The Benefits and Challenges of Leveraging Existing and Secondary Data for Pragmatic Research

David M. Vock, PhD

Through a handful of case studies, we explore some of the benefits and drawbacks of leveraging existing and secondary data in pragmatic research. Existing/secondary data present an important resource which can be used at many points in the lifecycle of pragmatic research including for planning and trial design, participant recruitment, endpoint ascertainment, calibration of treatment effects, among others. However, we argue that existing/secondary data should be interrogated for not only what it includes but also what it systematically does not capture. When possible, the limitations of existing/secondary data should be ameliorated in the design and analysis plan. Finally, we argue that many perceived weaknesses of existing and secondary data such as patient heterogeneity, measurement error of covariates, etc. should be reframed as strengths for pragmatic research.

Key Thought Questions:

1) What are some key barriers to using existing and secondary data in your research? How can they be overcome?
2) How can the limitations of existing and secondary data be rephrased as relative strengths of the sources?
3) What can methodologists do to improve the suite of available methods to make using existing and secondary data more

Key Points

1) Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research (e.g., planning, participant recruitment, endpoint ascertainment, calibration of treatment effects, etc.)
2) Any source of data should be interrogated for the not only what it includes but also what it does not capture.
3) Using existing and secondary data requires a data integration and security plan.
4) The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
5) Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Resources

1. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017 Jan;106(1):1-9.
2. Sidebottom AC, Sillah A, Vock DM, Miedema MD, Pereira R, Benson G, Lindberg R, Boucher JL, Knickelbine T, VanWormer J. (2018) Assessing the effect of the Heart of New Ulm Project: a population-based program to reduce cardiovascular disease. *Preventative Medicine* **112**:216-221. doi: 10.1016/j.ypmed.2018.04.016 PMID: 29634974
3. Sidebottom AC, Sillah A, Boucher J, Vock DM, Pereira R, Benson G, Knickelbine T, Miedema MD, VanWormer J. (2016) Changes in cardiovascular risk factors after 5 years of implementation of a population-based program to reduce cardiovascular disease: The Heart of New Ulm Project. *American Heart Journal* 175:66-76. doi: 10.1016/j.ahj.2016.02.006 PMID: 27179725

The Benefits and Challenges of Leveraging Existing and Secondary Data for Pragmatic Research

David M. Vock, PhD

4. Wolfson J, Vock DM, Bandyopadhyay S, Vazquez-Benitez G, Johnson PE, Adomavicius G, O'Connor PJ. (2017) Use and customization of risk scores for predicting cardiovascular events using electronic health record data. *Journal of the American Heart Association* 6(4):1-11. doi: 10.1161/JAHA.116.003670 PMID: 28438733 PMCID: PMC5532984
5. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Vazquez-Benitez G, O'Connor PJ, Johnson PE. (2016) Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics* **61**:119-131. doi: 10.1016/j.jbi.2016.03.009 PMID: 26992568 PMCID: PMC4893987

[Notes]