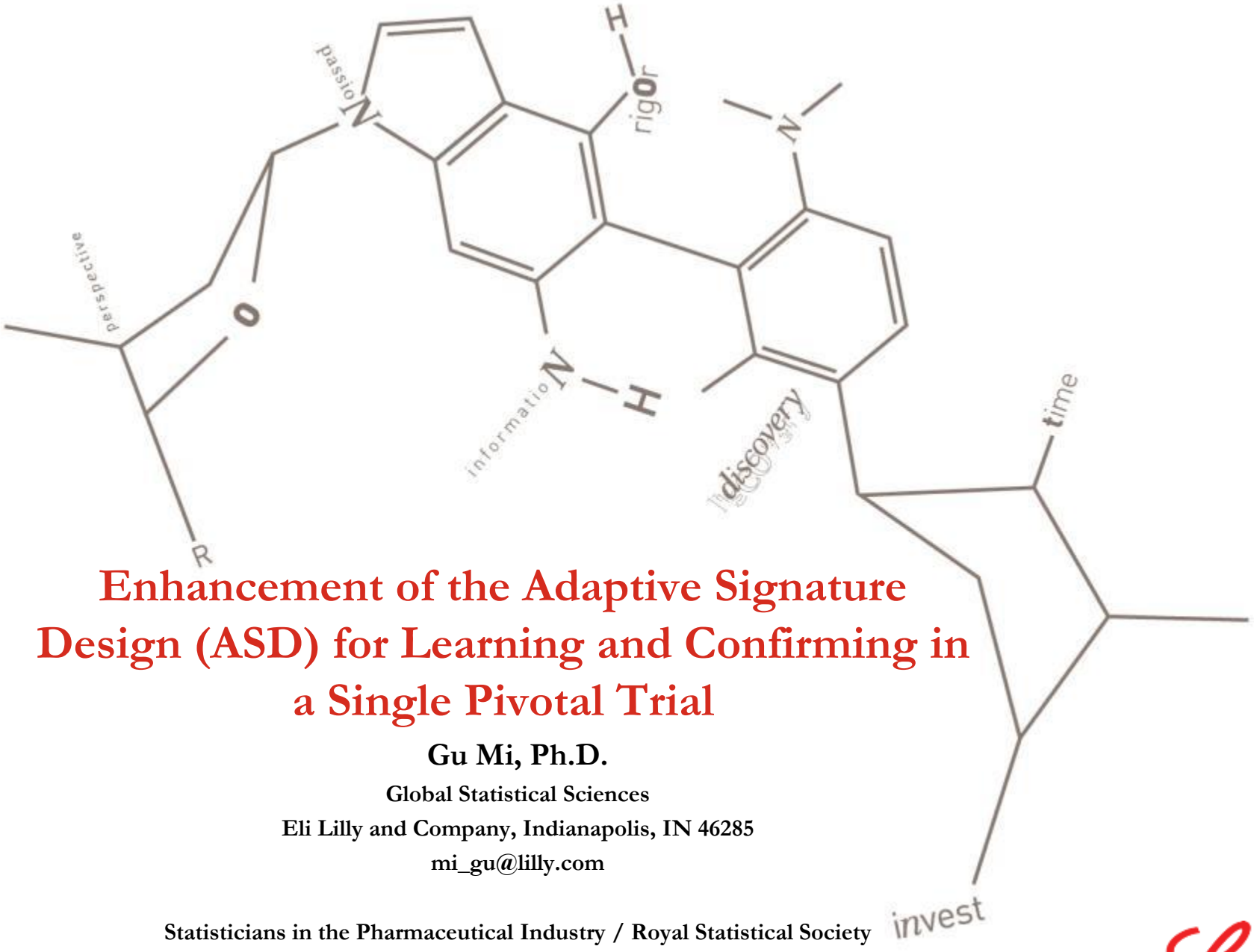


Enhancement of the Adaptive Signature Design (ASD) for Learning and Confirming in a Single Pivotal Trial

Gu Mi, Ph.D.

Global Statistical Sciences

Eli Lilly and Company, Indianapolis, IN 46285

mi_gu@lilly.com

Statisticians in the Pharmaceutical Industry / Royal Statistical Society Quarterly Journal Club

July 12, 2017

Outline

- Background
- Introduction of Adaptive Signature Design (ASD)
- Objectives
- Simulation Results
- Discussions
- Future Works

Received: 22 August 2016 | Revised: 8 March 2017 | Accepted: 3 April 2017

DOI: 10.1002/pst.1811

MAIN PAPER

WILEY

Enhancement of the adaptive signature design for learning and confirming in a single pivotal trial

Gu Mi

Eli Lilly and Company, Indianapolis, IN, USA

Correspondence
Gu Mi, Department of Global Statistical Science, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA.
Email: mi_gu@lilly.com

Because of the complexity of cancer biology, often the target pathway is not well understood at the time that phase III trials are initiated. A 2-stage trial design was previously proposed for identifying a subgroup of interest in a learn stage, on the basis of 1 or more baseline biomarkers, and then subsequently confirming it in a confirmation stage. In this article, we discuss some practical aspects of this type of design and describe an enhancement to this approach that can be built into the study randomization to increase the robustness of the evaluation. Furthermore, we show via simulation studies how the proportion of patients allocated to the learn stage versus the confirm stage impacts the power and provide recommendations.

KEYWORDS
adaptive design, biomarker, clinical trial, power evaluation

http://onlinelibrary.wiley.com/doi/10.1002/pst.1811/full

Background

- **Identify biomarkers for targeted population** in oncology trials
  - Target pathway is not well understood
  - Limited clinical data available from phase 2 studies
  - A predictive biomarker to identify sensitive subjects (M+) often unavailable at planning stage of a phase 3 study
  - Exploratory "*post hoc*" analyses not supporting registration: a new study to confirm
- **Adaptive Signature Design (ASD)**
  - Prospectively identify M+ group and test overall trt effect in a single trial
  - Learn stage: develop a "classifier" on *pre-specified* partition of overall pop.
  - Confirm stage: confirm classifier does indeed identify M+; test for overall trt effect
- **Practical relevance** of this type of design
  - KEYNOTE-001: evaluated efficacy and safety for PD-1 inhibition w/ pembrolizumab in NSCLC pts
  - Sponsor sought to define and validate an expression level of PD-1 ligand 1 (PD-L1) associated with the likelihood of clinical benefit
  - 495 pts receiving pembrolizumab assigned to either a learn set (182 pts) or a confirm set (313 pts)

Adaptive Signature Design (ASD)

Friedlin and Simon, *Clinical Cancer Research*, 2005

- Three components in two stages
  - Learn stage: develop a "classifier" on a *pre-specified* sub-population (learn set)
  - Confirm stage: two tests on all-comers and M+ (confirm set)
- Final analysis consists of comparisons of treatment arms
  - (1) in all-comers (both learn & confirm sets) at a significance level α_1
  - (2) in M+ pts in the confirm set at significance level α_2

  Study considered "positive" if either of the two tests is positive
- Original proposed allocation of patients
  - Equal allocation of learn/confirm sets (1:1) by the authors
  - Rule of thumb: 2/3 into training and 1/3 into test set to minimize MSE of prediction
  - Key constraint in the two-stage design: # pts not used in learn stage needs to be large enough for testing trt effect in confirm stage to be statistically significant
- Split of α b/t all-comers testing and M+ subgroup testing
  - Recommended $\alpha_1 = 0.04$ (80% of α) for all-comers, and $\alpha_2 = 0.01$ (20% of α) for M+

Objectives

- **Extensive "realistic" simulation studies for investigating**
  - Optimal allocation b/t learn and confirm sets
  - Alpha allocation for all-comers and M+ subgroup tests
  - Number of candidate biomarkers considered in the learn stage
- **Practical considerations**
  - Allocation of patients – two-stage randomization
  - Guidance/tools on performing similar simulations prior to implementation of any phase III study
  - By such simulations, we aim to **maximize the overall test power across plausible alternatives**

Data Generation Specifications

Virtual Data Generation Parameter Specifications		
Parameters	**Possible Values or Levels**	**Comments**
total sample size	700; 1400; 2100	randomization ratio 1:1
number of biomarkers	3; 10	only x_1 is the true biomarker
predictive effect ("scenario")	*moderate*	HR in M+ = 0.71 HR in M- = 1.00
	strong	HR in M+ = 0.60 HR in M- = 1.11
	strongest	HR in M+ = 0.54 HR in M- = 1.20

- Across all scenarios:
  - M+ and M- defined by a step function with cutoff 0.40 (i.e., M+ if $x_1 \leq 0.40$, and M- if $x_1 > 0.40$)
  - Hazard ratios (HRs) for M+ and M- populations give rise to an overall (mixed population) HR of 0.87
- In each scenario: 100 datasets simulated, with piecewise exponential as baseline survival function
- About 70% subjects experienced an event

Test Implementation Specifications

Test Implementation Specifications		
Parameters	**Possible Values or Levels**	**Comments**
learn/confirm allocation (%)	30/70; 40/60; 50/50; 60/40; 70/30	
α allocation (all-comer/M+)	0.025/0.025; 0.03/0.02; 0.035/0.015; 0.04/0.01	allocation at confirm stage
biomarker cutoff values	0.25 quantile, median, 0.75 quantile of x_i	=1 if $x_i <$ cutoff; 0 otherwise

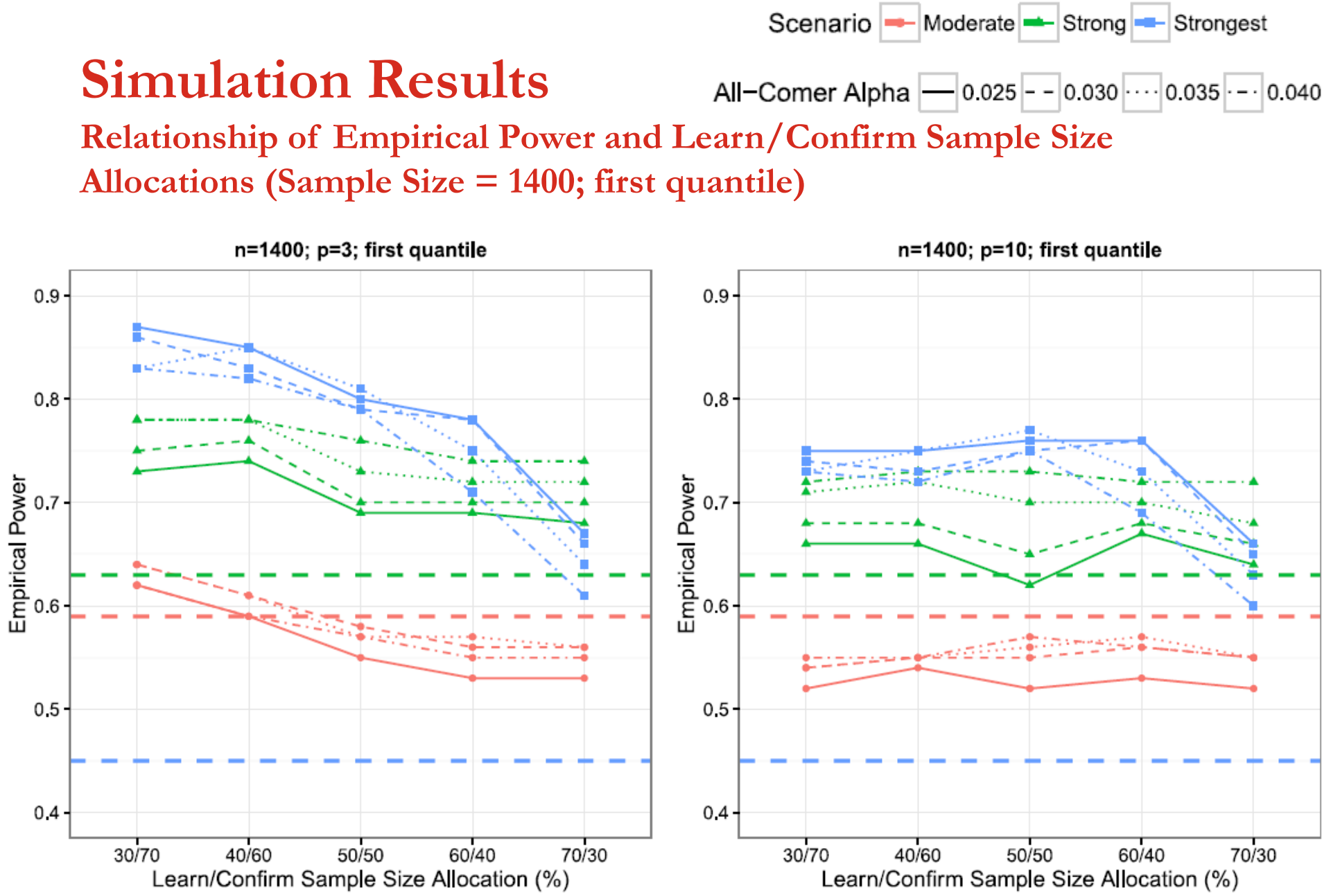
- Consider five learn/confirm allocations and four α allocations
- For simplicity: three fixed quantiles (0.25, median, and 0.75) for the biomarker cutoffs
- Cox PH model fit (single-marker analysis using x_i only):

$$h(t, \boldsymbol{X}) = h_0(t) \cdot \exp(trt + x_i + trt \cdot x_i)$$

  - Set α = 0.05 without multiplicity adjustment (exploratory stage)
  - If significant interaction: x_i is selected as a potential predictive marker
  - In case multiple markers have sig. interactions, the one w/ the most sig. interaction effect is chosen
  - To ensure one and only one marker is selected, choose the marker w/ the smallest interaction *p*-value in case of non-sig. interaction
  - At the end of learn stage: one single marker identified with a cutoff value is guaranteed

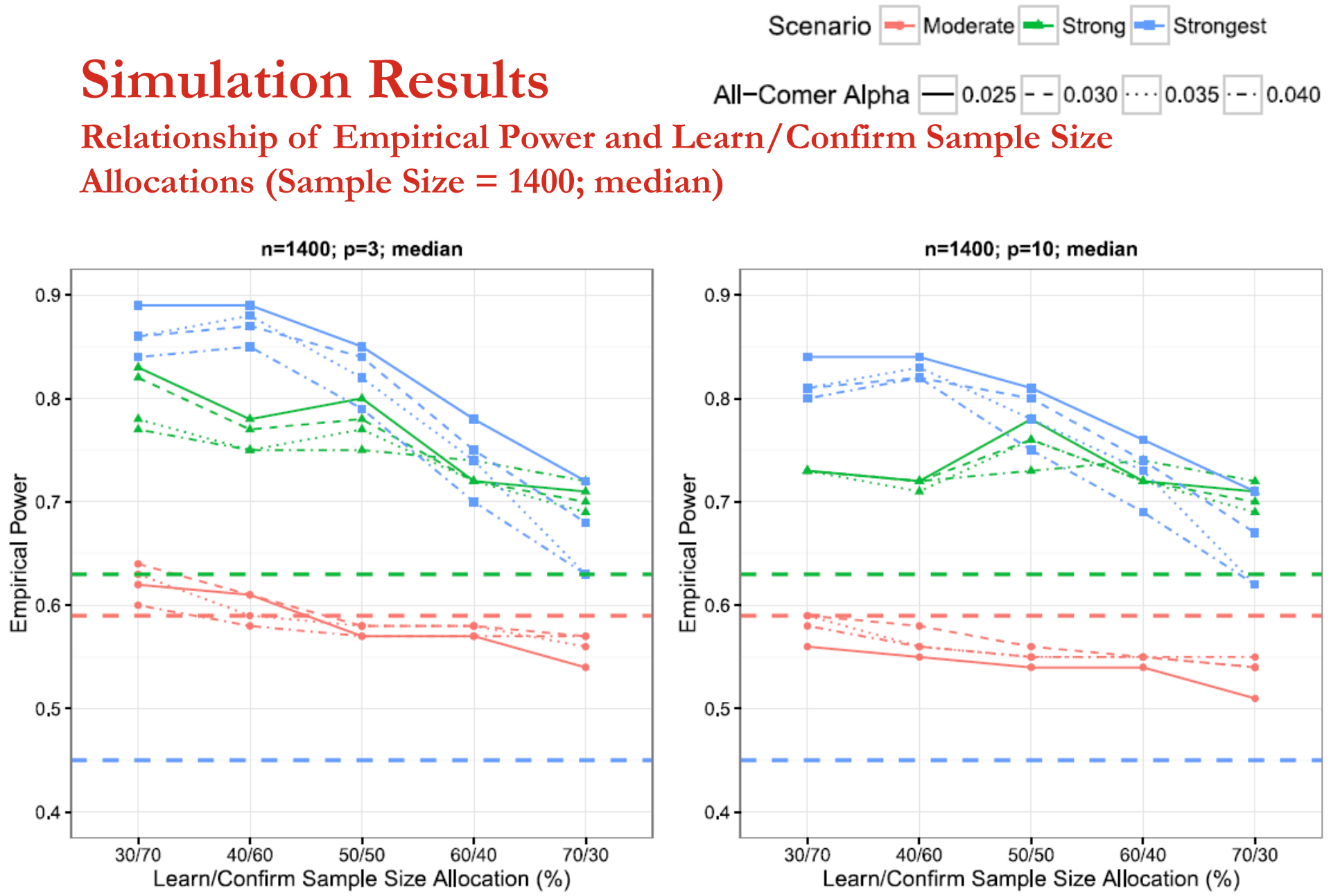
Simulation Results

Relationship of Empirical Power and Learn/Confirm Sample Size Allocations (Sample Size = 1400; first quantile)



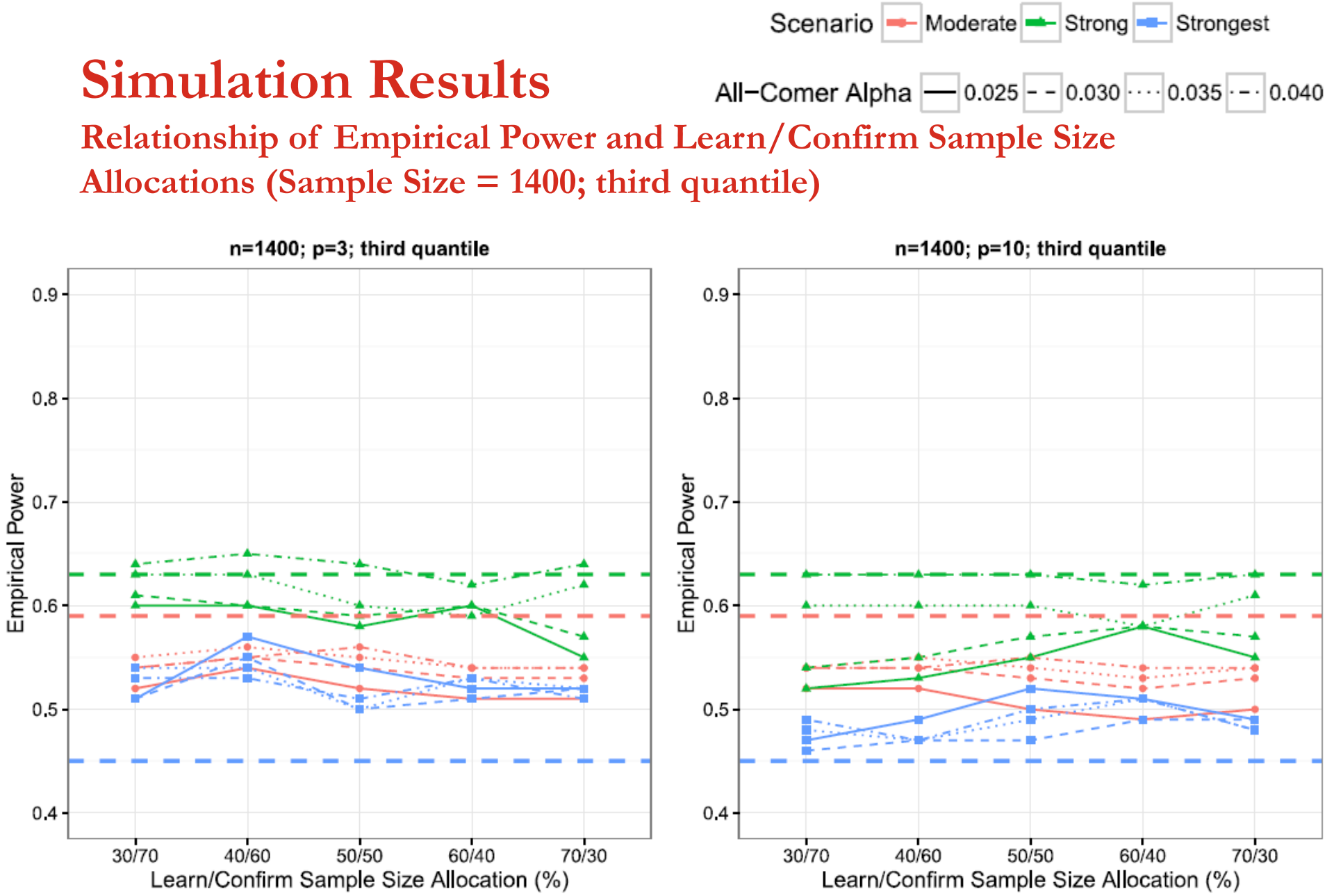
Simulation Results

Relationship of Empirical Power and Learn/Confirm Sample Size Allocations (Sample Size = 1400; median)



Simulation Results

Relationship of Empirical Power and Learn/Confirm Sample Size Allocations (Sample Size = 1400; third quantile)



Impact of Learn/Confirm Patient Allocation

- As the allocation changes from 30/70 to 70/30 (%)
  - A general decreasing trend of the empirical power when biomarker cutoff values are closer to the truth of 0.40 (i.e., first quantile and median), and when # biomarkers is small (3)
  - This trend is generally consistent across different α splits
- With an increased # biomarkers (from 3 to 10)
  - The same decreasing trend holds but the powers are reduced
  - This pattern becomes blurred when biomarker cutoff value is far from the truth of 0.40 (i.e., third quantile, 0.75)
- A learn/confirm allocation of 30/70 or 40/60 (%) offers the greatest power advantage irrespective of predictive effect strength
  - With a sample size of 1400 in a large phase III oncology trial
  - A relatively accurate estimate of the biomarker cutoff
  - A restricted number of biomarkers

Impact of Different Splits of α

- With sample size of 700 or 1400
  - Highest power is generally observed under "even split" of α (i.e., 0.025/0.025) in strongest scenario
  - "Even split" of α often results in lowest power in moderate/strong scenarios
- No evident pattern that a particular α split dominates across all situations
  - In some cases, especially at the largest sample size, the power difference is negligible across four α allocations (e.g., n = 2100, median cutoff)
  - In other cases, the power difference is evident (e.g., n = 700, median cutoff)
  - In general, the difference becomes smaller as sample size increases

Impact of # biomarkers in learn stage

- Power decreases as # biomarkers increases (due to an increased chance of falsely identifying a non-predictive biomarker)
- Even with an increased # biomarkers
  - Power gain of two-staged design over one-stage design is evident when the predictive effect of the biomarker is strong or strongest and the cutoff utilized is not too far from the truth (i.e., first quartile or median)
  - This is illustrated by the two-stage design power curves lying consistently above the corresponding horizontal dashed lines that represent one-stage design powers

Comparison of two-stage and one-stage design

- For strongest predictive effect, two-stage design dominates one-stage design in almost all scenarios
- With appropriate biomarker cutoff and optimal values of other parameters, power increase in two stage design over one stage design is substantial
  - 0.59 vs. 0.21 when n = 700
  - 0.89 vs 0.45 when n = 1400
  - 0.98 vs. 0.68 when n = 2100
- Minimal power gain or even power loss in case of
  - Limited predictive effect
  - Some predictive effect, but poorly selected biomarker cutoff
- Power loss because of
  - All-comer test being done at a reduced significance level
  - Sensitive patients in confirm stage being incorrectly selected

Practical Considerations - Allocation of Patients

- Patients cannot be allocated in a way that introduces systematic differences between the learn and confirm sets: allocation must be random and not associated with any time trend, or geographical for instance
- Both learn/confirm sets internally balanced w.r.t. treatment arm, and important prognostic factors
- A two-stage randomization:
  - Patients initially randomized to a trt, stratified by key prognostic factors as per usual
  - A subsequent randomization then takes place, in which patients are allocated to the learn or confirm set, with randomization once again stratified by the same key prognostic factors and additionally by treatment
- The randomization scheme, along with the key aspects of the two-stage design, should be described prospectively in study protocol and/or SAP to ensure data integrity and avoid any ambiguity in implementation after database lock

Discussions

- Identifying M+ subgroup should never be a hurdle in a two-stage design
  - Preclinical experiments (*in vitro* and *in vivo*) with an exploratory phase of biomarker development
  - Early-phase clinical development (e.g., phase II proof of concept)
  - Existing results from the same indication targeting the same pathway (e.g., FLEX, EGFR IHC H-score of 200, pre-specified in SQUIRE study)
  - Data-driven methods such as fitting GLM or machine learning techniques
- ASD has an important practical advantage
  - At end of learn stage sponsor can take practical steps: assay refinement, regulatory meetings, or identification of a partner to develop a diagnostic kit
  - This would enable the sponsor to develop a fully-validated assay, and test final samples using a market-ready version of the diagnostic assay, with an analysis plan that has gained regulatory approval
- An R package "simASD" is available at https://github.com/gu-mi/simASD

Future Works

- Impact of multiplicity control at learn stage (e.g., strong vs. weak control)
- Leveraging advanced strategies, such as graphical testing to allow α-propagation in a more sophisticated manner
- Choice of biomarker/subgroup identification methods at learn stage (examples):
  - Novel recursive partitioning procedure: SIDES
  - Tree-based method: GUIDE
  - With key considerations of multiplicity adjusted p-values and bias-corrected estimates of effect sizes
- Correlations among biomarkers
- Multiple markers allowed to enter into the confirm stage and the associated multiplicity issues (we only allowed for a single marker)

Acknowledgements

- **Jonathan Denne** and **Hollins Showalter** for critical reviews of the manuscript, and **Adarsh Joshi** and **Eric Nantz** for helpful discussions.

- Two anonymous reviewers and **Dr. Thomas Permutt**, Editor-in-Chief of *Pharmaceutical Statistics*, for their insightful suggestions.

Key References

- Mi G. Enhancement of the Adaptive Signature Design for Learning and Confirming in a Single Pivotal Trial. *Pharmaceutical Statistics* 2017 (http://onlinelibrary.wiley.com/doi/10.1002/pst.1811/full).
- Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 2005; 11:7872-7878.
- Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non–small-cell lung cancer. *New England Journal of Medicine* 2015; 372:2018-2028.
- Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 2011; 53:894-913.
- Lipkovich I, Dmitrienko A, Denne J, Enas. G. Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine* 2011; 30:2601-2621.
- Loh W, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine* 2015; 34:1818-1833.
- Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clinical Cancer Research* 2010; 16:691-698.

Thank You!

Questions and Comments

Back-up Slides

Simulation Results (Figures)

- **Figure 1 (sample size = 700)**

- **Figure 3 (sample size = 2100)**

(Please refer to the paper online for the two figures:
http://onlinelibrary.wiley.com/doi/10.1002/pst.1811/full)