

# 14th Annual University of Pennsylvania Conference on statistical issues in clinical trials/subgroup analysis in clinical trials: Opportunities and challenges (afternoon panel discussion)

Clinical Trials

1–11

© The Author(s) 2022

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/17407745231169681

[journals.sagepub.com/home/ctj](https://journals.sagepub.com/home/ctj)

Kosuke Imai<sup>1</sup> , Michael Rosenblum<sup>2</sup> and Mark Rothmann<sup>3</sup>

**Pam Shaw:** It is my great pleasure to introduce our discussant panelists. I'll start with Dr. Mark Rothmann, who is the Director of the US Food and Drug Administration (FDA) Division of Biometrics 2 within the Center for Drug Evaluation and Research, which reviews cardiology, nephrology, diabetes, lipids, obesity, and general endocrinology products. He has been with the FDA since 1999, I believe, and before that, he spent several years in academia. He's done a lot of work in the areas of subgroup analysis. We heard the morning speakers citing some of that work, including the Drug Trials Snapshots Program of the US Food and Drug Administration, heterogeneous treatment effects, and diversity in precise medicine.

Dr. Kosuke Imai is a professor in the Departments of Government and Statistics at Harvard University. He's also an affiliate in the Institute for Quantitative and Social Science. Before coming to Harvard in 2018, Dr. Kosuke taught at Princeton for 15 years where he was the founding director of the Program in Statistics and Machine Learning. He's also a visiting professor in the Faculty of Law and Graduate School of Law and Politics, University of Tokyo.

Our third panelist is Dr. Michael Rosenblum, Professor of Biostatistics at Johns Hopkins Bloomberg School of Public Health. His research focuses on improving the design and analysis of randomized trials. He's developed robust methods for improving precision and power by adjusting for prognostic baseline variables. He's also worked in adaptive trial designs with a focus on adaptive enrichment.

**Mark Rothmann:** Good afternoon, everyone, and thank you for inviting me to speak at this conference. Heterogeneous treatment effects have been an interest of mine for a while. I have written on it and have helped organize a workshop and symposium at the FDA, which I'll mention later.

Topics discussed in this meeting included finding the subgroup in which there's meaningful benefit (or any benefit), or that subgroup in which you would have the greatest power to do a clinical trial. We heard that subgroup analysis using statistical learning is not really a multiple testing problem, and I agree with that. And certainly, we do want to figure out which patients benefit, and we want to be able to tell patients what the effects and risks may be for them.

I think the FDA is recognizing more and more that treatment effects are often heterogeneous treatment effects. For a given patient, we don't observe their specific treatment effect, as we'd have to know what the difference in the outcomes would be if they were randomized to the experimental group versus the control group. But we only observed one of these. Baseline attributes, disease severity, condition, and genetics may affect the size of the treatment effect or treatment difference.

What are we actually testing versus what do we wish to test? From what we write mathematically, it seems we're testing for a common treatment effect. But we are truly testing for a positive average treatment effect. And it is more than preserving alpha when drawing conclusions. We do need to pay attention to what is actually demonstrated. And we should do better at informing patients of benefit risk that apply to patients like them and characterize the treatment effect with appropriate uncertainty that they (individually) may expect.

<sup>1</sup>Harvard University, Cambridge, MA, USA

<sup>2</sup>Johns Hopkins University, Baltimore, MD, USA

<sup>3</sup>US Food and Drug Administration, Silver Spring, MD, USA

## Corresponding author:

Kosuke Imai, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA.

Email: [Imai@Harvard.Edu](mailto:Imai@Harvard.Edu)

The traditional way of looking at clinical trials when I first joined the FDA 23 years ago was as follows: You needed to enroll a fairly homogeneous patient population in the clinical trial to minimize the variability in outcomes; and use unadjusted analysis, as there were not going to be any imbalances from the randomization. We also assumed that treatment effects don't vary unless proven otherwise. The overall estimated treatment effect applies to everyone or applies at least to everyone who could have been in the clinical trial.

While I still run into that traditional thought process, more and more there's a recognition that we should be doing clinical trials with fairly diverse patient populations, and follow everyone to the endpoints. Prognostic factors should be accounted for in the analysis to achieve variance reduction, which leads to more precise estimates. It's not about adjusting for imbalances. We should evaluate for heterogeneous treatment effects and provide the best information on treatment effects based on single factors, multi-factors, and accounting for correlation or confounding between factors.

When I became the statistical team leader for Metabolism and Endocrinology products, the first thing I did was ask the clinical leadership about factors that affected the treatment effect size on hemoglobin A1C (HbA1C) change. What appeared most influential was baseline HbA1c. We would have clinical trials in type 2 diabetes where baseline HbA1c varied quite a bit, providing greater power to show an interaction between treatment effect and baseline HbA1c. We would also have clinical trials in which baseline HbA1c didn't vary that much. The randomized trials in type 2 diabetes were fairly large. In fact, they tend to be oversized by a factor of 2 to 4, when you consider what you ultimately get for the width of the 95% confidence interval for the treatment effect versus what you needed for that width to have the desired power. Usually, we would get a "statistically significantly"  $p$ -value ( $<0.05$ ) when testing for an interaction effect between treatment and baseline HbA1c. Lack of variability in baseline HbA1c was sometimes the apparent reason for not achieving a  $p$ -value  $<0.05$ . It became clear to me that it may be fruitless to continuously test for that interaction. Instead, what's important is to characterize the effect modification of baseline HbA1c on 6 months HbA1c change.

Whether a product works in subgroup A is a different question from whether it works in subgroup B. But they are likely related questions that need to be considered. At the FDA Advisory Committee meeting, we had in December 2020 for cardiovascular disease, committee members discussed whether the product could first be shown to work in patients with low left ventricular ejection fraction (LVEF), and then see if the target population can be broadened.

We do have one product approved with a label that does give the estimated treatment effect for cardiovascular risk reduction and the confidence limits according to LVEF. That's important because the treatment effect does vary by baseline LVEF. Not everyone accepts LVEF as an effect modifier but many do. We do notice situations where the estimated treatment effect does vary a lot, and in a monotone fashion, by LVEF.

We need to be careful about multiplicity in interpreting subgroup results where an overall average treatment effect is demonstrated to be positive and then the population is partitioned into two subgroups. If we test each subgroup individually and don't demonstrate a positive treatment effect for either subgroup, we know we either made a type 1 error (for the overall population) or a type 2 error (for one or both subgroups).

Part of the problem is that when we evaluated the subgroups, we didn't make use of the fact that we had a statistically significant result for the overall average treatment effect being positive. We start the subgroup analysis by assuming no treatment effects in both subgroups.

If an overall average positive treatment effect is demonstrated and in every subgroup of interest, the 95% confidence interval rules out no treatment effect in a favorable direction, we are confident the product works in all those subgroups. We don't have to worry about what type of multiplicity adjustment needs to be made.

It's important to remember that a formal claim listed in the product label is different from treatment decisions made by individual patients. We may need more evidence that one therapy is better than another therapy for a claim than we do for an individual patient to make a treatment decision between the therapies.

Shrinkage estimation may be used in estimating subgroup treatment effects. When I discuss shrinkage estimation for a given application with my medical colleagues, the first question I ask is whether there are any known effect modifiers. We would want to put any effect modifiers and potential effect modifiers in the model. The remaining treatment effects would be modeled as exchangeable. The reader is directed to links to recent symposiums and Workshops co-sponsored by the FDA on Heterogeneous Treatment Effects.<sup>1,2</sup>

**Michael Rosenblum:** I first want to thank the organizers for the outstanding work on this workshop. I've learned a lot. And I also want to thank all the presenters and the other panelists. It's an honor to be included in this group.

I'll be talking about something that came up in Dr. Rothmann's talk, which addressed improving precision and power in randomized trials by leveraging baseline variables without making any additional assumptions. The connection to what has been presented so far is that you can do this for the overall population

treatment effect, that is, the average treatment effect. You can also do it for subpopulations as long as the subpopulations are large enough. If you only have a few people in the subpopulation, you can't do that. But if the populations are not too small, then you can use covariate adjustment for prognostic baseline variables to improve precision in estimating subgroup effects.

Covariate adjustment is underutilized; the improvements you can get from it can be quite substantial. Dr. Rothmann mentioned an example involving HbA1c, where you can get a large improvement in precision simply by using a covariate-adjusted analysis in the primary analysis. I've seen that as well across multiple disease areas.

The gain in precision depends on the population, the disease you're looking at, and the outcome of interest. What I typically see is between 5% and 25% reduction in the sample size to achieve a desired power just by using covariate adjustment rather than the unadjusted estimator. There have been some exciting recent developments in covariate adjustment.

I will focus on covariate adjustment where the goal is to estimate the marginal treatment effect, also called the average treatment effect, in a randomized trial. It is also possible to focus on conditional treatment effects as presented by Harrell et al.<sup>3</sup> but I won't discuss that here.

The covariate-adjusted analysis has to be preplanned in the primary efficacy analysis. Also you can use it equally well for subgroups, as long as they're not too small.

There's a common misconception about the goal of covariate adjustment. The misconception is that you're estimating something different from the unadjusted estimator (which ignores baseline variables), that is, that the estimand (the target of inference) is changing, but that's not correct. To illustrate this, consider the case where the estimand is the difference between means, the difference in proportions, or the restricted mean survival time for time-to-event endpoints. The target of estimation is exactly the same as when you use an unadjusted estimator. But the idea is you can get a better estimator than the unadjusted estimator when baseline variables are correlated with the outcome. And that can lead to improved precision and therefore a reduction in sample size.

The FDA just issued a draft guidance on covariate adjustment for randomized trials for drugs and biologics<sup>4</sup> that was updated last spring, and it is very well written. A quote from this guidance: "After suitably addressing the treatment effect definition, covariate adjustment using linear or nonlinear models can be used to increase precision." I love it!

You can find some resources about covariate adjustment on my website,<sup>5</sup> including a video recording of a training I gave on covariate adjustment for binary, ordinal, and time-to-event endpoints (where covariate adjustment is highly underutilized). There is also a

paper, led by Wang et al.<sup>6</sup> who derived a way to combine covariate adjustment with stratified randomization and get the benefits of both, by using a new method that is robust to model misspecification. I highly recommend his paper.

My current post-doc, Dr. Kelly Van Lanker, derived a way to incorporate covariate adjustment into group sequential designs with information adaptive monitoring so that the trial automatically adapts to how prognostic the baseline variables are. This approach addresses the problem that Dr. Rothmann brought up earlier: if you don't correctly account for how prognostic the baseline variables are, you might end up with a sample size that is too large or too small. Van Lanker et al.<sup>7</sup> present an adaptive method to get the correct sample size that imposes no penalty for doing the adjustment, asymptotically.

**Kosuke Imai:** It's been very interesting listening to these talks as a statistician working in the social sciences. The issues described below are also encountered in the clinical trials we implement.

I wanted to talk a little bit about statistical inference for subgroups that are discovered using machine learning algorithms. And before I get to that, I just wanted to review what's been talked about this afternoon (see Table 1).

Dr. Simon talked about adaptive experimental design, which is a very clever way of identifying a subgroup with a positive average effect by pre-specifying some strata and then dropping those with very little promise, trying to focus on the subgroups that have a higher chance of exhibiting a positive treatment effect.

Dr. Ivanova talked about multi-period cross-over trials and the inference that can be done based on the cross-validation and bootstrap. The goal was to identify the subgroup that maximizes the combination of the average treatment effect within that subgroup, but also the subgroup prevalence, so trying to identify a larger group in the population that has a large treatment effect.

Dr. Lipkovich discussed the estimation of conditional average treatment effects. This is closely related to the use of machine learning to estimate the conditional average treatment effect. That's a strategy to identify a subgroup with large conditional average treatment effect estimates. Using the modern machine learning algorithms, we might be able to do this more efficiently and effectively.

Finally, Dr. Schnell talked about non-exchangeable subgroups where the goal is to test the consistency or heterogeneity among subgroups. It's interesting to think about statistical tests of heterogeneity among subgroups and the challenges of multiple comparisons and how you make adjustments to p-values, and so on.

**Table 1.** Approaches to subgroup identification.

Approaches to subgroup identification		
1 Adaptive experimental design (Simon)	Goal: identify a subgroup with a positive average effect	Approach: pre-specify strata and then drop those with little promise
2 Multi-period crossover trial (Ivanova)	Goal: identify the subgroup that maximizes the product of the average treatment effect and prevalence	Approach: inference based on cross-validation and bootstrap
3 Estimation of the conditional average treatment effect (Lipkovich)	Goal: identify the subgroup that maximizes the product of the average treatment effect and prevalence	Approach: use machine learning to estimate the conditional average treatment effects (CATE) Identify a subgroup with large CATE estimates
4. Non-exchangeable subgroups (Schnell)	Goal: test consistency or heterogeneity among subgroups	Approach: challenges of multiple comparisons in subgroup analysis

All these talks raised one common theme: we need to think about how to conduct correct statistical inference when you use a subgroup identification based on an experiment and a machine learning algorithm.

I want to address subgroup identification when we use machine learning. What if we used a machine learning algorithm to identify subgroups? We can't assume that machine learning algorithms converge uniformly to the conditional treatment effects. Who knows, there are a lot of tuning parameters and it's often difficult to establish this type of theoretical property. Can we make proper statistical inference for discovered subgroups?

Once you discover these subgroups using machine learning methods, can we make statistical inferences? And how do we take account of the fact that these algorithms often are "black box" or just based on an ad hoc procedure. We also want to avoid a computationally intensive procedure for statistical inference because machine learning algorithms often tend to be very computationally demanding.

In joint work with Michael Li at the Massachusetts Institute of Technology, we address the issue of statistical inference when you use machine learning for subgroup identification. We focus on the conditional average treatment effect, which is the average treatment effect, conditional on some covariate values. Investigators can develop a scoring system which is used to sort the subgroups based on the conditional average treatment effect estimates and sort the values of  $X$  from most impacted by the treatment to least impacted by the treatment.<sup>8</sup>

The Sorted Group Average Treatment Effect takes the machine learning estimate of conditional average treatment effects and sorts them based on its estimate and divides them into a certain number of groups. You can imagine dividing them into four groups, ranging from the group that has the largest effect to the group with the smallest effect based on the machine learning estimate of conditional average treatment effect.

Now the question is, once we sort them, can we do statistical inference? It turns out the easiest thing to do

is, within each subgroup, to take a difference of means between the treatment and control groups. That gives you an unbiased estimate of the average treatment effect within that subgroup. What's interesting about this is that you can rewrite this subgroup estimate as an estimate under certain type of individualized treatment rules. You can think of this machine learning estimate to be like classifying each individual into one of the groups that you created.

And we show in our article that you can actually make statistical inferences based on Neyman's repeated sampling framework. What's nice about this is that you're only basing your inference on random assignment treatment and random sampling of the units, and nothing else. It's a design-based approach where, whatever the machine learning algorithms are, we can come up with the uncertainty estimates based on these design-based features.

We can also account for the random splits due to cross-fitting. Cross-fitting is often used when you're using a machine learning algorithm to do this type of analysis. You can get the standard errors and the confidence intervals essentially for each of these group-specific average treatment effect estimates when the subgroup is identified by the machine learning algorithm without making any assumption about the properties of the machine learning algorithm.

You can also conduct statistical hypothesis tests for subgroups. You can use the nonparametric test of treatment effect homogeneity. The null hypothesis here would be that all the subgroups have the same average treatment effect, and we can come up with a test statistic and derive their covariance matrix.

You can also do nonparametric test of rank consistency, which means that under the null the sizes of average treatment effect across subgroups are correctly ordered. We've derived this test statistic, again, based on just the random assignment of the treatment and random sampling of units—this is a weighted chi-square distribution. Shown in Figure 1 are some simulation studies using different estimators.

Simulation Study

Estimator	truth	$n_{\text{test}} = 100$		$n_{\text{test}} = 500$		$n_{\text{test}} = 2500$	
		bias	coverage	bias	coverage	bias	coverage
Causal Forest							
$\hat{\tau}_1$	2.164	0.034	93.8%	0.041	95.0%	0.007	96.0%
$\hat{\tau}_2$	4.001	0.011	93.7	−0.060	94.4	−0.002	95.3
$\hat{\tau}_3$	4.583	−0.018	94.0	−0.003	96.4	0.020	95.8
$\hat{\tau}_4$	4.931	−0.077	94.6	0.001	94.3	0.003	95.6
$\hat{\tau}_5$	5.728	−0.058	96.0	−0.010	95.0	−0.009	95.2
BART							
$\hat{\tau}_1$	2.092	0.016	94.0%	−0.014	96.2%	0.009	95.8%
$\hat{\tau}_2$	3.913	0.127	95.1	0.028	94.0	−0.003	95.3
$\hat{\tau}_3$	4.478	−0.077	94.3	−0.041	95.0	−0.001	95.1
$\hat{\tau}_4$	5.042	−0.154	94.2	0.014	95.8	0.015	95.4
$\hat{\tau}_5$	5.881	−0.019	94.7	−0.019	94.4	−0.000	95.0
LASSO							
$\hat{\tau}_1$	3.243	0.028	94.1%	0.049	95.1%	0.003	95.1%
$\hat{\tau}_2$	3.817	−0.012	93.6	−0.013	94.5	−0.000	95.4
$\hat{\tau}_3$	4.318	−0.013	94.2	−0.002	94.5	0.010	95.0
$\hat{\tau}_4$	4.788	−0.041	94.0	−0.015	94.6	−0.001	94.6
$\hat{\tau}_5$	5.241	−0.046	94.4	0.021	95.1	0.002	95.3

**Figure 1.** Simulation study.

Here I used causal forest, Bayesian additive regression trees (BART), and Least Absolute Shrinkage and Selection Operator (LASSO). We have five groups sorted by the conditional average treatment effect estimated by each machine learning algorithm. You can see that even when the sample size is only 100, bias is very small, coverage is reasonable. And then as you increase the sample size, the coverage converges to 95%. So even as small as 100 or 500 observations, these Neyman-based confidence intervals do a very good job of estimating uncertainty.

To wrap up, statistical inference for subgroups is challenging, especially when they are discovered by complex machine learning algorithms that we nowadays use. We show that no modeling assumption is required to estimate the uncertainty of these type of subgroup analyses. Any machine learning algorithm can be used. In fact, it doesn't have to be machine learning. It can be some arbitrary rules stipulated by a human. It's completely design-based, so you rely only on random sampling assignments, random sampling, and random splits to quantify the uncertainty, and it is applicable to the cross-fitting estimators that are very popular these days.

We show by simulation that there's good small sample performance. As an extension, we're also looking at dynamic treatment regime settings, which some of the speakers talked about, such as cross-over designs. We have a forthcoming Journal of American Statistical Association paper on this topic, and a related paper currently in *arXiv*.<sup>9</sup> The proposed methodology is implemented through an open-source R package, evalITR, which is freely available for download at the

Comprehensive R Archive Network (CRAN; <https://CRAN.R-project.org/package=evalITR>), if you want to try it out.

**Pam Shaw:** This is a great tour of our four talks. I'd like to ask the afternoon speakers if they have any responses to the remarks made on their talks.

**Noah Simon:** I want to engage in a real discussion of using machine learning for subgroup identification, because I feel like half of my day job is machine learning and the other half is engaging with clinical trials and biomarker design.

I have found it is very hard to use these really cool machine learning methods to develop biomarker signatures that actually work in clinical practice that are either meaningful or transferable, even when you have lots of similar trials and similar disease areas. We have lots of statistical papers on them. They are very cool. I've engaged with them. I like to think about them. When we think about what's worked in oncology it's basically all single point mutations or expression of a single gene, or copy number variation. I think it would be great to have more discussion on that issue.

**Pam Shaw:** Dr. Lipkovich, you described some very specific methods and I think you had some questions. Following up on Noah's question, I think it was Dr. Mehrotra who tapped into one of the aspects of why this is difficult, and I wanted to ask about the need for pre-specification of the covariates and how that influences your methods. Do you recommend a filtering step to somehow remove noise covariates?

**Ilya Lipkovich:** I agree that we need pre-specification. What I was trying to argue is that what should be pre-specified is a strategy for machine learning, not the subgroups or anything specific. So of course, we need to pre-specify covariates. In terms of filtering, I would say there's filtering that doesn't look at outcomes and filtering that does. The first one changes operating characteristics, but not type 1 error, because you don't look at the outcome, so you can probably filter out covariates with a very low variability or a lot of missing data that you don't want to impute, or there's categorical data with very sparse cells.

But if you (as often people do) look at the outcomes, this should be part of the strategy. And when you evaluate the operating characteristics of the strategy, you need to somehow factor in these looks. As I mentioned, very often people devise complex strategies, but they account for uncertainty only in the very last stage of the strategy and completely ignore multiple looks at the data that were in fact part of the strategy.

**Pam Shaw:** We've heard about some of the difficulties of applying these machine learning methods. Dr. Simon, you're referring to kind of the bias tradeoff here. It's either going to be impossible to power, or you're going to have some bias based on some unaddressed part of your assumptions.

**Noah Simon:** And just generalizability in general. You can fit something that seems great, and then somehow it just doesn't generalize to the next oncology trial. It's slightly different, of course, but seems like we should be learning underlying biology, say, about checkpoint inhibitors. And if the assay doesn't work when you move the assay to something very similar, that seems like a problem.

**Pam Shaw:** One of the most "liked" questions of the morning came in from Devan Mehrotra for Dr. Simon. Dr. Simon, with your aggressive enriched design, is the point estimate of the average treatment effect biased upward, that is, does it exaggerate the true treatment effect? And if so, can we fix this using bootstrapping or another approach? And what is the true treatment effect that you're getting at with these designs?

**Noah Simon:** It's a great question and it depends. There are a lot of different ways to get an estimate. You could use all of the data and at the end of the trial identify what subgroup you think most benefits from the intervention and then use a re-substitution estimator, and that's certainly going to be biased.

Or you can take the group moving into the second stage and use all of those people. The stage 2 piece of that estimate won't be biased. The stage 1 piece will add to bias, because you're selecting the most promising looking subgroup. So, yes, there is potential for

bias. But if you only use a subset of your data for estimation, you're going to get a lot of variability and you might get a point estimate that doesn't seem to agree with your statistical test.

One of our papers<sup>10</sup> considers a bootstrapping approach to try to account for over-optimism. One thing you have to be a little careful about here is the contrast between a conditional versus an unconditional estimate.

Often when we run a group sequential design, we don't want to condition on the stage where we stopped, because that just loses us power and efficiency. The bootstrap will tend to get us a good estimate of treatment effect, where if you do some sort of selective inference type procedure, you may have a lot of variability in your estimate. You may end up with a very wide confidence interval, especially if you're near a decision boundary.

In terms of what subpopulation, we are actually evaluating the treatment effect on, there are different ways to go. Generally, I would think you're evaluating the treatment effect on the subpopulation identified by that second stage of the trial. And there may be some variability and a little bit of bias in that. But frankly, when we transport from the clinical trial experience to the clinic experience, I think things are going to change anyway.

**Pam Shaw:** What are the implications for design and sample size, of some of these methods focusing on subgroups and subgroup analysis if any of these methods are going to be useful from the point of view of interpretable, adequately powered trials? And that looks to be an open question, as each of you have presented different methods for getting at subgroups. What were some of the things you consider relative to power? Are you all calculating power when using these methods or are you just treating them as exploratory findings?

**Patrick Schnell:** In my experience, when we've powered for a subgroup analysis, usually based on direct estimation of treatment effects, we assume something plausible about what the true treatment effect surface might be. I have several different scenarios, and actually just fall back on simulation using a pretty wide variety of circumstances to see just how well the trial would do.

Unfortunately, you really can't get away from making assumptions or estimating what the covariate distribution of your sample will be. So even if you're used to doing things like calculating the power for a key test and just saying, well, we're going to make some adjustments and expect the power to be a little bit better than that, things do not really work that way when you're doing subgroup analysis just because of how much the power will depend on the size of those subgroups,

whether you're doing analyses in pre-specified subgroups or something more data-driven for selecting subgroups. It's admittedly been sort of a headache, but I'm not really aware of any way around it.

**Ilya Lipkovich:** I agree it's important to use empirical data. It's easy to preserve distribution of covariates so you can take some reasonable realistic clinical trial and then add treatment effect to this data. Typically, people are over-optimistic about how much data they need, and the power is not typically as great as you expect. But I think it's important to have empirical simulations rather than simulations that are too dominated by some assumed distributions.

**Pam Shaw:** Dr. Lipkovich, just described simulations at the level of 1000. I'm thinking that some of these methods may only be relevant for fairly large trials. But a member of the audience has posted that its utility may vary by disease area, and in particular, cancer, which can have some small sample size issues. This is particularly difficult because it is also a complex disease with highly variable populations. How does that affect the success of some of these methods?

**Noah Simon:** I think we should continue trying to engage machine learning in a number of these scenarios, and I think it's important, especially when we think about extracting information from images and using that for treatment. Extracting information from images is very, very hard. But I guess I still struggle with any areas where we have a treatment where we're able to build a signature that's not based on a well-defined target.

Cystic fibrosis may be another area where it's complicated, but it's all largely in that cystic fibrosis gene, to my understanding, and new treatments target that. And there are biomarker signatures. But you have to ask, do you have a mutation that has this functional issue? I guess I'm really curious about this question. I would love to hear more thoughts on disease areas where this could work.

In my experience, the treatments that work well for a subset of people do it based on dysregulation that we already understand pretty well and it's a matter of fine-tuning based on a small number of features.

**Pam Shaw:** Dr. Simon, I would like to ask about your experiences using machine learning methods that haven't seemed to be working very well. Is it any different from any other exploratory analysis, as we heard from Dr. Fleming and others outlining multiple examples in the AM session, where many results of exploratory analyses are spurious? So, is that really the fault of machine learning or how it's being interpreted?

**Noah Simon:** That's a great point. And it could be with a really large sample, we could learn complex rules. What Dr. Fleming said, which I think is really appropriate, is we can't even learn simple rules well with the data we're engaging with, right. So it's really hard to learn complicated rules. Maybe with a huge data set we could.

When we measure things, who knows what we're really measuring biologically at the time. There are a zillion assays for measuring the same thing. There's a lot of noise. Maybe there's some latent feature that we're measuring a proxy of. And so I think there's the complication there where building a complicated rule based on the wrong features may get you in trouble because it's not generalizable in the way that a simple rule based on the slightly wrong features would be.

**Ilya Lipkovich:** Maybe there's too much emphasis about learning an exact signature that you really just believe. I would be inclined to think along the lines of what Dr. Imai presented. Maybe we should just try to solve the modest problem of rejecting the global null that there is no heterogeneity. Then, if there is heterogeneity, you can look at variable importance and do some kind of procedure to select some predictors that may be important; then a future trial can be designed to explore those potential predictive variables. It's unrealistic to imagine that you can right away learn some signature and be able to prove it. I think it's a gradual process; the interaction test with one variable at a time that people have used should be replaced by something more clever based on a global hypothesis about the presence of heterogeneity using machine learning. There's no conclusive evidence about how powerful these tests are, unfortunately. But I think that some research in this area is very promising.

**Pam Shaw:** I think those are some great points in terms of setting realistic goals for these tools.

**Michael Rosenblum:** I'm generally in agreement with what Dr. Simon said about the challenges of using machine learning, especially his most recent point that machine learning can be very powerful when you have enormous sample sizes, enormous training data sets, like Amazon or search engines that have billions (or more) of data points. But if the sample size is what we see in trials, it's in some ways an open question how useful machine learning can be, because of the relatively smaller sample sizes. Consider the problem of trying to select which set of baseline covariates are most prognostic for an outcome where the goal is improving precision for estimating the average treatment effect for the overall population. In some cases, machine learning can do better than simpler approaches, but there's often a tradeoff, that is, if you change to a scenario where the baseline variables are useless, they're all noise,

the methods that do the best in finding good predictors also get tricked by noise pretty easily. There is ongoing work on how to optimize this tradeoff, for example, by Williams et al.<sup>11</sup> at NYU, who is making good progress.

**Noah Simon:** But with regard to using machine learning for prognostic signatures and improved precision, which seems like a great idea, are you also helped a little bit because you actually don't need your findings to be generalizable to another study, where for a predictive signature, you actually need it to work with the next group as opposed to just adding precision in the current study?

**Michael Rosenblum:** I have to think about this question more to give a good answer, but I agree that covariate adjustment for estimating marginal effects is a fundamentally easier problem than trying to learn which subgroups benefit more or less from the treatment.

I think there is hope in the covariate adjustment case that you can use outside data, large databases of electronic health records that don't involve the treatment of interest in the trial, but that do have the same population and outcome definition, so it could essentially match what's in the control arm of a trial. There you might have a lot of information that you can leverage, while you wouldn't really have that necessary for the problem that's the focus of this conference of learning which subgroups benefit more or less from a treatment.

**Pam Shaw:** Since we are often in the position of looking for subgroups because someone has asked us to do so, do we have any bias robust practices in subgroup analyses where the primary motivation for including them is that the funder requires it? Specifically, considerations of race/ethnicity analyses in trials where there are many co-confounders with that specific variable that may not be captured in your clinical trial data. In your experience, if the goal is to at least include subgroup analyses, not necessarily based on questions of biological evidence, but rather some other reason, what is the best practice for perhaps doing it or interpreting it?

**Noah Simon:** I think this is both a really important question and a really hard question. We obviously care about equity, and we have a history of engaging in medicine in incredibly unequitable ways. Forcing us to think about these things as we engage with new research and new drugs is really fundamentally important.

Engaging with race in particular is hard because race is a social category that is complex to engage with and complex to even define and think about. There is a statistician/mathematician/philosopher, Hu et al.,<sup>12</sup> whose work I think is really incredible and foundational, and

I would encourage people to read it, though it's complicated, thinking about what counterfactuals for social categories mean.

While NIH and other funders are coming at this correctly and saying it's very important, I think there are potential confounders, there are all sorts of issues, and race and ethnicity don't really match very well with genetic ancestry, making categorization very error prone. Performing a really simple analysis where you just engage with a subgroup, test for a treatment effect, report that treatment effect and act like that has clear meaning, can get you in trouble. It can get you to a place where you find something that is very difficult to understand. So I don't know that that answered the question of how to do it, but I do think it is much more complicated than it looks on the surface.

**Pam Shaw:** I think many people probably agree with this. A member of the audience asks: Suppose you've gone through one of these subgroup analyses that was done in an exploratory manner, that hasn't perhaps been properly powered, or properly pre-specified. How will FDA reviewers respond to such a finding?

**Mark Rothmann:** We are currently revising our earlier FDA sex differences guidance. And certainly, we should always look at differences between males and females in studies.

There are certainly some issues. Dr. Fleming showed some pretty big differences in one study and then the follow-up study goes the other way or shows no difference. We have seen differences between males and females where it's been due to the females being older. We have seen differences where weight or Body mass index (BMI) seems to matter, in the distribution difference between two groups.

Investigators should look at and report statistically significant and clear interaction effect by sex. For example, males get a larger additional decrease in the percent of cholesterol than females. And there may be some actual reasons for that. But we don't know how that translates to cardiovascular risk. We just know what it is for low-density lipoprotein (LDL) cholesterol, and people probably care more about cardiovascular risk.

So how would an FDA reviewer respond to such finding? I think we would be interested in knowing whether this is something that happened before in this indication or for this product. I don't think we would necessarily say that it's real just by observing a difference, because we do know many subgroup analyses are performed, and there are going to be random highs and random lows.

**Pam Shaw:** The next audience question is: is there any rule of thumb to pre-specify the subgroup in which you are hoping to establish benefit?



**Patrick Schnell:** So the main consideration there for me is to make it as big as you can so that you're encompassing people that would actually be eligible for the treatment, but at the same time your counter-consideration to that is that the bigger you make it and the more covariate points that you include, the worse your power will be, just because you're increasing your multiplicity problem. You gain a lot by ruling out completely ludicrous points like 200-year-old patients.

You can choose the covariate space after you've been unblinded to treatment assignments, as long as you are not incorporating outcome information in the process. You can look at your covariate distribution, take something that encompasses the sample in your trial but doesn't include a lot of extra space beyond that. You can even do power calculations conditional on your observed covariate distribution and treatment assignments if you're willing to hypothesize treatment effects. You basically make your covariate space include points where you actually have reasonable power to detect effects. A paper in *Clinical Trials* in 2018 gives an example of how you can do this.<sup>13</sup>

**Pam Shaw:** Dr. Ivanova, it's interesting to see these methods in the context of an ongoing clinical trial and the thought processes for how you chose the design. It's really helpful to see what other statisticians struggled with and where they landed instead of the picture-perfect presentation in a journal paper.

My understanding is you had these biomarkers in the A positive group that you'll be doing the formal testing on, but you're enrolling a larger group of people in this trial. What do you feel that your trial will be able to say about those folks that aren't in the A positive formal testing group?

**Anastasia Ivanova:** Before the trial for each of the five treatment arms we specified a biomarker positive subgroup, a subgroup where we are hoping to see a treatment effect. Some of these subgroups were supported by existing data and some were based primarily on prognostic biomarkers. The next question was how to design the study, keeping in mind that we have biomarker-positive and negative subgroups identified for each treatment arm. The most efficient way is to investigate each intervention in its biomarker positive subgroup only. If the intervention has no activity in asthma in that subgroup, it probably does not work in asthma. After discussing with our funder, the National Heart, Lung and Blood Institute, we decided to enroll both biomarker-positive and negative participants to each intervention. The efficacy analysis for each intervention is performed in the corresponding biomarker-positive subgroup. The biomarker-negative subgroup is used to refine the biomarker cutoff and to identify

promising biomarker-defined subgroups in a post hoc machine learning analysis.

**Pam Shaw:** Do you think the complexity of your design will make it hard to summarize the findings?

**Anastasia Ivanova:** Our goal was to have an efficient design that allows answering as many scientific questions as possible regarding the five novel interventions with potential for severe asthma. As a result, the design is complex. It is a multi-period crossover adaptive trial with a precision-medicine component. In addition, we collect peak flow and survey data on each participant twice daily. This is done to detect deterioration events other than asthma exacerbations. Observing more events than just exacerbations allows reducing the follow-up time on each intervention from a year typically used in asthma to four months. And yes, the complexity will make it more challenging to summarize and interpret the results.

**Pam Shaw:** Dr. Simon, rather than dropping subgroups that may not be performing well, what about using minimization or setting a minimum percentage to lower a particular value and increase the percent of the subgroup benefiting from the intervention to improve generalizability and balance?

**Noah Simon:** I like that idea. I think it's quite nice, especially if you run this in multiple stages and you want to have the potential to enroll people who you think are maybe less likely to benefit. I guess there is a tradeoff between administrative attractiveness and optimality. As I engage more and more with clinical trials, I see the value of simplicity of many of these things more in terms of making sure patients are informed of what's going on when they enroll in the trial, making sure that, like the medical professionals, they have some idea of what the outcome actually means.

I like that idea as maybe a way of striking a balance between the two, rather than being super aggressive about excluding people. Maybe we should not be so aggressive and not run the risk of missing people. I do think it's a question of tradeoffs, and you're 100% right that that is something to be concerned about and in some cases may be more concerned than in others. There is the risk that you run a trial that's not successful because you don't have the resources to run a large enough trial to see that say 50% of people are not benefiting, and only this other 50% are. I think probably we need to make those decisions on a case by case basis weighing the fact that there are potential losses either way.

**Pam Shaw:** The tradeoff is between designing a more precise trial by enriching for those with a larger response rate versus wanting more generalizability.

**Patrick Schnell:** Even if you do an enrichment design or something adaptive and you identify a subgroup that you believe is benefiting from the treatment on average, when you look at the complement of that subgroup, you're not necessarily saying that there is no beneficial treatment effect there. It's usually just that we don't have enough evidence to conclude that there's a treatment effect there. So even if a group of people is excluded from the subgroup that you end up finalizing for the trial, it doesn't necessarily preclude further investigation for those people.

**Pam Shaw:** It's fascinating to hear from folks from other disciplines. Dr. Imai gave a lot of interesting food for thought from work in other areas. And all of you I think have provided tremendous perspective on this difficult topic. So thanks to everyone for all the great questions, and to all of the speakers and panelists.

**Mary Putt:** We've come to the end of our 14th annual conference and our second virtual conference. I have completely enjoyed hearing these multiple perspectives on subgroup analyses. It's a great pleasure, and also a little overwhelming, given all these great talks, to offer a few closing remarks.

We started out this morning with an overview of the issues and the problems, the importance, and some of the evolution of thinking that's gone into subgroup analyses. We started off thinking about the reasons that we do subgroup analyses. Drs. Kent and McShane reminded us that one of the key reasons for even thinking about these subgroups is this idea of personalized medicine.

Primary analyses involve average treatment effects, but how can these results be really translated into effective treatment for individual patients? Dr. Fleming very importantly reminded us of the many, many conflicts of interest that can come into play with these subgroup analyses. For various reasons, we all want our trials to succeed. I think it's always important to keep in mind some of the very misleading results that can emerge if we aren't careful about how we carry them out.

Both Drs. Fleming and Unger talked about the importance of using subgroup analyses to explore generalizability rather than to drive the overall assessment. Dr. McShane gave elegant illustrations of this issue in the context of biomarker-based trials in cancer. I thought her examples using hierarchical analyses were powerful illustrations of some of the problems that we can get into.

Dr. Kent reminded us of the importance of the risk-based assessments and gave us a framework for interpreting subgroup results in the context of the results from the primary analysis. All of our speakers emphasized the importance of reproducibility.

What I found enormously helpful were the ideas about how to think about pre-specifying subgroup analyses. How do we think in advance of what the biological plausibility of a particular subgroup might be and what are our goals? Are our goals exploration, the next trial? Do we really need a hypothesis test? And if so, what is our sampling framework?

In the afternoon, lots of exciting new methodologies were presented. I really enjoyed hearing the rigorous thought processes that have gone into some of the methodological developments. In particular, I thought Dr. Simon's introduction demonstrated this with his beautifully defined question about who benefits the most, who benefits at all, or who has some reasonable benefit.

Both Drs Simon and Ivanova talked about introducing subgroups of interest, specifically into the design phase rather than into these post hoc analyses. Dr. Simon presented interesting ideas about adaptive enrichment, and Dr. Ivanova gave an elegant example of the subgroup issues in the PrecISE trial in asthma.

I loved the discussion of machine learning, both the methods, the wonderful overview that Dr. Lipkovich gave us, along with the references and the software links, and also the rigorous discussion about some of the benefits and challenges of using machine learning in clinical trials where we have smaller sample sizes and where we don't necessarily always know exactly what we're measuring. The discussion this afternoon with all of our panelists, Drs Rothmann, Imai, and Rosenblum, offered valuable insights.

## Acknowledgements

The authors thank their professional society sponsors the Society for Clinical Trials, the American Statistical Association, and the National Institute of Statistical Sciences whose structural support has been invaluable for the success of this conference. Members of the program committee include Drs Susan S. Ellenberg, Jonas H. Ellenberg, Mary Putt, Pam Shaw (currently at Kaiser Permanente Washington Health Research Institute), Alisa J. Stephens-Shields, and James Lewis in the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding


The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: M.R. was supported by the Johns Hopkins Center of Excellence in Regulatory Science and Innovation, which is funded by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as

part of a financial assistance award (grant no. U01FD005942). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by any of the aforementioned organizations, the FDA/HHS, nor the U.S. Government. The other author(s) declared no receipt of financial support for the research, authorship, and/or publication of this article.

## Participants

Kosuke Imai: Harvard University  
 Anastasia Ivanova: University of North Carolina  
 Ilya Lipkovich: Eli Lilly  
 Mary Putt: University of Pennsylvania  
 Michael Rosenblum: Johns Hopkins University  
 Mark Rothmann: US Food and Drug Administration  
 Patrick Schnell: Ohio State University  
 Pam Shaw: University of Pennsylvania  
 Noah Simon: University of Washington

## ORCID iD

Kosuke Imai  <https://orcid.org/0000-0002-2748-1022>

## References

1. Symposium of assessing communicating heterogeneity of treatment effects for patient subpopulations: challenges opportunities, 2018, <https://www.jhsph.edu/research/centers-and-institutes/center-of-excellence-in-regulatory-science-and-innovation/news-and-events/Critical-Issues-in-Heterogeneity-of-Treatment-Effect.html>
2. Workshop on heterogeneity of treatment effects in clinical trials: methods innovations, 2020, <https://mrctcenter.org/news-events/heterogeneity-of-treatment-effects-in-clinical-trials-methods-and-innovations/#1602863324215-1289c9d5-a82a>
3. Harrell FE Jr, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15(4): 361–387.
4. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Adjusting for covariates in randomized clinical trials for drugs and biological products. *Guidance for Industry*, <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products> (2021, accessed 16 February 2023)
5. Michael Rosenblum's Projects, <http://rosenblum.jhu.edu> (accessed 16 February 2023).
6. Wang B, Susukida R, Mojtatabi R, Amin-Esmaceli M and Rosenblum M. Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Adjustment for Additional Baseline Variables. *Journal of the American Statistical Association*, Theory and Methods Section. 2021, <https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1981338>
7. Van Lancker K, Betz J and Rosenblum M. Combining covariate adjustment with group sequential, information adaptive designs to improve randomized trial efficiency, 2022, <https://doi.org/10.48550/arXiv.2201.12921>
8. Imai K and Li ML. Experimental evaluation of individualized treatment rules. *J Am Stat Assoc* 2021; 118: 242–256.
9. Imai K and Li ML. Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments, 2022, <https://arxiv.org/abs/2203.14511>
10. Simon R and Simon N. Inference for multivariate adaptive enrichment trials. *Stat Med* 2017; 36(26): 4083–4093.
11. Williams N, Rosenblum M and Diaz I. Optimising precision and power by machine learning in randomised trials with ordinal and time-to-event outcomes with an application to COVID-19. *J R Stat Soc Ser A*. Epub ahead of print 23 September 2022. DOI: 10.1111/rssa.12915.
12. Hu L. What's "Race" in Algorithmic Discrimination on the Basis of Race? *J Mor Philos*, IN PRESS.
13. Schnell PM, Müller P, Tang Q, et al. Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials. *Clin Trials* 2018; 15(1): 75–86.