Combining registers, surveys and case-control studies in causal inference

Juha Karvanen

Department of Mathematics and Statistics,
University of Jyvaskyla, Jyväskylä, Finland,
juha.t.karvanen@jyu.fi

February 26, 2021

We consider causal inference with with multiple data sources that have partially overlapping variables. For instance, we may have independently collected survey data and case-control data that suffer from different selection biases. In addition, we may have register data that provides population statistics for the background variables but do not contain the treatment or the response variables. We use Do-search [1, 2, 3], a recently developed algorithmic implementation of do-calculus, to check the identifiability of the causal effect of the treatment on the response from the available data sources. For an identifiable query, Do-search returns a non-parametric formula that can be used to construct an estimator for the average treatment effect. Combining data sources allows us to identify causal effects that are not identifiable solely from a single source.

References

[1] J. Karvanen, S. Tikka, and A. Hyttinen. Do-search: A tool for causal inference and study design with multiple data sources. *Epidemiology*, 32(1):111–119, 2020.

[2] S. Tikka, A. Hyttinen, and J. Karvanen. *dosearch: Causal effect identification from multiple incomplete Data Sources*, 2019. R package version 1.0.6.

[3] S. Tikka, A. Hyttinen, and J. Karvanen. Causal effect identification from multiple incomplete data sources: a general search-based approach. *Journal of Statistical Software*, Accepted (https://arxiv.org/abs/1902.01073), 2020.