Bayesian shrinkage methods for routine estimation of subgroup treatment effects

Workshop on the use of Bayesian statistics in clinical development; EMA Amsterdam, 17th June 2025

Björn Bornkamp (Novartis), Nicky Best, Daniel Bratton (GSK), Monika Jelizarow (UCB), Natalia Muhlemann (Cytel)

Commentary by Professor John McMurray (University of Glasgow)

EFSPI
EUROPEAN FEDERATION OF STATISTICIANS IN THE PHARMACEUTICAL INDUSTRY
Representing Statistical Associations in Europe

Subgroup treatment effect estimation

- Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement
  - Kent et al (2020)
  - *"… In medical care, treatment decisions made by clinicians and patients are generally based - implicitly or explicitly - on predictions of comparative outcome risks under alternative treatment conditions. …"*
  - *"… interest is growing in understanding how a treatment's effect can vary across patients …"*

- But: Subgroup treatment effect estimates often failed to get replicated/generalize
  - For example Yusuf et al (1991), Wallach et al (2017) document this on examples

- Why?
  - Insufficient sample size → trial(s) sample sized for testing overall treatment effect
    → Noise dominates subgroup treatment effect estimates (in particular for small subgroups)
  - Multiplicity → interest is usually in a medium number of subgroups/covariates; often focus on best/worst estimated subgroup treatment effects

How to resolve dilemma of (i) interest and (ii) inherent data limitations?

- EMA subgroup guideline (2019):
  - Consider external data to assess credibility of subgroup finding(s) (biological plausibility and replication)
  - Semi-quantitative/qualitative process and case-specific
- Bayesian shrinkage subgroup treatment effects estimation
  - Motivation: Provide ***more reliable information to patients and health care providers***
  - How? Stabilize subgroup estimates by borrowing information from the complete population to estimate treatment effect in subgroup of interest
    - For subgroup variables with no strong prior/external evidence for increased or decreased treatment effects
  - Tilt bias-variance trade-off towards a lower mean-squared error

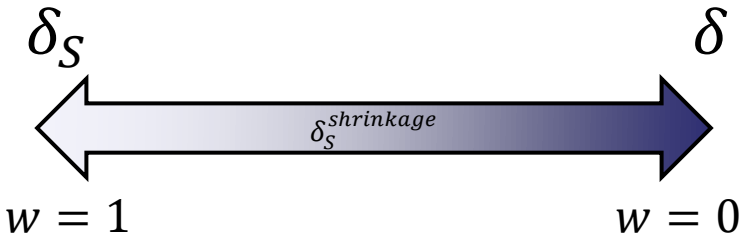
EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

31 January 2019
EMA/CHMP/539146/2013
Committee for Medicinal Products for Human Use (CHMP)

Guideline on the investigation of subgroups in confirmatory clinical trials

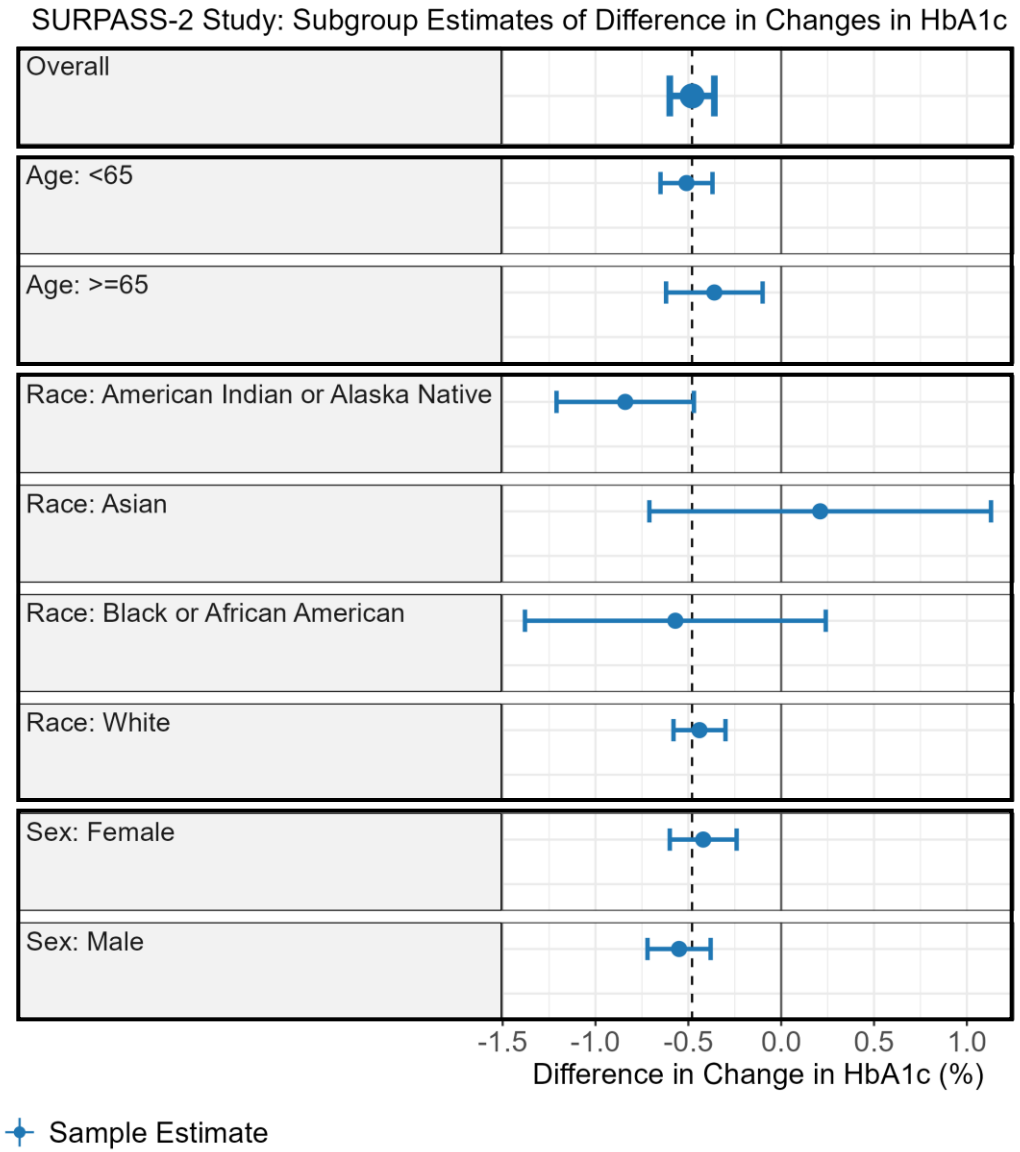
Bayesian hierarchical model

$$\delta_S^{shrinkage} = w\delta_S + (1-w)\delta$$



Example forest plot using Bayesian shrinkage

- Example from SURPASS-2 study
  - Tirzepatide vs Semaglutide in Type 2 Diabetes
- **Conventional** subgroup-specific **sample** treatment effect estimates, δ_j



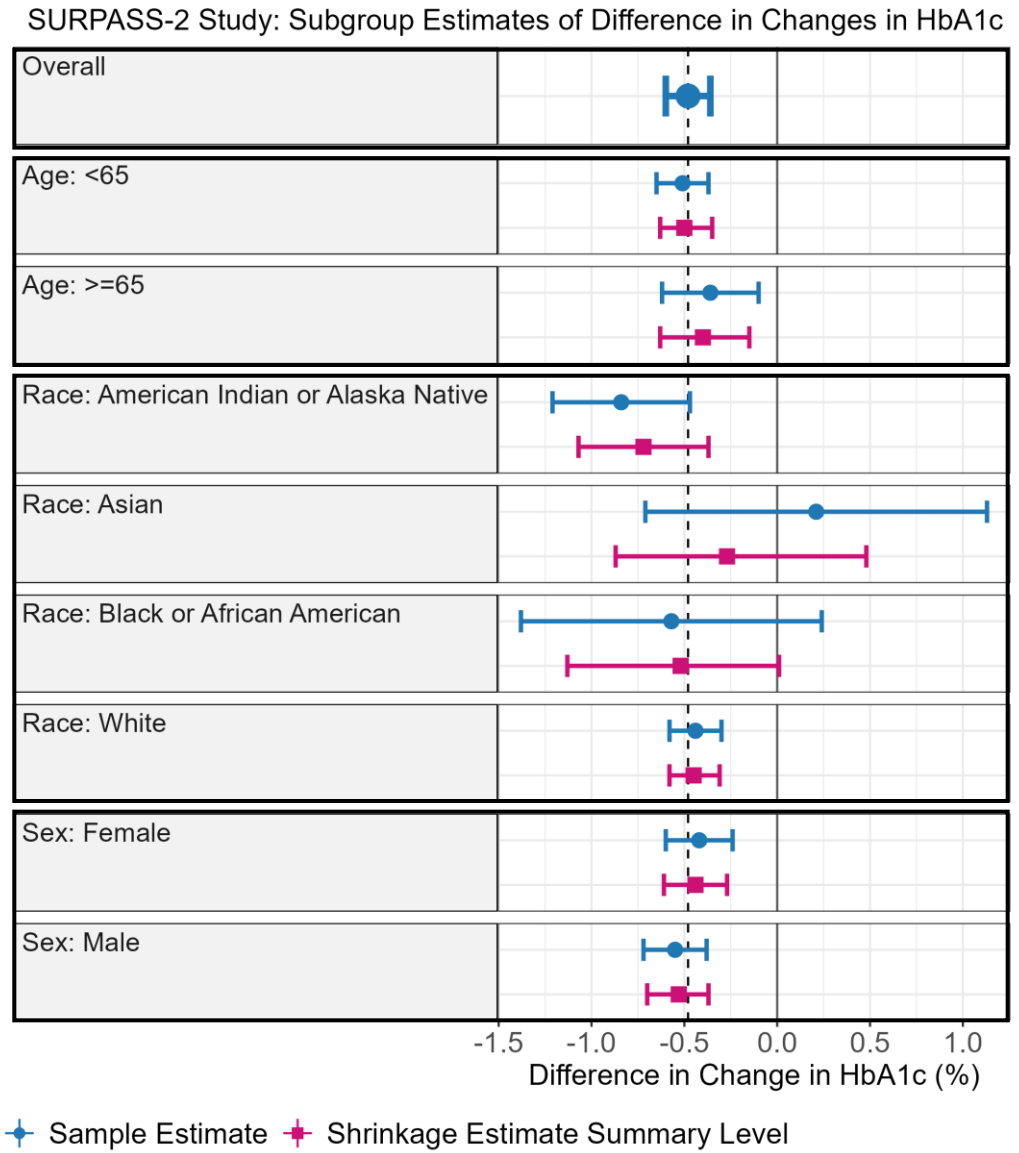
Example forest plot using Bayesian shrinkage

- Example from SURPASS-2 study
  - Tirzepatide vs Semaglutide in Type 2 Diabetes
- **Conventional** subgroup-specific **sample** treatment effect estimates, δ_j

Bayesian Hierarchical Modelling

$$\delta_j \mid \mu_j, \sigma_j \sim N(\mu_j, \sigma_j^2)$$
$$\mu_j \mid \mu, \tau \sim N(\mu, \tau^2)$$
$$\mu \sim p(\mu), \tau \sim p(\tau)$$

Subgroup-specific **shrinkage** treatment effect estimates

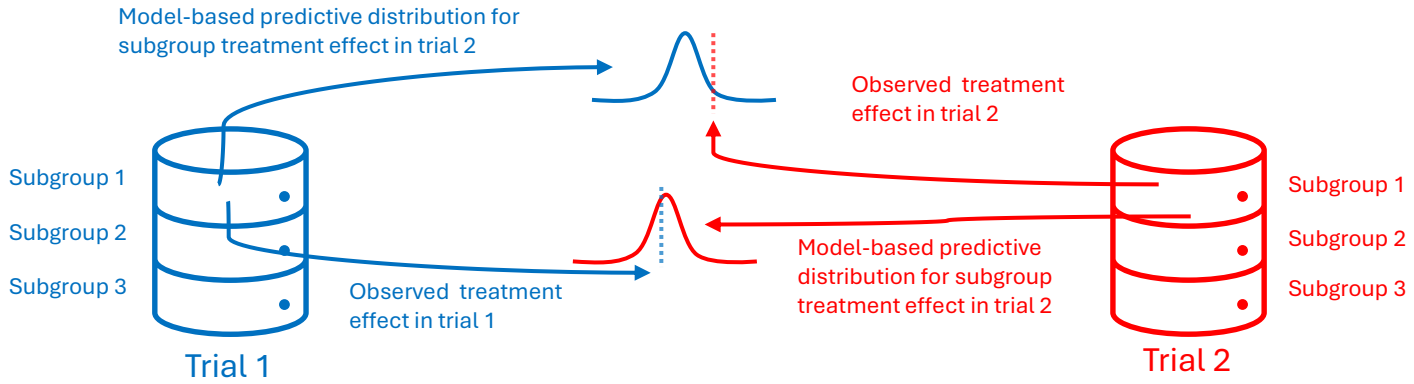


Non-Stats feedback and perspectives on pre-defined case-example

- Clear need to reduce the risk of spurious findings in subgroup analysis, but…
- “Shrinkage” is a loaded term – may be seen as purposefully trying to show consistency
- Two estimates for the same effect can be problematic – which to use in which situation?
- Sometimes confusion over exchangeability assumption – not the same as assuming all effects are equal
- Results of subgroup analyses based on a small dataset (e.g. phase 2 trial) may lead to doubt over exchangeability, even if CIs are very wide and overlapping (heavy focus on point estimates!)

How do we know it works?

- Gold standard: Simulation studies varying the degree of heterogeneity
  - Wolbers et al (2025) demonstrate that even under heterogeneous treatment effects across subgroups, shrinkage methods (based on global regression models) can outperform conventional subgroup estimates in terms of mean squared error
- Complementary approach: Out of sample predictions
  - Assess predictive performance based on similar (e.g. twin) Phase 3 trials



Benchmarking on study data (joint work with Sebastian Weber and David Ohlssen)

- Data
  - Continuous & time-to-event endpoint: Twin Phase 3 trials on same compound & control
  - Binary endpoint: Four similarly design Phase 3 trials on same compound & control
  - Always use one trial for model fitting and the remaining trial(s) for evaluation
- Compared methods
  - Overall treatment effect & conventional subgroup estimates
  - Hierarchical model in a version with “low” and “high” shrinkage
  - Further shrinkage models were also compared (see back-up slides)
- Assess predictive distributions based on proper scoring rules
  - Average scores over all subgroups of interest and prediction direction

Results (preliminary, averaged across 10 replicates)

Model (shrinkage)	Average scoring rule (larger is better) & SE		
	Case 1	Case 2	Case 3
Hierarch. model (high)	**-2.77 (0.01)**	**-4.64 (0.02)**	-4.56 (0.02)
Hierarch. model (low)	-2.79 (0.01)	-4.65 (0.02)	**-4.51 (0.03)**
Conventional	-2.99 (0.01)	**-5.90 (0.04)**	-5.00 (0.04)
Overall	**-3.65 (0.03)**	-4.82 (0.03)	**-5.89 (0.04)**

- Shrinkage models outperform conventional subgroup estimates and overall estimate (rather consistently across shrinkage methods)

Conclusions

- Well-known that conventional subgroup treatment effect estimates are unreliable
  - due to limitations in terms of sample size and multiplicity
- Model-based Bayesian shrinkage estimates for subgroup treatment effect estimates tend to generalize better
- Propose to complement standard subgroup treatment effect estimates (for example in forest plots in primary study publications & label) with estimates based on Bayesian shrinkage

References

- Kent, D. M et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. Annals of Internal Medicine, 172(1), W1-W25
- Wallach, J. D. et al (2017). Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Internal Medicine, 177(4), 554-560.
- Wang, Y. et al. (2024). Bayesian hierarchical models for subgroup analysis. Pharmaceutical Statistics, 23, 1065-1083.
- Yusuf, S. et al (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA, 266, 93-98.
- Jones, H. E. et al (2011). Bayesian models for subgroup analysis in clinical trials. Clinical Trials, 8, 129-143.
- 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

Full results (preliminary, averaged across 10 replicates)

Model (shrinkage)	Average Rank (across 3 cases)	Average SCRP (larger is better) & 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)
Conventional	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)**