Reproducibility, Confidentiality, and Open Data Mandates

Kimberlyn McGrail

There are a number of things to consider when contemplating reproducibility and open data mandates in the context of research uses of confidential and often highly-sensitive data. This note outlines a number of these considerations, and associated challenges, as well as a few potential responses to those.

# The research context

The specific context of focus here is quantitative analysis of routinely-collected data at the intersection of health economics, health services and policy research, and population health. The context is important because it implies three important features that are relevant to reproducibility. First, routinely-collected data refer to administrative and other data sources that are highly relevant sources for applied and policy-relevant research. By definition, however, they are not collected by the researcher and therefore do not “belong to” the researcher or the researcher’s institution. Second, these data sources can often be linked, which creates very rich information, and also means they can come through different data stewards; these stewards may not all have precisely the same rules and expectations around data use. Third, these data are highly sensitive, both because they are about sensitive subjects (e.g. health and socioeconomic status) and because they provide complex detail about individuals’ interactions with health and social services.

There are many existing resources that can help researchers navigate access to and use of these complex linked data, including for example [Health Data Research Network Canada](https://hdrn.ca/dash), and the [Canadian Research Data Centre Network](https://crdcn.ca/). In these cases, and many other similar services, it is not possible to deposit data in an open access framework. Again, this has to do with ownership, but also because these data are broad resources that can support hundreds of projects annually. It would likely not be efficient or effective to archive all of the resulting permutations from the same base data resources.

*Key definitions*

Related to this general context is that this area of research often focuses on questions that are about the functioning of complex systems. This means that projects will use lots of variables and assess many different relationships among them. There is a need for deep knowledge of the policy systems from which data are generated and any dynamics in those across sites or geography and over time. In addition, many of the variables that might be of interest are measures of emergent properties, which are [“…properties that manifest themselves as the result of various system components working together, not as a property of any individual component.”](https://sciencetrends.com/what-are-emergent-properties-definition-and-examples/) In other words, these are measures that do not exist in the data in their original form, but need to be developed, often with complex algorithms through some sort of validation study. The implication is that the greater variety of data that are linkable, the more opportunity there is to develop and then study these kinds of emergent properties. Some simple examples might be aspects of health system performance (safety, efficiency, etc.) or the concept of resilience among children.

All of this might help shift focus from the deposit of data to the sharing of code, algorithms, metadata and other resources that might help with reproducibility, as well as with replicability, robustness and generalizability. As a quick overview, the definitions used here for these concepts are:

**Reproducibility** is using the same data and same analysis and getting the same result.

**Replicability** is using different data and the same analysis and getting the same result.

**Robustness** is using the same data and different analysis and getting the same result.

**Generalizability** is using different data and different analysis and getting the same result.

All of these are important, with the emphasis depending in part on how unique or new a research question is, the existing evidence, and the importance of, or difference in, context when studying the phenomenon of interest.

# The challenges and possible solutions

Given the context and ambitions outlined above, there are three distinct challenges for research. First is accommodating sensitive or private data, second is the limited space in journal articles for complex methods for data development (i.e. data wrangling and shaping as opposed to statistical analytic methods), and the third is transferability of code across different systems and data generation processes.

**Sensitive and private data:** While some data cannot be placed in open access repositories, it is still possible to meet other principles of [open science](https://science.gc.ca/site/science/en/open-science-helping-make-science-accessible-all-canadians). Specifically, the data used can comply with the [FAIR principles](https://www.go-fair.org/fair-principles/) of being findable, accessible (as in, clarity on how they can be accessed), interoperable and reusable. Very much aligned with that, researchers can develop metadata for the final data set used for analysis. Good metadata will include both provenance information, such as the population covered, the time period, the purpose of the data set creation and so on, and specific details on the variables and the data sets and computational methods used to derive them. It can be helpful to think in terms of a “data set genealogy”, meaning a figure or diagram to show original data sets and how they come together to create the analytic data set. This ideally would include information on the starting population (N= ), inclusions and exclusions, and the resulting data set.

**Limited space in journal articles for data development methods**: The ability to reproduce or replicate studies relies on transparency. The required transparency includes both the process used to move from the real-world, routinely-collected data to an analysis-ready data set and the specific analysis or modelling approach used. Analysis and modelling is often the focus for scrutiny in peer review. The construction of the analytic-read data set, in contrast, is often described in enough detail to generally understand what was done but short of what is needed to replicate or reproduce the study.

One way to address this is to develop concept dictionaries such as the one run by the [Manitoba Centre for Health Policy](http://mchp-appserv.cpe.umanitoba.ca/search.php), and related code repositories that can help operationalize those concepts. Another possible response is to develop guidelines and a receptive publication venue for detailed methods, as a form of a “deposit paper” that could accompany new entries to a concept dictionary or new contributions to a code repository. This type of paper would provide transparency of the process used to create an analysis ready data set, and therefore the ability for other researchers to assess the process and the decisions made. A standard format and expectations for these methods papers would help researchers to structure the content and peer reviewers with their assessments of quality and completeness. Health Data Research Network Canada is pursuing these options, so may be a source of more information about these ideas in the future.

**Transferability of code:** Code repositories provide a place to share, but it is also important to think about the transferability of code. For example, it may be important to document the context of analysis (lightly in the code, with direction to metadata) and specifically how the context might alter the analysis, the findings, and potential for generalizability. Another possibility is to use [data harmonization](https://academic.oup.com/ije/article/46/1/103/2617181?login=false) or [data standardization](https://www.ohdsi.org/data-standardization/) practices to develop [common data models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6437693/). The attraction of common data models is that they provide pre-specified definitions for variables of interest, and the ability to access readily-usable and analysis-ready data from multiple jurisdictions. In other words, common data models address data and concept transferability once, which then can be used many times by any researcher with access to those data. The drawback is that these models may not have all the variables or emergent properties of interest, but this opens the possibility for extensions to existing models. If these are in place, and are well-documented, they create transparency, efficiency and the potential for reproducibility and replicability (and potentially robustness and generalizability) in research.

# Conclusion

Research using sensitive data raises some additional challenges for meeting the mandates of open science, including reproducibility. These challenges are solvable, but will require additional attention and work from research teams to ensure that the fundamental ingredients of transparency in data availability, access, manipulation and analysis are all present.