



PSI 2021 SUBGROUPS SIG SESSION

Some classical and novel approaches for assessing subgroup consistency

Speaker: David Svensson¹

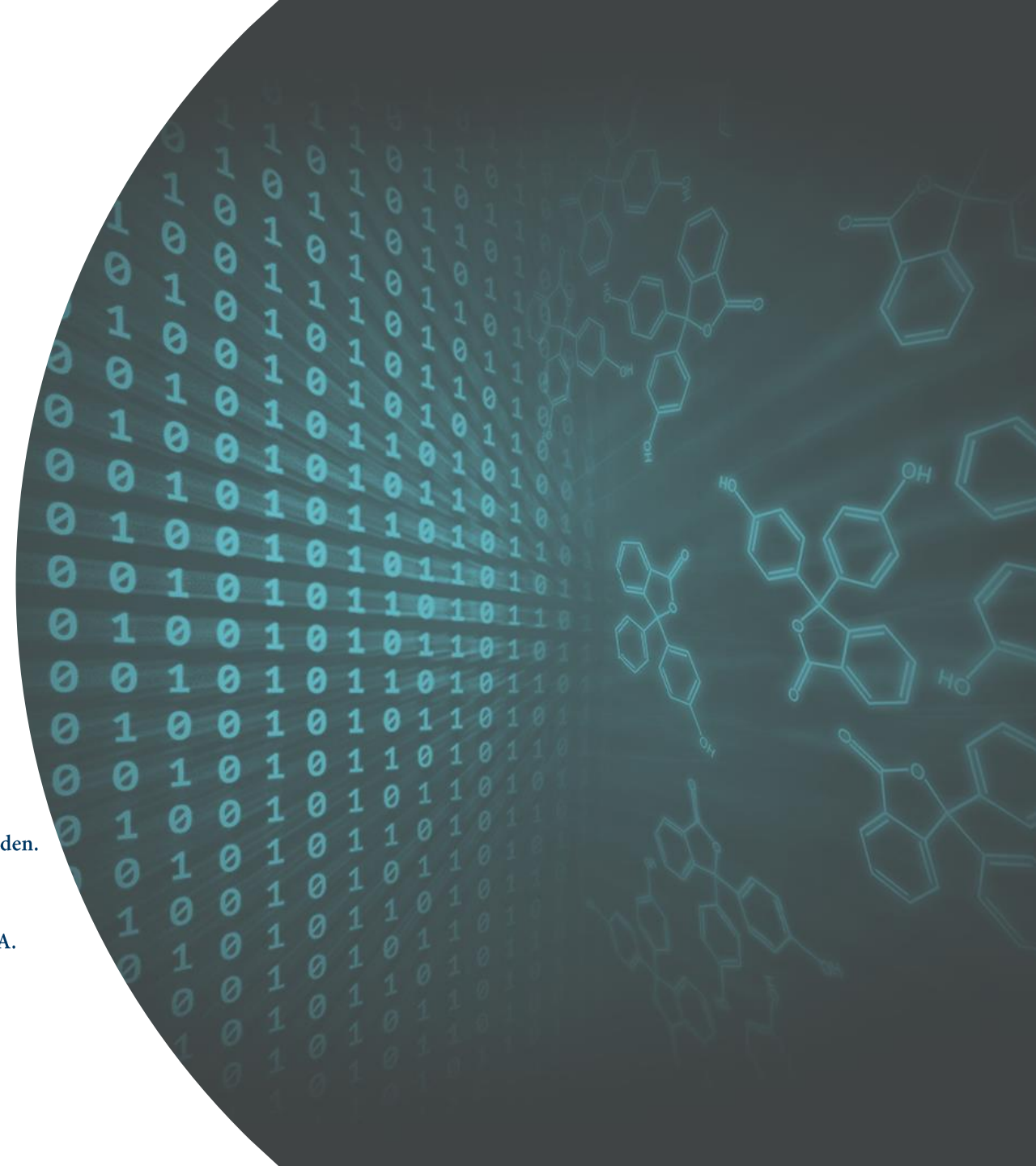
Joint with Dan Jacksson², Michael O'Kelly³, and Fredrik Öhrn¹

1: Statistical Innovation, BioPharmaceuticals R&D; Data Science & AI, Gothenburg, AstraZeneca R&D, Sweden.

2: Oncology Data Strategy & Network, Central Cambridge, AstraZeneca R&D, United Kingdom.

3: Center for Statistics in Drug Development, Decision Sciences Data Sciences, Safety and Regulatory, IQVIA.

22th June 2021



Content / Scope

SNAPSHOT OF SOME ONGOING WORK UNDER THE SIG UMBRELLA

- **Treatment Effect Heterogeneity & Subgroup Consistency Assessment**
 - *Recap: the typical statistical issues ...*
- Some alternative approaches...
 - *E.g., the graphical SEAMOS ... and tweaks of it.*
- But performance? Tweaks better? Or classical methods best?
 - *Some simulation results: work in progress*

Acknowledgement: Jonathan Bartlett, Aaron Dane, Amy Spencer, Andy Stone

DISCLAIMER: the **opinions** expressed in this presentation are those of the **authors**, and do **not** necessarily reflect the official policy of AstraZeneca/IQVIA.



Notorious subgroup problems in Pharma

- 1. **Confirmatory** testing of subgroups.
- 2. **Consistency Assessment** of treatment effects.
- 3. **Selection** of best (pre-specified) subgroup.
- 4. **Data-driven Discovery** of 'best subgroup' (ML/ Causal Inference)



Notorious subgroup problems in Pharma

- 1. **Confirmatory** testing of subgroups. ← strong α -error-control
- 2. **Consistency Assessment** of treatment effects.
- 3. **Selection** of best (pre-specified) subgroup.
- 4. **Data-driven Discovery** of 'best subgroup' (ML/ Causal Inference)



Notorious subgroup problems in Pharma

- 1. **Confirmatory** testing of subgroups.
- 2. **Consistency Assessment** of treatment effects.

- 3. **Selection** of best (pre-specified) subgroup.
- 4. **Data-driven Discovery** of 'best subgroup' (ML/ Causal Inference)

← Exploratory, discovery:

Enhanced effect
somewhere?



Notorious subgroup problems in Pharma

- 1. **Confirmatory** testing of subgroups.
- 2. **Consistency Assessment** of treatment effects.
- 3. **Selection** of best (pre-specified) subgroup.
- 4. **Data-driven Discovery** of 'best subgroup' (ML/ Causal Inference)

THIS TALK!

Exploratory but strict benefit-risk assessment, regulatory, labelling, ~ no excessive subgr. search pls

[see e.g., Koch & Framke 2014 for discussions]

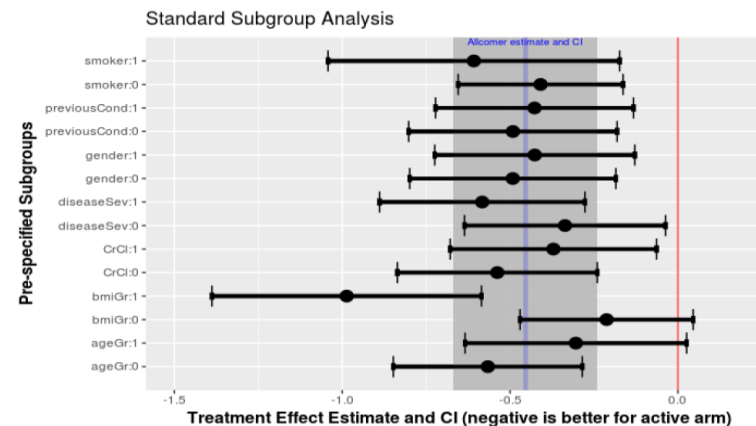
'Meaningful to talk about the overall effect across different sub-populations'?



Consistency Assessment is notoriously difficult ...

- Consistency yes = 'if subgroups look similar'... lacks consensus strict definition!
- Core assessment of pivotal trials, but **lots of inherent issues**:
- Done without type-1-error control... Actually, not a testing problem
- 'lack of reject' is not 'evidence of homogeneity'
- Standard is **Regulatory Review of a Forest Plot**:
 - biological plausibility, contextual knowledge. I.e., not merely an statistical inference.
 - But interaction p-values are often computed ...
 - The eye notes deviating point estimates ...

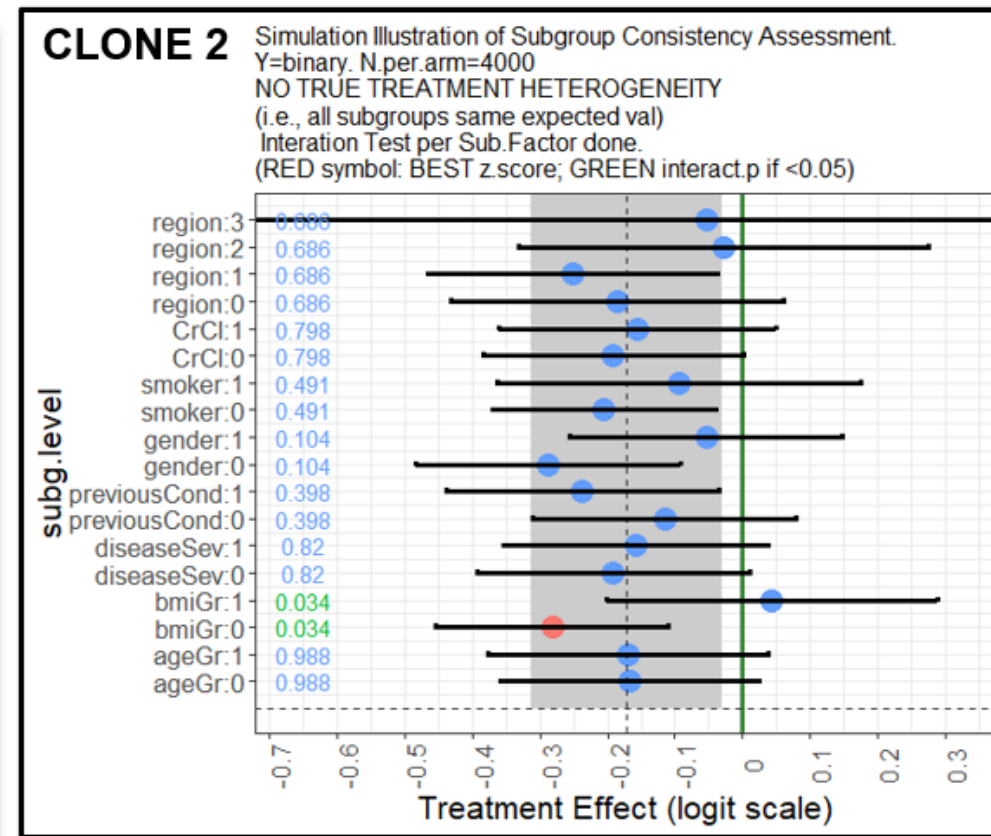
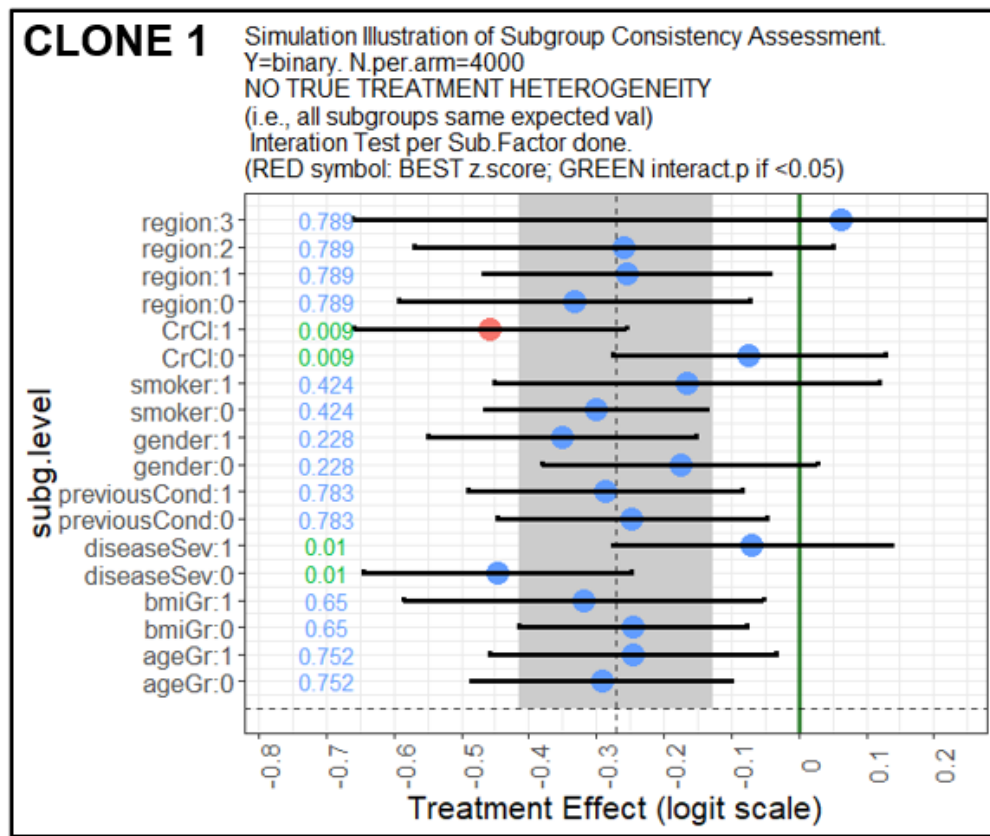
HOW MUCH DEVIATION FROM THE OVERALL IS EXPECTED?



Any statistician knows this: inherent issues

Simulation with $E[\text{subgroups}] = \text{constant}$.

Two replicates



$V(\text{estimated HTE}) \approx 4 \times V(\text{estimated overall treatment effect})$. *Pharmaceutical Statistics*. 2020;1-13.



Recap: construction of a FOREST plots

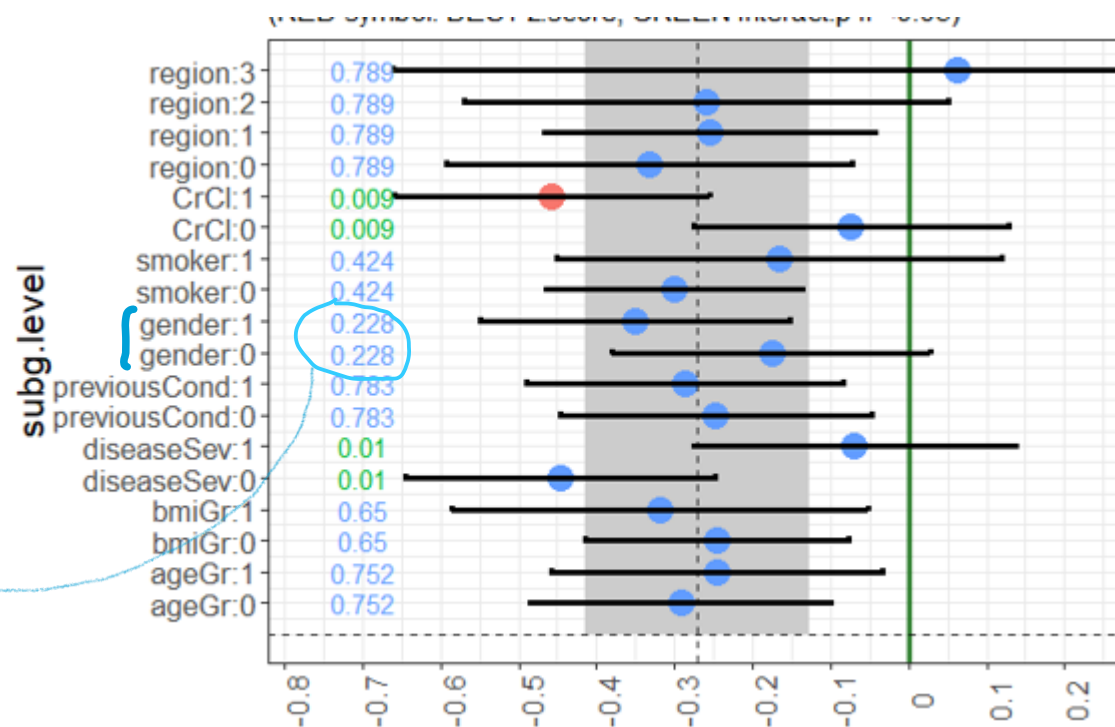
Let x_1, \dots, x_p be subgroup factors. Standard practice: one model per factor.

E.g., for x_j =GENDER, estimate 'male' and 'female' from a model in the style of (1) or (2)

(1) $GLM(Y \sim x_j + TRT + x_j \cdot TRT)$

(2) $GLM(Y = TRT + x_1 + x_2 + \dots + x_p + x_j \cdot TRT)$

GIVES: point est., CI & interaction ($x_j \cdot TRT$) p-value



(Notation 'GLM' here: refers to 'some appropriate model for your endpoint', e.g., could be Cox prop haz, NegBin. More later on this).



The debates go on – a snapshot of issues/views/ideas

'The hardest problem there is', [Ruberg 2021]

'Avoid dichotomization', 'a very bad idea', [Keene; Altman; Royston et. al, & many others]

Panel debate at recent DIA/FDA Biostat. Industry and Regulatory Forum, [Rothmann et. al 2020]

- '**we can do more and better**', 'role of **prior** evidence?'

- 'what is a good estimate for a 65-year, bald, Caucasian-American, male, patient?'

Novel definitions of consistency... [Kent et. al], Japanese guidelines, $D_{\text{japan}} > \pi * D_{\text{all}}$

Type II error rate more important than Type I? [Koch Schwartz], [Koch Hemmings]

Shrinkage? '*perhaps helpful*' [Alosh], [Rothmann et. al], [Jones et. al], [Varadhan et. al]

Difficulties due to confounding (overlap) [Varadhan et. al]

Heterogeneity Cochran Q, I^2 ; Equivalence Testing of trt effects; [e.g., Wellek, Koch & Framke]

Worst/best subgroup: Bootstrap Bias reduction [Rozenkranz] / Model-averaging shrinkage [Bornkamp et. al]

New effective graphical ideas ... [Muysers et. al, Ballarini et. al]

"Always do [disciplined] subgroup identification" !!! (e.g., Tree-based/ML ITE-approaches) [Ruberg in Rothmann 2021]



General consensus? Well, use 'holistic' assessments ...

Guideline on the investigation of subgroups in confirmatory clinical trials

EMA 2014/2019: Discuss potential subgroups at design stage, *Pre-specify 3 tiers & discuss biol. plausibility of results; use Forest plots (=avoid presenting isolated subgr. results).*

A formal rule for the interpretation of subgroup findings presented in a Forest plot that is both sensitive to detect inconsistency in treatment effects and specific to avoid false-positive findings is not available.

Subgroup Analysis and Interpretation for Phase 3 Confirmatory Trials: White paper of the EFSPi/PSI Working Group on Subgroup Analysis

Recommended e.g., graphical & **SEAMOS** (+ some other methods)

- e.g., 'assumption-free', handles overlaps, all endpoints ... appeared useful
- But performance of SEAMOS not yet entirely understood? (Beyond linear case?)
- EFSPi SIG ongoing work (Dane A., et. al.)
- PSI O'Kelly M. 2020: preliminary simulations presented (linear case): '*be careful*'



This talk: focusing on SEAMOS vs Classical methods

- Private communication under the EFSPI umbrella after M. O'Kelly's PSI talk 2020 led us to more investigations ...
 - *e.g., what happens beyond the linear case*
- The intention of this talk is to **tell that story**, rather than put forward a single method
- Evidence-based: 'What do we see' - at this time - regarding performance?



Original SEAMOS idea: track deviations from overall

What is being tracked? Almost the **yellow**:

Actually this one:

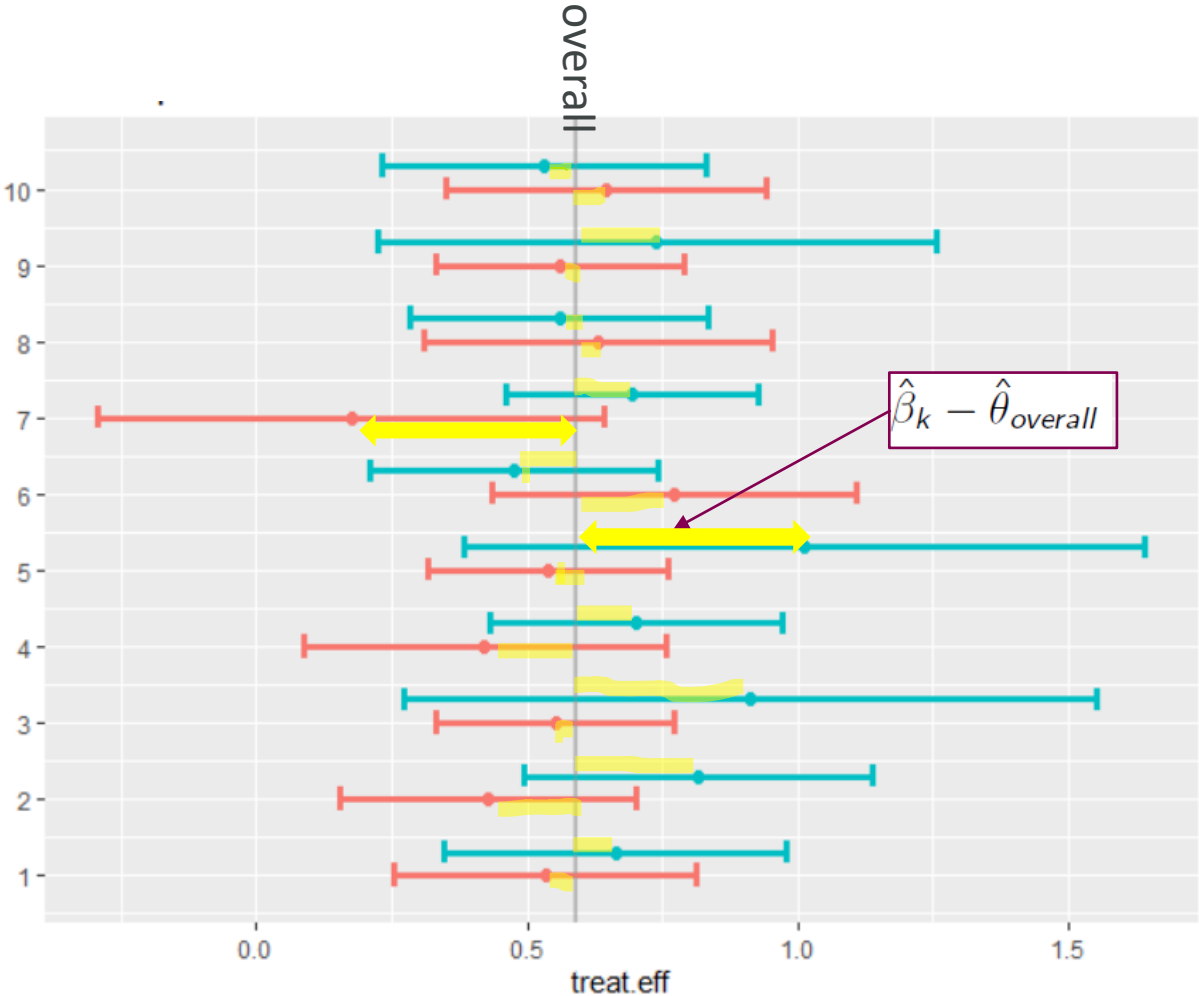
$$z_k = \frac{(\hat{\beta}_k - \hat{\theta}_{overall})}{\hat{\sigma}_k}$$

k enumerates the subgroups, and $\hat{\sigma}_k = SE(subrg_k)$ (e.g., k=1,...,20 if ten binary x).

Note: other scores don't fit the purpose:

Usual z-scores don't capture **deviation** from overall.

E.g., $\hat{\beta}_k - \hat{\theta}_{overall}$ doesn't **punish** small subgroups. (=too much attention to artefacts, see yellow)



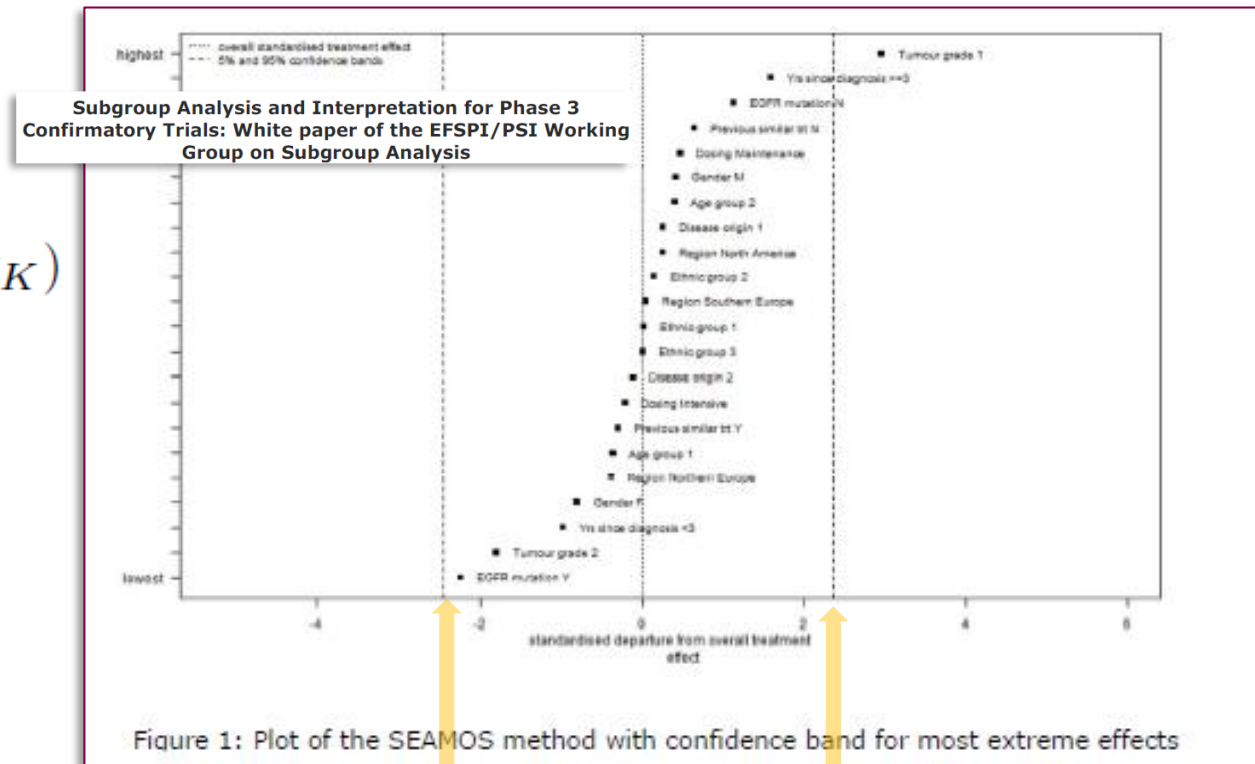
SEAMOS (cont)

You can order your observed $z_{(obs)} = (z_1, \dots, z_K)$

In each permutation, do the same:

- get 1000 new such ordered vectors under NULL
- graph this in Forest style:

Extract 10th and 90th percentiles from it and add as vertical reference lines in the plot
(example from Pharm.Stat. vol 18, issue 2, 2018, Dane et. al.)



If permuted values tend to be 'nicer' than your observed values, then evidence of H.T.E.

Formal test: p-value = $\frac{1}{1000} \sum_{h=1}^{1000} I(z_h^{(max)} \geq Q^{(obs)})$, where $z_h^{(max)} = \max(|z_{1h}|, |z_{2h}|, \dots, |z_{Kh}|)$, $h=1, \dots, 1000$
 and $Q^{(obs)} = \max(|z_{1,obs}|, |z_{2,obs}|, \dots, |z_{K,obs}|)$,



Note: SEAMOS is primarily a 'graphical aid'

- The EFSPI SEAMOS is a graphical method ('aiding the Eye')
 - but a natural question must be: '*is it any good*'?

In principle, it is conceivable to plot data 'nicely but lacking good theoretical properties'.

- Type 1 error: how often will it make us think 'there is something'?
- Power: if true H.T.E. present, 'will it discover it'?

(HTE= Heterogeneous Treatment Effects)

In the lack of formal proofs, **simulations are needed**. (Formal proofs? Ongoing research)

- And it is **NOT** apriori obvious that SEAMOS should be ok – see next slides.



Presence of prognostic effects? Be careful.

SEAMOS takes out both prognostic and predictive effects. Does it matter?

- (Permuting rows of X? e.g., see [Foster et. al]).

Generally, **be careful with prognostic variables**: for some endpoints, the trt estimates become biased (!) if the model is ignoring a truly prognostic variable

- [papers by GAIL, Hauck]. Also tendency in ML to go wrong on prognostic effects [Sechidis et. al].

Foster et. al. discussed linear case: various ways of permuting to 'make $g()$ =constant'

- for model $Y = h(\mathbf{x}) + g(\mathbf{x}) \cdot (TRT - \pi) + \epsilon$ with $TRT = \text{rand.trt}$, \mathbf{x} covariates, and $\mathbf{P}(TRT=1)=\pi$
- **TAKEAWAY MESSAGE: quite complex, and e.g., permute(TRT) is too simplistic.**



Co-authors had similar ideas: keep prognostic terms useful?

- Idea: rather than simulate NULL data, can we instead **“NULLIFY the analysis”**?

THINKING:

- Generally hard to generate NULL data (= keeping marginal properties, correlations, overall effect, prognostic effects – but removing covariate-trt-interactions)
- Already complex in linear case (see Foster et al).

What if the interaction term instead is permuted (when 'making the Forest plot')?

$$Y = \beta_0 + \gamma \cdot TRT + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots \beta_p \cdot x_p + \delta \cdot z_j \cdot TRT$$

(z_j = permuted version of subgr. factor x_j)



We explored via simulations – and beyond linear case.

We looked at **rejection rates** (10% nominal level): observed $\max(abs(z_k)) > \text{perm.based_critical.value}$
(percentile of permutation-based NULL reference distribution).

Endpoints: linear, binary, counts

Notation:

- SEAMOS1 = the standard 'EFSPI SEAMOS' ("NULLIFYING the data": X-matrix permuted)
- SEAMOS2 = our tweaked with progn. effects ("NULLIFYING the cov-trt-interactions")
- ZEAMOS2 = another tweak, SEAMOS2 but tracking 'full body of deviation' $\sum_k abs(z_k)$
- GIT = classical Global Interaction Test (=LR test, inclusion of interaction-terms add value?)
- Bonferroni {Subgr.factor-specific p-values}



However, highs and lows discovered: (1)

You think it is trivial to set up simulations for this? Think again.

Bissonnette, Ickes, Bernstein, and Knowles (1990) conducted a simulation study to compare the dichotomization approach to the regression approach for examining moderator effects. When no moderator effect was present in the population, they found high Type I error rates under the dichotomization approach, indicating common occurrence of spurious interactions.

Personality Moderating Variables: A Warning about Statistical Artifact and a Comparison of Analytic Techniques

Victor Bissonnette, William Ickes, Ira Bernstein, Eric Knowles

First published: September 1990 | <https://doi.org/10.1111/j.1467-6494.1990.tb00243.x> | Citations: 54

Inflation of classical methods noted (!) in count setting (poisson regression):

- if simulate e.g. 'age', 'bmi', 'eos', continuously, and link as
$$= a + \gamma \cdot TRT + \beta_1 \cdot age + \beta_2 \cdot bmi + \beta_3 \cdot eos + \epsilon \text{ (just an example)}$$
- Then mimicing standard practice, dichotomization per protocol, Forest plot, etc.
 - **But this inflates GIT!**

$GIT(\text{continuous } x) = 10\% \text{ rej. rate, } GIT(\text{dich. } x) > 10\% \text{ (e.g., 29\% in one example)}$



However, highs and lows discovered: (2)

So we simulated links directly in terms of the **dichotomized** x:

$$\sim a + b_1 * \text{age.group} + b_2 * \text{bmi.group} \dots$$

So, is everything fine now?

No: we noted BONFERRONI surprising high POWER (the $Y = \text{count}$ case again): due to

- **T1E INFLATION** of subgr.specific models of this kind: $\text{GLM}(Y \sim x_j + \text{TRT} + x_j \cdot \text{TRT})$

- So 'all-main-effect-models' were needed: $\text{GLM}(Y = \text{TRT} + x_1 + x_2 + \dots + x_p + x_j \cdot \text{TRT})$



Obvious, but still:

Don't embark on comparing POWER without first making sure NO INFLATION...

Actually still non-trivial: (despite link in terms of dich. x)...

BINARY CASE: **GIT still inflated** (quite a lot) when simulating **moderate-sized** trials

- type-1-error rate inflation - now due to issues with *asymptotics*.

(And care if using R `lrtest()` with `glm(, family=gaussian)` – doesn't render F test – inflation!)



SEAMOS inflation unless:

We soon noted that EFSPi version of SEAMOS relied on 'all main effects' models for good reasons:

- INFLATION if overall estimated without 'all main effects' (linear case)

Also, practical discoveries:

SEAMOS has **long run times**, so fewer permutations were considered:

- but, empirically, spurious results with fewer permutations ($n_{perm}=100$)
- We noted that **$n_{perm} \geq 1000$** needed for stability. (Of course, it depends!)



Simulations displayed here:

Resampling from data set ACTG175 (RCT, in R package 'speff2trial') [Credit to J. Bartlett for idea]

Gave RCTs with total $n=400$, $p=10$ subgroup factors

- typically many prognostic effects present,
- sometimes with differential effects (ANTINULL), sometimes without (NULL)

In particular: varying \mathbf{b}_1 and \mathbf{b}_2 in LINK=

$$= w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT + \epsilon \text{ where } w = \hat{\alpha} + X\vec{\beta}, \text{ i.e., an expression in } 10 \text{ } x_j$$

(NULL iff $\mathbf{b}_1=0$ and $\mathbf{b}_2=0$).



(Y=continuous)

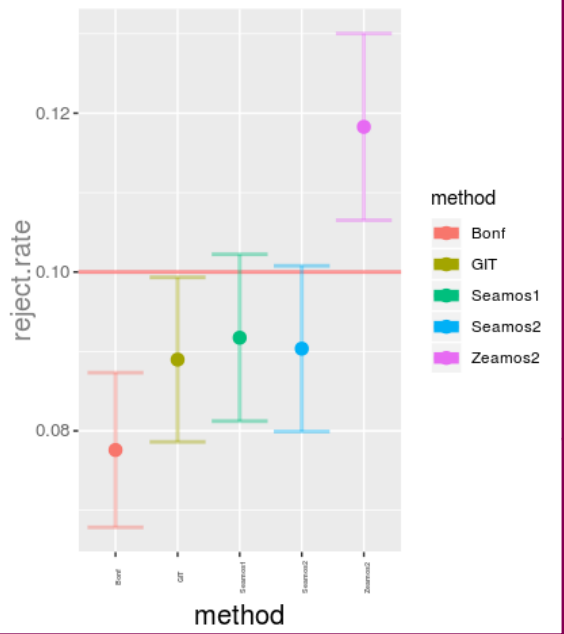
lin. model

key info

trial size = 400, resampling RCT.
 10 truly prognostic x, 2 predictive:
 x4 & x7 (pred. strength = b1, b2)
 Positive corr between x4 and x7

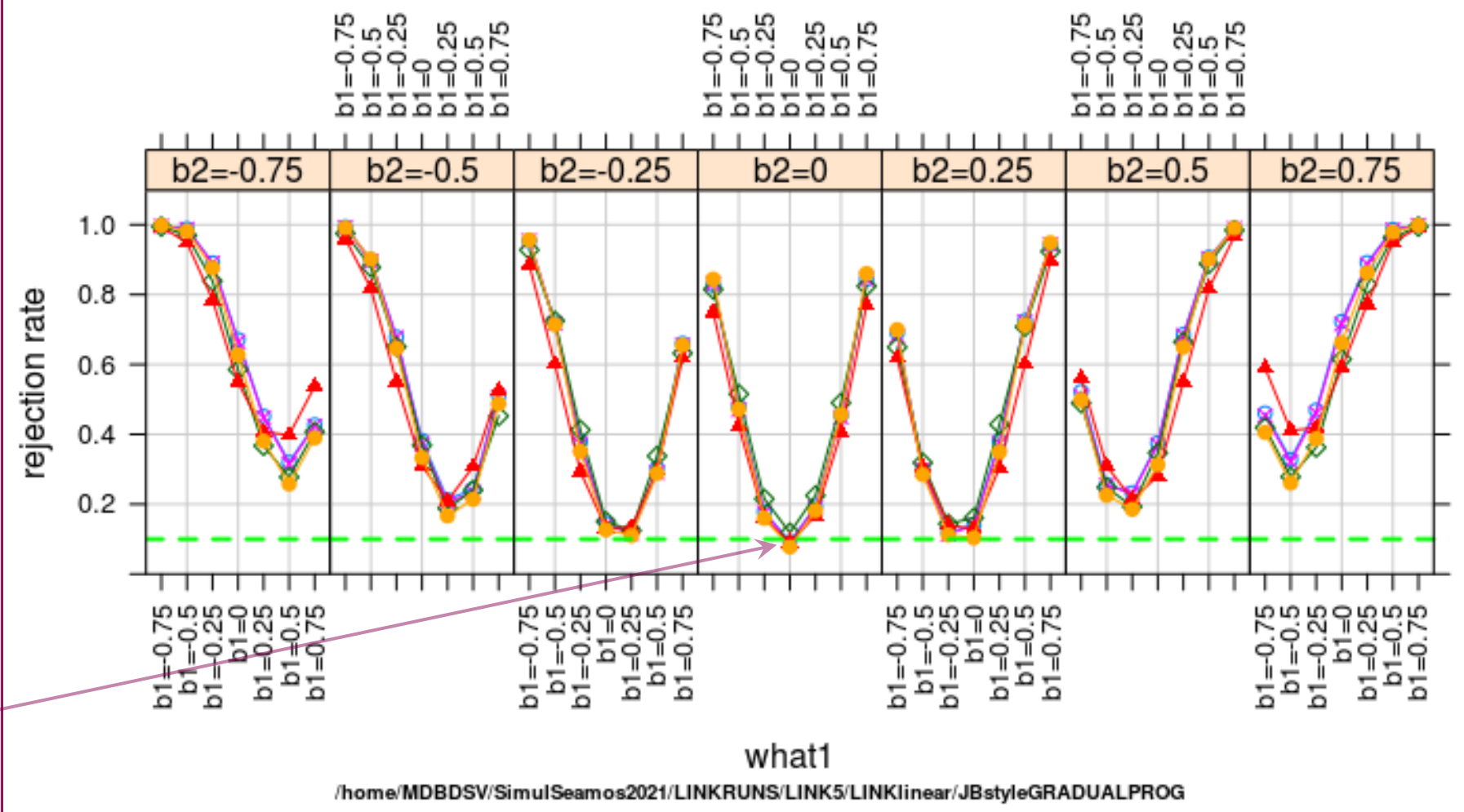
NULL (t1-error)

(ACTG175 resampl.approach) No.RCT.iters=2900, N.SEAM.PERM=1000
 (NULL: b1=0 and b2=0)
 Continuous endpoint, NULL case (b1==0 AND b2==0)



(ACTG175 resampl.approach)
 No.RCT.iters=2900, N.SEAM.PERM=1000
 NULL: b1=0 and b2=0

reject.rate.s1 ○
 reject.rate.s2 ✕
 reject.rate.z2 ◇
 reject.rate.git ▲
 reject.rate.bonf ●



$$= w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT + \epsilon \text{ where } w = \hat{a} + X\hat{\beta}, \text{ i.e., an expression in } 10 x_j$$

(Y=binary)

log.regr model

key info

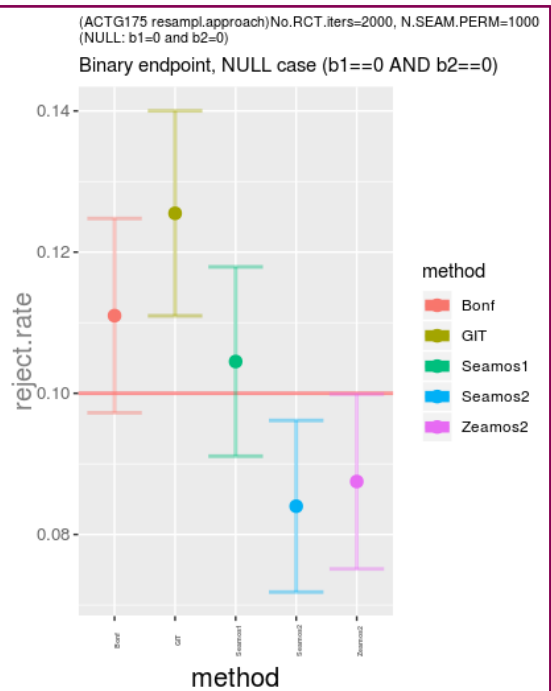
trial size = 400, resampling RCT.

10 truly prognostic x, 2 predictive:

x4 & x7 (pred. strength = b1, b2)

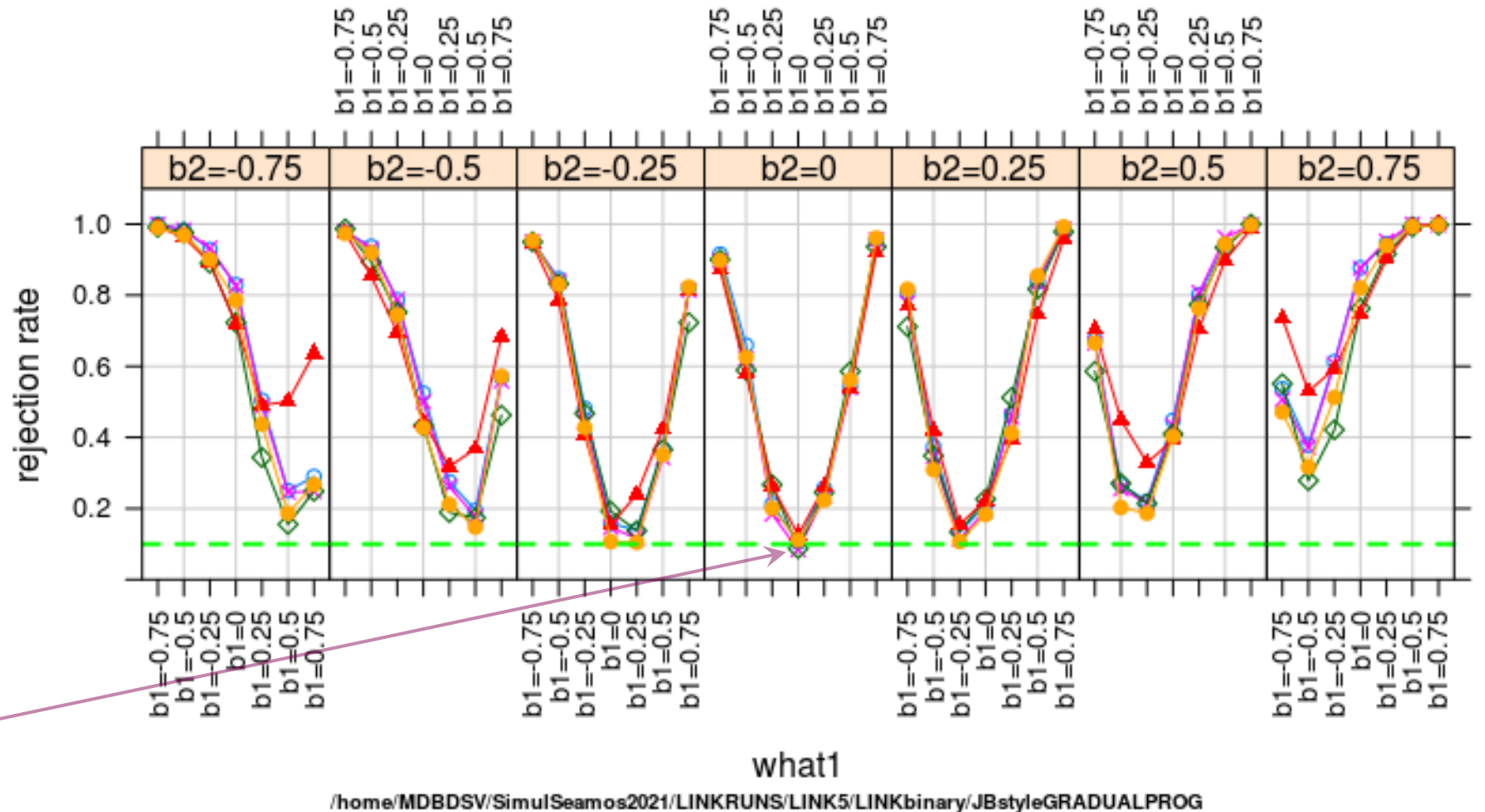
Positive corr between x4 and x7

NULL (t1-error)



(ACTG175 resampl.approach)
No.RCT.iters=2000, N.SEAM.PERM=1000
NULL: b1=0 and b2=0

reject.rate.s1 ○
reject.rate.s2 ✕
reject.rate.z2 ◇
reject.rate.git ▲
reject.rate.bonf ●



$$= w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT \quad \text{where } w = \hat{\alpha} + X\hat{\beta}, \text{ i.e., an expression in } 10 \ x_j$$

(Y=count)

Poisson-regr model

key info

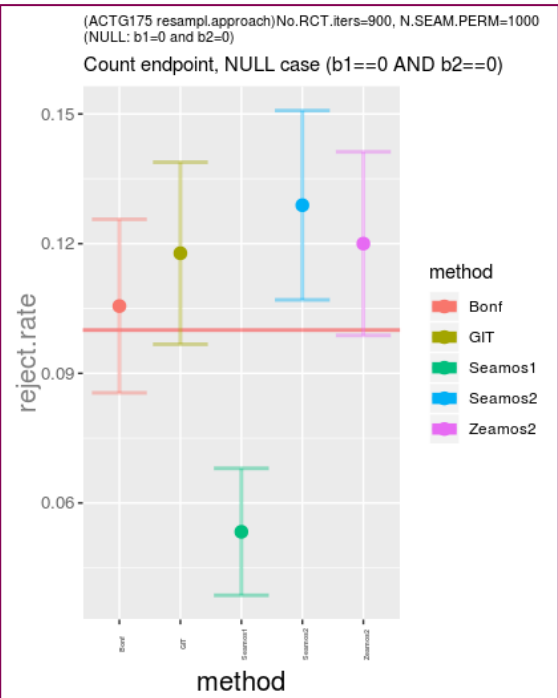
trial size = 400, resampling RCT.

10 truly prognostic x, 2 predictive:

x4 & x7 (pred. strength = b1, b2)

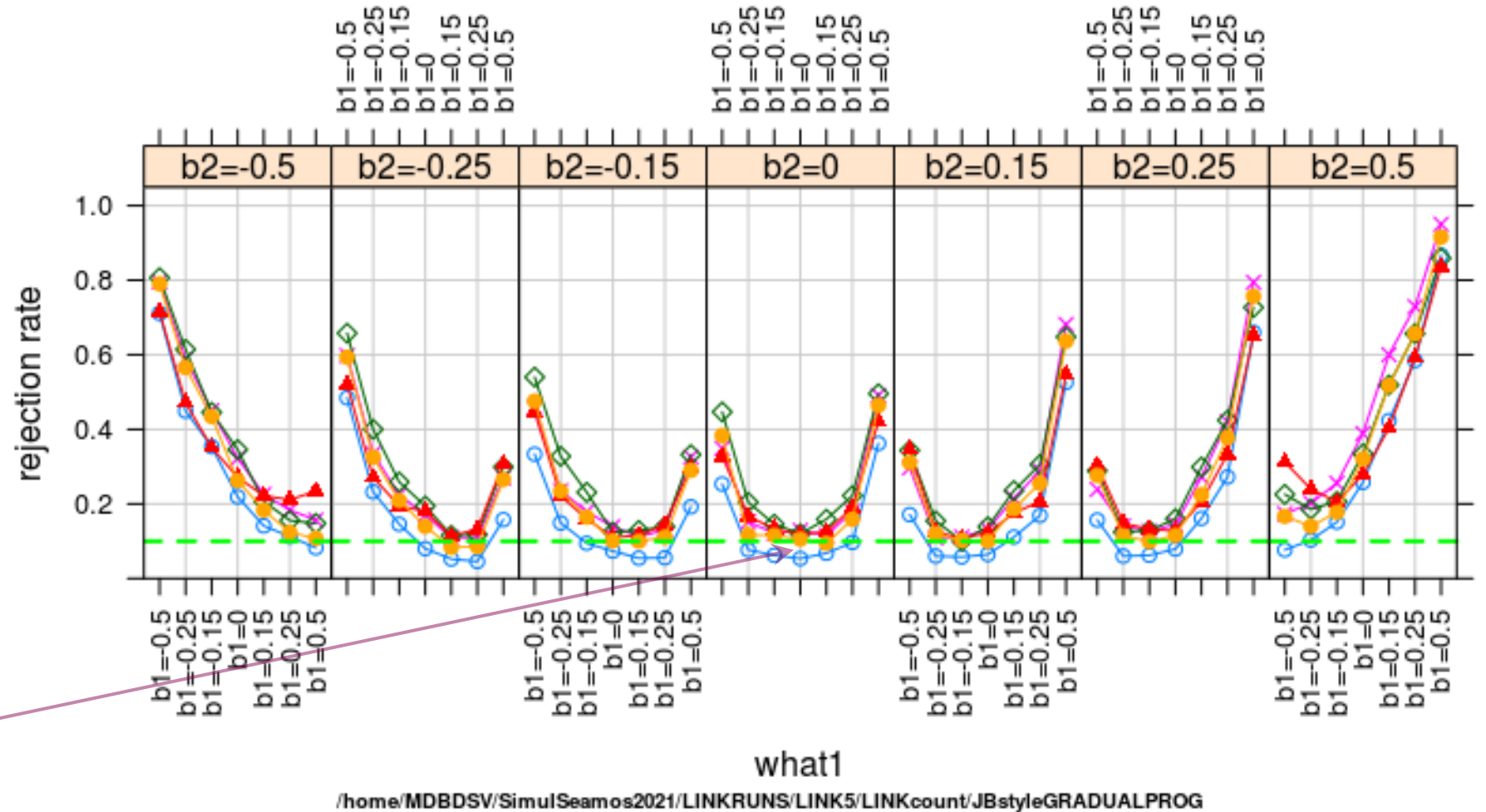
Positive corr between x4 and x7

NULL (t1-error)



(ACTG175 resampl.approach)
No.RCT.iters=900, N.SEAM.PERM=1000
NULL: b1=0 and b2=0

reject.rate.s1 ○
reject.rate.s2 ×
reject.rate.z2 ◇
reject.rate.git ▲
reject.rate.bonf ●



$$= w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT \quad \text{where } w = \hat{\alpha} + X\beta, \text{ i.e., an expression in } 10 \ x_j$$

CONCLUSIONS

- Ongoing research – snapshot of what the SIG does
- Too early for firm conclusions regarding SEAMOS:es. (JURY IS STILL OUT)
- Be careful with permutation methods – tricky animals!
- Dichotomization is here to stay – but it is bad in many ways (statistically speaking)



References (1):

- **Alosh M, Pennello G.** Statistical Considerations on Subgroup Analysis in Clinical Trials. Article in Statistics in Biopharmaceutical Research 2015.
- **Alosh, M.** Statistical Considerations on Subgroup Analysis: Interpretation of clinical trial findings and study design for targeted subgroup. Conference paper, FDA/DIA Statistics Forum, At North Bethesda, Maryland, US, April 2014.
- **Altman D, Royston P.** The cost of dichotomising continuous variables. *BMJ* 2006}; 332:1080.
- **Ballarini N., Chiu Y, König F., Posch F, Jaki T.** A critical review of graphics for subgroup analyses in clinical trials. *Pharmaceutical Statistics*. 2020;1–20.
- **Bissonnette V, Ickes W, Bernstein I, Knowles E.** Personality Moderating Variables: A Warning about Statistical Artifact and a Comparison of Analytic Techniques. 1990 <https://doi.org/10.1111/j.1467-6494.1990.tb00243.x>
- **Bornkamp, B., Ohlssen D., Magnusson B., Schmidli, H.;** Model Averaging for Treatment effect estimation in subgroups. *Phar. Statistics*, 2017, vol 16.
- **Brookes ST., Whitely E., Egger M., Smith GD., Mulheran PA., Peters TJ.;** Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol*. 2004, Vol 57(3).
- **Battioui C, Shen L, Ruberg SJ.** A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment.
- **Chen J, et. al.** Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceut. Statist*. 9: 242–253 (2010)
- **Cui L, Hung JHM, Wang SJ, Tsong Y.** Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics* 2002, Vol 12.
- **Dane, A., Spencer, A., Rozenkranz, G., Lipkovich, I., Parke, T.** Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the EFSP/PSI working group on subgroup analysis. *Pharmaceutical Statistics* 18(2), Dec 2018.
- **EMA Guideline** on the investigation of subgroups in confirmatory clinical trials http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf
- **FDA and Johns Hopkins University Center of Excellence in Regulatory Science and Innovation (JHU-CERSI).** *Assessing and Communicating Heterogeneity of Treatment Effects (HTE) for Patient Subpopulations: Challenges and Opportunities*. Public Symposium November 28, 2018.



References (2):

- **Foster, J., Kaciroti, N.** Permutation Testing for Treatment–Covariate Interactions and Subgroup Identification. *Statistics in Biosciences* · March 2015}.
- **Gail M, Wieand S, Piantadosi S.** Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. *Biometrika*, Vol. 71, No. 3. (Dec., 1984), pp. 431-444.
- **Hauck W, Anderson S, Marcus S.** Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? *Controlled Clinical Trials* 19:249–256 (1998)
- **Hemmings R., Koch A.** Commentary on: Subgroup analysis and interpretation for phase 3 confirmatory trials: White Paper of the EFSP/PSI working group on subgroup analysis by Dane, Spencer, Rosenkranz, Lipkovich, and Parke. *Pharmaceutical Statistics*. 2019;18:140–144.
- **Higgins JPT, Thompson.** Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2020, 21.
- **Jones H, Ohlssen B,** Neuenschwander B., Racine A., Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 8.
- **Keene O., Bratton O.** Subgroup Analysis: a view from the Industry. *Design and Analysis of Subgroups with Biopharmaceutical Applications*. Springer 2020, book chapter.
- **Kent D., Rothman P., Ioannidis J., Altman D., Hayward R.** Assessing and reporting heterogeneity intreatment effects in clinical trials: a proposal. *Pharmaceutical Statistics*}. 2020;1–13.
- **Koch A., Framke T.** Reliably basing conclusions on subgroups of Randomized Clinical Trials. *J. of Biopharmaceutical Statistics*. 2014.
- **Lipkovch I., Dmitrienko A., Muysers C., Ratitch B.** Multiplicity issues in exploratory subgroup analysis. *J. of Biopharmaceutical Statistics*. 2018.
- **Lipkovich I., Dmitrienko A., Denne, J., Enas G.** Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment inpatient subpopulations. *Statist. Med.*} 2011, 30 2601–2621.
- **Li Z, Chuang-Stein C, Hoseyni C.** The probability of observing negative subgroup results when the treatment effect is positive and homogenous across all subgroups. *Drug information Journal* 2007, 41.
- **Loh W., Man M., Wang S.** Subgroups from Regression Trees with adjustment for Prognostic effects and post-selection inference. *Statistics in Medicine*, 2018



References (3)

- **Muysers C, et. al.** A Systematic Approach for Post Hoc Subgroup Analyses With Applications in Clinical Case Studies. *Therapeutic Innovation & Regulatory Science* 2020, Vol. 54(3) 507-518 <https://doi.org/10.1007/s43441-019-00082-6>
- **Marchenko O., Katenka N.** *Quantitative Methods in Pharmaceutical Research and Development.* Springer 2020.
- **O’Kelly M.** Subgroup analysis: a look at the SEAMOS approach. PSI presentation 18th Aug, 2020.
- **Pocock S., Assmann S., Enos L., Kasten L.** Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist. Med.* 2002; 21:2917–2930.
- **Rothmann M., Crown W., Louis T., Permutt T., Ruberg S., Segal J., Scott J.** Assessing and communicating heterogeneity of treatment effects for patient subpopulations: Panel discussion on considerations in design and analysis. *Pharmaceutical Statistics*. 2020;1–13.
- **Rozenkranz G.,** *Biom. J.* 2016. Exploratory subgroup analysis in clinical trials by model selection.
- **Ruberg SJ.** Assessing and communicating heterogeneity of treatment effects (HTE) for patient subpopulations: the hardest problem there is. *Pharm Stat* 2020.
- **Royston P., Altman D., Sauerbrei W.** Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. in Med.*, 2006, 25 · March 2015.
- **Russek-Cohen, E.** Comments from the FDA Working Group on Subgroup Analyses. Presentation 2014.
- **Sechidis K., Papangelou K, Metcalfe P., Svensson D., Weatherall J., Brown G.** Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19), 2018, 3365–3376
- **Wellek, S.** (1997). Testing for absence of qualitative interactions between risk factors and treatment effects. *Biometrical Journal* 39(7):809–821



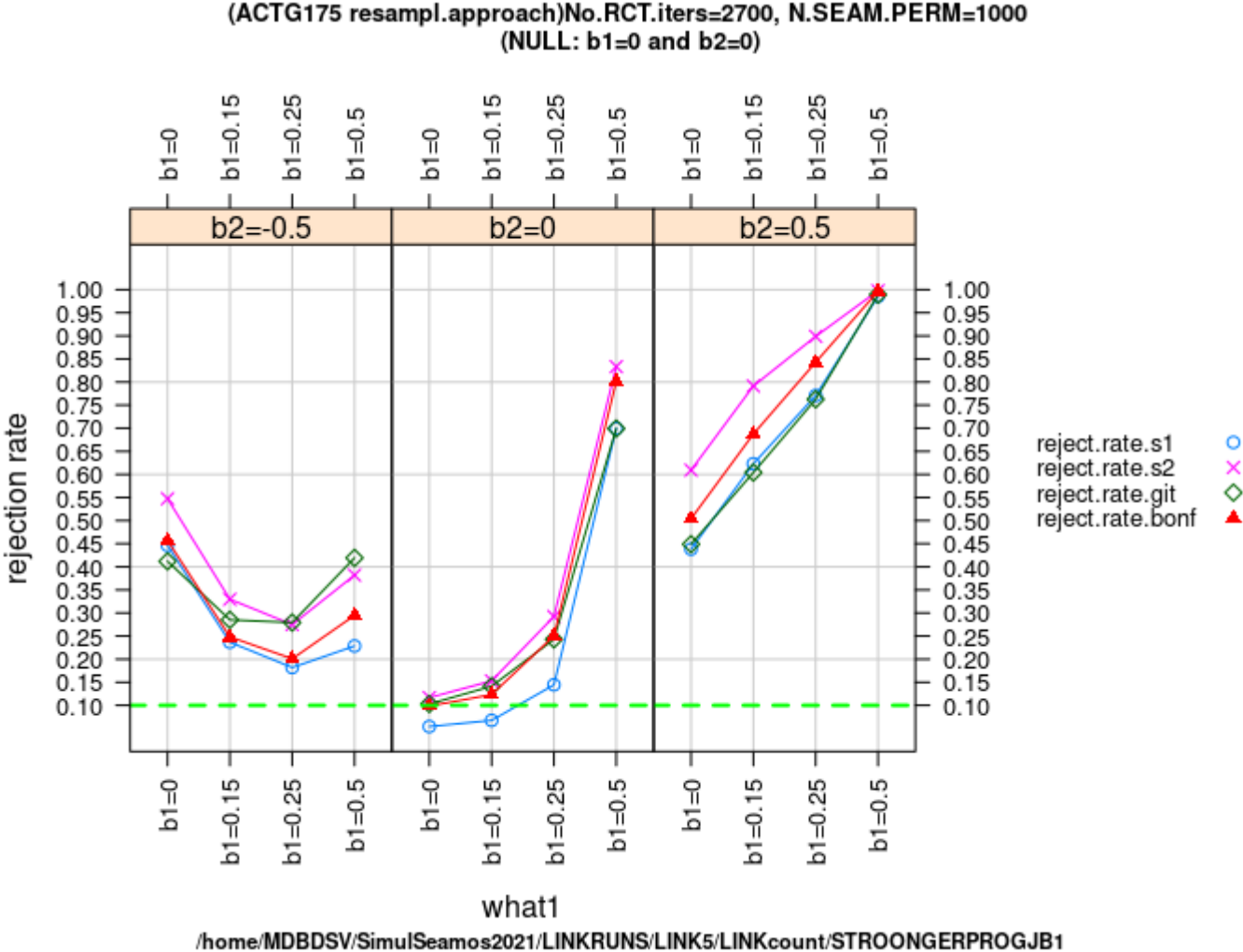
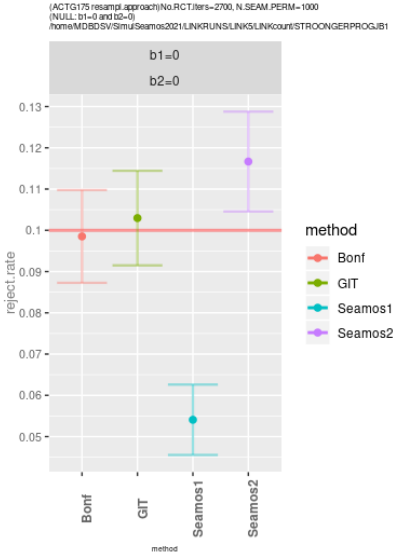
BACKUP



Some further simulation results

Again simulation via resampling and relying on the link expression below (red), but this time with stronger prognostic effects (all main effects were 0.25).

Examplifies that for some settings, the upgraded SEAMOS ("s2") did perform well?



$$\log(\lambda) = w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT$$



Some further simulation results (2)

Again simulation via resampling and relying on the link expression below (red), but this time with stronger prognostic effects (all main effects were 0.25).

Exemplifies that for some settings, the upgraded SEAMOS (with main effects & nullifying the interaction terms) did perform well.

$$\log(\lambda) = w + b_1 \cdot x_4 \cdot TRT + b_2 \cdot x_7 \cdot TRT$$


(ACTG175 resampl.approach)No.RCT.iters=2700, N.SEAM.PERM=1000
(NULL: b1=0 and b2=0)

/home/MDBDSV/SimulSeamos2021/LINKRUNS/LINK5/LINKcount/STROONGERPROGJB1



Pharmaceutical Statistics. 2019;18:140–144.

Commentary on: Subgroup analysis and interpretation for phase 3 confirmatory trials: White Paper of the EFSPi/PSI working group on subgroup analysis by Dane, Spencer, Rosenkranz, Lipkovich, and Parke

Robert Hemmings¹ | Armin Koch² 

¹Medicines and Healthcare Products Regulatory Agency, London, UK

²Institut für Biometrie, Medizinische Hochschule Hannover, Hannover, Germany

Correspondence

Armin Koch, Institut für Biometrie, University of Hannover, Hannover, Germany.

Email: koch.armin@mh-hannover.de

For this reason, we tend to disagree with the proposed approach in the White Paper that in the field of exploratory analyses control of type-1-error needs to be exercised through statistical methods. Rather, improved methods might better support signal generation and intelligent assessment. Arguably, power should be prioritised over Type I error where the objective is to generate signals for further inspection. Whilst sponsors might fear that this will lead to regulators over-interpreting results from one of multiple subgroup analyses, the guideline outlines that any signal will be assessed for its credibility considering whether it is replicated in other relevant data sources or has biological plausibility. These careful considerations mitigate the risk for regulatory action based on the trial data alone. It is important to remember



Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com

