MIRROS: Phase 3 trial with time-to-event endpoint, a cure proportion, and a futility interim analysis using response

Kaspar Rufibach
Methods, Collaboration & Outreach Group, Department of Biostatistics, Roche Basel
Joint work with Dominik Heinzmann and Annabelle Monnet (Roche, Basel)
SMi Adaptive Designs in Clinical Trials Conference London, 2nd April 2019

Generated 2019-04-01 at 22:42:47.

Agenda

1. Acute Myeloid Leukemia
2. Clinical development plan
3. Key questions of MIRROS
4. Implementation features
5. Health authority feedback
6. Conclusions

What MIRROS is NOT:

What MIRROS is NOT:

An adaptive trial.

What MIRROS is NOT:

An adaptive trial.

A seamless phase 2/3 trial.

What MIRROS IS:

What MIRROS IS:

A phase 3 trial with a futility interim.

Acute Myeloid Leukemia

Acute Myeloid Leukemia

Rare malignant blood disease.

Most common leukemia, lowest survival rate in adults: **median survival $\leq$ 1y**.

Recurrent **life-threatening infections**.

Chemotherapy: modest benefit without cure.

Stem cell transplant:

- "Bridge-to-transplant": Goal of any therapy. Needs **complete response** (CR) to initial therapy.
- Only way to survive AML.

Standard of care

No standard regimen for relapsed or refractory (R/R) AML. Breems et al. (2005)

No new drug approved for treatment of AML in **over 50 years**! Bose et al. (2017)

Standard of care

No standard regimen for relapsed or refractory (R/R) AML. Breems et al. (2005)

No new drug approved for treatment of AML in **over 50 years**! Bose et al. (2017)

THIS is unmet medical need!

Idasanutlin

p53: Tumor suppressor, many mechanisms of anticancer function.

Mouse double minute 2 homolog (MDM2): Negative regulator of p53 tumor suppressor.

Idasanutlin: binds to MDM2 $\Rightarrow$ prevents p53 - MDM2 interaction $\Rightarrow$ (re-)activation of p53 $\Rightarrow$ **reinstalls anti-tumor capacity of p53**.

Clinical development plan

Clinical development plan for Idasanutlin

Need for acceleration:

- Very high unmet medical need in R/R AML.
- Early phase results with Idasanutlin encouraging.
- Competitive landscape and economic constraints: Lean program only way to receive internal approval for pivotal trial.
- Willingness to trade-off risk reduction from randomized P2 against increased speed.

Skip or integrate Phase 2?

Assume we have **successful P1**.

Purpose of futility interim: optimize **P(stopping @ interim | H_0)**.

Hunsberger et al. (2009):

- **Integrate** P2 into P3: futility interim based on **intermediate** endpoint.
- **Skip** P2: futility interim based on **P3 primary** endpoint.

If trial

- stops at futility interim: basically performed randomized P2.
- passes futility interim: P3 pivotal trial well on its way.

Key advantage of setup: Decision to proceed to full P3 part based on randomized comparison. Parmar et al. (2008)

Mirros

MDM2 **I**dasanutlin in **R**elapsed **R**efractory AML for **OS**.

- **Population**: R/R AML.
- **Comparison**: **Idasanutlin** + cytarabine vs. placebo + cytarabine.
- **Phase III, 2:1 randomized, double-blind, placebo-controlled** clinical trial.
- Primary endpoint: **overall survival**.
- Planned recruitment: 374 patients (wild-type sample, + 66 mutant patients).

`https://clinicaltrials.gov/ct2/show/NCT02545283`

Key questions of MIRROS

Key questions of MIRROS

Key questions of MIRROS

1. Primary endpoint OS. Sample size with **cure proportion** in both arms?

Key questions of MIRROS

1. Primary endpoint OS. Sample size with **cure proportion** in both arms?
2. Base **interim** on OS or something else? If the latter, what?

Key questions of MIRROS

1. Primary endpoint OS. Sample size with **cure proportion** in both arms?
2. Base **interim** on OS or something else? If the latter, what?
3. How to compute **operating characteristics** of interim analysis?

Cure proportion model

See e.g. Sun et al. (2018).

Let

- S_i^*, f_i^*: survival and density functions of **uncured** patients.
- p_i: proportions of patients cured.

Survival and hazard function in each treatment arm ($t \geq 0$):

$$\begin{aligned} S_i(t) &= p_i + (1 - p_i) S_i^*(t), \\ h_i(t) &= \frac{(1 - p_i) f_i^*(t)}{p_i + (1 - p_i) S_i^*(t)}. \end{aligned}$$

Ratio of hazard functions:

$$\theta(t) = h_2(t)/h_1(t) = \left(\frac{1 - p_2}{1 - p_1}\right) \frac{f_2^*(t)}{f_1^*(t)} \left(\frac{p_1 + (1 - p_1) S_1^*(t)}{p_2 + (1 - p_2) S_2^*(t)}\right).$$

Even if both S_i^* exponential $\Rightarrow$ $\theta(t)$ **depends on time** (if ≥ 1 p_i is > 0).

Cure proportion model – assumptions

What if we ignored cure proportions and simply computed necessary events d using Schoenfeld's formula?

Cure proportion model – assumptions

What if we ignored cure proportions and simply computed necessary events d using Schoenfeld's formula?

- Study will (typically) be **underpowered**.
- Time to clinical cutoff will be **underestimated**.

Cure proportion model – assumptions

What if we ignored cure proportions and simply computed necessary events d using Schoenfeld's formula?

- Study will (typically) be **underpowered**.
- Time to clinical cutoff will be **underestimated**.

Control arm, based on **historical data**, H_0:

- Median OS 6m.
- Cure: 0.080.

Cure proportion model – assumptions

What if we ignored cure proportions and simply computed necessary events d using Schoenfeld's formula?

- Study will (typically) be **underpowered**.
- Time to clinical cutoff will be **underestimated**.

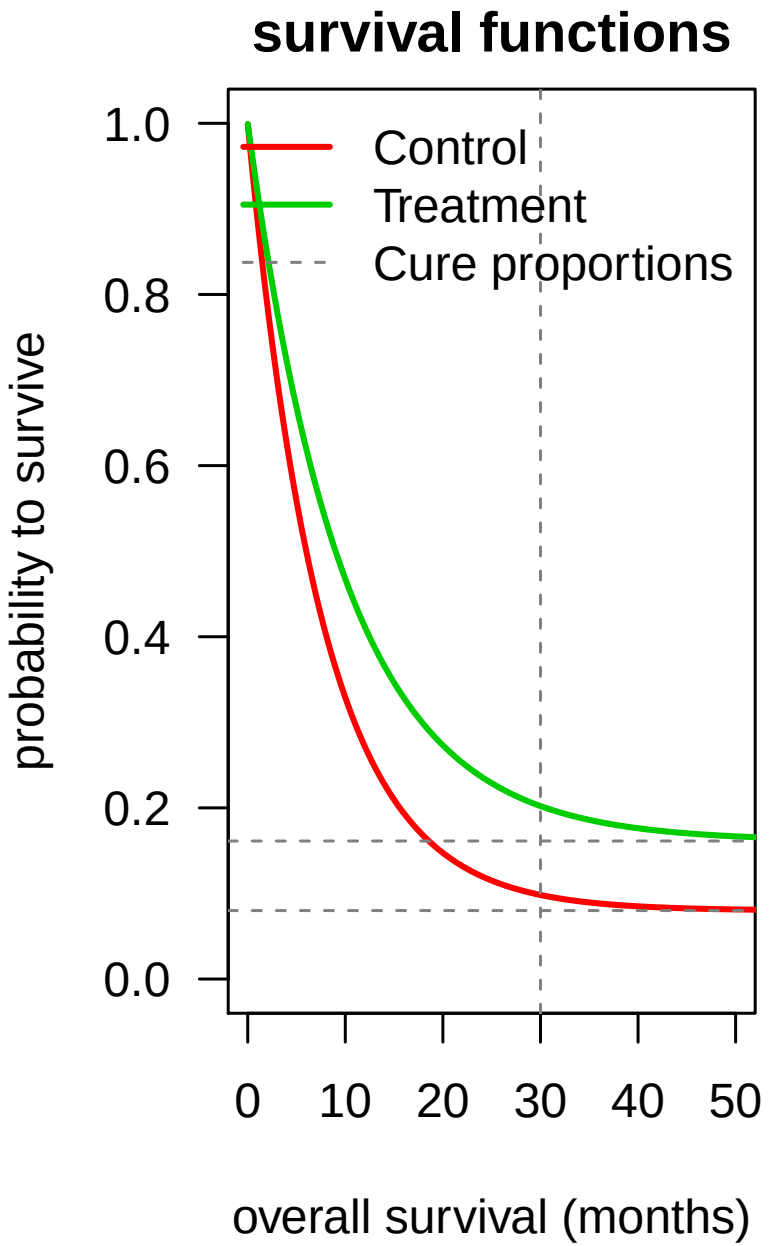
Control arm, based on **historical data**, H_0:

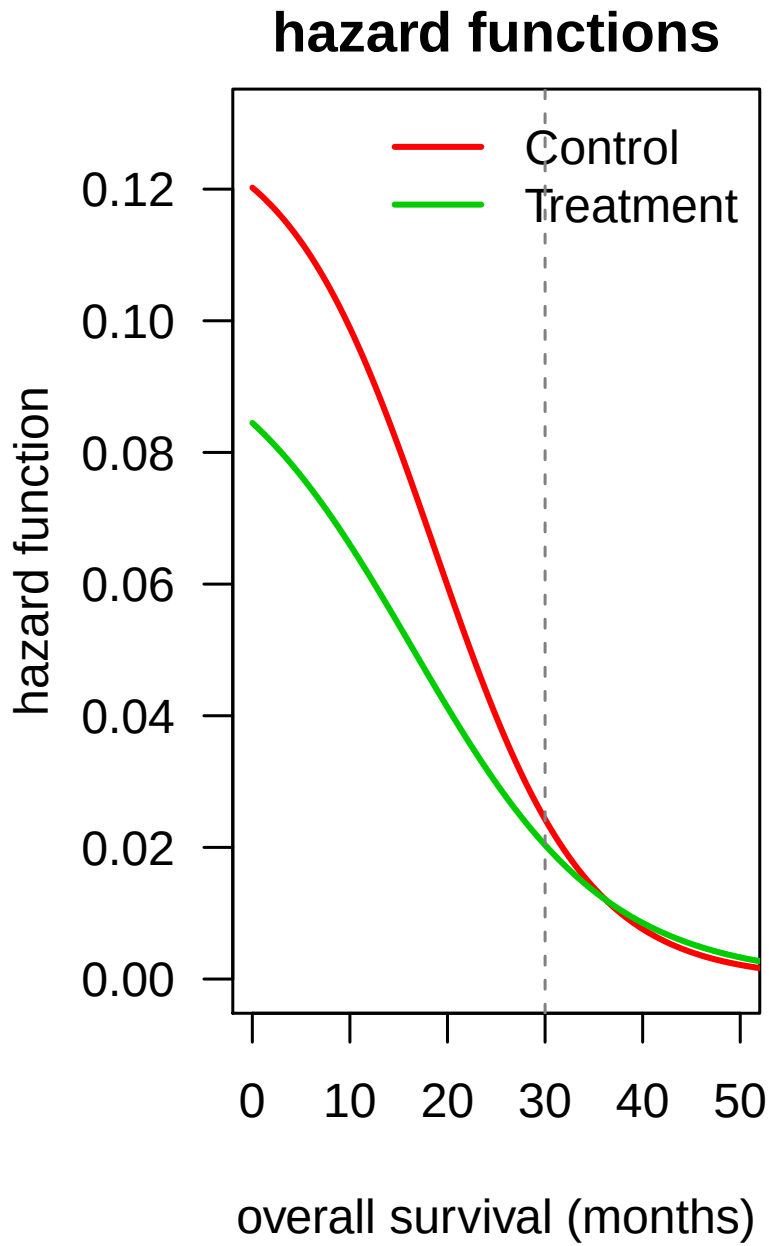
- Median OS 6m.
- Cure: 0.080.

Targeted effect size treatment arm (for 85% power, H_1):

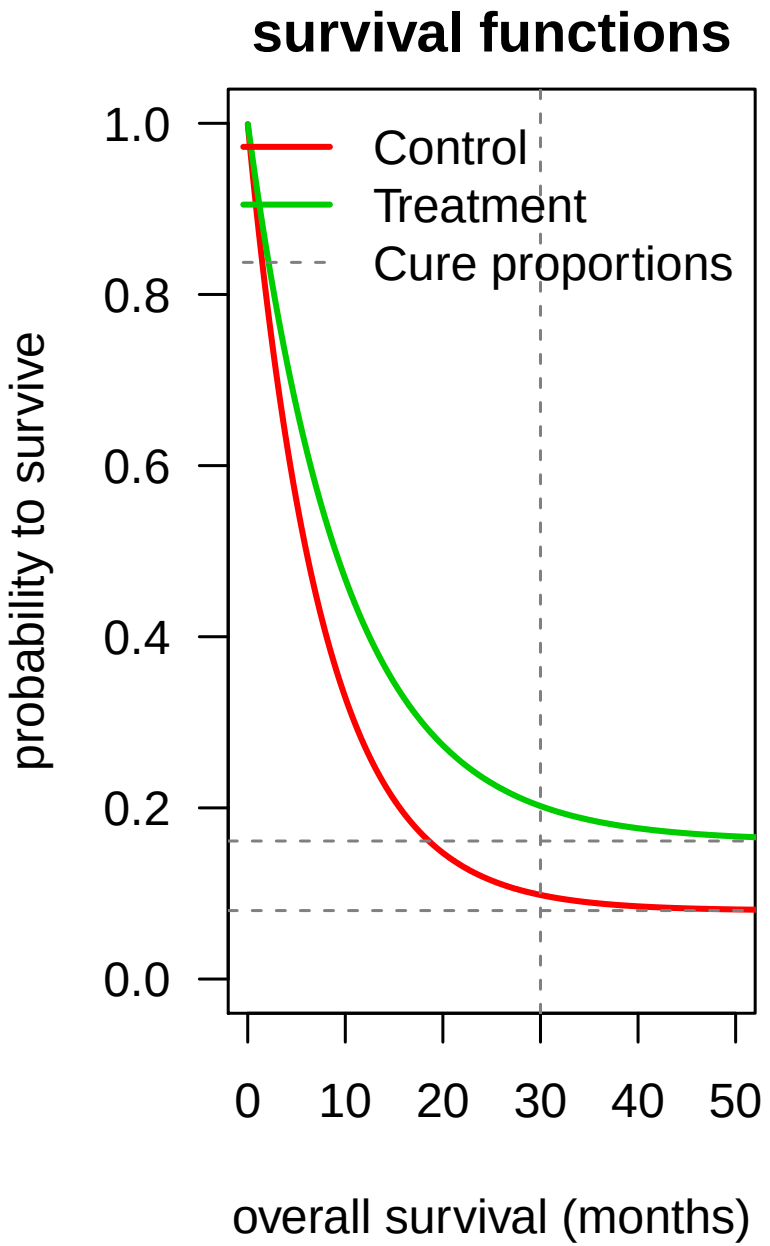
- Median OS 9m.
- Cure: 0.161 (see later for justification).

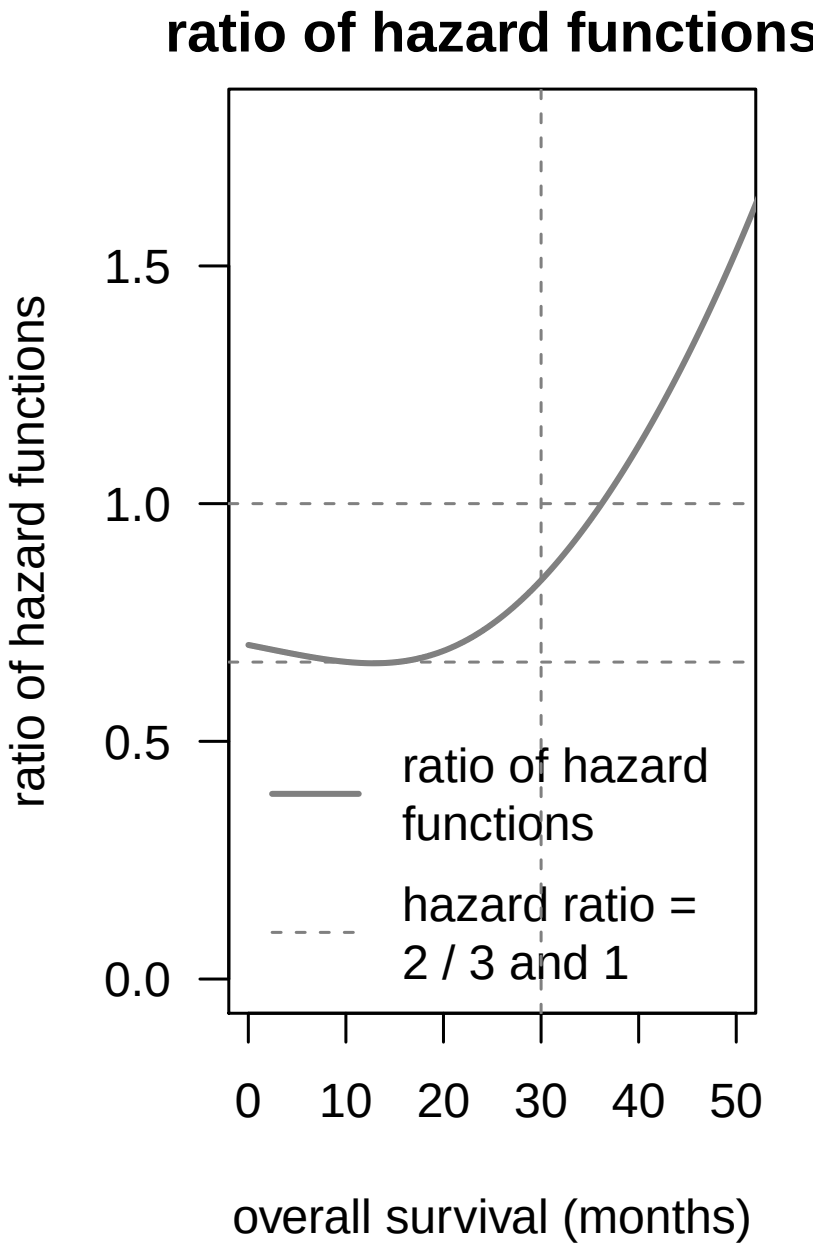
Cure proportion model – assumptions





Cure proportion model – assumptions





Cure proportion model – sample size

To find sample size:

- Compute necessary events d_0 using Schoenfeld's formula.
- **Simulate** from assumed S_i's, compute power for grid of $d = d_0, \ldots, d_1$.
- Choose d such that (unweighted) logrank test gives targeted power.

MIRROS: 2-sided $\alpha = 0.05$, $\beta = 0.15$, some accrual and drop-out assumption.

Assumption	$S_1^{-1}(0.5)$	$S_2^{-1}(0.5)$	p_1	p_2	d	power	time
MIRROS	6.0	9.0	0.080	0.161	275	0.852	**38.8**
PH, no cure	6.0	9.0	0	0	246	0.858	**29.2**
MIRROS with #events for PH, no cure	6.0	9.0	0.080	0.161	246	**0.810**	33.7

Cure proportion model – effect quantification

Cure proportion model – **no proportional hazards**. Unweighted logrank...

- ...**not most powerful** test, but loss modest (see above).
- ...**still valid** test, i.e. protects type I error.

How to quantify effect?

- **Kaplan-Meier** estimates provide entire information in data.
- Desire to summarize effect in one number.
- Hazard ratio and logrank test: if NPH, estimand depends on censoring distribution!
- Regulatory environment: typically accepted to reject H_0 using valid test, and then quantify effect differently.
- MIRROS: violation of PH only very late. Give hazard ratio, and estimate of cure proportion difference.

Rufibach (2019) has extended discussion in **estimand** context.

Futility interim analysis

Mitigate risk if drug does not work (sufficiently).

Planned after **120** patients are recruited.

Why not use OS for interim decision?

- 53 (under H_0) and 46 deaths (under H_1) expected at interim. Substantial uncertainty.
- Cures have not happened yet at the interim.
- Confounding by early (mainly safety-related) deaths.

Bottom line: interim is **too early for OS** to be meaningful endpoint.

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Futility interim is **non-binding**.

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Futility interim is **non-binding**. Why do we need to model it at all?

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Futility interim is **non-binding**. Why do we need to model it at all?

- How to choose interim boundary on CR?

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Futility interim is **non-binding**. Why do we need to model it at all?

- How to choose interim boundary on CR?
- Decision-makers want to be able to trade-off

$$\textbf{False Positive} = \mathrm{P}(\text{continue @ interim} \mid H_0)$$

vs.

$$\textbf{False Negative} = \mathrm{P}(\text{stop @ interim} \mid H_1).$$

Intermediate endpoint

Complete response:

- Sufficiently associated with OS.
- CR **necessary** for good OS / cure: Patient needs CR to have chance for cure, via bridge-to-transplant.
- Odds ratio as effect measure.

Futility interim is **non-binding**. Why do we need to model it at all?

- How to choose interim boundary on CR?
- Decision-makers want to be able to trade-off

$$\textbf{False Positive} = \mathrm{P}(\text{continue @ interim} \mid H_0)$$

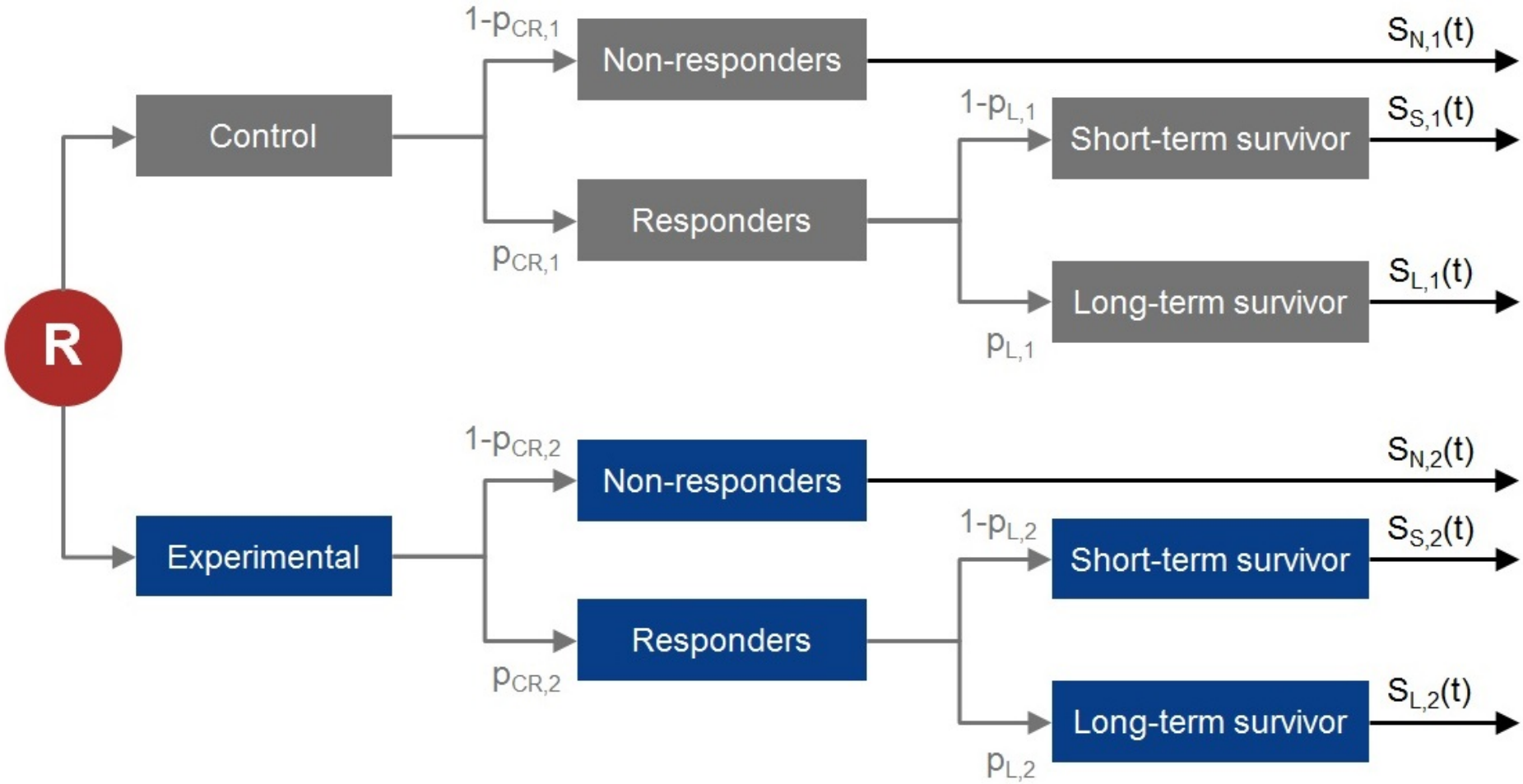
vs.

$$\textbf{False Negative} = \mathrm{P}(\text{stop @ interim} \mid H_1).$$

If futility based on OS $\Rightarrow$ conditional power.

If CR is intermediate endpoint: **mechanistic simulation model**.

Mechanistic simulation model



Mechanistic simulation model

Connects CR to OS.

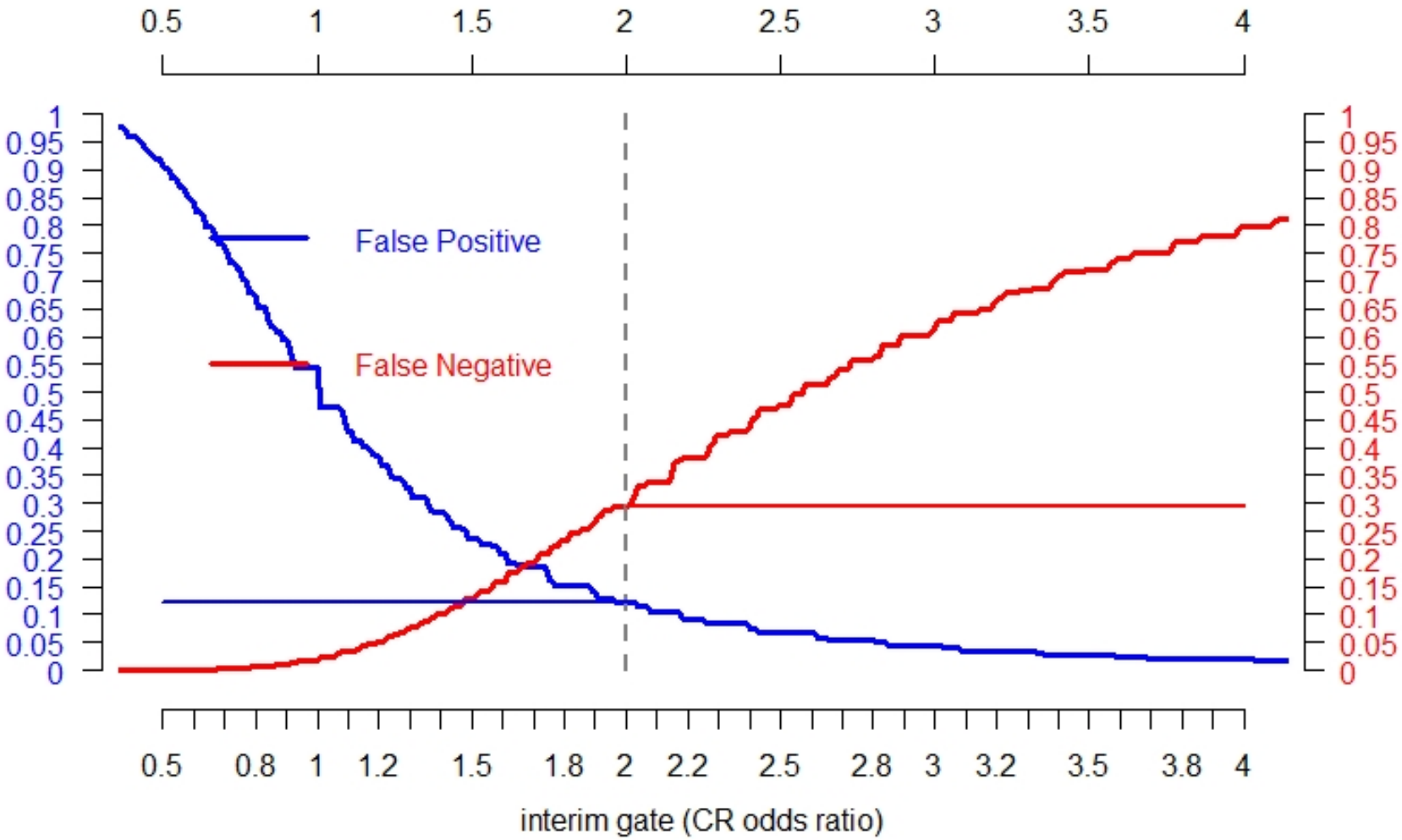
Need to inform all assumptions:

Quantity	Control arm	Treatment arm
Survival function of non-responders	$S_{N,1}$	$S_{N,2}$
Probability to have CR	$p_{CR,1}$	$p_{CR,2}$
Probability to be long-term responder | CR	$p_{L,1}$	$p_{L,2}$
Survival function of short-term responders	$S_{S,1}$	$S_{S,2}$
Survival function of long-term responders	$S_{L,1}$	$S_{L,2}$
#patients recruited per month	n_{1j}	n_{2j}
Months of recruitment	$j = 1, \ldots, N$	
Total #patients recruited	$n_1 = \sum_{j=1}^{N} n_{1j}$	$n_2 = \sum_{j=1}^{N} n_{2j}$
Drop-out rate per month	τ_1	τ_2

Align parameters such that **mechanistic simulation model can reproduce sample size**!

P(CR) control: 0.16. Assume OR = 2.5 to improve on this with treatment $\Rightarrow$
P(CR tmt) = 0.323. P(longterm survivor) = 0.5. This gives cure proportions.

Operating characteristics of various interim boundaries



False Positive = P(continue @ interim | no effect)
False Negative = P(stop @ interim | alternative used for powering)

Operating characteristics of various interim boundaries

Sweet spot: **odds ratio of 2**,

- False Positive = P(continue @ interim | no effect) $\approx$ 12%,
- False Negative = P(stop @ interim | alternative assumed for powering) $\approx$ 30%.

Interim decision:

- Based on independent data monitoring committee (iDMC) recommendation, i.e. sponsor **blinded**,
- **non-binding**,
- included safety criterion (molecule class toxicity) and criteria for early deaths $\Rightarrow$ OS component.

Power loss of adding futility interim

Can easily get that from simulations.

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim:

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim: **63%!**

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim: **63%!**
- Illustrates risk-appetite. Futility interim somehow becomes "informal efficacy interim".

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim: **63%!**
- Illustrates risk-appetite. Futility interim somehow becomes "informal efficacy interim".
- Do we always compute the power loss when adding futility interims? Do we **increase number of events** to account for it?

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim: **63%!**
- Illustrates risk-appetite. Futility interim somehow becomes "informal efficacy interim".
- Do we always compute the power loss when adding futility interims? Do we **increase number of events** to account for it?

Who cares anyway $\Rightarrow$ interim **passed**!

Implementation features

Implementation features

A (industry) clinical trial is **not a pre-specified static** undertaking!

- Not clear whether p53 mutant patients (≈15%) also benefit from Idasanutlin.
  - Still included, as evidence unclear and high unmet medical need.
  - But testing too late for randomization, i.e. could not stratify for p53 status.
  - Adds uncertainty to recruitment assumptions.
- Decision-makers sceptical about interim gate based on CR only. Additionally engineered EFS criterion (not discussed here).

Implementation features

A (industry) clinical trial is **not a pre-specified static** undertaking!

- Biomarker development: typically in Phase 2! Recommendation on biomarker development by iDMC.
- Seamless designs in general: sponsor does not get to see data for a **long time**. Unease for decision-makers.
- No accrual suspension for interim $\Rightarrow$ data cleaning and decision needs to come fast.

Health authority feedback

Health authority feedback

FDA:

- **Preferred randomized P2**.
- Challenged lack of stratification on p53 mutation status.
- Companion Diagnostic component with blinded P2 data $\Rightarrow$ not clear how to decide on development.
- **Challenged assumptions**, asked for additional sensitivity analyses.
- Concerns of early events driving interim analysis. OS not part of futility decision, but early tox deaths are.
- US sites only opened after passing the IA.

EMA:

- Agreed to accelerated development due to high unmet need.
- PH assumption discussed, support hazard ratio as appropriate effect measure.

Conclusions

Current status of MIRROS

Interim analysis passed on 17th Sept 2017.

Final analysis cutoff projected for **Q4 2019**.

Conclusions

Conclusions

- Account for power loss and timing delay if you have cure proportions.

Conclusions

- Account for power loss and timing delay if you have cure proportions.
- Think about how to quantify effect.

Conclusions

- Account for power loss and timing delay if you have cure proportions.
- Think about how to quantify effect.
- Skipping / integrating P2 into P3 allows for **acceleration** and **risk-mitigation**. If you stop at interim not much is lost in fact.

Conclusions

- Account for power loss and timing delay if you have cure proportions.
- Think about how to quantify effect.
- Skipping / integrating P2 into P3 allows for **acceleration** and **risk-mitigation**. If you stop at interim not much is lost in fact.
- Mechanistic simulation model allows to associate binary intermediate to time-to-event primary endpoint and explore interim analysis **operating characteristics**.

Conclusions

- Account for power loss and timing delay if you have cure proportions.
- Think about how to quantify effect.
- Skipping / integrating P2 into P3 allows for **acceleration** and **risk-mitigation**. If you stop at interim not much is lost in fact.
- Mechanistic simulation model allows to associate binary intermediate to time-to-event primary endpoint and explore interim analysis **operating characteristics**.

We have implemented this in a REAL trial!

Resources

Paper (under revision): Rufibach et al. (2019), available at `https://arxiv.org/abs/1901.01308`.

Code to reproduce simulations and **plan your own trial** on github: `https://github.com/numbersman77/integratePhase2.git`.

BBS seminar on synthetic controls

Basel Biometric Section spring seminar: **Synthetic controls - what do we need and how far can we go?**

- Basel.
- **May 10th**, 2019, 9:00-16:00.
- Speakers from industry, Flatiron, European regulators.
- Rejoinders by regulators: Norbert Benda (BfArM), Jan Müller-Berghaus (PEI), Anja Schiel (Norwegians Medicine Agency & Chair BSWP), Kit Roes (UMC Utrecht MEB and EMA BSWP), Meinhard Kieser (University of Heidelberg).
- Panel discussion with all speakers.

`http://bbs.ceb-institute.org`

Thank you for your attention.

References I

- Bose, P., Vachhani, P. and Cortes, J. E. (2017). Treatment of relapsed/refractory acute myeloid leukemia. *Current Treatment Options in Oncology* **18** 17. https://doi.org/10.1007/s11864-017-0456-2
- Breems, D. A., Van Putten, W. L., Huijgens, P. C., Ossenkoppele, G. J., Verhoef, G. E., Verdonck, L. F., Vellenga, E., De Greef, G. E., Jacky, E., Van der Lelie, J., Boogaerts, M. A. and Lowenberg, B. (2005). Prognostic index for adult patients with acute myeloid leukemia in first relapse. *J. Clin. Oncol.* **23** 1969–1978.
- Cantor, A. B. and Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat Med* **11** 931–937.
- Goldman, B., LeBlanc, M. and Crowley, J. (2008). Interim futility analysis with intermediate endpoints. *Clin Trials* **5** 14–22.
- Hunsberger, S., Zhao, Y. and Simon, R. (2009). A comparison of phase II study strategies. *Clin. Cancer Res.* **15** 5950–5955.
- Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-term Survivors*. John Wiley & Sons, United States.
- Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79** 731–739. http://www.jstor.org/stable/2337228

References II

- Parmar, M. K., Barthel, F. M., Sydes, M., Langley, R., Kaplan, R., Eisenhauer, E., Brady, M., James, N., Bookman, M. A., Swart, A. M., Qian, W. and Royston, P. (2008). Speeding up the evaluation of new agents in cancer. *J. Natl. Cancer Inst.* **100** 1204–1214.
- Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints – estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*, *to appear*. `https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1917`
- Rufibach, K., Heinzmann, D. and Monnet, A. (2019). Integrating phase 2 into phase 3 based on an intermediate endpoint while accounting for a cure proportion. Preprint: `https://arxiv.org/abs/1901.01308` Code available on github: `https://github.com/numbersman77/integratePhase2.git`.
- Schemper, M., Wakounig, S. and Heinze, G. (2009). The estimation of average hazard ratios by weighted cox regression. *Stat. Med.* **28** 2473–2489.
- Sun, S., Liu, G., Lyu, T., Xue, F., Yeh, T. M. and Rao, S. (2018). Design considerations in clinical trials with cure rate survival data: A case study in oncology. *Pharm Stat* **17** 94–104.
- Tsodikov, A. D., Ibrahim, J. G. and Yakovlev, A. Y. (2003). Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models. *J Am Stat Assoc* **98** 1063–1078.
- Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1** 423–439.

Backup slides.

Cure proportion model – estimation

Numerous parametric and nonparametric estimates of relevant quantities:
Cantor and Shuster (1992), Maller and Zhou (1992), Maller and Zhou (1996), Tsodikov et al. (2003).

Obvious nonparametric estimate of p, with $\widehat{S}$ Kaplan-Meier:

- $\widehat{S}(t_0)$ for some $t_0 > 0$.
- Maller and Zhou (1992): Kaplan-Meier evaluated at largest observed time, censored or event, consistently estimates p_0 under "sufficient follow-up" condition Tsodikov et al. (2003).
- Finite sample: likely not use latest observed time to evaluate the Kaplan-Meier estimate at. Rather **trade-off bias to reduce variability** of estimate.
- Choose milestone t_0 where clinically, cure seems very plausible.

Why two models?

We have two models:

- Cure proportion model to derive sample size,
- mechanistic simulation model to explore interim operating characteristics.

Why?

Reasons:

- Futility interim analysis has no implication on type I error $\Rightarrow$ independent of key design characteristic.
- Cure proportion model:
  - Simple,
  - depends on less assumptions than mechanistic model,
  - Robust model to plan sample size.
- Mechanistic simulation model:
  - Interim setup has potential to be changed before or while study is running. Prefer not to have these changes interfere with sample size.
  - Only used for (internal) decision-making via iDMC, no filing relevance $\Rightarrow$ can "afford" more modeling.

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 3.5.1 (2018-07-02)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: reporttools / xtable / biostatKR / survival

This document was generated on 2019-04-01 at 22:42:47.