Global Biostatistics and Data Sciences, BMS

Using surrogate endpoints and win statistic to improve decision-making in oncology

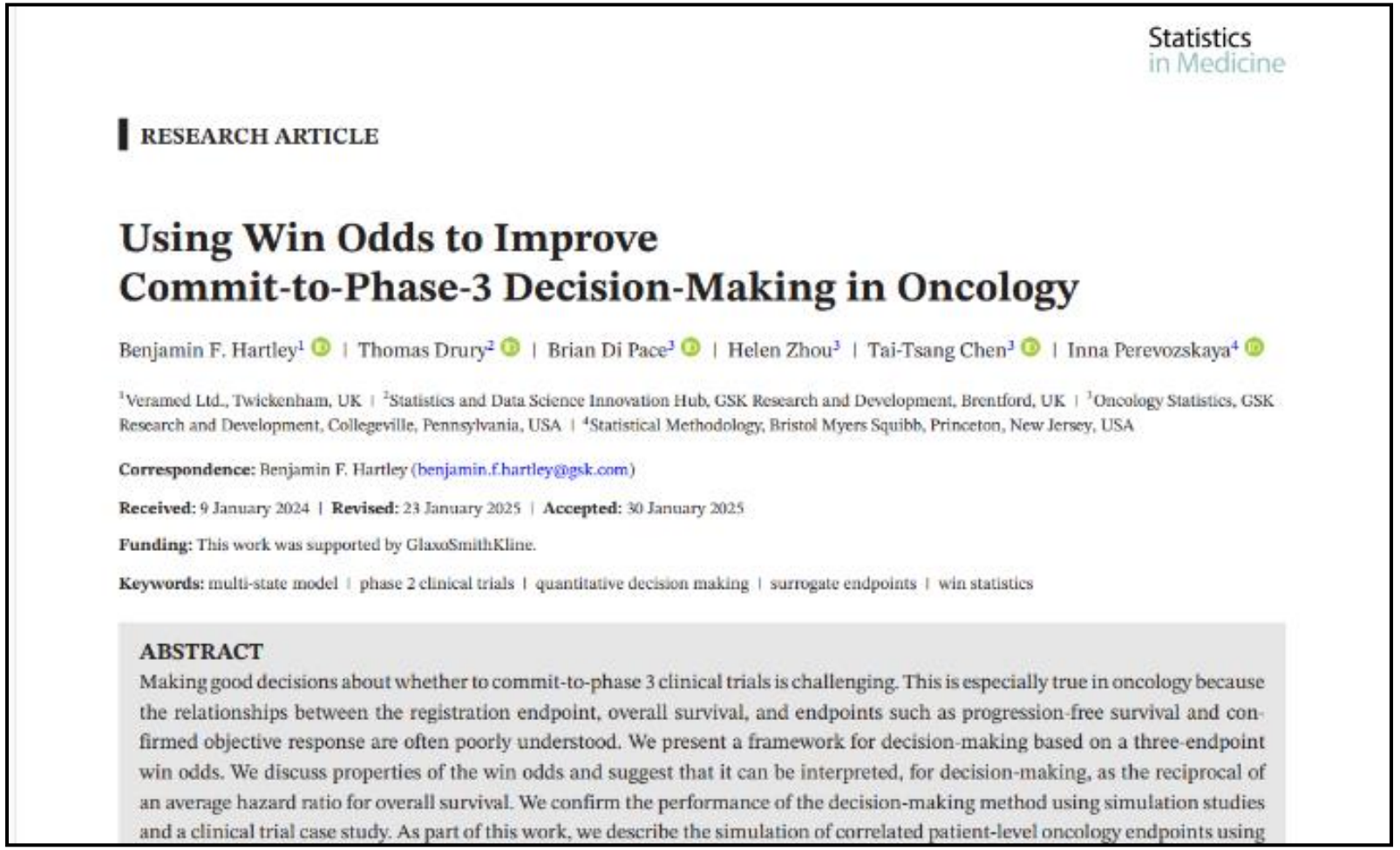
Inna Perevozskaya,
Senior Biometrics Fellow and Head of Statistical Methodology at BMS

Bristol Myers Squibb®

Disclaimer

- The views expressed in this presentation are of my own and do not represent my past or current employers
- I am currently employed by BMS, but the work presented here was done during my employment at GSK

Based on:



Statistics
in Medicine

RESEARCH ARTICLE

Using Win Odds to Improve Commit-to-Phase-3 Decision-Making in Oncology

Benjamin F. Hartley[1] | Thomas Drury[2] | Brian Di Pace[3] | Helen Zhou[3] | Tai-Tsang Chen[3] | Inna Perevozskaya[4]

[1]Veramed Ltd., Twickenham, UK | [2]Statistics and Data Science Innovation Hub, GSK Research and Development, Brentford, UK | [3]Oncology Statistics, GSK Research and Development, Collegeville, Pennsylvania, USA | [4]Statistical Methodology, Bristol Myers Squibb, Princeton, New Jersey, USA

Correspondence: Benjamin F. Hartley (benjamin.f.hartley@gsk.com)

Received: 9 January 2024 | **Revised:** 23 January 2025 | **Accepted:** 30 January 2025

Funding: This work was supported by GlaxoSmithKline.

Keywords: multi-state model | phase 2 clinical trials | quantitative decision making | surrogate endpoints | win statistics

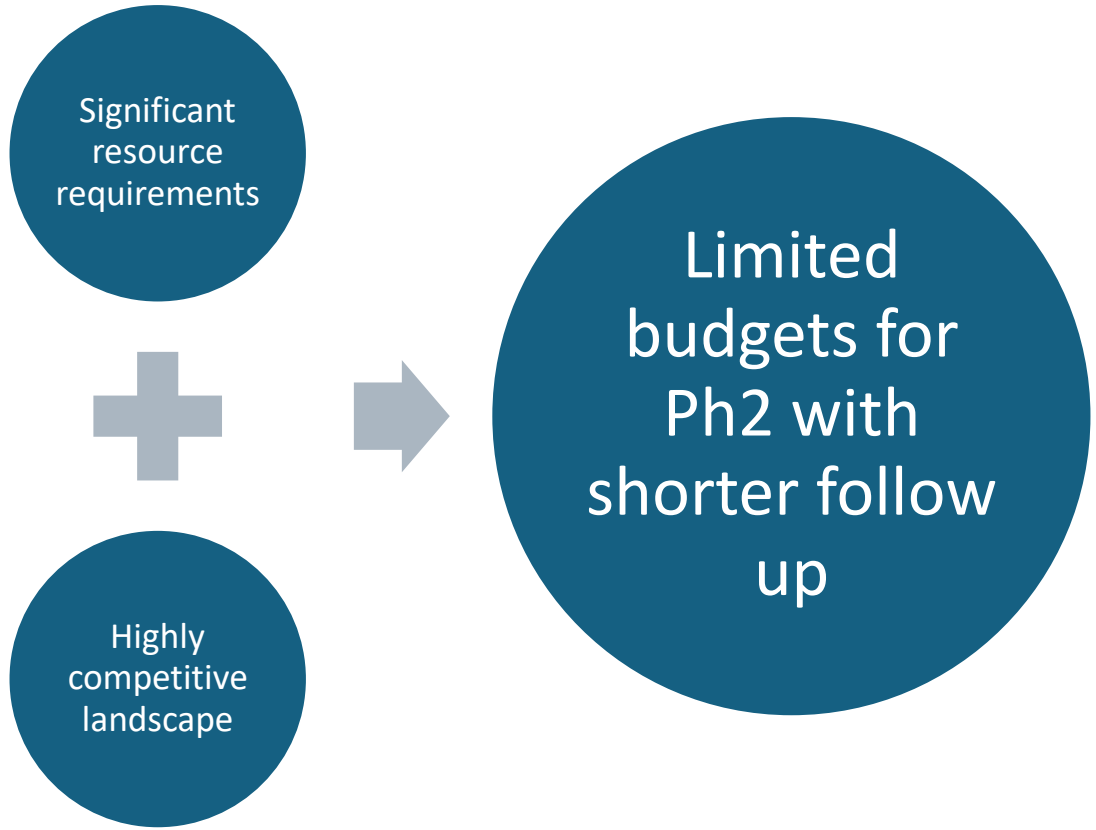
ABSTRACT

Making good decisions about whether to commit-to-phase 3 clinical trials is challenging. This is especially true in oncology because the relationships between the registration endpoint, overall survival, and endpoints such as progression-free survival and confirmed objective response are often poorly understood. We present a framework for decision-making based on a three-endpoint win odds. We discuss properties of the win odds and suggest that it can be interpreted, for decision-making, as the reciprocal of an average hazard ratio for overall survival. We confirm the performance of the decision-making method using simulation studies and a clinical trial case study. As part of this work, we describe the simulation of correlated patient-level oncology endpoints using

How translational statistics can enhance the design and analysis of clinical trials ?

By bridging the gap between intermediate signals and long-term outcomes of interest

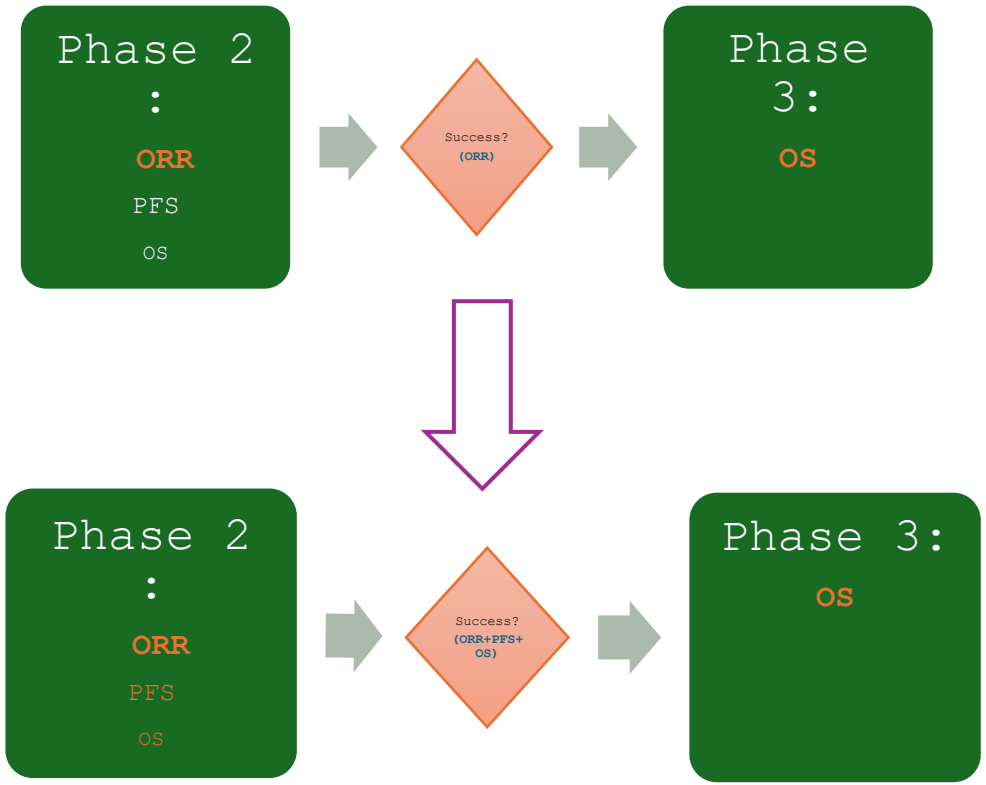
- Oncology trials require significant resources
- Conducted in highly competitive landscape
- Resulting in increased pressure to screen conduct PoC faster
- Most Phase 2 trials are designed based on short term response-Objective Response Rate (ORR)
- While Phase 3 trials are designed based on long term endpoint-overall survival (OS)
- Creating the need to link these endpoint via translational



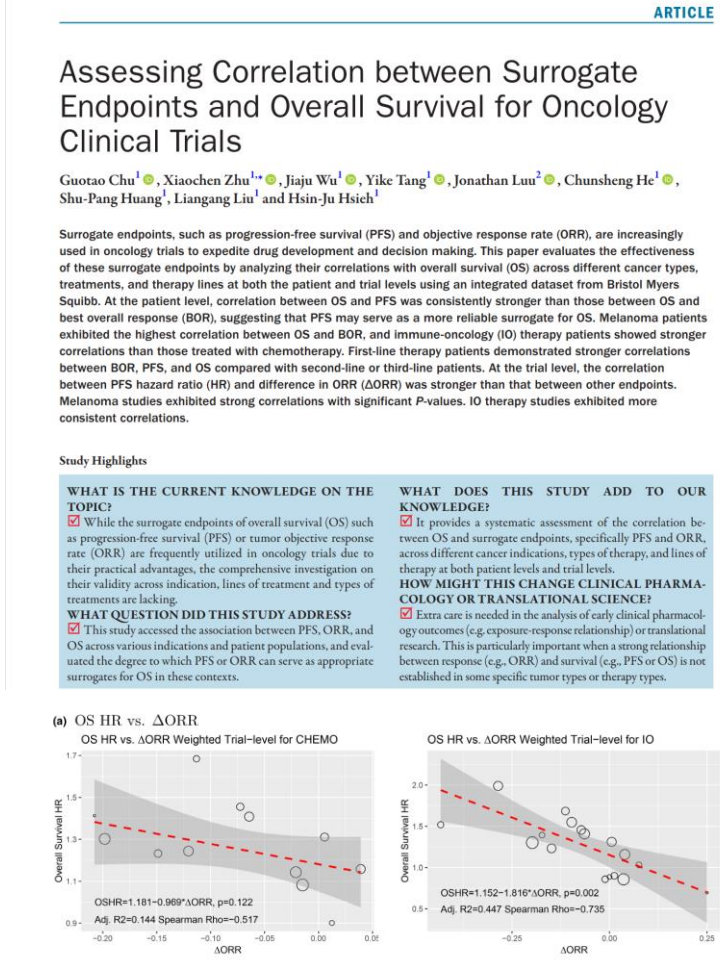
Decision making in oncology program development

- Typical decision-making scheme in oncology drug development starts with phase 2 based on ORR as primary endpoint
  - due to shorter follow-up, very limited PFS or even OS available
  - go/no-go decision at the end of phase 2 typically is based on ORR only
- But the success of phase 3 is based on OS
  - OS is hard to predict based on ORR and very limited PFS/OS information available when we finish Ph2 and start designing Ph3
- The main challenge is making informed decisions after Phase 2 trials, where only short-term endpoints like ORR and limited PFS or OS data are available.
- This can lead to costly failures in Phase 3 if decisions are based on imperfect surrogates rather than robust, combined evidence from all endpoints.

- The goal: shift from end of Ph2 decision-making based on 1endpt to decision based on totality of data
  - Win-ratio/win-odds provide convenient instrument to make decision combining multiple endpoints



ORR/PFS are useful surrogates for OS but not perfect



ARTICLE

Assessing Correlation between Surrogate Endpoints and Overall Survival for Oncology Clinical Trials

Guotao Chu[1], Xiaochen Zhu[1,*], Jiaju Wu[1], Yike Tang[1], Jonathan Luu[2], Chunsheng He[1], Shu-Pang Huang[1], Liangang Liu[1] and Hsin-Ju Hsieh[1]

Surrogate endpoints, such as progression-free survival (PFS) and objective response rate (ORR), are increasingly used in oncology trials to expedite drug development and decision making. This paper evaluates the effectiveness of these surrogate endpoints by analyzing their correlations with overall survival (OS) across different cancer types, treatments, and therapy lines at both the patient and trial levels using an integrated dataset from Bristol Myers Squibb. At the patient level, correlation between OS and PFS was consistently stronger than those between OS and best overall response (BOR), suggesting that PFS may serve as a more reliable surrogate for OS. Melanoma patients exhibited the highest correlation between OS and BOR, and immune-oncology (IO) therapy patients showed stronger correlations than those treated with chemotherapy. First-line therapy patients demonstrated stronger correlations between BOR, PFS, and OS compared with second-line or third-line patients. At the trial level, the correlation between PFS hazard ratio (HR) and difference in ORR (ΔORR) was stronger than that between other endpoints. Melanoma studies exhibited strong correlations with significant *P*-values. IO therapy studies exhibited more consistent correlations.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
☑ While the surrogate endpoints of overall survival (OS) such as progression-free survival (PFS) or tumor objective response rate (ORR) are frequently utilized in oncology trials due to their practical advantages, the comprehensive investigation on their validity across indication, lines of treatment and types of treatments are lacking.

WHAT QUESTION DID THIS STUDY ADDRESS?
☑ This study accessed the association between PFS, ORR, and OS across various indications and patient populations, and evaluated the degree to which PFS or ORR can serve as appropriate surrogates for OS in these contexts.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
☑ It provides a systematic assessment of the correlation between OS and surrogate endpoints, specifically PFS and ORR, across different cancer indications, types of therapy, and lines of therapy at both patient levels and trial levels.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?
☑ Extra care is needed in the analysis of early clinical pharmacology outcomes (e.g. exposure-response relationship) or translational research. This is particularly important when a strong relationship between response (e.g., ORR) and survival (e.g., PFS or OS) is not established in some specific tumor types or therapy types.

- The link between the treatment effects on ORR , PFS vs. OS sometimes is known, as shown in this paper:
  - Exploring how well surrogate endpoints—specifically PFS and ORR—predict overall survival in oncology trials.
  - The study uses a large integrated dataset from Bristol Myers Squibb (BMS) and evaluates correlations at both the patient level and trial level across various cancer types, therapy types, and lines of treatment.
  - Similar assessment can be performed based on literature-based meta-analysis
  - Their study shows that effects can vary considerably in different cancers and by treatment class
  - PFS/ORR may not always be predictive of OS, especially in a larger, more diverse patient populations
- But what to do if data/meta-analysis is not readily available?
  - **We propose a decision to commit to a phase 3 clinical trial be based on novel statistical methods combining all these endpoints at the end of Phase 2 (C2Ph3 decision)**
  - Even if information for the OS and PFS is not completely mature
  - This is **not** a proposal to change a registrational endpoint, just to enhance internal decision-making

Proposed approach overview

Simulation-based with case study application

Goal:

- Evaluate the impact of novel analysis methods in Ph2 on "quality" of C2PH3 decision
- Methods considered leverage totality of data: ORR + partial PFS and OS data to enhance Ph2 Go/No-Go decisions

Methods:

- Build a patient-level data generation model for ORR , PFS and OS
  - Multi-State-Model (MSM)
  - Validate it leveraging real study data available (from later line therapy for solid tumor)
- Use generated data in Quantitative Decision-Making (QDM) algorithm to assess performance of novel methods combining ORR , PFS , and OS endpoints via simulation
  1. Win ratio/win odds methods (Pocock et al. 2012)
  2. Joint Frailty Models (Rogers et. al 2016)
  3. Conventional methods (based on single endpoint or "gating")
- Apply findings to at least 1 case study
- All methods assume two-arm randomized Phase 2 trial (active vs control). We do not change the design, just methods to analyze data

Simulation structure at a glance

We use different data generation and analysis models, for robustness assessment

Data Generation:

Multi-State Model (MSM)

- Generates dependent (ORR, PFS, OS) data at patient level.
- Parameters of transitions estimated from case study data
- Created 10 scenarios of treatment effect and durability of response (with clinical input)

Analysis Models:

- Win Ratio / Odds
- Joint Frailty
- ORR or PFS or OS Only
- ORR gating OS

Simulation:

Method	Pr Go to Ph3	POS (ph3|ph2)
Win Ratio/Odds	XX.X%	YY.Y%
Joint Frailty	XX.X%	YY.Y%
ORR or OS only	XX.X%	YY.Y%
ORR gating OS	XX.X%	YY.Y%

When using non-survival methods for C2PH3 decision-making we used assumptions about how those analysis metrics relate to HR_{OS}

- HR_{PFS} and Δ_{OR} related to HR_{OS} according to a provided meta-analysis
- WO equivalent to $1/HR_{OS}$

Commit to Phase 3 Decision Criteria (C2P3)

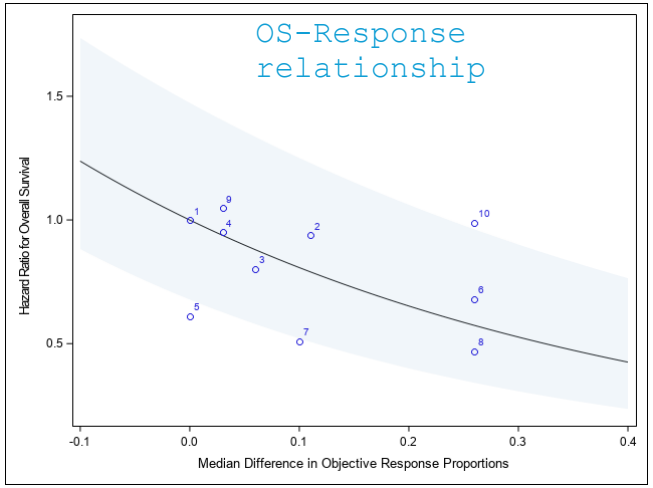
Unified Quantitative Decision-Making (QDM) framework for all analysis methods

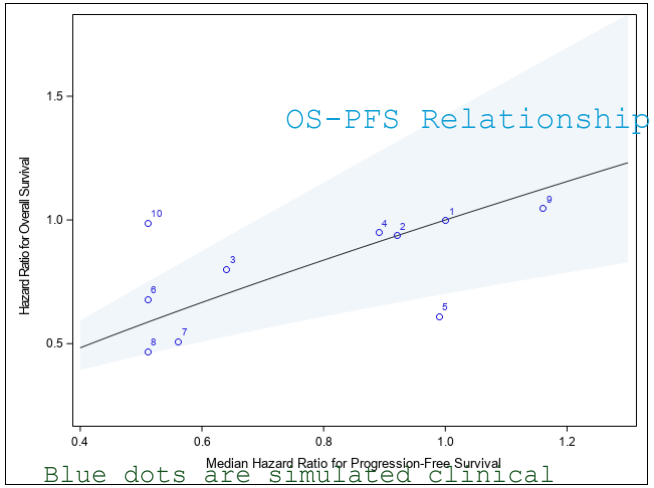
- Decision criteria at the end of Phase 2 (based on Lalonde et a. 2007)

$$P\left[HR_{OS\ Ph3} \leq MV\right] < p \qquad P\left[HR_{OS\ Ph3} \leq MV\right] \geq p$$

$$C2Ph3 = \begin{cases} \text{No-Go} \\ \text{Go} \end{cases}$$

- $p = 0.75\,; MV = 0.9$ used as an example, can be changed
- For OS as single endpoint and joint frailty model, $P[HR_{OS} < MV]$ can be estimated directly
- For methods which do not directly estimate a hazard ratio for OS, we used assumptions about how those analysis metrics relate to HR_{OS}
  - HR_{PFS} and Δ_{OR} related to HR_{OS} according to meta-analysis
  - WO equivalent to $1/HR_{OS}$





Blue dots are simulated clinical scenarios,
curves are based on unpublished meta-analysis

Decision-Making Rules for Various Methods

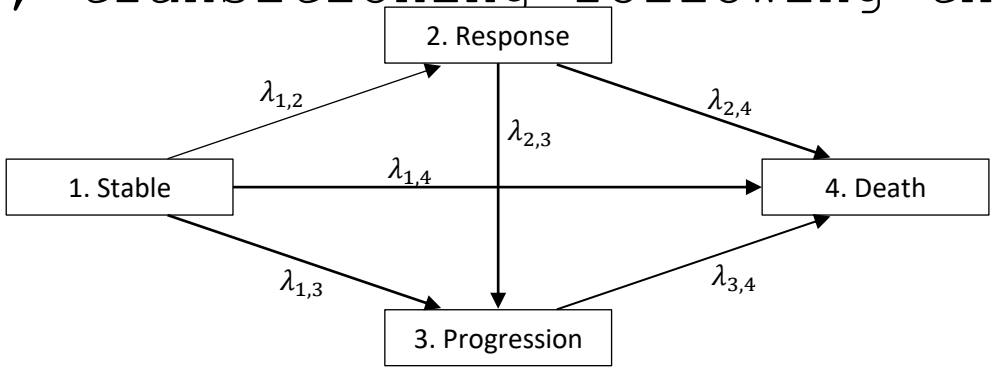
QDM Application example: If $MV_{OS} = 0.9$ then, $MV_{PFS} = 0.86$ and $MV_{ORR} = 0.05$

	Method	C2Ph3 "Go" Rule (else "No-Go")
Combined endpoint methods	Win Odds (OS, PFS, ORR)	$P[1/WO \leq MV] \geq 0.75$
	Joint frailty model (OS,ORR)	$P\left[HR_{OS\,Marg} \leq MV\right] \geq 0.75$
Single endpoint Methods	OS	$P[HR_{OS} \leq MV] \geq 0.75$
	PFS	$P[HR_{PFS} \leq MV_{PFS}] \geq 0.75$
	ORR	$P[\Delta_{ORR} \geq MV_{ORR}] \geq 0.75$
Gating methods	Gating PFS then OS	if $P[HR_{PFS} \leq MV^{-}_{PFS}] \geq 0.75$ else if $(P[HR_{PFS} \geq MV^{+}_{PFS}] < 0.75$ * and $P[HR_{OS} \leq MV] \geq 0.75)$
	Gating response then OS	if $P[\Delta_{ORR} \geq MV^{+}_{ORR}] \geq 0.75$ else if $(P[\Delta_{ORR} \leq MV^{-}_{ORR}] < 0.75$* and $P[HR_{OS} \leq MV] \geq 0.75)$

ORR endpoints to control for false positive rate

Multi-State Model (MSM*) for Data Generation

- Subjects move through the natural disease progression states during the study, transitioning following the arrows:

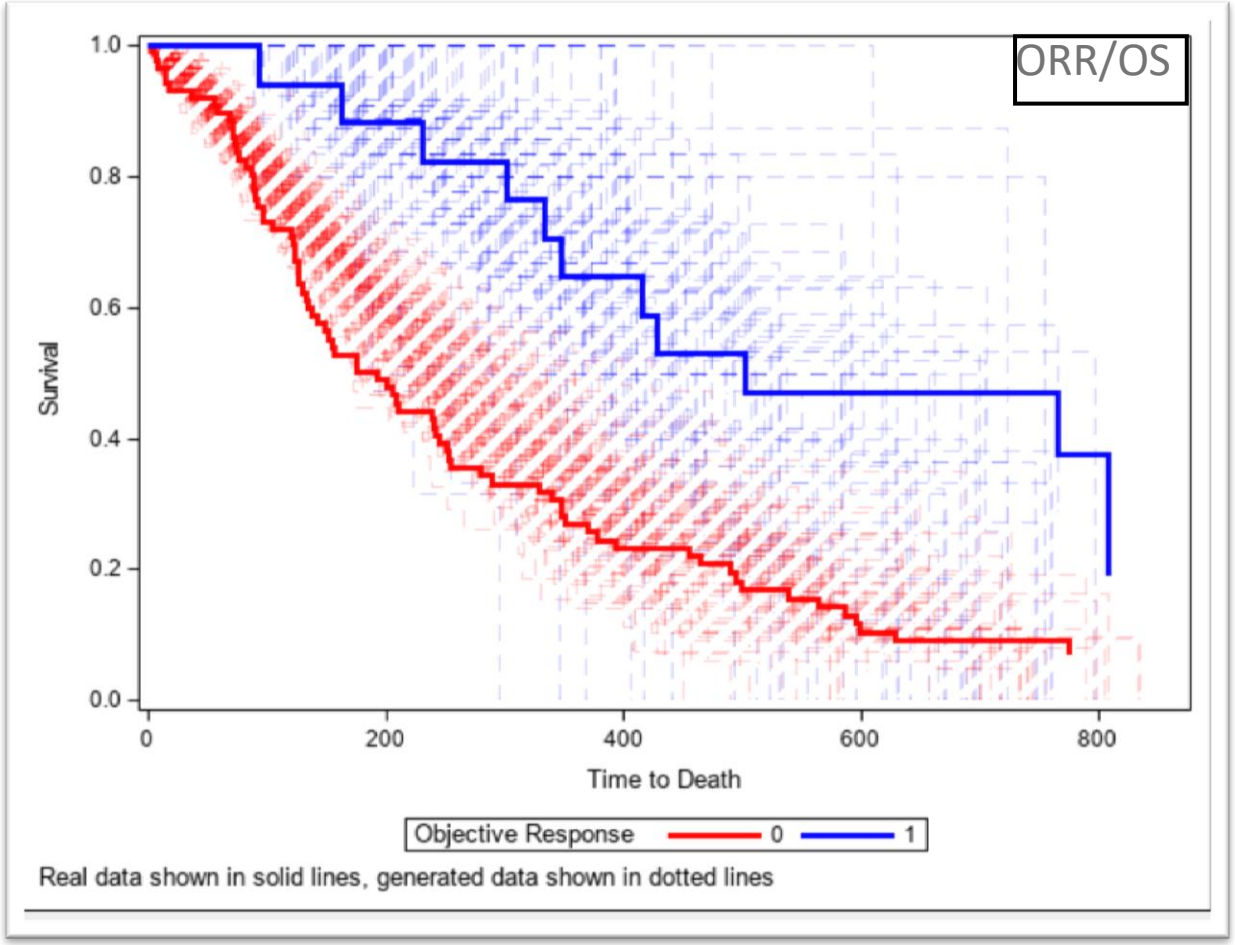


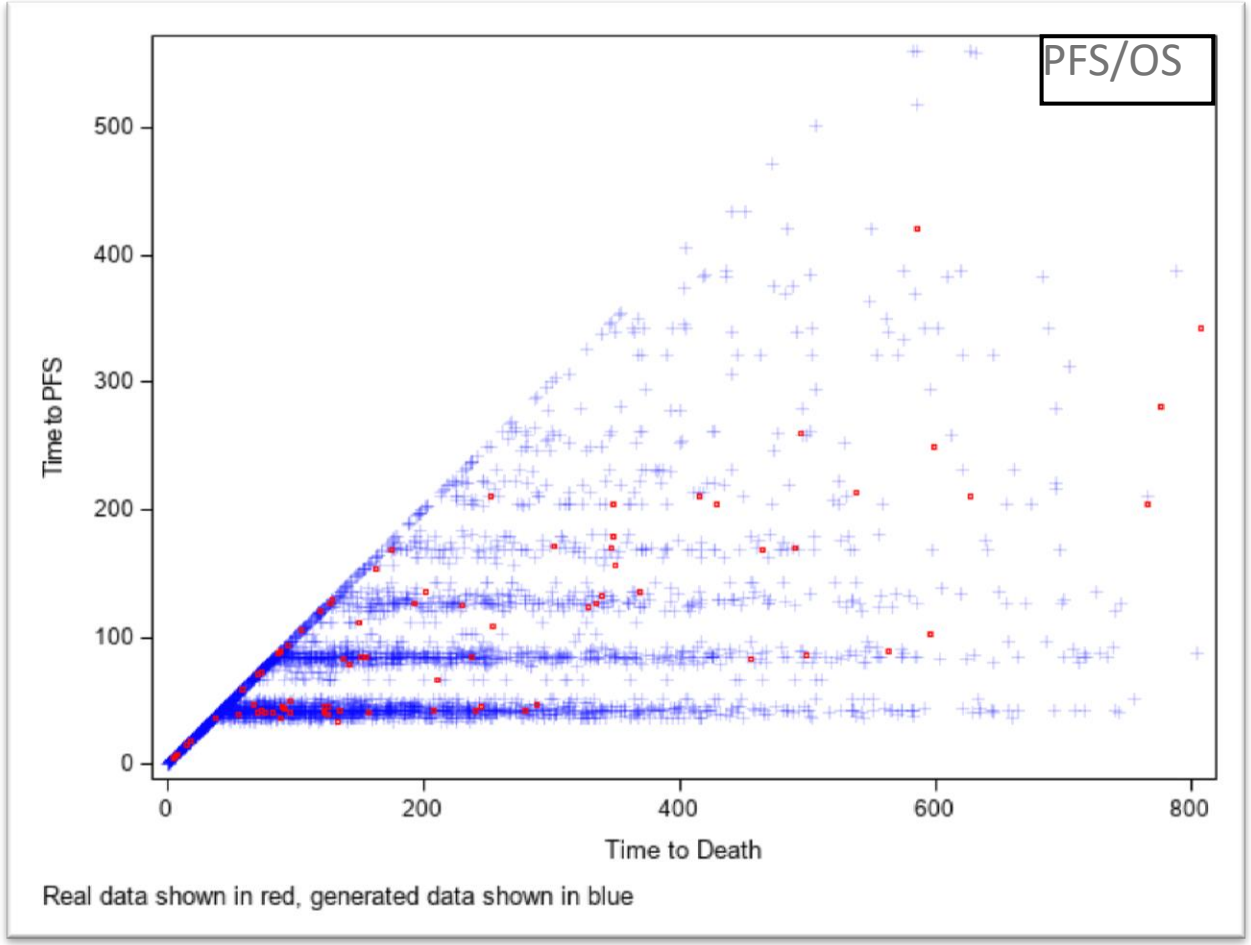
- When all transitions inactive ⇒ "global null" case
- We fitted historical data to MSM to obtain transition hazards for control arm (Slide 11)
- Then modified transition hazards $\lambda_{i,j}$ for the active treatment arm, to create treatment effects (with clinician's input)
- This was done for 2 case studies , resulting in 2 simulation studies

*Refs: (Meller, Beyersmann, & Rufibach, 2019); (Beyer, Dejardin, Meller, Rufibach, & Burger, 2019)

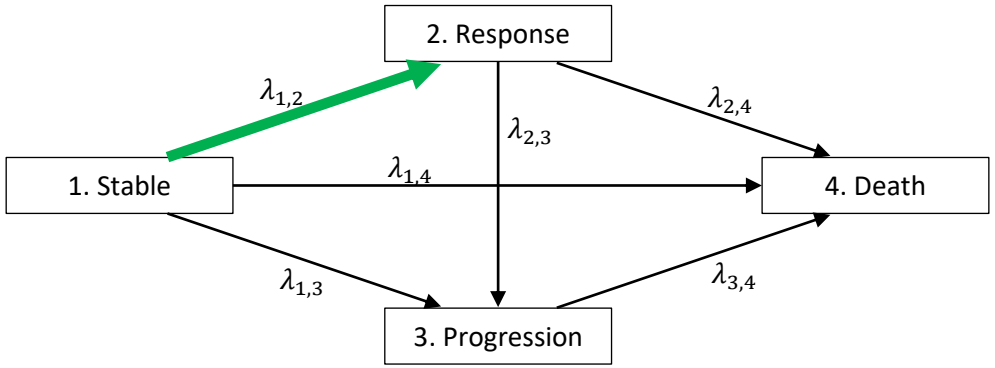
Illustration: MSM model fit in simulations study 1

Observed Data and 200 Simulated Datasets

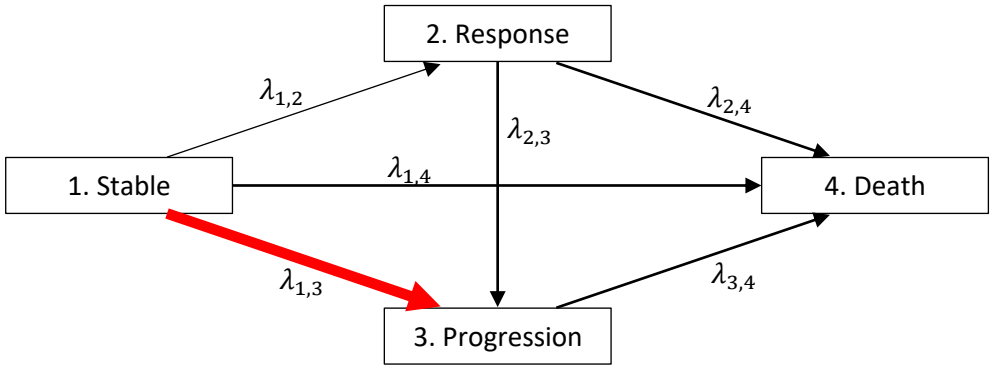




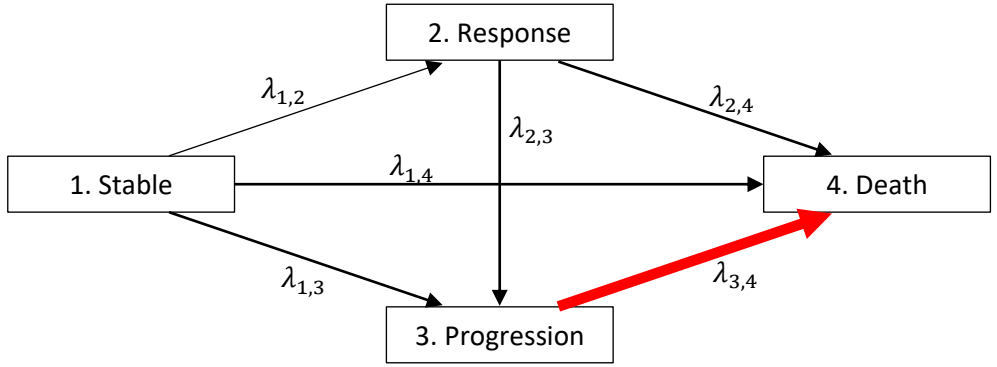
Single Transition Treatment Effects*



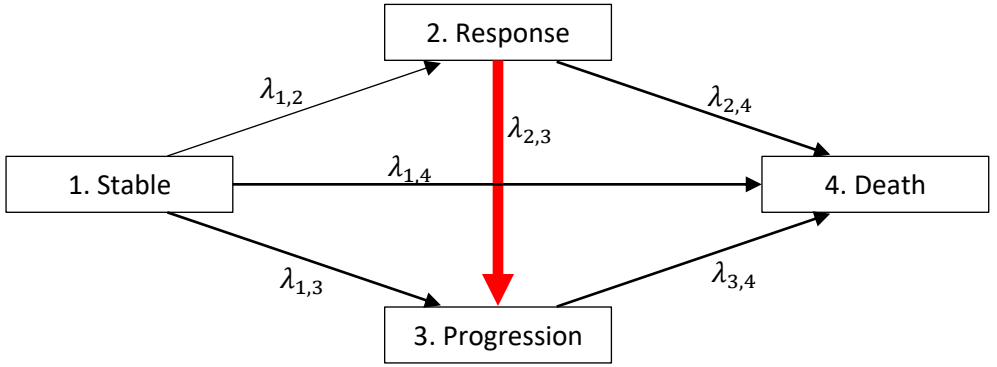
2. Increase Responders



3. Decrease Progression from Stable



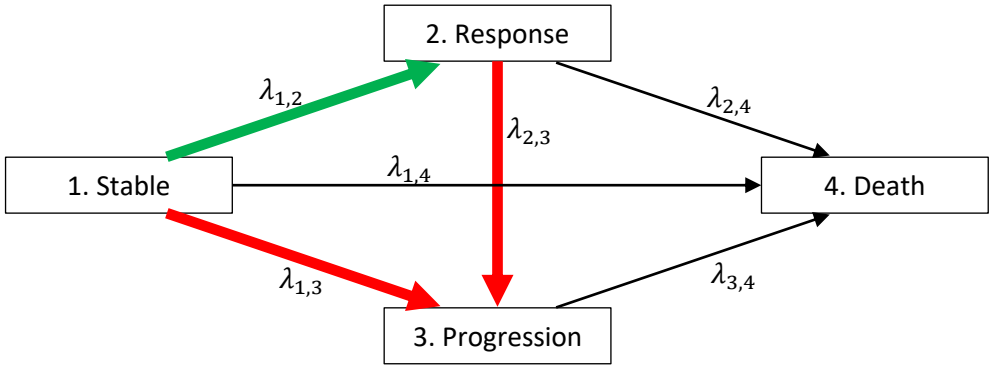
5. Decrease Post-Progression Mortality



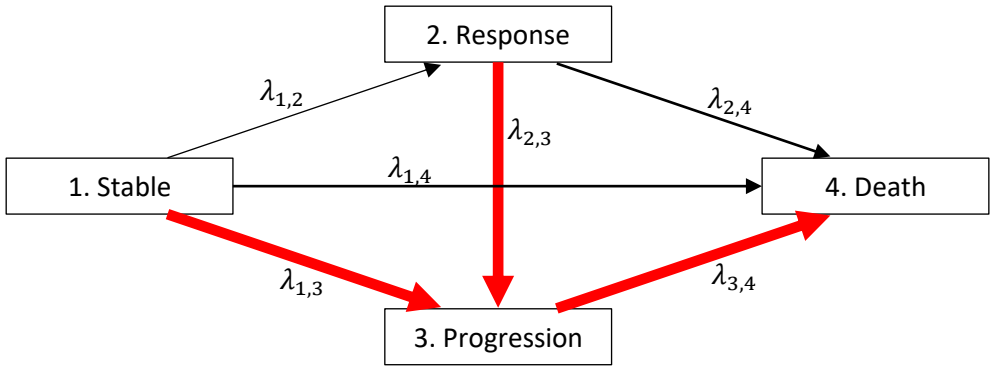
4. Decrease Progression from Response

- *hypothesized effects, not real studies

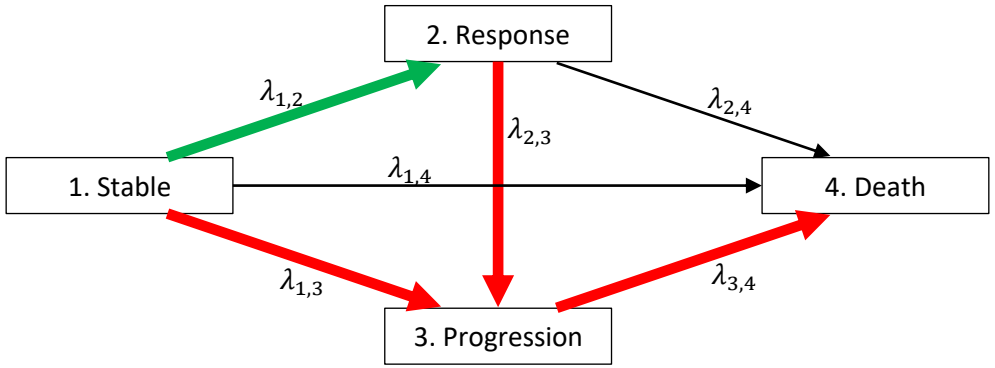
Positive Multiple Transition Treatment Effects*



6. Increase Response, Decrease Progression



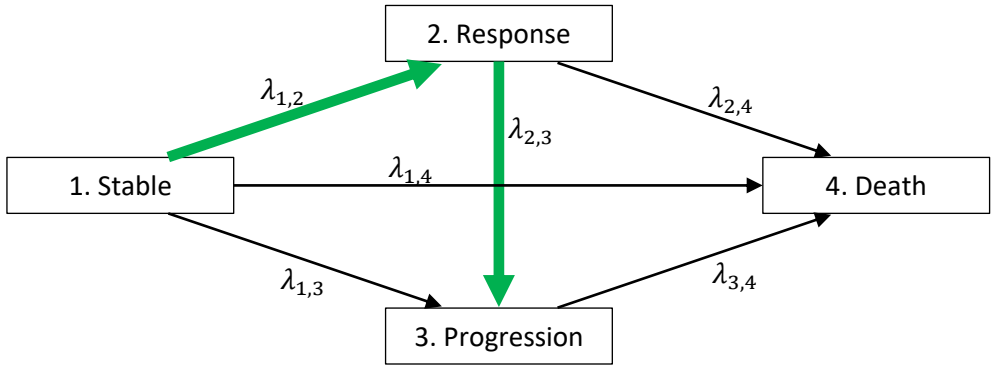
7. Decrease Progression, Post-Progression Mortality



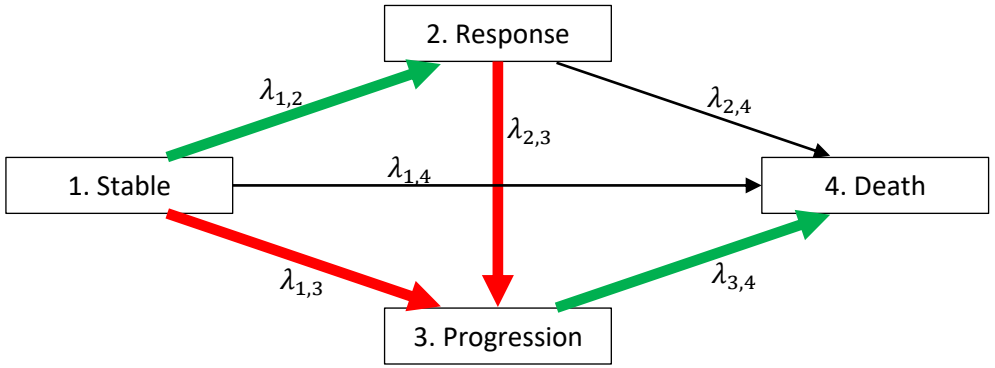
8. Super Drug: Increase Response and Decrease Progression, Post-Progression Mortality

*hypothesized effects, not real studies

Mixed Multiple Transition Treatment Effects*



9. Increase Responders with Rebound Progression



10. Increase Responders, Decrease Progression with Rebound

*hypothesized effects, not real studies

Simulation recap:

Goal: to evaluate the impact of novel Ph2 methods on C2PH3 decision

Methods

"Novel" bi-variate methods

- Win-ratio/win-odds
- Joint frailty

(All leverage OR + partial PFS/OS)

vs.

"Conventional" methods

- ORR only
- OS only
- PFS only
- Gating: ORR then OS
- Gating: PFS then OS

Data

Metrics:

- Pr (GO to Ph3| Ph2 data)

- **Scenarios of "Truth":**
  - 10 treatment effects for (OS,PFS,ORR) and their dependence structure
  - correlated at patient level
  - via Multi-State Model (MSM)
  - goes beyond simple trial-level correlation assumption
  - survival maturity (from ~50% to ~10% on control)

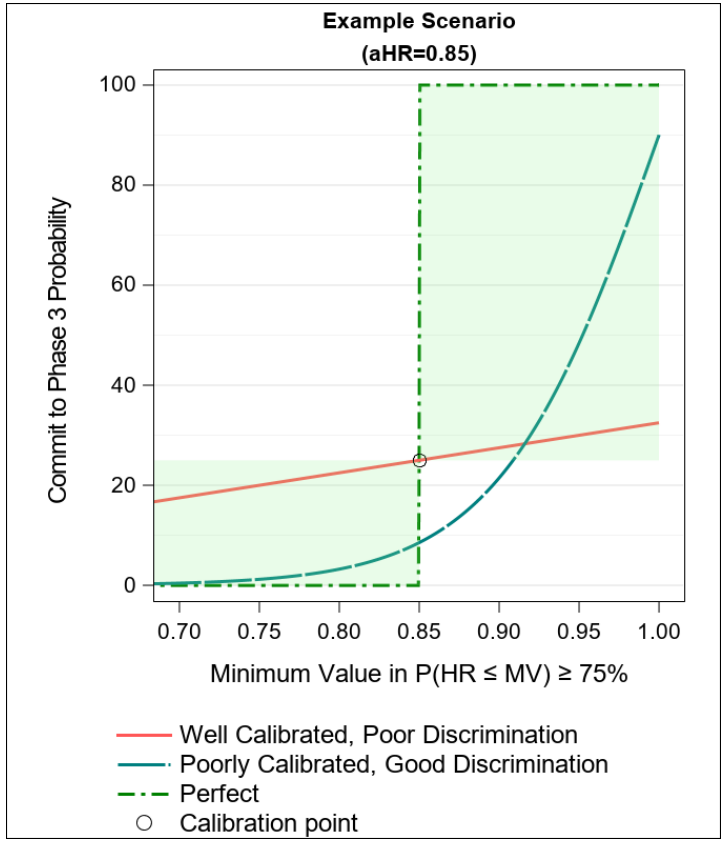
Decisions

- 1) Do novel methods incorporating surrogates +OS give better Ph3 GO/No-GO decision than benchmark conventional methods?

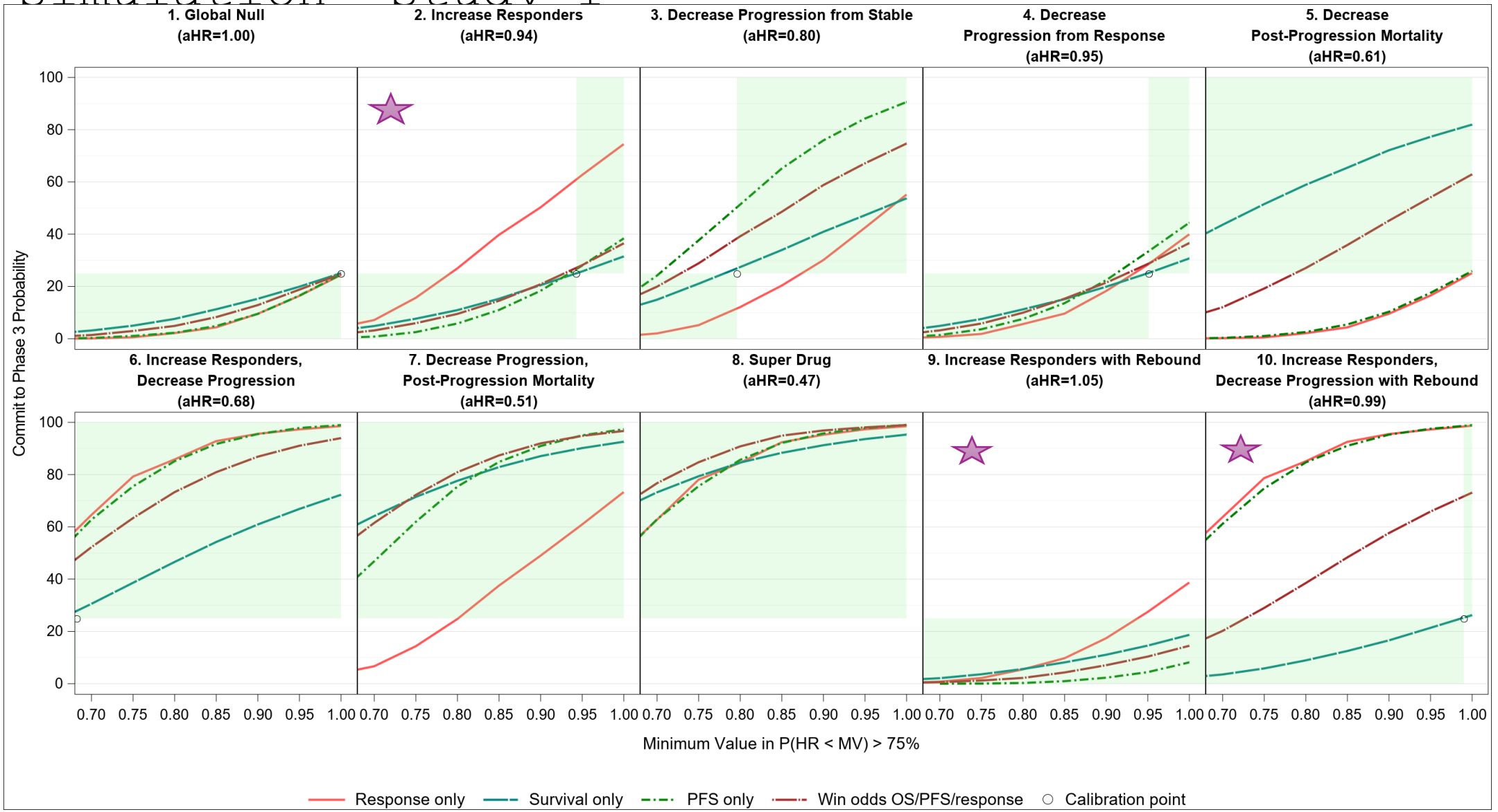
- 2) If yes, under what conditions?

Calibration: a "perfect" decision (assists with interpretation of results)

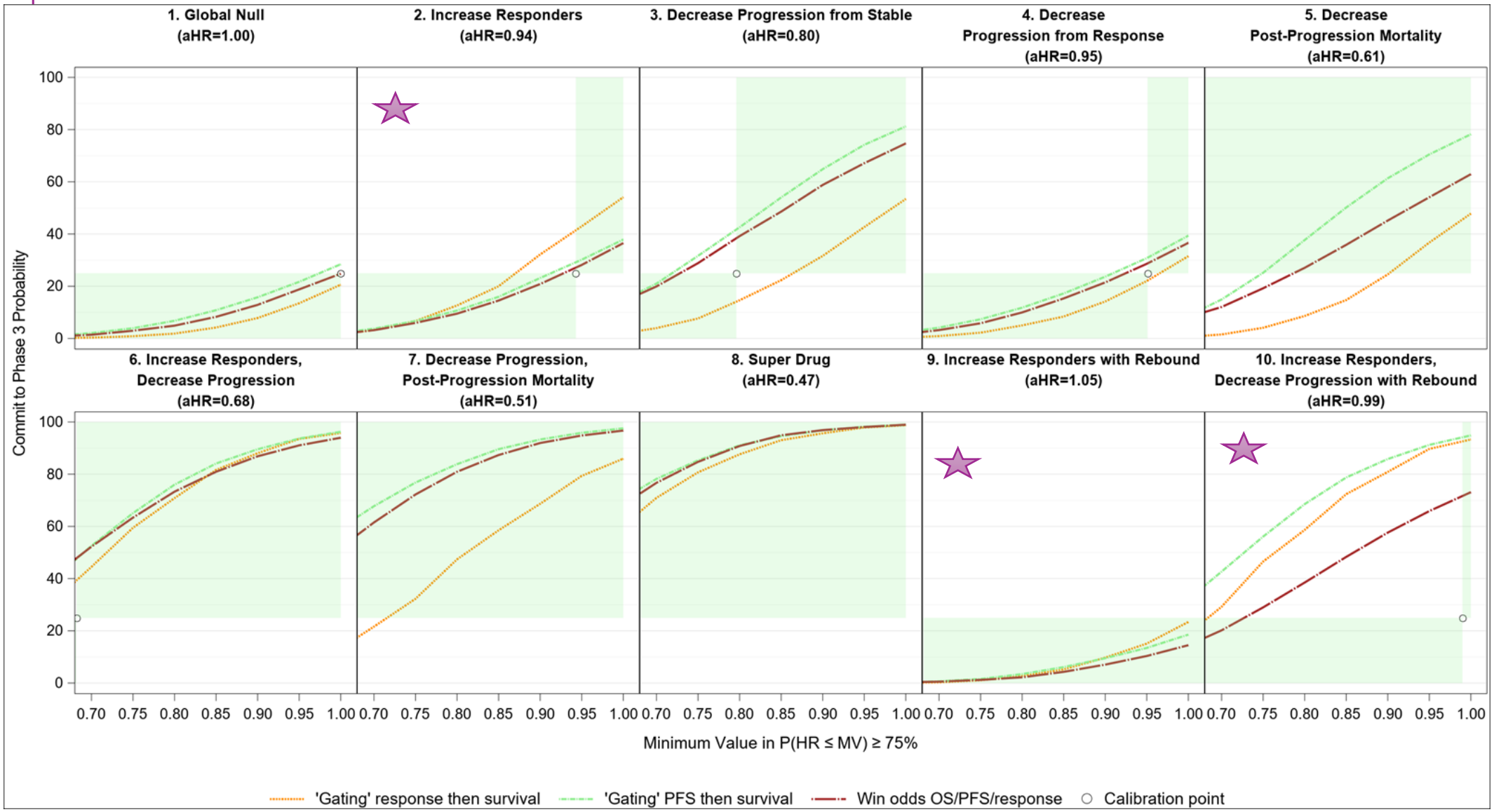
- C2Ph3 with 100% probability if true HR_{OS} < MV
- C2Ph3 with 0% probability if true HR_{OS} > MV
- C2Ph3 with 25% probability if true HR_{OS} = MV
  - This is the calibration of "type 1 error" when the target confidence threshold p in our quantitative decision-making rule is 0.75
  - This is how the " ideal" decision curve should look like
  - Never achievable in practice but we will look for closeness to a similar pattern in the simulation results that



Results: win odds vs single endpt. methods, simulation study 1

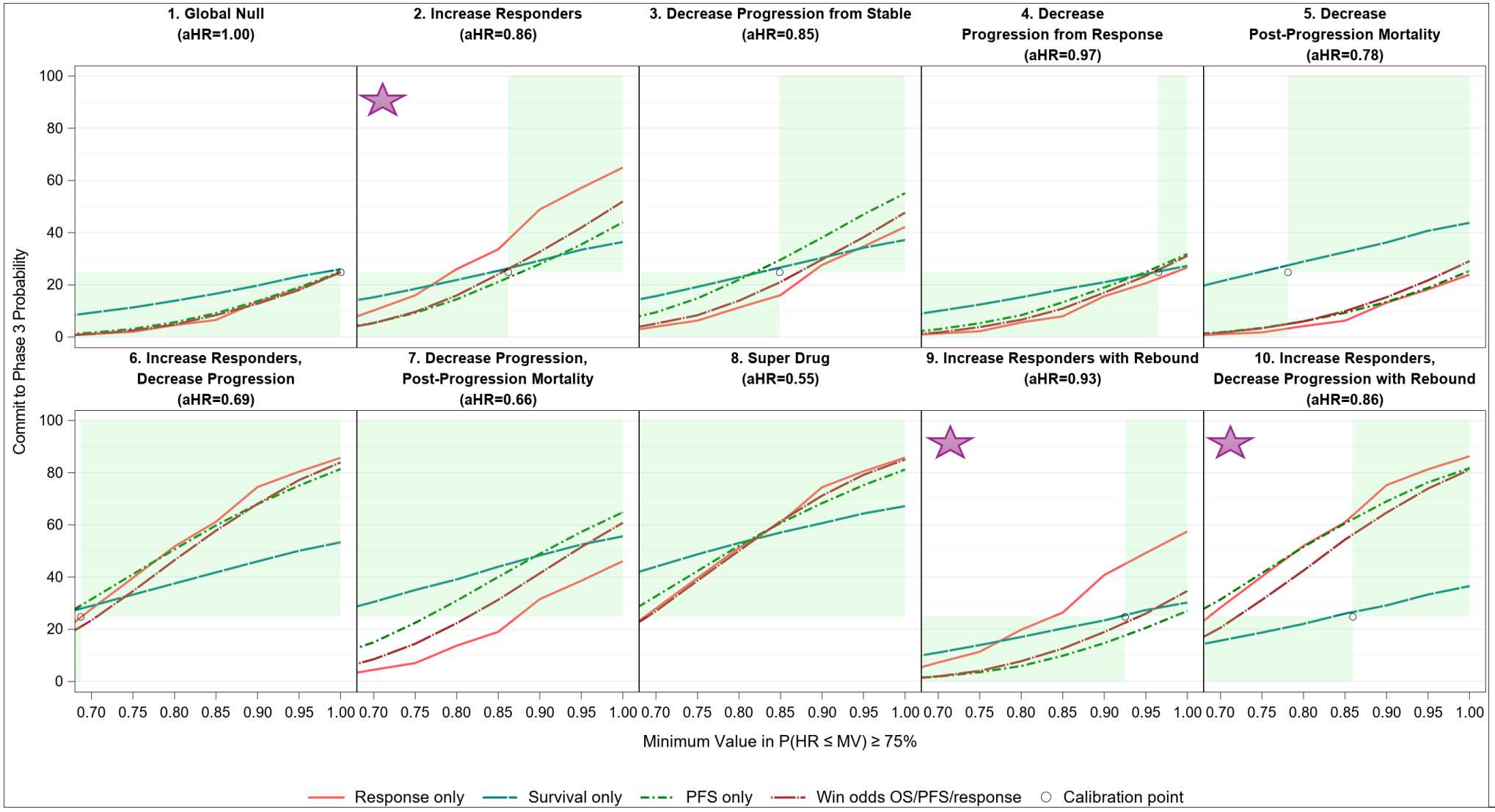


Results: win odds vs gating methods, simulation study 1



Simulation Results – case study 2; less mature OS

Consistent pattern, less dramatic difference between methods



Simulation Results Summary:

Win odds vs Other Methods

Based on our set of comprehensive (although not exhaustive) scenarios, simulations suggest that the **3-endpoint win odds** (ORR/PFS/OS) of Phase 2 data is **a robust C2Ph3 metric** and we **recommended it** as method of choice for internal decision-making:

- OS methods are well calibrated (unbiased) but poor discriminants (lack of precision)
- ORR and PFS methods are precise enough but may not translate into OS (bias)
- Gating C2Ph3 methods are significant improvement over single endpoint methods but they rely on a meta-analysis (or some intuition) to fix the position of the "gate".
- Win odds/ratios prioritise OS, and only move to PFS, and ORR in absence of more important events; can be interpreted without a meta-analytic relationship; improves precision by adding more information
- Win ratios suffer from instability in the presence of ties, win odds improve on this
- Joint frailty model estimates don't (generally) improve on marginal models of survival;
  - they are hard to implement

Case Study: Illustrating decision making

Treatment effect estimates on various endpoints

ENDPOINT	SUMMARY METRIC	ESTIMATE (95% CI)	FAVORS
Overall Survival	HR for OS	1.38 (0.87, 2.20)	Control
Progression-Free Survival	HR for PFS	0.82 (0.52, 1.28)	Active
Confirmed Objective Response	Difference in ORR	0.071 (-0.092, 0.204)	Active
Three Endpoints	Loss Odds (Reciprocal of Win Odds)	1.14 (0.71, 1.84)	Control
Three Endpoints	Loss Ratio (Reciprocal of Win Ratio)	1.15 (0.71, 1.89)	Control

(cont.)

Summary of Various C2Ph3 Methods and Decisions they Invoke

Method	Go Condition(s) Used For $P[HR_{OS} < 1]$	Probability of Go Condition(s)	Probability Level for C2Ph3 Decision		
			$p = 0.75$	$p = 0.65$	$p = 0.50$
Overall Survival	$P[HR_{OS} < 1]$	0.07	No-Go	No-Go	No-Go
Progression-Free Survival	$P[HR_{PFS} < 1]$	0.80	Go	Go	Go
Confirmed Objective Response	$P[\Delta_{ORR} > 0]$	0.65	No-Go	Go	Go
Gating PFS then OS	$P[HR_{PFS} < 0.8665]$ (else $P[HR_{OS} < 1]$)	0.59 (else 0.07)	No-Go	No-Go	Go (PFS)
Gating Confirmed Objective Response then OS	$P[\Delta_{ORR} > 0.0617]$ (else $P[HR_{OS} < 1]$)	0.55 (else 0.07)	No-Go	No-Go	Go (response)
Three Endpoint Win Odds	$P[1/WO_{OS,PFS,ORR} < 1]$	0.28	No-Go	No-Go	No-Go

Caveats

We have assumed Win Odds can be interpreted as the reciprocal of HR

- It may be a strong assumption; remains to be seen how it holds in practice
- Our 2 case studies were in same tumor type
- In principle, our methods are not histology-specific, but it needs to be tested in different cancers/populations

The MSM can be useful beyond just data generation:

- Visualisation of the mode of action of a drug
- Illustration of non-proportionality of OS hazards
- Can be used to predict OS ⇒ use for decision-making (further work)
  - MSM is a simplified categorical model of longitudinal tumour burden
  - Can be compared with more complex fully continuous models of tumour kinetics to understand its potential value as a predictive tool

Other related work in this area

Incorporating tumor growth trajectory into OS prediction

1. Struemper et al. (2025) "*Development of a Joint Tumor Size-Overall Survival Modeling and Simulation Framework Supporting Oncology Development Decision-Making*"
   - Developed a treatment-agnostic model linking tumor size dynamics to overall survival
   - Model developed using data of 786 NSCLC patients from seven trials.
   - Using tumor growth rate and patient covariates, the model accurately predicts survival across various therapies
   - It aids early decision-making, trial design, and regulatory support but requires further validation, especially for targeted therapies and early-stage NSCLC.
2. Wang et al. (2009): *"Elucidation of Relationship Between Tumor Size and Survival in Non-Small-Cell Lung Cancer Patients Can Aid Early Decision Making in Clinical Drug Development"*
   - NSCLC survival is influenced by baseline tumor size and early treatment response, with larger tumors linked to worse outcomes. Early tumor shrinkage (PTRwk8) is a strong predictor of overall survival and serves as a useful surrogate endpoint in trials. The validated model incorporates tumor size changes and survival factors to accurately forecast survival and support early-phase trial decisions, aiding drug development and Phase III trial design

- Both use idea similar to ours (predict OS based on earlier endpoint and improve decision-making) but leverage full TG trajectory rather that its binary "derivative" (ORR)
  - Intuitively, that should lead to more accurate prediction, but validation of that claim is needed comparing the methods side-by side.
  - Both works above rely heavily on specific model/disease assumptions
  - Computing full TG trajectory (and collecting data for it) is very expensive . MSM offers a "compromise" by capturing just a few discrete states of disease progression rather than full curve but aiming to make same decision (OS prediction from where the patient is in their trajectory)

Other related work in this area

Word of caution against using win statistics as an endpoint

- Thompson et al. (2025) discuss *"Interpretational challenges of the win ratio in analyzing hierarchical composite endpoints in chronic kidney disease using the Win Ratio (WR) in chronic kidney disease"*
  - It a collaborative work of multiple statisticians across different companies and academia
  - Main advise is to use WR as a discriminatory measure (to establish presence of treatment effect
  - But avoiding using it as an estimate of such effect due to various challenges with its interpretation
    - To mitigate these challenges, they advise to complement it with component analysis, avoid mixing outcomes of different severities, use fixed follow-up periods, stratify analyses, and be cautious in regulatory settings.
  - They also suggest considering alternatives like net benefit and win-odds
  - WR has been on the rise in popularity recently due to its ability to boost "power" in decision making. But as we bring in more endpoints (for power), we often use interpretability.
  - WR has been used in FDA-approved trials (e.g., Tafamidis, Acoramidis) but more clear guidelines are needed and authors do not recommend to use it as the sole basis for regulatory decisions
- The messaging is consistent with our work (although setting is different-CKD), in a sense that we advocate use win-odds

Other related work in this area (cont.)

Don't stop at Ph2, follow-up Ph2 patients and further enhance decision making with seamless ph2/3 and multiple endpoints

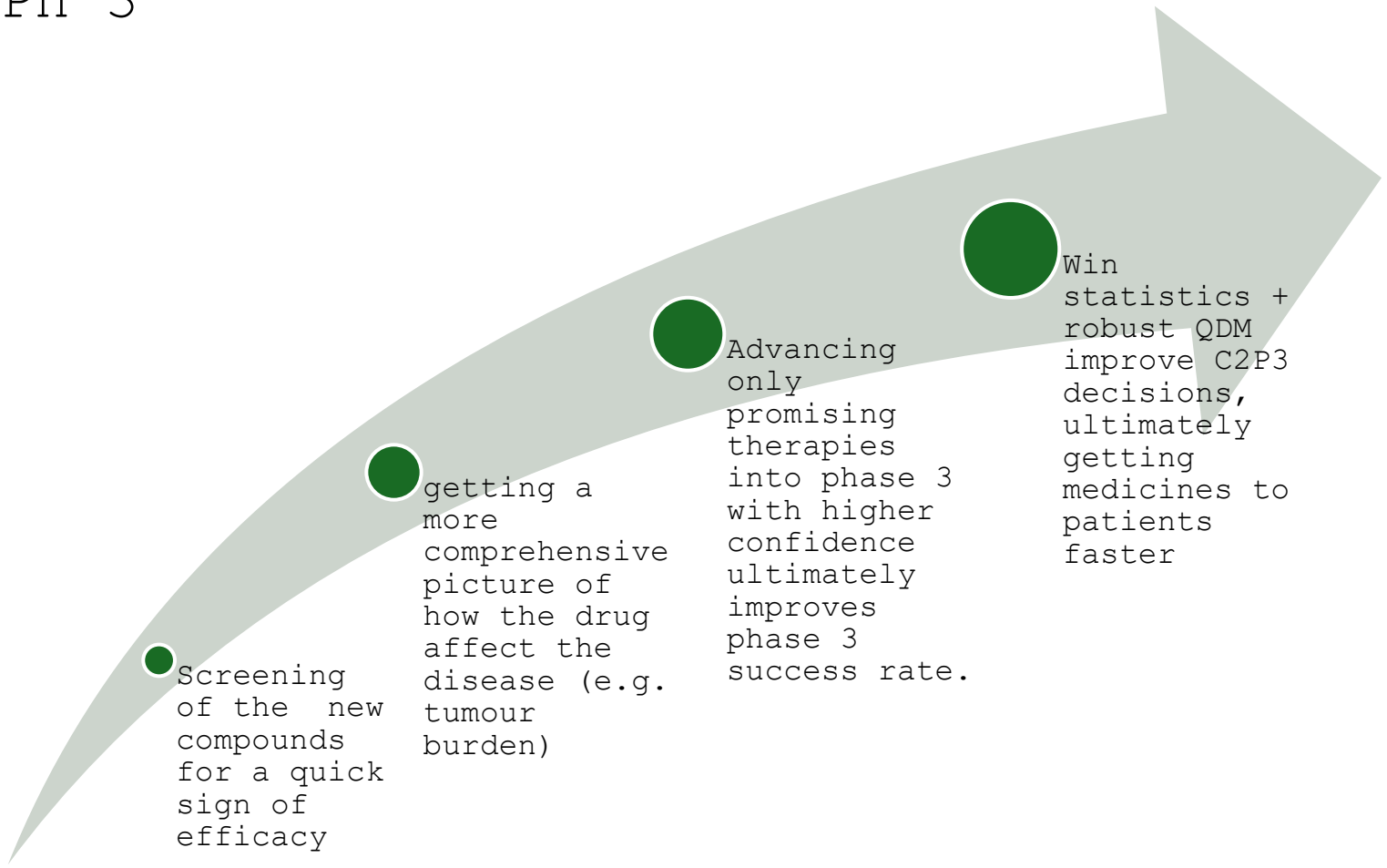
- Gotte et al. (2020) *"Optimal decision-making in oncology development programs based on probability of success for phase III utilizing phase II/III data on response and overall survival. "*
  - Same goal as our work: focus on improving probability of success (PoS) in Phase III trials by optimizing ph2 decisions and leveraging multiple endpoints
  - Incorporate formal QDM (decisions based on estimated PoS)
  - Integrate short-term (e.g., response) and long-term (e.g., overall survival) endpoints
  - Focus on ORR and OS relationship but model it differently
  - Extend the concept of decision-making in separate ph2 and ph3 trials to *Ph2+ design*:
    - Initial go/no-go decision based on Phase II data
    - Continued follow-up of Phase II patients to get more mature OS
    - Interim analysis during Phase III using updated Phase II data (including extra follow-up)
    - Option to stop Phase III early if updated data indicates low PoS
    - Can be extended to fully seamless Ph2/3 designs (future work)
  - As a result, their simulation show reduced risk of premature or costly Phase III investments, particularly in cases when the drug affects response but does not affect survival.
- Different methods, consistent messaging and approach rooted in robust quantitative decision-making

Summary and Conclusions

- Making good C2P3 decisions in oncology drug development is a challenging problem:
  - Relationships between the OS (registration endpoint), and PFS/ORR (Ph2 endpoint) are often poorly understood
  - OS is not mature enough in ph2 to make decisions solely on OS
- We have presented a novel framework for Ph2 decision-making based on combining OS with surrogate endpoints:
  - 3-endpoint win-odds: OS→PFS→ORR
  - Quantitative decision-making based on OS achieving clinically meaningful level (MV) with pre-specified probability (e.g. 75%)
- And assessed comparative performance of win odds vs conventional methods using data generated with MSM in a comprehensive simulation study
  - Parameters of MSM for the control arm were derived from real data (2 case studies with varying level of mortality)
  - 10 treatment effects were created based on clinician's input reflecting varying mechanism of action and magnitude of trt effect
- Our approach of using MSM to jointly simulate correlated patient level RECIST endpoints helps to understand the relationships between treatment effects on ORR, PFS and OS and creates clinically realistic data for testing the performance of various methods being assessed. Additional benefits MSM beyond data generation include visualization of non-proportional hazards and predictive potential
- Multiple related approaches exist in the literature echoing our key messaging:
  - Combining OS with surrogate endpoints enhances Phase 2 decision-making, potentially reducing Phase 3 failures and accelerating drug development
  - Assumptions about win odds interpretation and tumor type generalizability require further validation. It's advisable to use win-odds as decision-making tool, not estimation
  - Other related work includes tumor growth modeling and Phase 2+plus design that models OS/ORR relationship differently (via mixture distribution). Both claim increase precision of decision-making after Phase by incorporating more information.

Final Remarks

Evolution of Ph 2 decision-making to reduce risk of failure in Ph 3



References

1. Abellan, J. J., Nicolas Bonnet, A. C., Frewer, P., Gotte, H., Lawo, J.-P., Madsen, J., . . . Saint-Hilary, G. (n.d.). Implementation and practical aspects of quantitative decision-making in clinical drug development. *arXiv*.
2. Beyer, U., Dejardin, D., Meller, M., Rufibach, K., & Burger, H. U. (2019). A multistate model for early decision-making in oncology. *Biometrical Journal*.
3. Brunner, E., Vendemeulebroecke, M., & Mütze, T. (2021). Win odds: An adaptation of the win ratio to include ties. *Statistics in Medicine, 40*.
4. Dong, G., Hoaglin, D. C., Qiu, J., Matsouaka, R. A., Chang, Y.-W., Wang, J., & Vandemeulebroecke, M. (2020). The Win Ratio: On Interpretation and Handling of Ties. *Statistics in Biopharmaceutical Research, 12*(1).
5. Dutang, C., Goulet, V., & Pigeon, M. (2008). actuar: An R Package for Actuarial Science. *Journal of Statistical Software, 25*(7).
6. Eisenhauera, E., Therasse, P., Bogaerts, J., Schwartz, L., Sargent, D., Ford, R., . . . Verweij, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 228 – 247.
7. Entree Lung Study Report. (n.d.).
8. Erdmann, A., Beyersmann, J., & Rufibach, K. (2023). Oncology clinical trial design planning based on a multi-state model that jointly models progression-free and overall survival endpoints. *ArXiv*.
9. FDA. (2021). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products: Guidance for Industry, Draft Guidance Revision 1.
10. Ferreira, J. P., Jhund, P. S., Duarte, K., Claggett, B. L., Solomon, S. D., Pocock, S., . . . McMurray, J. J. (2020). Use of the Win Ratio in Cardiovascular Trials. *JACC: Heart Failure, 8*(6).
11. Hernán, M. A. (2010). The Hazards of Hazard Ratios. *Epidemiology, 21*(1).
12. Jackson, C. (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software, 38*(8).

References(cont.)

13. Meller, M., Beyersmann, J., & Rufibach, K. (2019). Joint modelling of progression-free and overall survival and computation of correlation measures. *Statistics in Medicine*, 4270-4289.
14. Miettinen, O., & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine, 4*(2).
15. Oakes, D. (2016). On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika, 103*(3).
16. Pembrolizumab Study Report. (n.d.).
17. Pocock, S. J., Ariti, C. A., Collier, T. J., & Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal, 33*.
18. Wang, D., & Pocock, S. J. (2016). A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharmaceutical Statistics, 15*(3).
19. Wreede, L. C., Fiocco, M., & Putter, H. (2011). mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*.
20. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. Biometrics 2004; 60:747–756
21. Rogers JK, Yaroshinsky A., Pocock SJ, Stokard D., Pogod J. Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. Statistics in Medicine, 2016, 35 2195–2205
22. Struemper, H., Rathi, C., Muliaditan, M., Goulooze, S.C., Franzese, R.C., Mantero, A., Melhem, M., Post, T.M. and Visser, S.A.G. (2025), Development of a Joint Tumor Size–Overall Survival Modeling and Simulation Framework Supporting Oncology Development Decision-Making. CPT Pharmacometrics Syst Pharmacol
23. Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock E, Gobburu J. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. Clinical Pharmacology & Therapeutics. 2009 Aug;86(2):167-74.
24. Henrik F. Thomsen, Samvel B. Gasparyan, Julie F. Furberg, Christoph Tasto, Nicole Rethemeier, Patrick Schloemer, Tuo Wang, Niels Jongs, Yu Du, Tom Greene. Interpretational challenges of the Win Ratio in analyzing Hierarchical Composite Endpoints in Chronic Kidney Disease. arXiv:2504.05909 (2025)
25. Götte H, Xiong J, Kirchner M, Demirtas H, Kieser M. Optimal decision-making in oncology development

Thank you

Bristol Myers Squibb®


Bristol Myers Squibb®