Data-driven evaluation of treatment effect heterogeneity

B. Bornkamp[1] K. Sechidis[1] D. Svensson (SIG Lead)[2] A. Venkatasubramaniam [3] on behalf of the Treatment Effect Heterogeneity European Special Interest Group

[1] Novartis Pharma AG, Basel, Switzerland

[2] AstraZeneca, Gothenburg, Sweden

[3] GSK, Stevenage, United Kingdom

10th EFSPI Regulatory Statistics Workshop, Basel, 2025

Primary Focus

The main focus of the **Treatment Effect Heterogeneity SIG** is on methods for assessing heterogeneity of treatment effects (HTE) across a population of patients. *Heterogeneity* can be defined as a nonrandom explainable variability in the direction and magnitude of individual treatment effects, including both beneficial and adverse effects [5]. Assessing the presence of treatment effect variability and explaining it by patient and disease characteristics may help to identify patient subgroups that benefit more or less from a given treatment.

Goals of HTE assessment

- Rigorously assessing available data for evidence against an a-priori assumption of treatment effect homogeneity.
- Supporting decisions on potential population enrichment or informing inclusion criteria for future clinical development studies.
- Generating hypotheses after a failed study.
- Supporting reimbursement applications.
- Providing a comprehensive characterization of treatment effects in scientific publications.
- Supporting development of precision medicine solutions.

Types of subgroup analysis

- *Pre-specified subgroups*: Treatment effects can be evaluated across a (typically small) number of patient subgroups pre-specified in advance based on existing clinical knowledge or hypotheses (e.g., males vs. females).
- *Data-driven subgroup analysis*: HTE can be assessed in a manner where the analysis methodology is pre-specified but relevant patient subgroups and characteristics driving their formation are discovered without a specific hypothesis.

The SIG is currently focused primarily on principled methods for **data-driven subgroup analysis**.

Methodological Challenges and Opportunities

- **Causal Inference** provides a framework for the underlying goal of identifying Conditional Average Treatment Effect (CATE) $\Delta(x_i) = E[Y_i(1) - Y_i(0)|X = x_i]$, i.e., the expected difference in outcomes Y had the patient i, characterized by baseline variables x_i, been treated with treatment 1 versus 0. The fundamental challenge is that in most settings we can observe only $Y_i(1)$ or $Y_i(0)$ but not both. One solution is to model and predict the unobserved outcomes using flexible, high-performing Machine Learning (ML) models f as building blocks $\Delta(x) = f(1, x) - f(0, x)$.
- **Multiple Hypothesis Testing** considerations for controlling false positive rate and quantification of uncertainty about findings are indispensable because the analysis involves multiple stages with many statistical decisions at each step. Resampling techniques for multiplicity control are more complex and computationally intensive in this setting.
- **"Honest" inference** is required to answer questions like: Is there any treatment effect heterogeneity? What is the magnitude of the treatment effect in discovered subgroups? What is the uncertainty about the estimates of CATE? Unbiased ("honest") inference is especially challenging if data-driven discovery and inference have to be done on the same data set.
- **Robustness of findings** needs to be assessed given that methods rely on stochastic algorithms and spurious results can be expected in finite samples. Important questions include: Does the method consistently choose the same variables as predictive of HTE? Does the method consistently classify patients into subgroups in the same way?
- **Semi-supervised Machine Learning** tackles a problem where the true target $\Delta(x)$ is unobservable, but the nuisance functions $f(1, x)$ and $f(0, x)$ can be estimated from data. However, some ML aspects, e.g., model selection and overfitting control, become more challenging. Nevertheless, off-the-shelf ML methods can be tailored and made more robust for the task of HTE assessment.
- **Statistical Power** Clinical trials are often designed to estimate the ATE and are not adequately powered to detect treatment effect heterogeneity. Can supplementary trials / external datasets be leveraged to improve the accuracy of CATE estimates?

Inside SIG scope

a. Regulatory consistency assessment of pre-specified subgroups
b. Promoting theoretical understanding of methodology for data-driven subgroup identification;
c. Recommendations for practical ways of working with the data: how to best approach the assessment of HTE
d. Conducting joint work on e.g, benchmarking methodology via simulations
e. Following research advances on HTE also in other industries (such as econometrics/financial applications)
f. Inspiring statisticians/data scientists working in the HTE field

Outside SIG scope

Confirmatory testing of Subgroups under strong T1E control.

Methodology Examples

a. Subgroup Detection in Dose Finding using MOB
b. Bias Reduction using Model Averaging
c. Subgroup Detection under complexity constraints using SIDES
d. Predictive biomarker discovery controlling type-I error using knockoffs
e. Meta Learning approaches for estimating CATE based on Machine Learning
f. STEPP: model based visualisation of effect as function of a predictive biomarker

About the Special Interest Group

Creation and Affiliation

The **Treatment Effect Heterogeneity SIG** was formed in 2015 as an industry response to the new EMA Guideline on the investigation of subgroups in Confirmatory Clinical Trials, which primarily focused on methodology and interpretation of regulatory consistency assessments, i.e., exploratory pre-specified subgroups. Since then, the area of HTE assessment developed significantly, and the scope of the SIG has correspondingly become wider. It currently consists of 32 members from 14 companies.

Activities

The members of the SIG meet approximately monthly to discuss methodological aspects, recent publications, share case studies, and listen to invited speakers. Members also self-organize into work streams to collaborate on publications, presentations, and research. E.g, the SIG arranged a dedicated conference session at PSI every year since 2021.

Topics of Interest

- **Processes** for setting objectives, engaging with stakeholders, analysis planning, data preparation, interpretation of findings, and reporting
- **Methods** for data-driven HTE discovery, inference, and robustness assessment
- **Software tools** and **practical applications**
- Identification of **methodological challenges** and pitfalls, **benchmarking** of methods

Selected recent collaborative work by the SIG

During 2024-2025, several papers were published under the umbrella of the SIG (cross-industry collaborations) and they are listed below:

- ENAR tutorial, New Orleans, March 24 (Lecturer: Ilya Lipkovich; co-authored by David Svensson, Bohdana Ratitch, and Alex Dmitrienko). Title: *Modern approaches for identifying heterogeneous treatment effects from experimental and observational data*, based on our 2024 tutorial paper in Statistics in Medicine [1].
- A recent collaboration among many SA-SIG members and other experts involved resulted in a practical process-oriented guideline for a structured approach to of HTE assessment, including preparation of data, cross-functional communication with stakeholders, and interpretation of findings.[2]
- Another SIG collaboration was devoted to exploring the intersection of methodology for causal effects and Interpretable Machine Learning, resulting in a tutorial-style overview. [4]
- Another publication compares the various methods for assessing treatment effect heterogeneity, but also evaluates their performance in simulation scenarios that mimic real clinical trials. Furthermore it introduces an R package (benchtm), which can simulate synthetic biomarker distributions based on actual clinical trial data and create interpretable scenarios to benchmark methods for identifying and estimating treatment effect heterogeneity. [3].

Method	Modeling type (1)	Application type (2)	Dimensionality (3)	Results type (4)	Inferential support (5)	Software (6)
Global outcome modeling						
Virtual Twins[61]	Freq/NP	RCT, OS	High	CATE, S	No	R [aVirtualTwins]
S-, T- X-learner[55]	Freq/SP, NP	RCT, OS	High	CATE, S	No	R [rlearner][a]
Global treatment effect modeling						
Interaction Trees[62]	Freq/NP	RCT	Medium	S	No	B
GUIDE[63]	Freq/NP	RCT	Medium	B,S	Yes	B
Model-based trees and forests[65]	Freq/NP	RCT	High	B, S	Yes	R [model4you]
Causal forests[71,72]	Freq/NP	RCT, OS	High	B, CATE	Yes[G, CATE]	R [grf]
Bayesian causal forests[160,230]	Bayes/NP	RCT, OS	High	CATE	Yes[CATE]	R [bcf, bartCausal]
Bayesian linear models[177,179]	Bayes/P	RCT	Low	CATE	Yes[CATE, S]	R [DSBayes, beanz]
Modified loss methods[78]	Freq/P, SP, NP	RCT, OS	High	CATE	No	R [personalized]
R-learner[83]	Freq/P, SP, NP	RCT, OS	High	CATE	No	R [rlearner][a]
Direct modeling of ITR						
AIPW estimator[96]	Freq/SP	RCT, OS	Medium	ITR	No	R [DynTxRegime]
OWL,[95] RWL,[100] AOL[101]	Freq/P, SP, NP	RCT, OS	High	ITR	No	R [DTRlearn2]
Tree-based ITR[99,116,118]	Freq/SP, NP	RCT, OS	Medium	ITR	No	R [T-RL,[b] policytree]
Direct subgroup identification						
SIDES and SIDEScreen[126,127]	Freq/NP	RCT	Medium	B, S	Yes[G, S]	B, R [SIDES, rsides]
TSDT[134]	Freq/NP	RCT	Medium	B, S	Yes[S]	R [TSDT]
PRIM[124]	Freq/NP	RCT	Medium	S	No	R [SubgrpID]
Sequential-Batting[131]	Freq/NP	RCT	Medium	S	Yes[S]	R [SubgrpID]
CAPITAL[120]	Freq/NP	RCT	Medium	S	No	R [policytree]
Bayesian Model Averaging[181]	Bayes/NP	RCT	Low	S	Yes[S]	R [subtee]

[a] Available at GitHub **xnie/rlearner**.
[b] Available at GitHub **Team-Wang-Lab/T-RL**.

References

[1] I. Lipkovich, D. Svensson, B. Ratitch, and A. Dmitrienko. "Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data". In: *Statistics in Medicine* n/a.n/a (2024).

[2] K. Sechidis, S. Sun, Y. Chen, J. Lu, C. Zang, M. Baillie, D. Ohlssen, M. Vandemeulebroecke, R. Hemmings, S. Ruberg, and B. Bornkamp. "WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors". In: *Under review* (2024). DOI: https://arxiv.org/abs/2405.00859.

[3] S. Sun, K. Sechidis, Y. Chen, J. Lu, C. Ma, A. Mirshani, D. Ohlssen, M. Vandemeulebroecke, and B. Bornkamp. "Comparing algorithms for characterizing treatment effect heterogeneity in randomized trials". In: *Biometrical Journal* 66.1 (2024), p. 2100337.

[4] D. Svensson, E. Hermansson, N. Nikolaou, K. Sechidis, and I. Lipkovich. *Overview and practical recommendations on using Shapley Values for identifying predictive biomarkers via CATE modeling*. 2025. arXiv: 2505.01145 [stat.ME].

[5] R. Varadhan, J. Segal, C. Boyd, A. Wu, and C. Weiss. "A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research.". In: *J Clin Epidemiol.* 66(8) (2013), pp. 818–825.