Case Selection

CITF-Postdoc

Zack Batist

2025-01-07

ℹ Note

This document is still a work in progress.

Case Study Research

In case-study research, cases represent discrete instances of a phenomenon that inform the researcher about it. The cases are not the subjects of inquiry, and instead represent unique sets of circumstances that frame or contextualize the phenomenon of interest (Stake 2006: 4-7).

Cases usually share common reference to the overall research themes, but exhibit variations that enable a researcher to capture different outlooks or perspectives on matters of common concern. Drawing from multiple cases thus enables comprehensive coverage of a broad topic that no single case may cover on its own (Stake 2006: 23). In other words, cases are contexts that ascribe particular local flavours to the activities I trace, and which I must consider to account fully for the range of motivations, circumstances and affordances that back decisions to perform activities and to implement them in specific ways.

Moreover, the power of case study research derives from identifying consistencies that relate cases to each other, while simultaneously highlighting how their unique and distinguishing facets contribute to their representativeness of the underlying phenomon. Case study research therefore plays on the tensions that challenge relationships among cases and the phenomenon that they are being called upon to represent (Ragin 1999: 1139-1140).

It should be noted that case study research limits my ability to define causal relationships or to derive findings that may be generalized across the whole field of epidemiology. This being said, case study research allows me to articulate the series of inter-woven factors that impact how epidedemiological researchers coordinate and participate in data-sharing initiatives, while explicitly accounting for and drawing from the unique and situational contexts that frame each case.

Key Factors

To reiterate, this project investigates the social and collaborative apparatus that scaffold data-sharing initiatives in epidemiology. Through analysis of data obtained through interviews with various relevant stakeholders attached to data-sharing initiatives, the project will ascertain the actions taken and challenges experienced to mediate the varied motivations, needs and values of those involved. In effect, the project aims to articulate the collaborative commitments that govern the constitution and maintenance of epidemiological information commons, and to relate these to technological, administrative and epistemic factors.

In other words, I aim to make certain under-appreciated social and collaborative commitments that underlie data-sharing initiatives more visible and to draw greater attention to certain sensibilities, attitudes, and apprehensions that are relevant to contemporary discourse on the nature of epidemiological data and ongoing development of information infrastructures designed to support data integration and re-use.

Here I outline some key factors that will guide the selection of cases so as to ensure that the project meaningfully addressses its goals.

1. Longevity

Initiatives that have existed for different durations of time will have different capacity to reflect on their practices. Younger projects will not have had as much of a chance to produce any research outcomes, but may be valuable sources for insight on expectations. More established projects will be able to reflect on unexpected challenges they may have experienced.

It will be good to have at least one younger project representing an initiative still “in flux”, one or two “legacy” projects (no longer active), and one or two at intermediate stages (extracting data for meaningful analysis, expanding the initiative’s scope, etc).

2. Community composition

The size and composition of the community, degree of familiarity among its members, and the mechanisms through which connections are managed constitute additional important factors to consider. Communication and decision-making may take different forms when teams are either smaller and locally-concentrated or larger and dispersed. Decision-making may also be significantly impacted by diffferent governance models and degrees of community participation. It would be interesting to identify how leaders are differentiated from other participants, norms and expectations for getting involved in leadership positions, and considerations that are made when making decisions that impact the community.

3. Support structures

Data-sharing may be supported by diverse funding models or tech stacks to support the work, which may significantly impact how the work progresses. Comparing sources of support for data-sharing will help me to explore how data-sharing is either integrated into or supplemented as a distinct outgrowth of “normal” science.

Specifically, it will be interesting to compare the extent to which projects are left to cobble together their own data-sharing infrastucture, and how this impacts attitudes and norms regarding the curation and nature of research data. I wonder whether lack of government support fosters creative, entrepeneurial, experimental or community-led models, how funding is provided to supporting the development of collaborative research networks, and how these feed back into norms and attitudes regarding the independence of individual research projects and the formation of collectively-maintained information commons.[1]

I expect a tendency for cases to be supported by limited-term, federally-funded grants, though it might be worth exploring how supplementary funding provided by non-government agencies, including private firms (through MITACS, for instance) and philanthropic organizations (such as the Gates Foundation) impact the work. I would therefore like to included cases funded through these kinds of initiatives in this project.

4. Disciplinary trends

Data-sharing is undoubtably impacted by attitudes concerning the nature of data and their roles in scientific knowledge production, and it is therefore necessary to account for different perspectives. Although I am still somewhat unfamiliar with the diversity of thought on such matters in epidemiology, I intuit that much of the open science movement is driven by rather positivist attitude. I would like to include cases that take on alternative approaches to science.

5. Historical or contextual factors

Science is beholden to political trends, which impact ability to obtain funding and collaborate accross borders (e.g. Brexit's impact on trans-European funding, including initiatives to attract and retain talent). Moreover, certain events, such as the Covid-19 pandemic, trigger responses in the scientific community. Even if these events are not the focus of the research, they must still be accounted for due to their presumed impacts.

[1] There is some precedent for this in the social sciences and humanities, which are fields that open science policies and infrastructures are not really designed to handle. This marginizaliation had contributed to experimentation with community-based governance models (as per the Radical Open Access Collective) and broader community involvement in policy decisions concerning how the rich diversity of social science and humanities data should be curated.

6. Kinds of data

The nature of the data will surely impact how they are shared. In epidemiology specifically, there are ethical limitations on sharing precise patient records. This may be especially salient in studies focusing in health in Indiginous populations, which may involve additional consideration in contexts of data-sharing.[2] Moreover, controls on data collection procedures, including limited or controlled scope or decisions to account for specific factors (such as race, which is prevalent in American datasets but largely ignored elsewhere) may significantly impact what can be done with them when integrated at scale.

Selecting Cases

Since a significant aspect of this work is to compare different approaches to data-sharing that have not yet been systematically articulated, it will be necessary to loosely define the parameters through which each case will be initially characterized. I will rely on structured consultations with the research community to make sense of the data-sharing landscape and select cases accordingly. By consulting with key stakeholders, I will arrive at a consensus about which cases are worth approaching while documenting the rationale behind these selections.

The consultation process is meant to ensure that case selection adheres to community will and reasoning, while also ensuring that cases are logistically feasible. I will therefore ask for input from leading members of epidemioligical data-sharing initiatives who are familiar with the goals of the this project, and who are involved with the Maelstrom Project which establishes logistical boundaries around the scope of the project.

Fixed cases

Maelstrom will serve as a “fixed point” that limits the scope of the cases’ breadth, while also ensuring that participants (and

[2] The Data Governance and Management Toolkit for Self-Governing Indigenous Governments https://indigenousdatatoolkit.ca may be helpful for exploring these concerns, but I am still looking for epidemiologically-oriented resources on such matters.

myself) have a common frame of reference. Moreover, the practices and values that support Maelstrom's operations have already been documented to a certain extent by its leaders (cf. Doiron et al. 2017; Fortier et al. 2017; Fortier et al. 2023; Bergeron et al. 2018), by its partners (cf. Doiron et al. 2013; Wey et al. 2021; Bergeron et al. 2021) and by scholars of scientific practice (cf. M. J. Murtagh et al. 2012; Demir and Murtagh 2013; Madeleine J. Murtagh et al. 2016; Tacconelli et al. 2022; Gedeborg et al. 2023). This prior work will serve as valuable resources supporting this project.

Additionally, the fact that all cases interact with Maelstrom for their technical infrastructure will greatly simplify the interviews by reducing the "overhead" of having to learn or be told about the technical systems, which may distract from the primary themes I seek to address during interviews.

CITF will also serve as a fixed case. This is partly for logistical reqasons, since the grant is meant to support the CITF Databank, and this project will align with concurrent research on user experiences pertaining to CITF specifically. At the same time, CITF is relevant to the project's objectives in its own right, and will contribute meaningful insight in comparison with other cases.

Logistical Constraints and Sources of Bias

After identifying potential cases, I will reach out to project leaders to invite them to participate. I will prepare a document outlining this project's objectives and the roles that cases will play in the work. I will also set up a meeting prior to them deciding whether they would like to participate so I can ascertain whether they understand the project and to help determine who may serve as people who can sit for interviews (I expect to hold 12-15 interviews ranging between 60-90 minutes in duration).

I may prioritize local connections, which provide favourable conditions for holding interviews (i.e., people are more willing to show things that can not be conveyed through a screen, and the pre- and post-interview phases provide meaningful insight).

This may introduce bias in that I may obtain more in-depth and nuanced information from local initiatives than those occuring abroad. This can be mitigated by travelling to conduct interviews in person, however the costs of travel may introduce their own biases favouring cases that are easier to reach.

Bergeron, Julie, Dany Doiron, Yannick Marcon, Vincent Ferretti, and Isabel Fortier. 2018. "Fostering Population-Based Cohort Data Discovery: The Maelstrom Research Cataloguing Toolkit." *PLOS ONE* 13 (7): e0200926. https://doi.org/10.1371/journal.pone.0200926.

Bergeron, Julie, Rachel Massicotte, Stephanie Atkinson, Alan Bocking, William Fraser, Isabel Fortier, and the ReACH member cohorts' principal investigators. 2021. "Cohort Profile: Research Advancement Through Cohort Cataloguing and Harmonization (ReACH)." *International Journal of Epidemiology* 50 (2): 396–97. https://doi.org/10.1093/ije/dyaa207.

Demir, Ipek, and Madeleine J. Murtagh. 2013. "Data Sharing Across Biobanks: Epistemic Values, Data Mutability and Data Incommensurability." *New Genetics and Society* 32 (4): 350–65. https://doi.org/10.1080/14636778.2013.846582.

Doiron, Dany, Paul Burton, Yannick Marcon, Amadou Gaye, Bruce H. R. Wolffenbuttel, Markus Perola, Ronald P. Stolk, et al. 2013. "Data Harmonization and Federated Analysis of Population-Based Studies: The BioSHaRE Project." *Emerging Themes in Epidemiology* 10 (1): 12. https://doi.org/10.1186/1742-7622-10-12.

Doiron, Dany, Yannick Marcon, Isabel Fortier, Paul Burton, and Vincent Ferretti. 2017. "Software Application Profile: Opal and Mica: Open-Source Software Solutions for Epidemiological Data Management, Harmonization and Dissemination." *International Journal of Epidemiology* 46 (5): 1372–78. https://doi.org/10.1093/ije/dyx180.

Fortier, Isabel, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, et al. 2017. "Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization." *International Journal of Epidemiology* 46 (1): 103–5. https://doi.org/10.1093/ije/dyw075.

Fortier, Isabel, Tina W. Wey, Julie Bergeron, Angela Pinot de Moira, Anne-Marie Nybo-Andersen, Tom Bishop, Madeleine J. Murtagh, et al. 2023. "Life Course of Retrospective Harmonization Initiatives: Key Elements to Consider." *Journal of Developmental Origins of Health and Disease* 14 (2): 190–98. https://doi.org/10.1017/S2040174422000460.

Gedeborg, Rolf, Wilmar Igl, Bodil Svennblad, Peter Wilén, Bénédicte Delcoigne, Karl Michaëlsson, Rickard Ljung, and Nils Feltelius. 2023. "Federated Analyses of Multiple Data Sources in Drug Safety Studies." *Pharmacoepidemiology and Drug Safety* 32 (3): 279–86. https://doi.org/10.1002/pds.5587.

Murtagh, M. J., I. Demir, K. N. Jenkings, S. E. Wallace, B. Murtagh, M. Boniol, M. Bota, et al. 2012. "Securing the Data Economy: Translating Privacy and Enacting Security in the Development of DataSHIELD." *Public Health Genomics* 15 (5): 243–53. https://doi.org/10.1159/000336673.

Murtagh, Madeleine J., Andrew Turner, Joel T. Minion, Michaela Fay, and Paul R. Burton. 2016. "International Data Sharing in Practice: New Technologies Meet Old Governance." *Biopreservation and Biobanking* 14 (3): 231–40. https://doi.org/10.1089/bio.2016.0002.

Ragin, C C. 1999. "The Distinctiveness of Case-Oriented Research." *Health Services Research* 34 (December):1137–51. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089057/.

Stake, Robert E. 2006. *Multiple Case Study Analysis*. New York: Guilford Press.

Tacconelli, Evelina, Anna Gorska, Elena Carrara, Ruth Joanna Davis, Marc Bonten, Alex W. Friedrich, Corinna Glasner, et al. 2022. "Challenges of Data Sharing in European Covid-19 Projects: A Learning Opportunity for Advancing Pandemic Preparedness and Response." *The Lancet Regional Health – Europe* 21 (October). https://doi.org/10.1016/j.lanepe.2022.100467.

Wey, Tina W., Dany Doiron, Rita Wissa, Guillaume Fabre, Irina Motoc, J. Mark Noordzij, Milagros Ruiz, et al. 2021. "Overview of Retrospective Data Harmonisation in the MINDMAP Project: Process and Results." *J Epidemiol Community Health* 75 (5): 433–41.

https://doi.org/10.1136/jech-2020-214259.