The 10^{th} Annual Conference on Digital Experimentation @ MIT
November 10-11, 2023, MIT Samberg Conference Center

Should I Stop or Should I Go: Early Stopping with Heterogeneous Populations

Hammaad Adam[1], Fan Yin[2*], Mary Hu[2], Neil Tenenholtz[3], Lorin Crawford[3], Lester Mackey[3], Allison Koenecke[4]
[1]MIT, [2]Microsoft, [3]Microsoft Research, [4]Cornell

Randomized experiments are the gold-standard method of determining causal effects, whether in clinical trials to evaluate medical treatments or in A/B tests to evaluate online product offerings. Randomized experiments often need to be stopped prematurely—whether due to ethical or financial reasons—if the treatment yields an unintended harmful effect. While many existing methods determine when to stop an experiment early, these methods are typically applied to the collected data in aggregate and do not account for treatment effect heterogeneity with diverse patient or user populations. We first establish that current methods often fail to stop experiments when the treatment harms a minority group of participants. We then use causal machine learning to develop the first broadly-applicable method for heterogeneous early stopping, Causal Latent Analysis for Stopping Heterogeneously (CLASH). Finally, we demonstrate that CLASH's performance outperforms baselines on simulated and real data in early stopping for clinical trials with minority patient groups, as well as A/B tests with varied user devices.

Stopping homogeneously is harmful There are a variety of statistical methods that determine when to stop an experiment for harm [3, 1]. Investigators in both clinical trials and A/B tests will often choose to use a subset of these methods—collectively referred to as "stopping tests." Stopping tests not only identify harmful effects from early data, but also limit the probability of stopping early when the treatment is not harmful. However, stopping tests are typically applied to the population in aggregate (i.e., "homogeneously") and do not account for heterogeneous populations. For example, a drug may be safe for younger patients, but harm patients over the age of 65. In a hypothetical situation where younger and older patients are equally sized groups with equal but opposite treatment effects, the true Average Treatment Effect (ATE) is zero, so a traditional homogeneously-applied stopping test with H_0: ATE ≤ 0 is designed to continue to completion at least $(1-\alpha)\%$ of the time. Unfortunately, this failure to stop means that half of the trial participants will be harmed.

Consider a clinical trial for a drug such as warfarin, which has no harmful effect on the majority of the population but increases the rate of adverse effects in elderly patients [2]. Using a simple simulation (Fig 1), we demonstrate that if elderly patients comprise $\leq 20\%$ of the trial population, then applying a stopping test homogeneously would rarely stop the trial for harm. For example, if elderly patients comprise 10% of trial participants, then the probability of ending the trial early using homogeneous stopping tests is less than 20%, even if the treatment has a very large harmful effect. Thus in most cases, the trial continues to recruit elderly patients until its scheduled end, many of whom will be harmed by their participation. This outcome violates the bioethical principle of non-maleficence and is clearly undesirable.

CLASH method for stopping heterogeneously While a growing body of literature has studied how to infer heterogeneous effects [4], little work has studied how to adapt common stopping tests to respond to heterogeneity. Here, we develop CLASH, the first broadly applicable tool for heterogeneous early stopping. CLASH does not require prior knowledge of the source of heterogeneity, makes no parametric assumptions, and works with any data distribution.

[*] Current affiliation: Amazon. Work performed while at Microsoft (prior to joining Amazon).

At each interim checkpoint of the trial, CLASH operates in two stages. In Stage 1, CLASH uses causal machine learning to estimate the probability that each participant is harmed by the treatment. Then in Stage 2, it uses these inferred probabilities to reweight the test statistic of any chosen stopping test (adapting the existing stopping test to better detect treatment harms on participants). CLASH allows a practitioner to use their stopping test of choice: it is thus flexible and easy-to-use. We theoretically establish that, for sufficiently large samples, CLASH stops trials faster than the homogeneous approach if the treatment harms only a subset of trial participants. CLASH also does not stop trials unless a group is harmed: it thus leads to faster early stopping without stopping unnecessarily.

CLASH outperforms baselines in early stopping In an extensive series of simulation experiments, we demonstrate that CLASH outperforms existing baselines (Fig 2). If the minority group is harmed, CLASH (red) significantly increases the stopping probability over the homogeneous approach (blue) and SUBTLE (purple; a recently-developed heterogeneous-stopping baseline method [5] that only builds upon the often homogeneously-applied mSPRT stopping test). For large effect sizes, CLASH is as effective as an oracle that has prior knowledge of the harmed subgroup (green). Crucially, if the treatment has no harmful effect, CLASH does not stop the trial more often than either baseline. CLASH's improvements over baseline methods are robust across parameters including the stopping test used, majority and minority group sizes, treatment effect size, and the number of covariates.

We further illustrate CLASH's performance using real-world data from a technology company running a large-scale A/B test to evaluate the effect of a software update on user experience with 500,000 participants (Fig 3). CLASH is able to effectively detect user harms in certain geographic regions (wherein the software update was determined to have a significantly negative impact on relevant metrics), and appropriately stop early. Overall, we show that our method leads to effective heterogeneous early stopping across a range of randomized experiments, outperforming baselines, and nearing oracle-level performance at large sample sizes.

We emphasize that early stopping is a nuanced decision. For example, if a treatment harms only a subset of participants, it may be desirable to stop the experiment only on the harmed group but continue it on the rest of the population. In other situations, it may make sense to stop the trial altogether. Such decisions are influenced by the treatment's potential benefit, the nature of harm, and other ethical considerations; while our method is a useful aid for practitioners to make difficult decisions on early stopping, it is not intended to replace discussion on trial ethics.

References

[1] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[2] Aditi Shendre, Gaurav M Parmar, Chrisly Dillon, Timothy Mark Beasley, and Nita A Limdi. Influence of age on warfarin dose, anticoagulation control, and risk of hemorrhage. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 38(6):588–596, 2018.

[3] A Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 1945.

[4] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[5] Miao Yu, Wenbin Lu, and Rui Song. Online testing of subgroup treatment effects based on value difference. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1463–1468. IEEE, 2021.

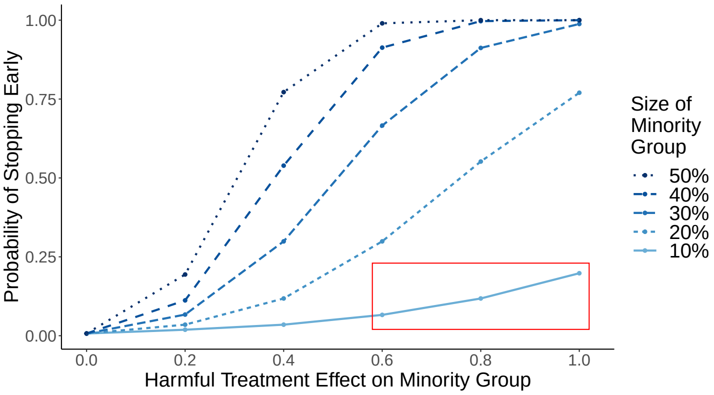

Figure 1: Shortcomings of applying stopping tests homogeneously (simulated using the O'Brien-Fleming stopping test).

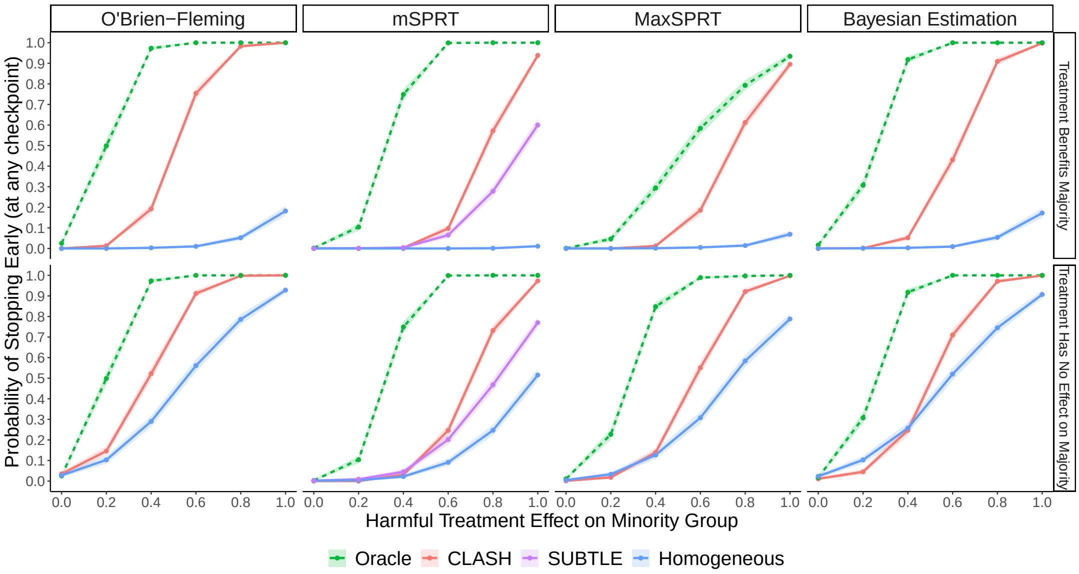

Figure 2: Performance of CLASH in a simulation experiment with normally distributed outcomes.

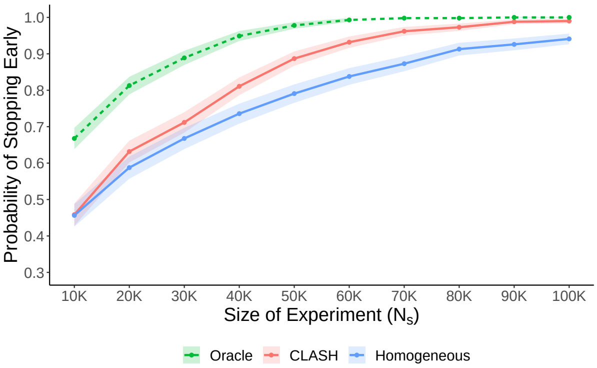

Figure 3: CLASH's performance by sample size with real-world data from an A/B test on 500,000 participants.