**ENAR Proposal Submission**

6/20/23

**Session Title:** Integrating Data from Multiple Sources to Estimate Causal Effects

**Motivation:** Reliable estimation of causal effects is vital to make informed treatment decisions. There has been a burgeoning interest in utilizing multiple data sources to improve causal effect estimation throughout the causal inference literature. Combining datasets has strong potential to improve the estimation and generalizability of causal effects; however, there are many intricacies that must be effectively accounted for to ensure accurate estimation. For example, combining randomized controlled trials (RCTs) together can improve the power for identifying treatment effect moderation, but heterogeneity across trials can complicate this assessment. Similarly, bringing in observational data can effectively supplement trial data, but the observational and trial data should be sufficiently comparable and population differences need to be adjusted for. This session will include three talks that explore approaches for integrating data from different sources to estimate causal effects. We will reflect on the benefits and challenges that come with combining data and explain methods that researchers can use to best leverage the multiple datasets. Two of the presentations focus on combining RCTs to estimate heterogeneous treatment effects, while the third explores using real world data to supplement an RCT and construct a hybrid control arm. Finally, the session will include a discussant with expertise in this area of data integration, specifically surrounding the combination of observational and trial data and the importance of taking into account the level of overlap in the two populations. We feel that this session would bring significant value to ENAR attendees looking to learn more about causal inference and the role that data integration can play in effectively estimating treatment effects.

**Carly Brantner**

**Title:** Combining Trials to Estimate Heterogeneous Treatment Effects in a Target Sample

**Abstract:** Clinicians are often motivated to determine which treatment would work best for an individual based on their observed characteristics, but reliably doing so requires large amounts of data and adjustment for confounders. One approach to estimate these effects can be through combining randomized controlled trials (RCTs). However, methods that combine RCTs often yield estimates that are conditional on trial membership, so applying these models to a target sample is not straightforward. This presentation introduces approaches for estimating conditional average treatment effects (CATEs) for a target sample, based on a model derived from multiple RCTs. The approaches draw from meta-analytic prediction intervals to create 95% intervals for the CATEs in the target sample. We examine simulations based on real data that compare meta-analytic prediction intervals with resampling-based intervals from causal forests. We discuss the trade-offs of parametric and non-parametric approaches and how researchers might choose between them. These approaches allow future researchers to effectively leverage multiple RCTs to estimate treatment effects in a target sample of interest.

**Cathy Shyr**

**Title:** Multi-study R-learner for Heterogeneous Treatment Effect Estimation

**Abstract:** Estimating heterogeneous treatment effects is crucial for precision medicine. While multiple studies can improve the generalizability of results, leveraging them for estimation is statistically challenging. Existing methods assume identical treatment effects across studies, but this may be violated due to various sources of between-study heterogeneity. To this end, we propose a unifying framework for heterogeneous treatment effect estimation robust to between-study heterogeneity in the nuisance functions and treatment effects. Our approach, the multi-study R-learner, extends the R-learner to obtain principled statistical estimation with machine learning (ML) in the multi-study setting. It incorporates ML for estimating heterogeneous treatment effects, nuisance functions, and membership probabilities, which borrow strength across studies. Our method achieves robustness in confounding adjustment through its loss function and can leverage both randomized and observational studies. We provide asymptotic guarantees in the case of series estimation and illustrate using cancer data that it has the lowest error compared to existing methods in the presence of between-study heterogeneity.

**Benjamin Ackerman**

**Title:** Augmenting Randomized Trials With Real-World Data: A Simulation Study Evaluating Methods For Hybrid Control Arm Analyses

**Abstract:** Randomized trials (RCTs) are considered the gold standard for estimating causal effects of new therapeutics, yet statistical challenges remain in detecting effects in rare disease populations. Innovative approaches exist to supplement trials with real-world data (RWD) by constructing a hybrid control arm to increase study sample size and power. RWD suitability for this use case is critical and dependent on RCT design, study population and outcome measure. Even with high-quality RWD, differences in study populations must be properly accounted for and eligibility criteria must be aligned to ensure comparability, particularly for characteristics that are prognostic of the outcome. We present propensity score weighting and frequentist borrowing methods to align study populations when conducting hybrid control arm analyses. We simulate studies under varying degrees of covariate imbalance and illustrate conditions where use of these RCT-augmentation methods yield greater precision than RCT-only analyses while maintaining the type-1 error. We present criteria when selecting suitable RWD, and highlight considerations when implementing the proposed methods for hybrid control arm analyses.

**Irina Degtiar**

**Discussant**

**Trang Nguyen**

**Chair**