019). Nested combination tests with a time-to-event endpoint using a short-term endpoint for design adaptations. *Pharmaceutical Statistics* 18:329–350.

# Second extended edition is out!

---



# Selected References - Thank You!

---

- ▶ Brückner M, Burger HU, Brannath W (2018). Nonparametric adaptive enrichment designs using categorical surrogate data. *Statistics in Medicine* 37:4507–4524.
- ▶ Beyer U, Dejardin D, Meller M, Rufibach K, Burger HU (2020). A multistate model for early decision-making in oncology. *Biometrical Journal* 62:550–567.
- ▶ Brannath W, Posch M, Bauer P (2002). Recursive combination tests. *Journal of the American Statistical Association* 97:236–244.
- ▶ Liu Q, Pledger G (2006). On design and inference for two-stage adaptive clinical trials with dependent data *Journal of Statistical Planning and Inference* 136:1962–1984.
- ▶ 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* 23:1333–1335.
- ▶ 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.
- ▶ Magirr D, Jaki T, Koenig F, Posch M (2016). Sample Size Reassessment and Hypothesis Testing in Adaptive Survival Trials. *PLoS ONE* 11(2):e0146465.
- ▶ Jörgens S, Wassmer G, König F, Posch M (2019). Nested combination tests with a time-to-event endpoint using a short-term endpoint for design adaptations. *Pharmaceutical Statistics* 18:329–350.