Empirical benchmarking of two recent approaches for augmenting clinical trials via external data

PSI 2024, Amsterdam

Erik Hermansson [1] Fredrik Öhrn [3] Stefan Franzén [2] David Svensson [1]

[1]Respiratory & Immunology & Statistical Innovation, AstraZeneca, Gothenburg, Sweden

[2]Medical & Payer Evidence Statistics, AstraZeneca, Gothenburg, Sweden

[3]Statistics and Decision Sciences, Global Development, Johnson and Johnson Innovative Medicines, Gothenburg, Sweden

June 4, 2024

Outline

Introduction

- **Bayesian and Frequentist methods**
  - Two potential ways to do this is Bayesian Dynamic Borrowing (BDB) [1, 6, 8] and Prognostic Score Methodology (PSM) [2, 9], but many more available [8]
- **Increased acceptance by regulatory agencies**
  - BDB have been approved by the FDA in some settings where recruitment is difficult and PSM recently went through the process of an EMA qualified opinion [5]
- **Setting: Phase 2B with mostly sponsors risk**
  - One historical trial to borrow from
  - We borrow only the control arm and have a continuous endpoint

Robust Mixture Prior

- Many different ways to do Bayesian dynamic borrowing, such as power prior and hierarchical prior [8]. We will focus on Robust Mixture prior (RMP)
  - RMP borrows dynamically via a likelihood weighting. If there is a large prior data conflict the informative prior wont influence the posterior as much [1, 6]
  - The borrowing can be expressed as Effective Sample Size (ESS) [4] and can be seen as *expected* additional enrolled control patients
  - However, BDB can *potentially* inflate the Type 1 Error and have a lower power.
- **How to select W?**
  - We choose the weight such that the prior contribute by no more than 50% of the future control arm measured with ESS

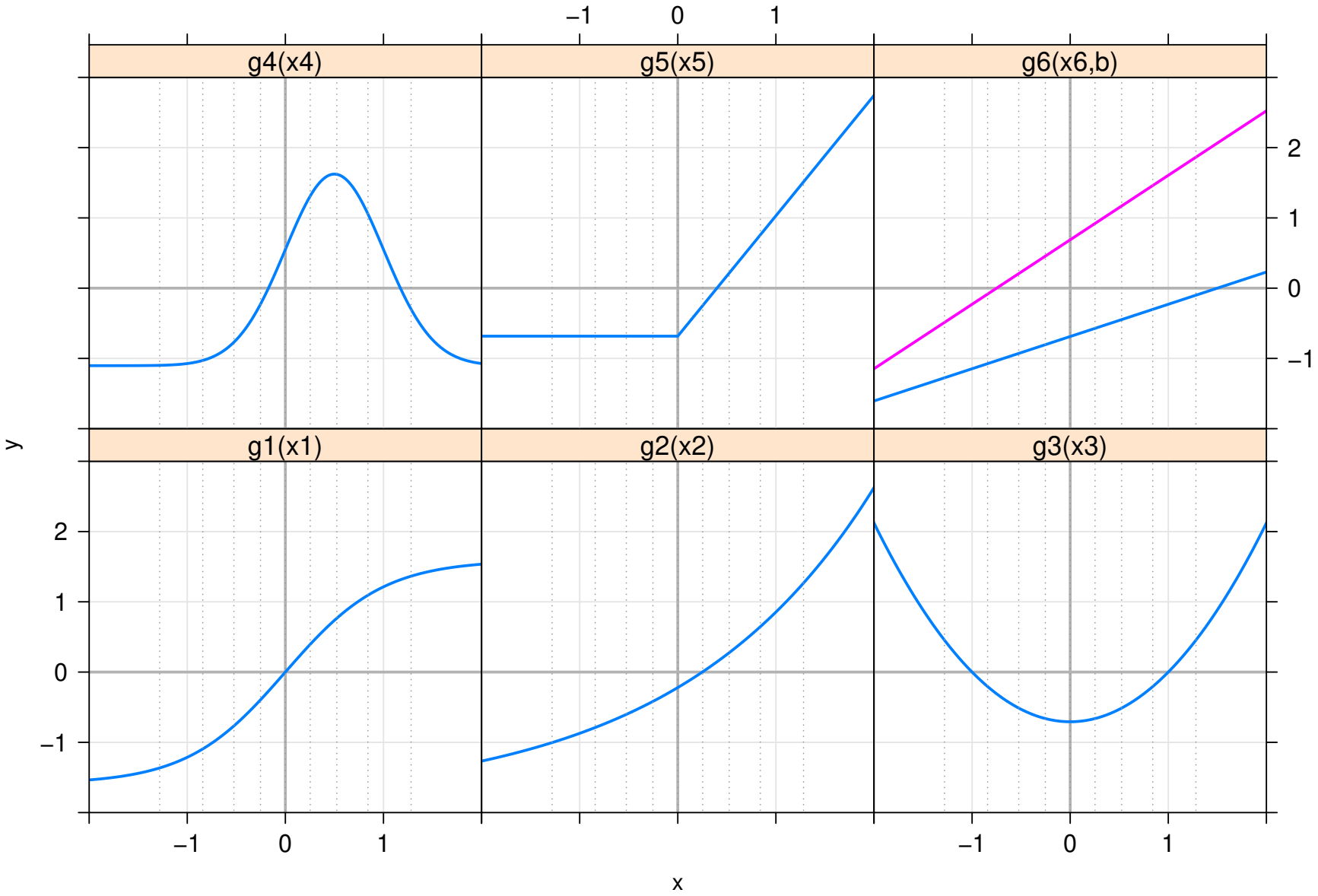
Prognostic Score Methodology

- With relevant historical data one can build a prediction model of the future control response [2, 9] which reduces the residual variance in proportion to the R^2 of the model (and protects the Type 1 Error!)
  - This can then be written as an ANCOVA such as $E[Y_k] = \alpha + \delta T_k + \beta^t x_k + \gamma \hat{z}^{(RCT)}$
  - *NOTE: For PSM to add value it need to improve the R^2 above what an ANCOVA would have*
- **Many different models can be used**
  - To improve upon an ANCOVA it probably needs to be able to detect potential non-linearity in the data
  - We have opted for XGBoost as it performs well on tabular data [3]

Dynamic Twins - A natural conclusion?

- We can combine the two methods by borrowing the control effect after adjusting with PSM. A similar approach have been suggested by Vanderbeek et al [7]
  - The adjusted control response is the intercept from a linear regression where centered covariates have been included i.e. $Y = \alpha + \beta(\boldsymbol{X} - \bar{\boldsymbol{X}})$ and then $\mathbf{E}[Y] = \alpha$
- This potentially increases the power by reducing the residual variance and by increasing the effective sample size [4]
  - However, we can now have a misspecification of the prognostic model and the future control response!

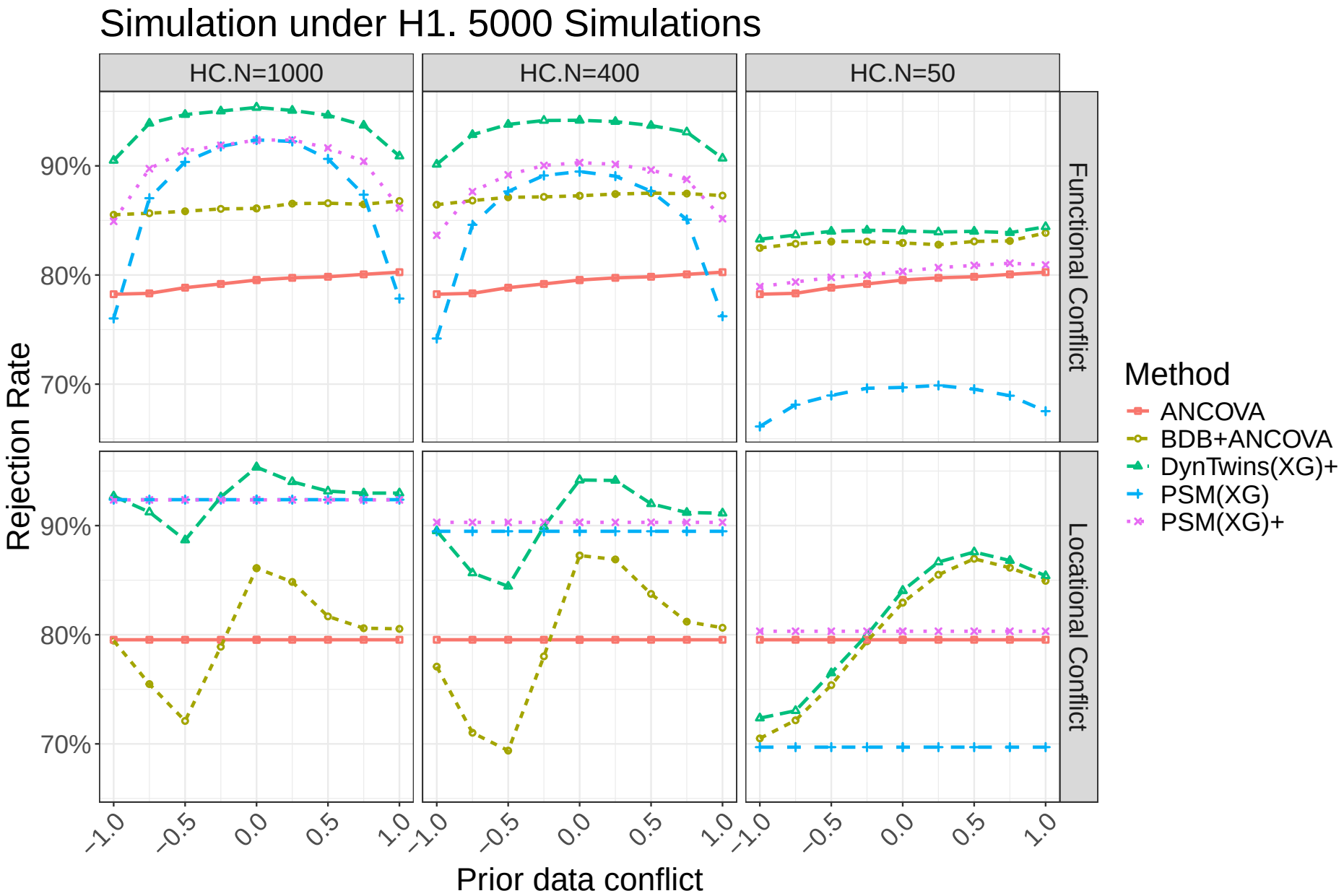
Simulation Model



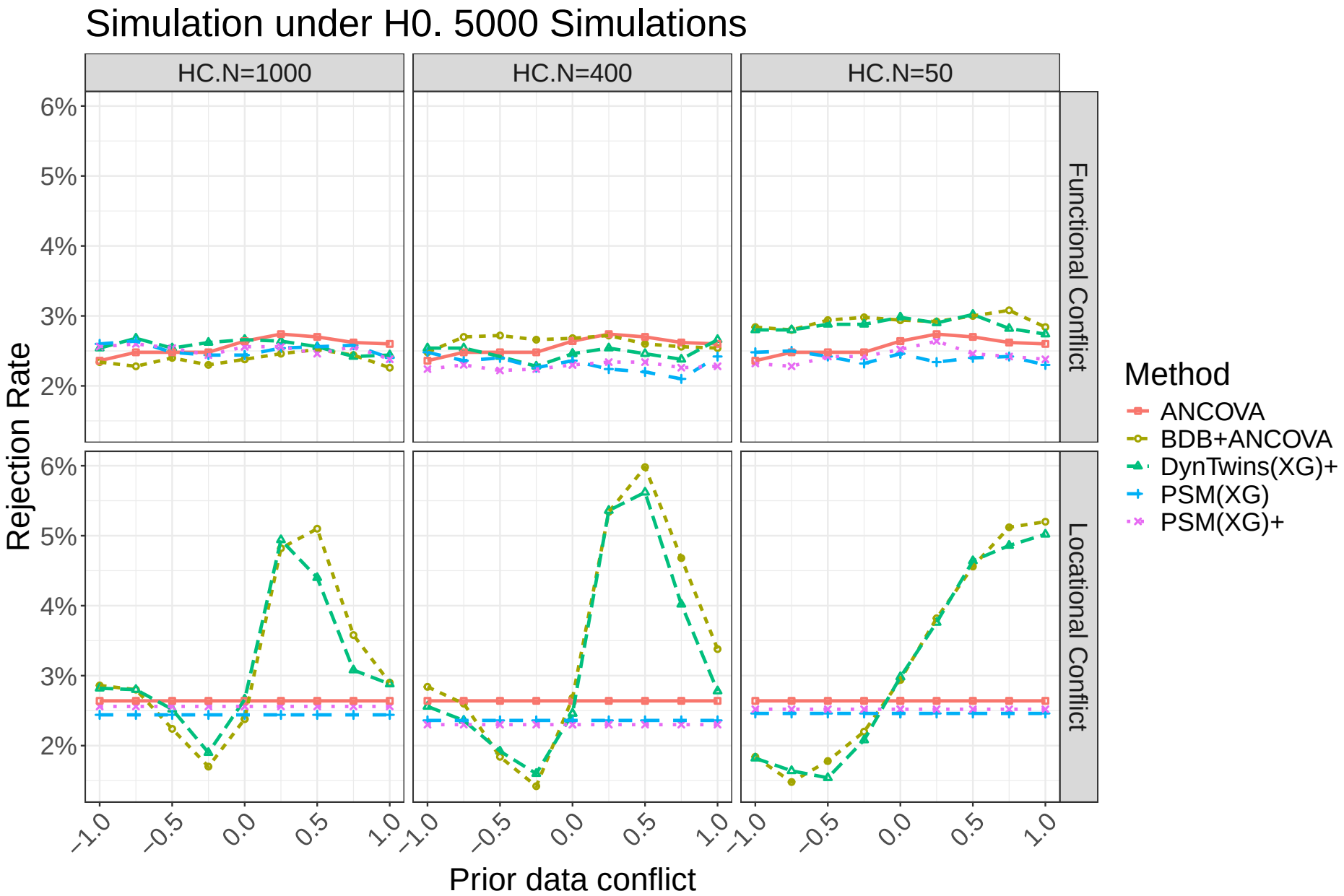
Simulation set up

- The historical control arm is seen as fixed and the future RCT is simulated from $y = \delta \cdot T + f_{prog}(\mathbf{x}, b; \mathtt{W}, \mathtt{w}, \mu) + \varepsilon$
  - Three different historical sample sizes, 50, 400 and 1000
  - The future study has 400 patients and have a 1:1 randomization
- **Two different types of prior data conflict**
  - By varying μ between -1 and 1, we can create *prior data conflict*. We call it `Locational Conflict`.
  - W^2 is altered between 0 and 2. This is to alter the *functional relationship of the future data* and is called `Functional Conflict`.

Alternative Hypothesis



Null Hypothesis



Discussion & Conclusion

- Both methods are vulnerable to data conflicts, although in different ways.
  - BDB can have higher type 1 error and lower power
  - While the downside in PSM is limited as the cost is one degree of freedom
  - BDB can be done with only summary level data while PSM requires individual level patient data
- The methods work in different ways to arrive at the same goal - higher power/reduced sample size
  - NB: BDB borrows in absolute numbers while PSM borrows in proportion to the future trial
- When little data is available, PSM does not add much value while BDB can still give meaningful gains.
- One can combine the two methods, but is sensitivity to Locational **and** Functional Conflict

References I

[1] Nicky Best et al. "Assessing efficacy in important subgroup in confirmatory trials: An example using Bayesian dynamic borrowing". In: (2021).

[2] Carl-Fredrik Burman et al. "Digital twins and Bayesian dynamic borrowing: Two recent approaches for incorporating historical control data". In: *Pharmaceutical Statistics* n/a (2024), pp. 1–19.

[3] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. *Why do tree-based models still outperform deep learning on tabular data?* 2022.

[4] Beat Neuenschwander et al. "Predictively consistent prior effective sample sizes". In: *Biometrics* 76.2 (2020), pp. 578–587.

[5] "Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)". In: (2022).

References II

[6] Heinz Schmidli et al. "Robust meta-analytic-predictive priors in clinical trials with historical control information". In: *Biometrics* 70.4 (2014).

[7] Alyssa M. Vanderbeek et al. *Bayesian Prognostic Covariate Adjustment With Additive Mixture Priors*. 2024.

[8] Kert Viele et al. "Use of historical control data for assessing treatment effects in clinical trials". In: *Pharm Stat* 13 (2014), pp. 41–54.

[9] David Walsh et al. "Using digital twins to reduce sample sizes while maintaining power and statistical accuracy". In: *Alzheimer's & Dementia* 17.S9 (2021), e054657.

Simulation Model

The historical and future control arms are simulated from $y = f_{prog}(\mathbf{x}, b; \mathbf{w}, \mu) + \varepsilon$ where

$$\begin{aligned} f_{prog}(\mathbf{x}, b; \mathbf{w}, \mu) = \mu + w_1 \cdot g_1(x_1) + w_2 \cdot g_2(x_2) + \\ w_3 \cdot g_3(x_3) + w_4 \cdot g_4(x_4) + w_5 \cdot g_5(x_5) + w_6 \cdot g_6(x_6, b) \end{aligned} \tag{1}$$

with $\varepsilon \sim N(0, \sigma)$, where $\sigma^2 = 4$ and $\sum_{j=1}^{6} w_j^2 = 6$. The historical data is kept fixed and where the configuration $\mathtt{W} = 1, \mathtt{w} = 1$ and $\mu = 0$ represents exchangability of the historical and future data. The future RCT is simulated from the following model:

$$y = \delta \cdot T + f_{prog}(\mathbf{x}, b; \mathtt{W}, \mathtt{w}, \mu_1) + \varepsilon, \tag{2}$$

Learner Comparison

HC.N	Method	Prior Mean R2	Post Mean R2	Drift in R2
HC.N=1000	PSM(LR)	0.33	0.32	0.01
HC.N=1000	PSM(RF)	0.45	0.43	0.02
HC.N=1000	PSM(XG)	0.55	0.54	0.01
HC.N=400	PSM(LR)	0.33	0.31	0.02
HC.N=400	PSM(RF)	0.39	0.38	0.01
HC.N=400	PSM(XG)	0.50	0.49	0.01
HC.N=50	PSM(LR)	0.40	0.25	0.15
HC.N=50	PSM(RF)	0.15	0.18	-0.03
HC.N=50	PSM(XG)	0.26	0.13	0.12

Table: 5000 simulations of the mean R^2 based on in sample predictions and out of sample predictions for different learners and different amount of historical data.

Functional Conflict