Questions and Answers (Q&As)

Mind the Gap Webinar – Toward Causal Inference in Cluster Randomized Trials: Estimands and Reflection on Current Practice

Fan Li, Ph.D.

November 3, 2022

Q: The within-cluster estimands are presented based on finite samples. Does this mean that you excluded cases where the cluster size can go to infinity? Why? And what are the implications of such choice? What may happen if you assume the cluster size can go to infinity?

A: In this presentation, we have restricted to the cases where we assume the cluster population is finite, and when we invoke asymptotic theory, we let the number of clusters (in this case number of units at the treatment assignment level) go to infinity. Generally, this is a technical decision to facilitate the derivation of analytical results. Simulations have confirmed that these results, while being derived from asymptotic analysis, can hold in relatively small samples.

One can certainly assume the cluster size to go to infinity, but the asymptotic results may be different. The exact forms of those results would need to be studied in future work.

Q: Could you explain the pros and cons of using an identity link function versus a logit link function for binary outcome in generalized estimating equation (GEE) analysis?

A: In the context of associational analysis (without the need to worry about well-defined causal estimands), the choice between link functions depends on the effect measure of interest (estimating risk difference versus odds ratio parameters). In cluster randomized trials where the interest lies in estimating, for example, the individual-average treatment effect, the same consideration can apply—that is, one can use independence GEE with an identity link to estimate the causal risk difference and use independence GEE with a logit link to estimate the causal odds ratio. An example of the latter can be found in this article: arxiv.org/abs/2205.05624

For comparing GEE with an identity link versus a logit link to estimate a common target estimand, the choice may depend on efficiency, and this is yet to be examined in future simulation studies.

Q: In case we're unsure about informative cluster size, is it safe to estimate the two average treatment effects, then decide about the cluster-dependence?

A: The decision to pursue one or both types of average treatment effect (cluster-average and individual-average) would depend on the scientific question of interest, and can be study specific. In cases where both effects are of interest, it would be recommended to use methods that are robust to informative cluster size to target each specific causal effect—for example, using independence GEE to estimate individual-average treatment effect and using independence GEE with inverse cluster size weights to estimate cluster-average treatment effect.

Not sure I fully understand "cluster-dependence" in the question, but I suppose it refers to the existence of informative cluster size (dependence of treatment effect on cluster size). At the moment, a formal test for equality between the cluster-average and individual-average treatment effects is still being developed and, when available, that test can aid in the identification of informative cluster size.

Q: If the effect of interest changes between clusters, what is the individual-average treatment effect actually measuring?

A: Individual-average treatment effect is considered a population average, where the population refers to the pooled population of individuals across all clusters. This estimand remains well-defined even if the true cluster-specific average treatment effect differs across clusters, and cluster treatment effect heterogeneity itself should not affect the interpretation of the individual-average treatment effect estimand. To further elaborate, this would be like individual treatment effect heterogeneity does not effect or invalidate the interpretation of the average treatment effect estimand in an individually randomized trial.

Q: Do any of the results and conclusions here apply to generalized linear mixed modeling (GLMM) with non-continuous outcomes?

A: Unfortunately, the results do not directly apply to GLMM with non-identity link functions. This is because the likelihood function has a complex integral representation, and the model-robustness aspects of the associated regression estimators are more challenging to study analytically. However, with a binary or count outcome, if the causal effect measure is on the difference or additive scale, one can still consider the linear mixed model as a working model to consistently estimate such an effect measure (the robustness to parametric working assumptions still applies). With ratio effect measures such as risk ratio or odds ratio, a g-computation estimator would be needed, and this development can be found in the following paper: arxiv.org/abs/2210.07324

Q: Is the random selection of clusters just a simplifying assumption? If that's the case, then I'm thinking a non-random selection of clusters is the more realistic scenario. Am I wrong to think so?

A: The random selection of clusters is considered as a simplifying assumption. This can be thought of as a parallel assumption to that implicitly assumed for individually randomized trials, where units of assignment (individuals) are often considered as random samples from some (even if imaginary) target population that effects can be directly generalizable to. For a cluster randomized trial, it is likely that the clusters are not randomly sampled from the target population of clusters. In that case, if the interest lies in the causal effect among the target population of all clusters, one needs to measure additional data in that larger population and invoke additional assumptions to estimate the desired causal effect. This is a point where the presentation may intersect with the literature on generalizability and transportability, and merit additional study.