Benchmarking Bayesian subgroup shrinkage methods on clinical data

Björn Bornkamp

PSI conference 2025, London
June 11, 2025

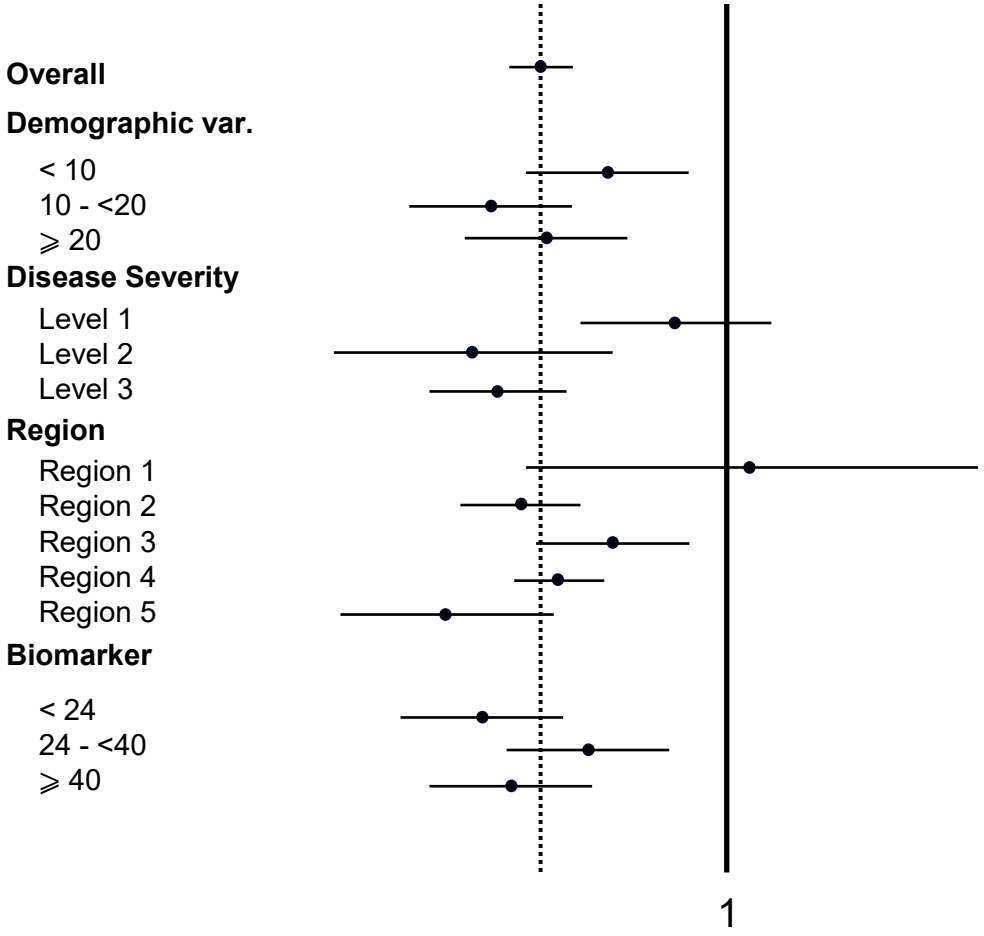
joint work with Sebastian Weber and David Ohlssen

NOVARTIS | Reimagining Medicine

Agenda

- Introduction
- Simple subgroup shrinkage models
- Regression-based shrinkage models
- Benchmarking of methods on twin studies
  - Continuous, time-to-event and binary outcome

Subgroup Analysis & Forest plots



- Estimation of subgroup treatment effects challenging
  - Limited sample size & multiplicity

Statistical Modeling, Causal Inference, and Social Science

Home Authors Blogs We Read Books Sponsors

You need 16 times the sample size to estimate an interaction than to estimate a main effect

Posted on March 15, 2018 9:11 AM by Andrew

- Idea of shrinkage methods
  - Shrink subgroup treatment effects towards overall treatment effect
    - For given subgroup: Smaller MSE (more reliable), accepting some bias (bias-variance trade-off)
    - For subgroups with extreme observed effects: Less biased inference

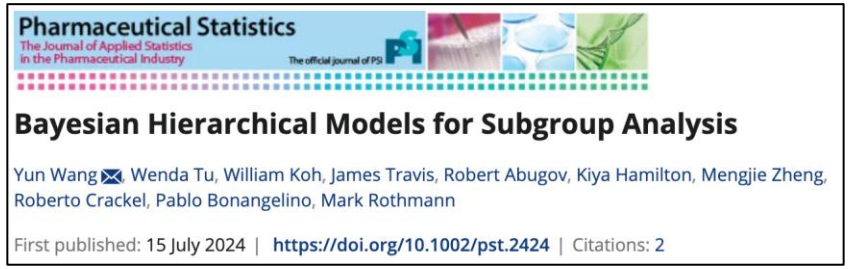
Simple subgroup shrinkage models

Overall and fully stratified subgroup models

- Overall model
  - Linear predictor for patient i: $\eta_{i,j,k} = \beta_1 + \beta_2 z_i + \boldsymbol{\beta}_3' \boldsymbol{x}_i$
  - z_i: Treatment indicator, $\boldsymbol{x}_i$: Additional covariates

- Fully stratified subgroup models
  - $\eta_{i,j,k} = \beta_{1,j,k} + \beta_{2,j,k} z_i + \beta_{3,j,k} \boldsymbol{x}_i$
    - Index j: Subgroup variable (e.g. gender)
    - Index k: Subgroup within subgrouping variable (e.g. female)
  - Weakly informative priors on all model parameters
  - Fitted separately for each subgroup variable with index j and each subgroup with index k

Simple shrinkage model (review: Wang et al 2024)



- $\eta_{i,j,k} = \beta_{1,j,k} + \beta_{2,j,k} z_i + \boldsymbol{\beta}_3' \boldsymbol{x}_i$
- Hierarchical priors for the treatment effect $\beta_{2,j,k} \sim N(\beta_{2,j}, \sigma_{2,j}^2)$
  - Subgroups within subgroup variable treated as exchangeable
  - adequate if no prior/external evidence that one of subgroups has differential treatment effect
- Prior for between-subgroup variance $\sigma_{2,j}^2$
  - Number of subgroups within a subgrouping variable small (2 - 5)
    → challenging to estimate the variance from data
  - Select half-normal $HN(\tau)$ prior for $\sigma_{2,j}$ based on prior distribution of $|\beta_{2,j,k} - \beta_{2,j,k'}|$
    - Quantiles of $\left|\beta_{2,j,k} - \beta_{2,j,k'}\right|$ as fractions of planned treatment effect δ_{plan}

τ	5%	25%	50%	75%	95%	
$0.5\delta_{plan}$	0.01	0.09	0.26	0.61	1.54	**High shrinkage**
δ_{plan}	0.02	0.17	0.52	1.22	3.09	**Low shrinkage**

Regression-based shrinkage models

Global regression model

- Multiple subgroup variables → Multiple partially overlapping subgroups
  - Simple shrinkage model requires non-overlapping subgroups
  - Fit multiple simple subgroup shrinkage models (one per subgroup variable)

- Alternative: Global regression model based on subgroup indicators
  - (Dixon & Simon 1991, Jones et al. 2011, Wolbers et al. 2025)
  - Only one model fit required; all subgroup estimates derived from the same model
  - Provide „adjusted“ parameter estimates (→ helps identify drivers of heterogeneity)

Global regression model

- $\eta_{i,j,k} = \beta_1 + \beta_2 z_i + \sum_l^L (b_l + g_l z_i) s_{i,l} + \boldsymbol{\beta}_3' \boldsymbol{x}_i$
  - $s_{i,l}$: is a binary subgroup indicator for subgroup $l = 1, \ldots, L$.
  - L is the overall number of subgroups evaluated for all subgroup variables.
  - b_l, g_l: Prognostic and predictive effect of subgroup indicator $s_{i,l}$
  - For β_1, β_2 and $\boldsymbol{\beta}_3$use weakly informative priors
  - For b_l, g_l use shrinkage prior distributions

- Note:
  - Shrinkage necessary: Standard regression would over-fit and no dummy coding for subgroup coefficients (parameters not identified in the frequentist sense)
  - Higher-order interactions across subgroups here not included

Shrinkage prior 1: Horseshoe
Piironen & Vehtari (2017)

- Idea of original horseshoe prior (Carvalho et al 2009)
  - Shrink small signals aggressively towards 0; don't shrink large signals
  - Normal$\left(0, \tau^2\lambda_i^2\right)$ prior with $\tau \sim Cauchy^+(0,1)$ and $\lambda_i \sim Cauchy^+(0,1)$
- Idea of regularized horseshoe (Piironen & Vehtari, 2017)
  - Large signals may be completely unpenalized for horseshoe
  - Use Normal$\left(0, \tau^2\tilde{\lambda}_i^2\right)$ with $\tilde{\lambda}_i^2 = \frac{c^2\lambda_i^2}{c^2+\tau^2c^2}$ with $c \sim Inv-Gamma(\frac{\nu}{2}, s^2/2)$ instead of λ_i^2
  - Prior for global scale can be derived based on expected proportion of non-zero vs zero coefficients
    - Later use 0.5 (**high shrinkage**) and 1 (**low shrinkage**)

Shrinkage prior 2: R2D2
Zhang et al (2022)

- Basic idea: Prior on $R^2 = \frac{explained\ variance}{explained\ variance + residual\ variance}$
- Prior for each coefficient: $\text{Normal}\left(0, \tilde{\lambda}_i^2\right)$
  - Global shrinkage (overall prior variance) determined by beta prior distribution on R^2
  - Local shrinkage (how to split prior variability across coefficients) determined by a Dirichlet prior
- Concentration parameter of Dirichlet: whether prior variability is evenly spread across all coefficients or concentrated on only a few
- In benchmarking later use a uniform distribution for R^2
  Use concentration parameter equal to 0.2 (**high shrinkage**) and 0.5 (**low shrinkage**)
- Notes
  - For non-normal data use „pseudo-variance“ (→ variance of intercept-only model on link scale)
  - Note: Unpenalized covariates formally don‘t enter R^2

Benchmarking on twin studies

Idea of benchmarking

Fit: Every model using data from trial 1 (2)

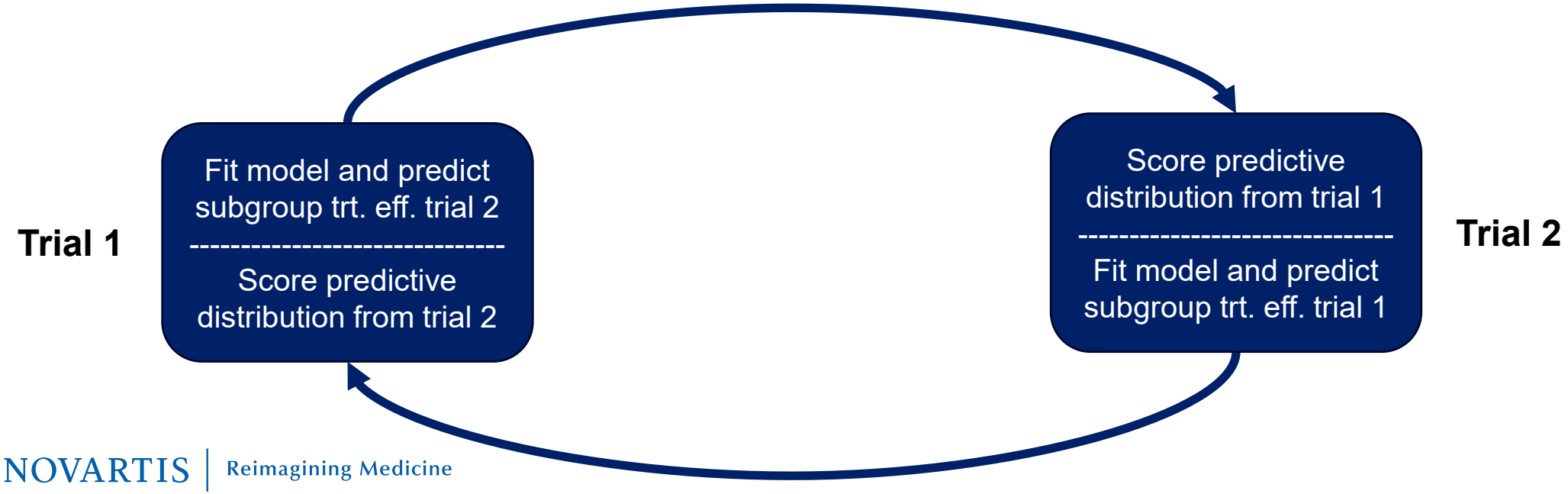
Out of sample prediction: For each subgroup, use the model to form predictive distribution of treatment effect in trial 2 (1)

Scoring: Predictive treatment effect distribution for each subgroup (and both directions) compared to observed treatment effect using scoring rule, rewarding low bias and uncertainty.

Ranking by case: Scores are calculated for each method and subgroup. Higher scores are better. Methods ranked according to average score (average across all subgroups and both directions of predicting).

Benchmarking data

- Continuous and time-to-event data
  - Utilize secondary endpoints from twin concurrent Phase 3 trials
  - Use each trial once for fitting and once for preciction (2 cycles of fit and predict)
- Binary data
  - Utilize primary endpoint from 4 similarly designed (& partially concurrent) Phase 3 trials
  - Here use 1 trial for model fitting and predict 3 trials (4 cycles of fit and predict)

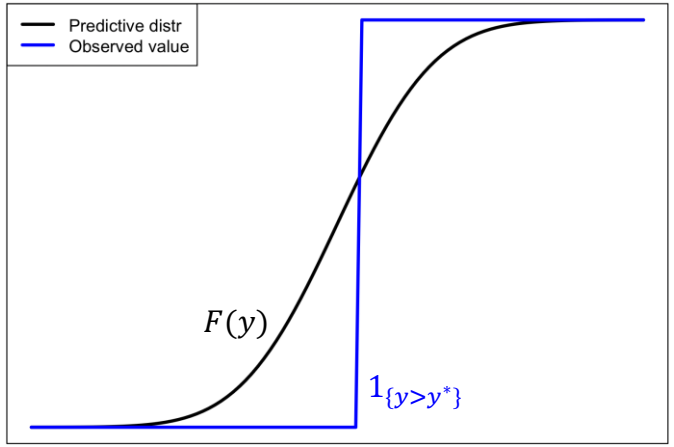


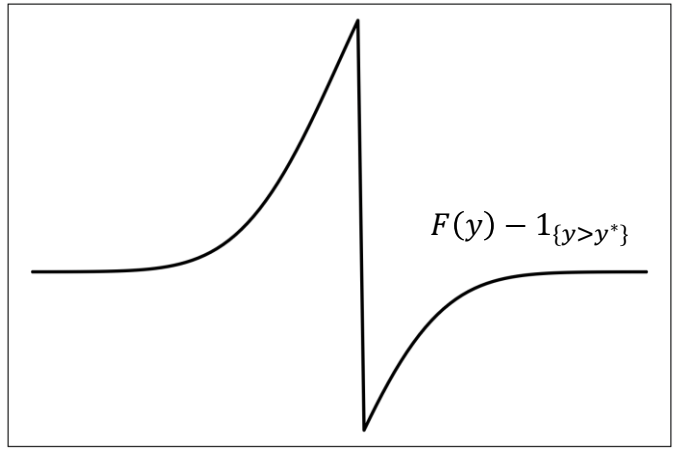
Case study methods – Continuous Ranked Probability Score and

- Scoring rules assess a predictive distribution vs an observed value
  - Here: Predictive distribution for subgroup from one trial → observed subgroup treatment effect in other trial
- The continuous ranked probability score (CRPS, Gneiting et al 2007) is given by

$$CRPS(F, y^*) = -\int_{-\infty}^{\infty} \left(F(y) - 1_{\{y>y^*\}}\right)^2 dy$$

  - $F(y)$ cdf of the predictive distribution for subgroup treatment effect in one trial
  - y^* subgroup treatment effect observed in other trial
- CRPS is not scale invariant
  - Problematic when averaging score across predictions with different predictive variance
  - **Scaled CRPS** (Bolin & Wallin, 2023) solves this issue





Results (preliminary, averaged across 10 replicates)

Model (shrinkage)	Average Rank (across 3 cases)	Average SCRP (SE)		
		Case 1	Case 2	Case 3
Simple shrinkage (high)*	3.00	-2.77 (0.01)	-4.64 (0.02)	-4.56 (0.02)
Simple shrinkage (low)*	3.00	-2.79 (0.01)	-4.65 (0.02)	-4.51 (0.03)
Horseshoe (high)*	3.33	-3.08 (0.01)	-4.40 (0.04)	-4.54 (0.03)
R2D2 (low)	3.33	-2.94 (0.01)	-4.64 (0.03)	-4.52 (0.04)
R2D2 (high)*	4.00	-2.98 (0.01)	-4.55 (0.04)	-4.61 (0.03)
Horseshoe (low)*	5.00	-3.09 (0.02)	-4.48 (0.03)	-4.62 (0.03)
Fully stratified	6.67	-2.99 (0.01)	-5.90 (0.04)	-5.00 (0.04)
Overall	7.67	-3.65 (0.03)	-4.82 (0.03)	-5.89 (0.04)

* Can lead to divergences in stan during model fitting

Discussion

- Prime-time for Bayesian shrinkage estimation (see also Wang et al 2024)
- Many options on how to perform subgroup shrinkage
- Benchmarking
  - Simulation Study: Challenging to be “truly” neutral
  - Alternative: Use concurrent, similarly designed studies to assess predictive ability
  - Limitations
    - There are always differences (known or unknown)
    - Small number of data-sets available
- Results
  - Outperformance of shrinkage versus standard methods
  - Shrinkage methods close together


BIOPHARMACEUTICAL REPORT VOLUME 31, NO. 4

2024 ASA BIOPHARMACEUTICAL SECTION REGULATORY-INDUSTRY STATISTICS WORKSHOP SESSION ON “BAYESIAN SHRINKAGE ESTIMATION FOR SUBGROUPS: IS IT READY FOR PRIME TIME?”

Talk 1: Mark Rothmann (FDA/CDER/OTS/OB): “Practical experiences with Bayesian subgroup shrinkage methods for drug trials snapshots”

Bayesian shrinkage estimation for subgroup analysis is ready for primetime. In 2019, the FDA posted

References

- Bolin, D., & Wallin, J. (2023). Local scale invariance and robustness of proper scoring rules. Statistical Science, 38, 140-159
- Carvalho, C. M. et al (2009). Handling sparsity via the horseshoe. Proceedings of Machine Learning Research, 5, 73-80
- Dixon, D. O., & Simon, R. (1991). Bayesian subset analysis. Biometrics, 871-881.
- Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., & Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. Clinical Trials, 8, 129-143.
- Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics 11, 5018-5051
- Wang, Y. et al. (2024). Bayesian hierarchical models for subgroup analysis. Pharmaceutical Statistics, 23, 1065-1083.
- Wolbers, M., et al (2025). Using shrinkage methods to estimate treatment effects in overlapping subgroups in randomized clinical trials with a time-to-event endpoint. Statistical Methods in Medical Research
- Zhang, Y. D. et al (2022). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. Journal of the American Statistical Association, 117, 862-874.

Björn Bornkamp
bjoern.bornkamp@novartis.com

Thank you

NOVARTIS | Reimagining Medicine