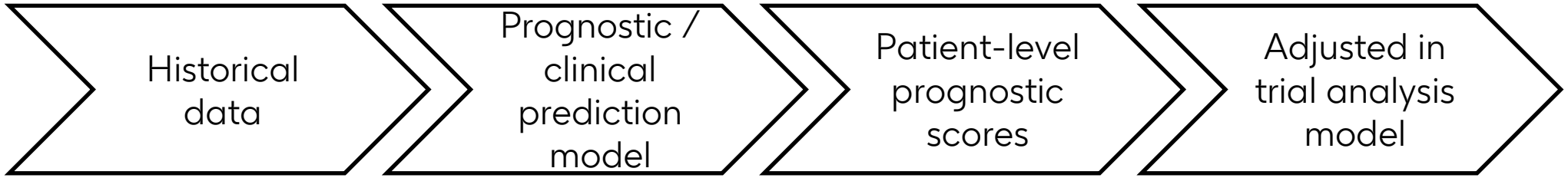
Incorporating prognostic scores in time-to-event analysis

Harry Parr, Doug Thompson, Tasos Papanikos, Aris Perperoglou

Why prognostic adjustment?

Prognostic adjustment can improve trial efficiency

- Adjusting for baseline prognostic information can help to explain some of the variability in clinical trial outcomes
- Namely to improve precision in estimating the treatment effect
- Prognostic scores combine multiple baseline features into one ‘composite’ derived score
- These prognostic scores are derived from historical data and act as a bridge to then include into prospective/contemporary trial analysis.



HP1

Regulatory positions – FDA & EMA

Relevant commentary from the regulators on covariate adjustment and its value

U.S. FOOD & DRUG ADMINISTRATION

Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products

Guidance for Industry

III. RECOMMENDATIONS FOR COVARIATE ADJUSTMENT IN CLINICAL TRIALS

*"Covariate adjustment leads to efficiency gains when the covariates are **prognostic** for the outcome of interest in the trial. Therefore, FDA recommends that sponsors adjust for **covariates that are anticipated to be most strongly associated with the outcome of interest**. In some circumstances these covariates may be known from the scientific literature. In other cases, it may be useful to **use previous studies** (e.g., a Phase 2 trial) to select prognostic covariates or **form prognostic indices**."*

U.S. FOOD & DRUG ADMINISTRATION

Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products

Discussion Paper and Request for Feedback

"At an even more personalized level, AI/ML can also be used in the context of **digital twins** of patients, an emerging method that could potentially be used in clinical research. To create digital twins of patients, **AI/ML can be utilized to build in silico representations or replicas of an individual** that can dynamically reflect molecular and physiological status over time (European Medicines Agency, 2022; Laubenbacher, Sluka, & Glazier, 2021; Schuler et al., 2021). [...] the digital twin could potentially provide a comprehensive, longitudinal, and computationally generated **clinical record that describes what may have happened to that specific participant if they had received a placebo**."

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

20 September 2022
EMADOC-1700519818-907465
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)

Draft agreed by Scientific Advice Working Party (SAWP)	10 February 2022
Adopted by CHMP for release for consultation	24 February 2022[1]

"**CHMP qualifies PROCOVA as prognostic score adjustment** and the proposed procedures, as described in a handbook for trial statisticians, **could enable increases in power or precision of treatment effect estimates** in controlled randomised clinical trials with continuous outcomes. ... Approaches with **non-linear models for analysis** and direct comparisons to such models, as well as models with treatment-by-covariate interactions **are out of scope of this qualification procedure.**"

GSK

Why prognostic adjustment is not straightforward for TTE

- For continuous outcomes, adjustment usually improves precision, not the estimand.
- For hazard ratios, adjusted Cox models estimate a **conditional HR**.
- The trial-level target is often a **marginal effect**.
- Therefore, apparent power gains from Cox adjustment are a consequence due to non-collapsibility.
  - Precision may decrease (i.e. SEs get larger)
  - Estimand shift away from the null due to non-collapsibility
- With HRs, adjustment can change both precision and interpretation. For the conditional cox, the estimate move further away from the null at a faster rate than the SE increase

Why g-computation and how's it calculated?

Monte Carlo g-computation - sample pseudo-populations from the fitted Cox baseline hazard

For each patient, use the fitted model to estimate survival under each treatment:

$$A_i \in \{0,1\}$$

$$h(t|A_i, X_i, PS_i) = h_0(t)\exp\{\beta_A A_i + \beta_X^T X_i + \beta_{PS} PS_i\}$$

Predict survival for each patient twice for each arm

$$\hat{S}_i^{A=0 \text{ then } 1}(t) = \hat{S}(t|A_i = 0 \,\&\, 1, X_i, PS_i)$$

$$\hat{S}^{(\mathrm{A}=0)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_i^{(0)}(t) \quad \& \quad \hat{S}^{(A=1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_i^{(1)}(t)$$

$$\widehat{\Delta}_s(t) = \hat{S}^{(A=1)}(t) - \hat{S}^{(A=0)}(t)$$

What is restricted mean survival time (RMST)?

RMST is the expected event-free survival time up to a fixed time horizon τ

RMST(τ) = difference in area under the survival curves from time 0 to τ, i.e.

$$\Delta\text{RMST}(\tau) = \int_0^{\tau} \hat{S}^{A=1}(t)\,dt - \int_0^{\tau} \hat{S}^{A=0}(t)\,dt$$

e.g.

RMST treatment = 10 ½ months

RMST control = 9 months

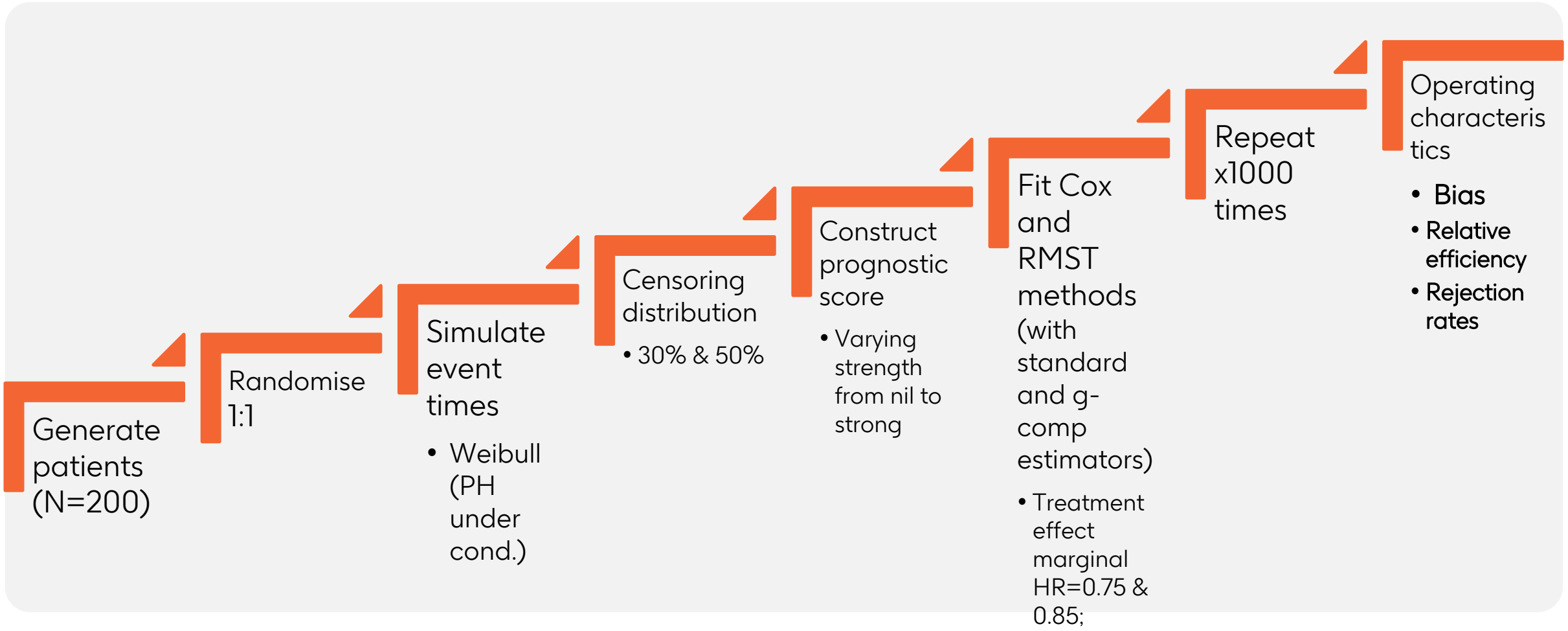
RMST difference = 1 ½ months of additional event-free time to tau

RMST is marginal and collapsible

It gives an absolute event-free time contrast

It avoids the HR non-collapsibility issue

Simulation study overview



DGM and prognostic score construction

Observed and latent prognosis

Each patient has a total prognostic index:

$$\eta_i = \eta_{\text{obs},i} + \eta_{\text{lat},i}$$

No treatment × covariate interactions are simulated (prognostic effects only)

Observed component:

$$\eta_{\text{obs},i} = X_i^{\top} \beta_X$$

Latent component:

$$\eta_{\text{lat},i} \sim N\left(0, \sigma^2_{\text{lat}}\right)$$

Prognostic score

$$PS_i = \eta_{\text{lat},i} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2_{PS})$$

A lower σ^2_{PS} means stronger/better recovery of latent prognosis

Simulation scenarios & method comparison

Scenario definitions and analysis comparison

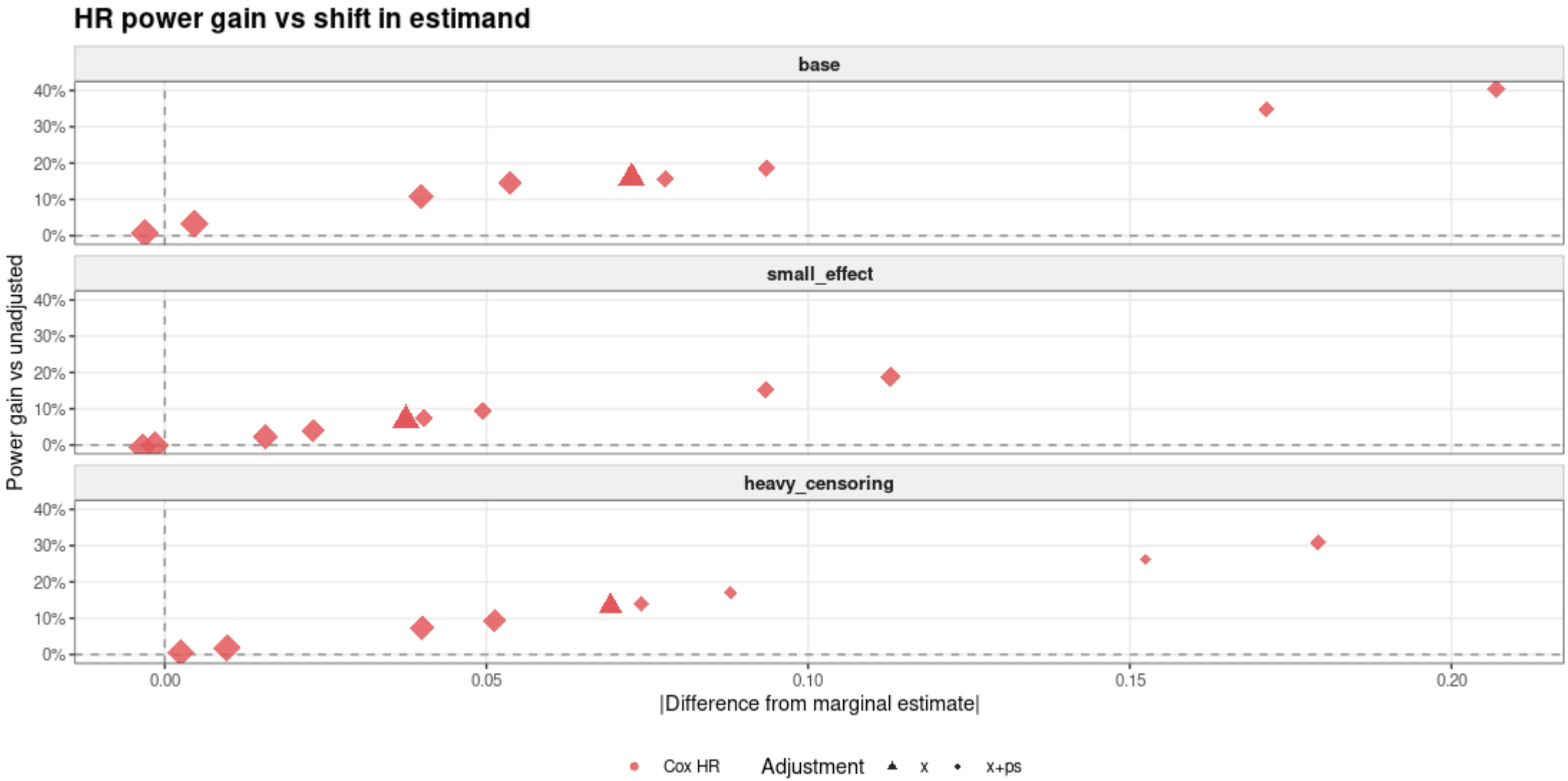
treatment-only, +x, +PS, +x+PS

Strength	$noise_{SD}$	Interpretation
nil	5.0	Random noise
weak	2.0	Faint prognostic signal
moderate	0.5	A useful but imperfect score
strong	0.1	A near-perfect recovery of 'latent' risk (i.e. the oracle)

Scenarios	Family / Models
Standard: 30% NI censoring, HR=0.75, $\widehat{RMST} = 2.3$	Cox HR g-comp
Small effect: 30% NI censoring, HR=0.85, $\widehat{RMST} = 1.3$)	RMST Kaplan-Meier
Heavy censoring: (50% NI censoring, HR=0.75, $\widehat{RMST} = 2.3$)	RMST g-comp

Results

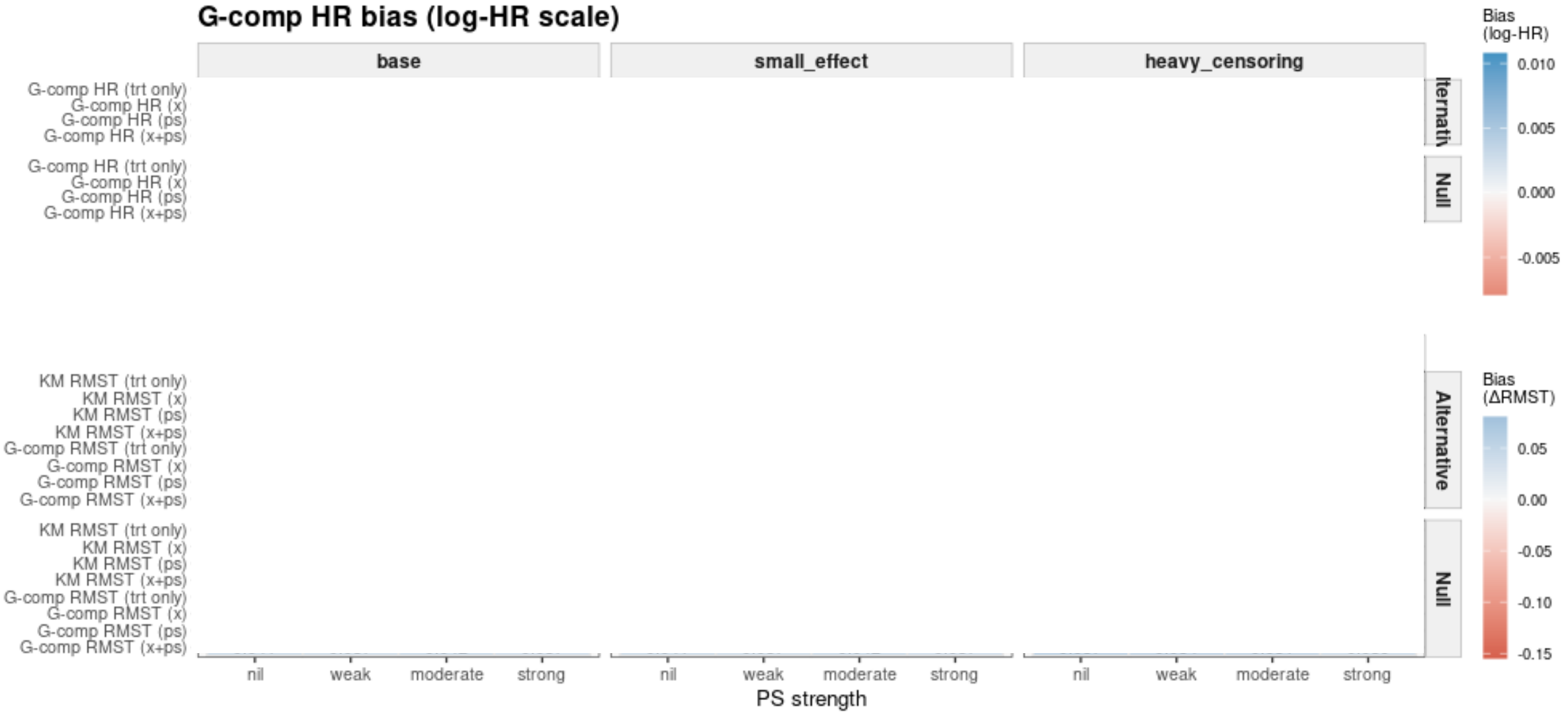
The power-bias-estimand trade-off for a HR



Bias

Bias (each method vs own estimand)

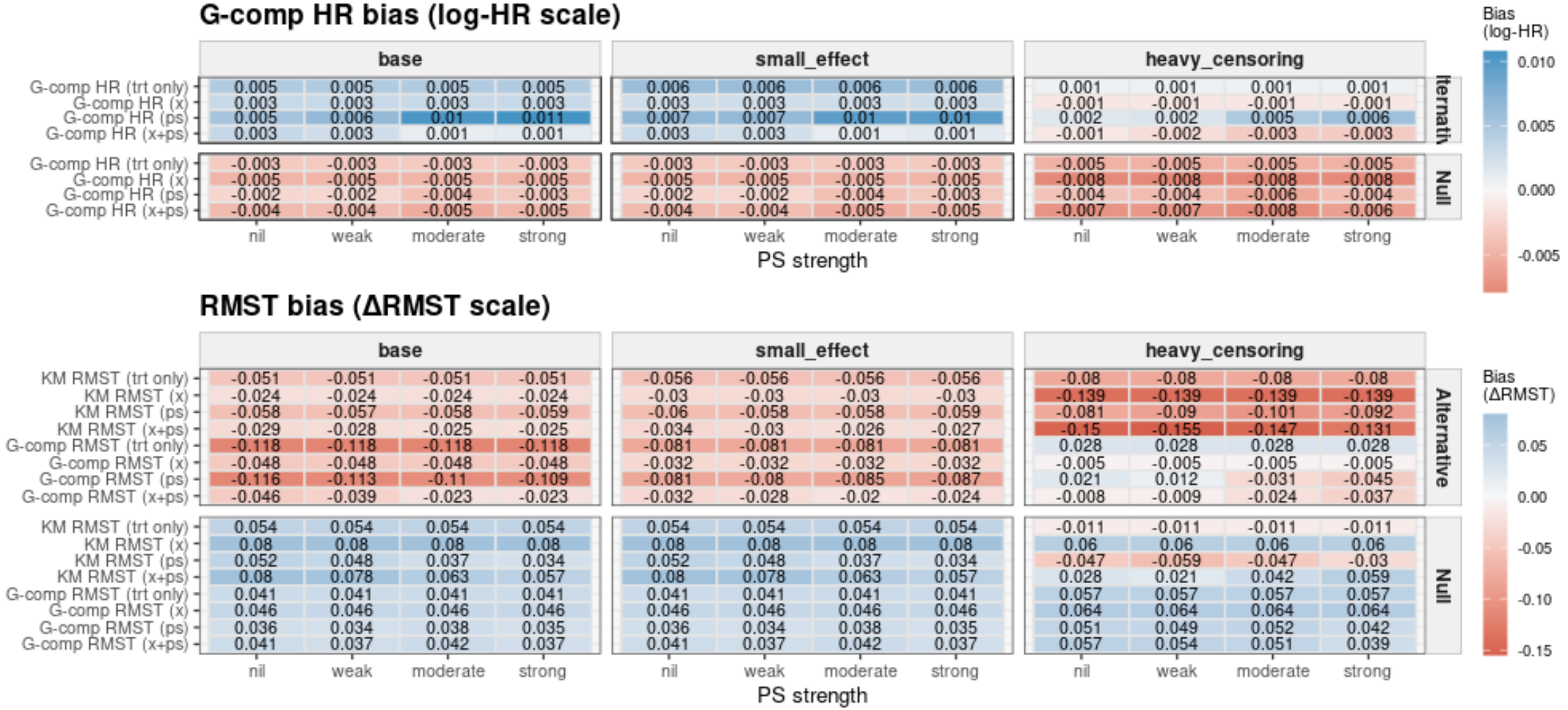
Red = underestimates effect; blue = overestimates | Separate scales for HR and RMST



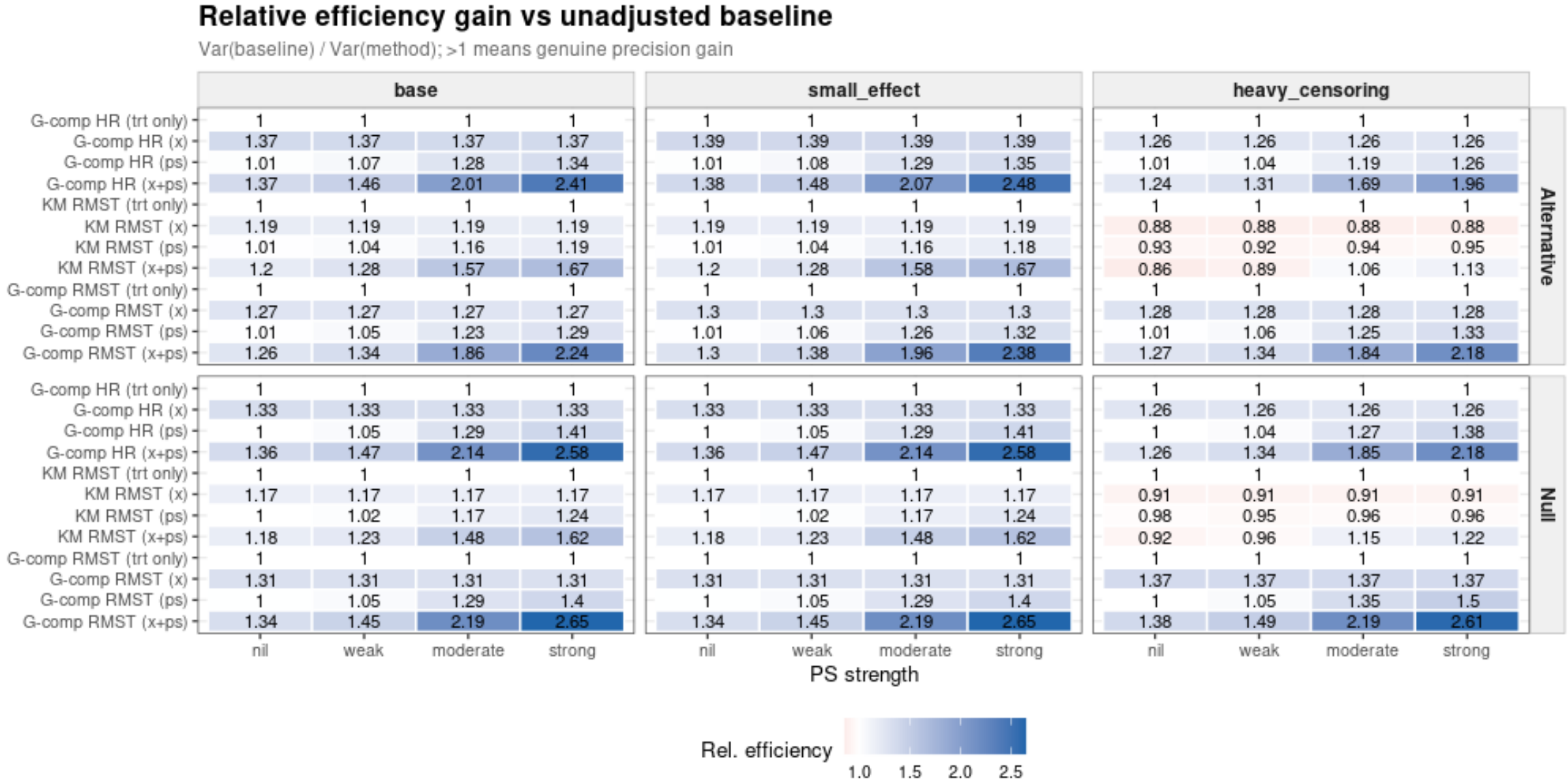
Bias

Bias (each method vs own estimand)

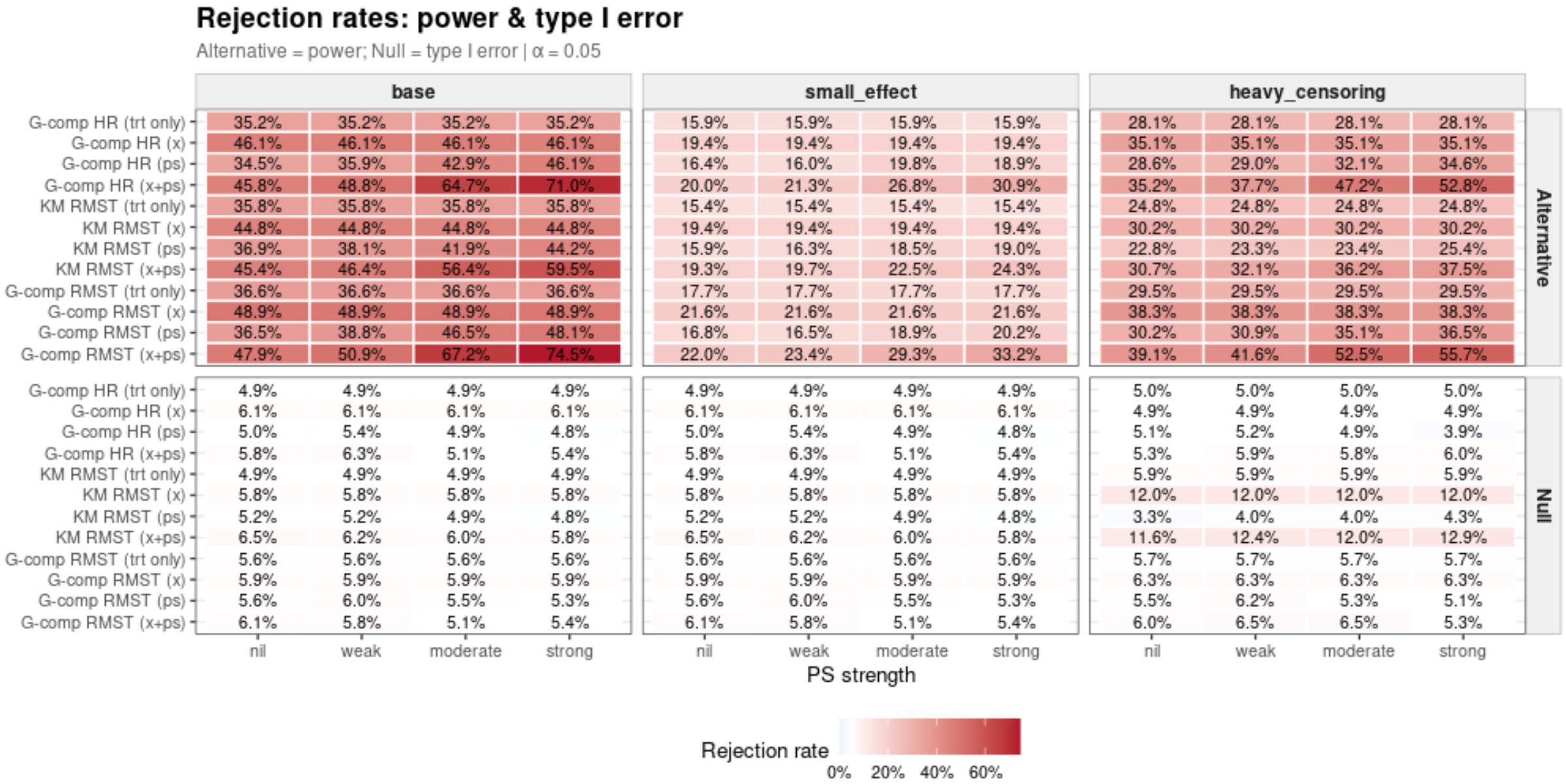
Red = underestimates effect; blue = overestimates | Separate scales for HR and RMST



Relative efficiency gain vs baseline



Rejection rates (power & type 1 error)



Headline results summary

1. G-comp HR gives good precision gains with a stable estimand

- Relative efficiency 2.0× (base) to 1.7× (heavy censoring) with moderate PS
- Power increase from 35% to 65% (base), 16% to 27% (small effect), 28% to 47% (heavy censoring) with mod. PS
- Bias ≤ 0.003 (log-HR) across all settings; type I error 5-6%, generally controlled (coverage ~95%, not seen)
- All gains come from variance reduction

Headline results summary

1. G-comp HR gives good precision gains with a stable estimand

- Relative efficiency 2.0× (base) to 1.7× (heavy censoring) with moderate PS
- Power increase from 35% to 65% (base), 16% to 27% (small effect), 28% to 47% (heavy censoring) with mod. PS
- Bias ≤ 0.003 (log-HR) across all settings; type I error 5-6%, generally controlled (coverage ~95%, not seen)
- All gains come from variance reduction

2. G-comp RMST is a good complementary estimand

- Relative efficiency 1.9× (base) to 1.8× (heavy censoring) with moderate PS
- Power from 37% to 67% (base); 18 to 29% (small effect), 30% to 53% (heavy censoring) with moderate PS
- Bias ≤ 0.024 with x+PS (~1% of true ΔRMST); coverage 94–95%; type I error 5.1–6.5%
- Need to include prog. covs. in outcome model: g-comp RMST trt only show attenuated bias (0.12) due to unmodelled heterogeneity

Headline results summary

1. G-comp HR gives good precision gains with a stable estimand

- Relative efficiency 2.0× (base) to 1.7× (heavy censoring) with moderate PS
- Power increase from 35% to 65% (base), 16% to 27% (small effect), 28% to 47% (heavy censoring) with mod. PS
- Bias ≤ 0.003 (log-HR) across all settings; type I error 5-6%, generally controlled (coverage ~95%, not seen)
- All gains come from variance reduction

2. G-comp RMST is a good complementary estimand

- Relative efficiency 1.9× (base) to 1.8× (heavy censoring) with moderate PS
- Power from 37% to 67% (base); 18 to 29% (small effect), 30% to 53% (heavy censoring) with moderate PS
- Bias ≤ 0.024 with x+PS (~1% of true ΔRMST); coverage 94–95%; type I error 5.1–6.5%
- Need to include prog. covs. in outcome model: g-comp RMST trt only show attenuated bias (0.12) due to unmodelled heterogeneity

3. KM-adjusted RMST caution needed under heavy censoring

- Type I error inflated to 12–13% for x and x+PS under heavy censoring, due to unstable IPCW weights and underestimating SE.
- G-comp doesn't have this issue as it handles censoring through Cox likelihood, not reweighting

GSK

Headline results summary

1. G-comp HR gives good precision gains with a stable estimand

- Relative efficiency 2.0× (base) to 1.7× (heavy censoring) with moderate PS
- Power increase from 35% to 65% (base), 16% to 27% (small effect), 28% to 47% (heavy censoring) with mod. PS
- Bias ≤ 0.003 (log-HR) across all settings; type I error 5-6%, generally controlled (coverage ~95%, not seen)
- All gains come from variance reduction

2. G-comp RMST is a good complementary estimand

- Relative efficiency 1.9× (base) to 1.8× (heavy censoring) with moderate PS
- Power from 37% to 67% (base); 18 to 29% (small effect), 30% to 53% (heavy censoring) with moderate PS
- Bias ≤ 0.024 with x+PS (~1% of true ΔRMST); coverage 94–95%; type I error 5.1–6.5%
- Need to include prog. covs. in outcome model: g-comp RMST trt only show attenuated bias (0.12) due to unmodelled heterogeneity

3. KM-adjusted RMST caution needed under heavy censoring

- Type I error inflated to 12–13% for x and x+PS under heavy censoring, due to unstable IPCW weights and underestimating SE.
- G-comp doesn't have this issue as it handles censoring through Cox likelihood, not reweighting

Future work: assess PS miscalibration, non-PH (for conditional), subgroup/predictive biomarker settings, and faster IF / delta-method SEs

GSK

References

- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2), 481–488.
- Schuler, A., et al. (2022). Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *Int J Biostat*, 18(2), 329–356.
- Daniel, R., et al. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors. *Biometrical Journal*, 63(3), 528–557.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period. *Mathematical Modelling*, 7(9–12), 1393–1512.
- Royston, P. & Parmar, M. K. B. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials. *BMC Med Res Methodol*, 13, 152.
- Uno, H., et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*, 32(22), 2380–2385.
- Ge, M., et al. (2011). Covariate-adjusted difference in event rates in a randomized trial. Pharmaceutical Statistics, 10, 1–22.

GSK

gsk.com