January 30, 2025

Subject: REB Submission #25-01-057

To whom it may concern,

I am writing to submit my application to the Research Ethics Board. Please find all required documents attached to this file. Please also note that I have completed the required TCPS 2: CORE-2022 ethics training (certificate #0001391231).

Due to the minimal risk projected for the study, I request that this application undergo an expedited review.

I hope that the information contained in this application is sufficient to proceed and to warrant my application's swift approval.

Zachary Batist, PhD
zachary.batist@mcgill.ca

Postdoctoral Researcher | CITF Databank
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health, McGill University

Cc: David Buckeridge
Professor and Faculty Supervisor
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health, McGill University
david.buckeridge@mcgill.ca

Research Protocol

Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Zachary Batist

2025-01-21

Study purpose and rationale

This study investigates the social, technical, administrative and epistemic factors that scaffold data-sharing initiatives in epidemiological research. It takes to heart the notions that data are media that facilitate communication across different research contexts, that data are created with specific intent, and that data are bounded by the social, practical and material circumstances of their creation. In light of these facts, the study approaches data-sharing as a means of reconciling the varied circumstances of datasets' creation — both among themselves, and in relation to contexts of reuse. It therefore frames data-sharing as efforts to foster a series of collaborative ties beyond a project's original indended scope.

Data harmonization is one means of data-sharing that draws multiple studies' recorded observations into a unified formal schema, whose structure is driven by specific objectives and more general underlying suppositions and values. Although these schemas are arrived at through discussion, compromise and consensus-building, with an eye toward practical outcomes afforded by alignment of complementary records, these socially-mediated interactions are generally under-recognized as important factors contributing to data-sharing initiatives' success relative to their potential impact. The goal of this study is to survey how various factors are prioritized in harmonization activities, the rationales behind these decisions, and the relative efficacy of different approaches to data-sharing.

The project's goal is to survey what factors are being priorized within various data harmonization initiatives, the rationales behind these decisions, and the relative efficacy of these different approaches. More specifically, the project seeks to adress the following research questions:

- What are the objectives of data-sharing initiatives, how were they established, and what progress has been made to achieve them?
- What strategies do data-sharing initiatives employ to ensure they are able to meet their objectives, and how effective are they?
- What values underlie these strategies, and can they be linked with effective outcomes?

To be clear, the intent is not to pit various approaches against each other, but rather to ascertain what actions specific strategies entail, the circumstances in which each is adopted, the value that they bring, and the trade-offs involved. In other words, the goal is to reveal the diverse ways in which data-sharing occurs, and how different approaches impact the outcomes in different ways.

Description of study population, inclusion and exclusion criteria

The study population comprises individuals who lead, support or participate in epidemiological data-sharing initiatives. These individuals include profesional researchers, research trainees and administrative and technical support staff affilited with epidemiological projects that coordinate, support or participate as a member of data-sharing initiatives.

Only data-sharing initiatives that partner with the Maelstrom Research project, which facilitates collaborative epidemiological research through rigorous data documentation, harmonization, integration, and co-analysis, will be considered to serve as cases for this study. Maelstrom is a well-established entity in this field and has established a broad network of partner projects to select cases from. Maelstrom serves as a "fixed point" which ensures that Dr. Batist and the study participants share a common frame of reference. Maelstrom established generalizable nomenclature and toolsets across all its partner projects, which reduces the overhead of matching different terms and and practices across cases. This is especially valuable in the context of interview-based research, wherein it is crucial to remain focused on obtaining information that is relevant to the themes the project seeks to address during the limited time alloted. Additionally, Maelstrom's principle investigators and partners have written extensively about the values and practical challenges concerning data-sharing in epidemiology, which provides a rich foundation upon which the analysis may be based (cf. Demir and Murtagh 2013; Fortier et al. 2017; Bergeron et al. 2018; Fortier et al. 2023).

Sample size and how it was determined

The study draws from semi-structured interviews with 12-15 individuals from 4-5 cases.

The number of cases reflects the capacity to draw adequate comparison across unique circumstances while also complementing the meaningful number of individuals who may serve as interview participants. The project seeks to interview 3-4 people from each case, including individuals who work in specific roles, such the leaders of data-sharing consortia, support staff, and leaders of contributing projects. Additionally, the number of individuals also reflects the practical constraints that this work affords, namely the time-consuming nature of transcribing interviews and conducting qualitative data analysis; the number of participants represents a careful balance between a meaningful sample size and the amount of work required to collect, process and analyze the data within the project's one-year timeframe.

Design and description of methodology

This study is informed by a set of theoretical and methodological frameworks formed within a more interdisciplinary "science studies" tradition, which contribute to a more sociological outlook on science as cultural practice (cf. Pickering 1992). In practical terms, the study documents the social and collaborative experiences involved in various research practices, which ultimately bind the many ways in which scientists do science.

The study will specifically focus on how people contribute to and extract from information commons, which comprise both formal documents and mutually-held and information-laden situated experiences. This involves examining the ways in which participation in disciplinary or even more specialized communities of practice fosters mutual understanding about the potential and limitations pertaining to other people's data; and how this communally-held knowledge is accessed and re-produced. This approach aligns with the situated cognition methodological framework for examining the improvised, contingent and embodied experiences of human activity, including science (cf. Suchman 2007; Knorr Cetina 2001).

The situated cognition framework prioritizes subjects' outlooks, which are contextualized by their prior experiences, and enables scholars to trace how people make sense of their environments and work with the physical and conceptual tools available to them to resolve immediate challenges. Situated cognition therefore lends itself to investigating rather fluid, open-ended and affect-oriented actions, and is geared towards understanding how actors draw from their prior experiences to navigate unique situations.

Situated cognition is especially salient in explorations of how people who are learning new skills learn how to work in new and possibly unfamiliar ways, and in this sense is closely related to Lave and Wenger's (1991) theory of situated learning (or 'communities of practice' approach), which focuses on how individuals acquire professional skills in relation to their social environments. In such situations, situated cognition enables observers to examine how people align their perspectives as work progresses, and to understand better how people's general outlooks may have changed under the guidance of more experienced mentors. In other words, situated cognition enables researchers of scientific practices to account for discursive aspects of work, including perceived relationships, distinctions or intersections between practices that professional or research communities deem acceptable and unacceptable, and the cultural or community-driven aspects of decisions that underlie particular actions.

In taking on this theoretical framework, the study frames epidemiology as a collective endeavour to derive a coherent understanding of population-level health trends, which involves the use of already established knowledge in the validation of newly formed ideas, and which relies on systems designed to carry information obtained with different chains of inference. These systems have both technical and social elements. The technical elements are the means through which information becomes encoded onto information objects so that they may form the basis for further inference. The social elements constitute a series of norms or expectations

that facilitate the delegation of roles and responsibilities among agents who contribute their time, effort and accumulated knowledge to communal goals.

As such, in constructing the arguments of this study and in carrying out the interviews that grounds it, the study will rely upon both realist and constructivist viewpoints. In one sense, the study relies on documenting how people actually act, including the longer-term and collaborative implications that their actions may have on other work occurring throughout the continuum of practice. To accomplish this, the study identifies research activities from the perspective of an outside observer. The study also ascribes meanings to things (such as physical or conceptual tools, or objects that captivate subjects' interests) in ways that conform to the analyst's own perspective as an investigator of scientific research practices. On the other hand, a constructionist perspective enables the author to consider how individual agents make components of information systems suit their needs to facilitate communication or interoperability among actors who hold different situated perspectives. By listening to participants' views about the systems with which they engage, including explanations as to why they act in the ways that they do, the principal investigator is able to trace the assumptions and taken-for-granted behaviours that frame their perspectives. Moreover, these insights are useful for developing a better understanding of how participants identify with particular disciplinary communities and their perception of their roles within broader collective efforts.

Ultimately, this study is about the social order of scientific research, i.e. the frameworks, mindsets or sets of values that humans adopt to carry out their work in specific ways. Human beings rely upon physical and conceptual apparatus to do this work but, in order to understand how they do science *in ways that conform to the epistemic mandates of the scientific enterprise,* it is necessary to prioritize attention to human intention, drivers and pressures. The study emphasizes the agency of human drivers — as opposed to tools and procedures — since humans are the ones who (a) identify problems that need to be resolved; (b) imagine, project or predict potential outcomes of various kinds of actions that they may select to resolve the challenges; and (c) learn from prior experiences and change their behaviours accordingly. By highlighting how pragmatic actions are conducted in relation to broader social and discursive trends and tendencies, the study considers scholarly practices in terms of potential, certainty and desire from the perspectives of practitioners themselves.

To this end, the study follows an abductive qualitative data analysis (QDA) methodology to construct theories founded upon empirical evidence, which relates to, but is distinct from, grounded theory. Grounded theory consists of a series of systematic yet flexible guideline for deriving theory from data through continuous and reiterative engagement with evidence (Charmaz 2014: 1). The approach taken for this study draws from what Charmaz (2014: 14-15) calls the "constellation of methods" associated with grounded theory that are helpful for making sense of qualitative data. However, it differs from grounded theory as it is traditionally conceived in that the principal investigator came to the project with well defined theoretical goals (as described above) and did not make a concerted effort to allow the theory to emerge through the analytical process. Proponents of a more open-ended or improvised approach, as grounded theory was originally applied, argue that researchers should be free to generate

theories in accordance with their own creative insights and their intimate engagements with the evidence. We can evaluate the quality of such work in terms of the dialogical commitments between researchers and their subjects, and between researchers and those who read their work (Glaser and Strauss 1967: 230-233). Others view grounded theory more as a means of clarifying and articulating phenomena that lie below the surface of observable social experiences (Strauss and Corbin 1990; Kelle 2005). Proponents of this approach are very concerned with ensuring that concepts, themes and theories are truly represented in and limited by the data, and therefore prioritize adherence to systematic validation criteria to ensure the soundness of their claims.

Another view, known as constructivist grounded theory, most resembles the approach taken for this study. It recognizes that it is impossible to initiate a project without already holding ideas regarding the phenomena of interest, and that the ways that one ascribes meanings to the data represent already established mindsets or conceptual frameworks (Charmaz 2014). It encourages reflection on the researcher's standpoint as they pursue an abductive approach rooted in their own preconceptions (Mills, Bonner, and Francis 2006).

All of these approaches rely on a core set of methods of coding and memoing. Coding, which involves defining what data are about in terms that are relevant to the theoretical frameworks that inform the research, entails rendering instances within a text as interpreted abstractions called codes (Charmaz 2014: 43). These methods, which are described in more detail below, are particularly useful for examining the broad assemblage of evidence comprising various kinds of media and spanning multiple case studies. The abstraction of specific instances as conceptual codes enables comparisons across documents that would otherwise prove difficult to compare, due either to the analyst's own preconceptions (drawn from internalized narratives or biases) that might have framed their attitudes, to disproportionate volumes of evidence that might obscure parallels between case studies, or to difficulties experienced when examining different kinds of documents that call for different lenses or perspectives.

Definition of end-points

This project will produce insights regarding the practical benefits and challenges involved in epidemiological data-sharing. It will identify how relevant stakeholders *actually* engage with the systems that scaffold data-sharing initiatives, which may differ from modelled behaviours specified in aspirational plans and procedural documents. In effect, by articulating how these systems succeed or fail to account for their users practical needs and disciplinary values, this study will provide constructive feedback that will inform their further development.

The study will produce three peer-reviewed articles. One of these will be published in a journal concerned with scientific practice or research data management (e.g. Computer Supported Cooperative Work; Scientific Data); Another will be published in a journal dedicated to advancing research practices in epidemiology (e.g. Epidemiologic Methods) A third paper will comprise a more practical set of guidelines deriving from the findings, as either an editorial or

as a "10 simple rules" style article. I will also present this work at conferences and workshops, as opportunities arise.

Moreover, this work is intended to be constructive, and it is therefore necessary to ensure that the findings may be put to practical use so as to enhance and improve data-sharing initiatives. Dr. Batist will therefore draft policy briefs and reach out to leading stakeholders at the helm of major data-sharing infrastuctures and policy frameworks (such as the ongoing revisions to federal open science policies) so that the findings may directly inform efforts to improve these systems.

Additionally, Dr. Batist will promote this work publicly. This will entail publishing an article in The Conversation, a website with a broad public following and which specializes in showcasing specialized research for the general public. Dr. Batist may also share the findings on podcasts about open science and science policy with broad interdisciplinary appeal. Moreover, owing to his broad pan-disciplinary background, Dr. Batist is plugged into a diverse network of scholars, and through active engagement on social media and posting regular updates on his professional blog his work will reach a very broad audience.

Measurements and study instruments

The study draws from around 12-15 semi-structured interviews with epidemiological researchers who lead and carry out initiatives to share and harmonize data. Interviews will be analyzed using qualitative data analysis methods following the procedures outlined in the following section.

Data analysis plan

Data

This project will draw from around 12-15 semi-structured interviews with epidemiological researchers who lead and carry out initiatives to share and harmonize data. Interviewees will comprise members of several projects that partner with the Maelstrom research group, a service that supports data-sharing harmonization efforts in the field of epidemiology.

Interviews will be oriented by the study's goal to document processes of reconciling different stakeholders' interests as they converge in the formation of a common data resource. Specifically, interviews will focus on motivations for their initiatives, the challenges they experience, how they envision success and failure, their perceptions of their own roles and the roles of other team members and stakeholders, the values that inform their decisions, how the technological apparatus they set up enables them to realize their goals and values, and ways in which they believe data-sharing could be improved. See the attached interview protocol for further details on the questions that will be asked and how participants' responses will contribute to the project's findings, as well as logistical considerations.

Interviews will be transcribed, and transriptions will be edited to optimize them for use in qualitative data analysis software. Secure and locally-hosted automated speech recognition software may be used to create preliminary transcripts, which will then be manually edited.

All data will be collected and curated in full compliance with the ethics protocol.

Methods

The study will implement qualitative data analysis (QDA) methods to highlight collaborative aspects data-sharing in epidemiology, as elicited in the corpus of transcribed interviews. QDA involves encoding the primary sources of evidence in ways that enable a researcher to draw cohesive theoretical accounts or explanations. This is done by tagging segments of a document (such as an interview transcript) using codes, and by embedding open-ended interpretive memos directlly alongside the data. Through these methods, a researcher is able to articulate theories based on empirical evidence that reflect the informants' diverse experiences.

Coding — which involves defining what specific elicitations are about in terms that are relevant to the theoretical frameworks that inform the research — entails rendering instances within a text as interpreted abstractions called codes (Charmaz 2014, 43). Codes can exist at various levels of abstraction. For instance, an analyst may apply descriptive codes to characterize literal facets of an instance within a text, and theoretical codes to represent more interpretive concepts that correspond with aspects of particular theoretical frameworks. This project will primarly implement an "open" coding protocol, which entails creating codes on the fly when prompted by encounters with demonstrative instances in the text. As new codes are generated in this manner, they are situated within a code system that affords greater taxonomic structure to encoded observations, thereby facilitating more effective queries. Coding in this manner involves synthesis of concepts that speak to the analyst's understanding of the phenomena of interest, while forcing the analyst to remain receptive to limits imposed by what is actually contained in the corpus. In other words, coding involves applying a precise language to segments of transcribed interviews that serve to bridge the gap between what participants said and the theoretical frameworks that the analyst applies to explore them as epistemic activities, interfaces and values (cf. Charmaz 2014; Saldaña 2011, 95–98).

Memoing entails more open-ended exploration and reflection upon latent ideas in order to crystallize them into new avenues to pursue (Charmaz 2014, 72). Constructing memos is a relatively flexible way of engaging with data and serves as fertile ground for honing new ideas. Memoing is especially crucial while articulating sensitizing concepts, which Charmaz (2003, 259) refers to as the "points of departure from which to study the data". Memoing allows the researcher to take initial notions that lack specification of well-defined attributes, and gradually refine them into more cohesive, definitive concepts (Blumer 1954, 7; Bowen 2006). Exploring the main features, relationships or arrangements that underlie a superficial view of a sensitizing concept through memoing helps the analyst to identify what kinds of things they need to locate in the data in order to gain a full understanding of the phenomena

of interest. Memoing is also very important in the process of drawing out more coherent meaning from coded data (cf. Charmaz 2014, 181, 290–93). By creating memos pertaining to the intersections of various codes and drawing comparisons across similarly coded instances, an analyst is able to form more robust and generalizable arguments about the phenomena of interest and relate them to alternative perspectives expressed by others.

Throughout the analysis, Dr. Batist will follow the approach that Nicolini (2009) and Maryl et al. (2020, para. 30) advocate, who suggest "zooming in to a granular study of particular research activities and operations and zooming out to considering broader sociotechnical and cultural factors." This involves "magnifying or blowing up the details of practice, switching theoretical lenses, and selective re-positioning so that certain aspects are fore-grounded and others are temporarily sent to the background" (Nicolini 2009, 1412). This approach is useful in the context of this study because the research projects that represent the cases start from different positions but share common practices and tendencies that vary according to those contextual circumstances. It is therefore possible to tactfully switch between those lenses to understand the interplay between circumstances and practical implementations, which vary across cases, which have their own histories, memberships, sets of tools, methods, and social or political circumstances.ships, sets of tools, methods, and social or political circumstances.

This work will be performed using computer assisted qualitative data analysis software that enables analysts to retrieve segments of interview transcripts and identify patterned distributions of codes from across the entire corpus. Querying the dataset in this way enables the analyst to articulate elaborated accounts of specific kinds of activities, decisions, values and sentiments that cut across various informants' perspectives.

Regarding statistics

Statistical methods will play a limited role in this study. Basic summary statistics (e.g. cross tabulation) will be used to represent the distribution of codings across individual interviews or ranges of interviews, which will help to identify trends and associations as they pertain to their limited scopes. This will be used to support theory-building but will not be used to infer generalizable causal relationships.

Details on confidentiality

The specific circumstances that frame each case are significant factors that will shape the findings, and the study will benefit from participants consenting to associate their identities with their interview responses. Nevertheless, participants will be provided with the option to render their interview responses confidential.

The specific circumstances that frame each case are significant factors that will shape the findings, and the study will benefit from participants' consent to associate their identities with their interview responses. However, they may choose to render their interview responses

confidential while maintaining their role a research participant. Participants may change their decision regarding whether or not to associate their identities with their interview responses up to one week after the interview, at which point the principal investigator will begin transcribing and analyzing the records pertaining to the interview. Participants will be reminded about this option immediately after the interview and one week following the interview via email.

The study engages with a relatively small community, and there is minimal social risk that others may be able to determine the identities of those whose research practices and professional relationships are being documented, even if their responses are rendered confidential. To address this issue, if any single participant from a case decides to render their responses confidential, the responses of all participants pertaining to that case will be rendered confidential as well, and the identify of the project that serves as the case will be obfuscated too.

In situations whereby a participant decides to render their responses confidential, or has their responses rendered confidential due to another member of their case deciding to do so, only the principal investigator will have access to records containing un-obfuscated information that may identify them. These un-obfuscated records, which may include audio and video records of interview sessions, as well as unedited transcripts and textual notes containing information that may reveal the participants' identities, will be kept in secure and encrypted media, and destroyed within five years of concluding the study, which provides sufficient time to revisit the data and produce additional research outputs. However, edited transcripts scrubbed of all information that may identify research participants may be kept, published and archived. If participants consent to maintaining association between their responses and their identities, un-obfuscated records and transcripts may be kept, published and archived.

The study is committed to adhering to fundamental data security practices, including those specified in McGill University's Cloud Directive which regulates the curation of sensitive research data. Physical records will be kept in a locked drawer in secure workspaces, either at McGill University's School of Public and Global Health or at the principal researcher's home office. Digital records will be stored on encrypted and password-protected drives.

Statement on ethical considerations

The study will be conducted according to ethical principles stated in the Declaration of Helsinki (2013). Ethics approval will be obtained before initiating the study. Consent forms will take into consideration the well-being, free-will and respect of the participants, including respect of privacy. The practices undertaken to ensure adherence to these principles are described in the ethics protocol.

References

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.

Blumer, Herbert. 1954. "What Is Wrong with Social Theory?" *American Sociological Review* 19 (1): 3–10. https://doi.org/10.2307/2088165.

Bowen, Glenn A. 2006. "Grounded Theory and Sensitizing Concepts:" *International Journal of Qualitative Methods* 5 (3): 12–23. https://doi.org/10.1177/160940690600500304.

Charmaz, Kathy. 2003. "Grounded Theory: Objectivist and Constructivist Methods." In *Handbook of Qualitative Research*, edited by Norman K. Denzin and Yvonna S. Lincoln, 2nd ed., 249–91. Thousand Oaks, California: SAGE.

———. 2014. *Constructing Grounded Theory*. 2nd ed. SAGE.

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.

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.

Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. https://doi.org/10.4324/9780203793206-1.

Kelle, Udo. 2005. ""Emergence" Vs. "Forcing" of Empirical Data? A Crucial Problem of "Grounded Theory" Reconsidered." *Forum: Qualitative Social Research* 6 (2): 133–56. https://doi.org/10.17169/FQS-6.2.467.

Knorr Cetina, Karin. 2001. "Objectual Practice." In *The Practice Turn in Contemporary Theory*, edited by Theodore R. Schatzki, Karin Knorr Cetina, and Eike von Savigny, 175–88. London; New York: Routledge.

Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge, United Kingdom: Cambridge University Press.

Maryl, Maciej, Costis Dallas, Jennifer Edmond, Jessie Labov, Ingrida Kelpšienė, Michelle Doran, Marta Kołodziejska, and Klaudia Grabowska. 2020. "A Case Study Protocol for Meta-Research into Digital Practices in the Humanities." *Digital Humanities Quarterly* 14 (3). https://www.digitalhumanities.org/dhq/vol/14/3/000477/000477.html.

Mills, Jane, Ann Bonner, and Karen Francis. 2006. "The Development of Constructivist Grounded Theory." *International Journal of Qualitative Methods* 5 (1): 25–35. https://doi.org/10.1177/160940690600500103.

Nicolini, Davide. 2009. "Zooming In and Out: Studying Practices by Switching Theoretical Lenses and Trailing Connections." *Organization Studies* 30 (12): 1391–1418. https://doi.org/10.1177/0170840609349875.

Pickering, Andrew. 1992. "From Science as Knowledge to Science as Practice." In *Science as Practice and Culture*, edited by Andrew Pickering, 1–26. University of Chicago Press.

Saldaña, Johnny. 2011. *Fundamentals of Qualitative Research*. Understanding Qualitative

Research. New York: Oxford University Press.
Strauss, Anselm, and Juliet Corbin. 1990. *Basics of Qualitative Research*. Sage Publications.
Suchman, Lucy. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge University Press.
World Medical Association. 2013. "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects." *JAMA* 310 (20): 2191. https://doi.org/10.1001/jama.2013.281053.

Interview Protocol

Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Zachary Batist

2025-01-21

Interviews are oriented by my goal to document values and attitudes concerning data harmonization efforts, as elicited by research participants in their responses. Participants will be asked to reflect on:

- the motivations for their initiatives;
- the challenges they experience;
- how they envision success and failure;
- their perceptions of their own roles and the roles of other team members and stakeholders;
- the values that inform their decisions;
- how the technological apparatus they set up enables them to realize their goals and values; and
- ways in which they believe data-sharing could be improved.

To this end, each interview will proceed following a strategic order:

1. Participants' goals and perspectives

Follow a life-history method to better understand participants' professional backgrounds and their roles within their projects. The goal is to obtain information about their paths, not the rehearsed origin story.

- To start, can you please tell me a little bit about your background?
- What is the project, and what is your role?
- How did you find yourself in this role?
- How has your previous experience prepared you for your role?

2. Projects' missions, purposes, motivations

This section is about the project in general, including its purpose, scope and value. Information about practices and procedures will be sought in a subsequent phase of the interview.

- What are the project's goals?
- What makes the project unique?
- What is the project doing that no other similar project is doing?
- Do you consider this project as similar to any other initiatives?
- What are they, and in what ways are they simiar or different?

- What are the expected outcomes?
- Have you achieved these goals and outcomes?
- If not Are you on track to achieving them?
- What are some challenges that the project experienced, and how have you worked to overcome them?

3. Practices, procedures, relationships

This section asks about specific actions and interactions that the participant engages in.

Roles and relationships

- What does your role entail?
- Can you provide a couple examples of things that you recently did in this capacity?

- Who else do you frequently rely on, and what are their roles?
- Can you describe what they do, and perhaps give a few examples drawn from their recent work?

The interview might proceed in different ways depending on their initial responses. Here are some questions I might ask, corresponding with the participants' role and area of expertise.

Maintaining the project community

- Please briefly describe the process through which you obtain new partners or users.
- Can you please recall a recent example?

- How well do you know each partner?
- Did you know them before their involvement?

- Would you describe the project as a tight knit community, or more open-ended?

- How do you communicate with partners and contributors?
- What kinds of media or platforms do you use, and are they targeted for specific purposes? i.e. email, newsletters, social media, skype, personal communication at conferences

- Are there particular people in each project who you communicate with more frequently than others?
- Who are they, and why are these the people who you connect with?

- What do you consider your role or responsibility vis-a-vis the development/growth of this community?
- How do you foster the community's development and growth?
- Do you consider these efforts to be effective?

- Does your role as someone who leads a data harmonization initiative differentiate you from other epidemiologists?
- How has your relationship with other epidemiologists changed after initiating this project and taking on this role?

Reflections on data's value

- How has the data been used or analyzed?
- Do you track how the data is used?
- Is this tracking formal or informal?

- What patterns or trends are apparent based on this tracking?
- In your view, has the data been used in productive ways?
- In what ways are people either maximizing or not fully utilizing the data's full potential?

- Can you tell me about any unexpected or alternative uses of the data?
- What made them significant to you?

- Which skills and competencies do you think researchers need to possess in order to be able to make proper use of the data in their work?

- Based on your experience, what are the main obstacles for the effective and widespread adoption of these skills?
- What are some positive factors, or drivers, that can make that prospect more tangible?

Data ownership

- Who has rights (in the legal sense or informally) over the information contained in the system, or in related documents and datasets?
- Can you tell me about any conflicts or tensions that emerged relating to expressions of propriety or ownership over data?

Collecting data

- Do projects collect data with future harmonization in mind?
- If so, how does this affect data collection procedures, and does this play a role in subsequent decision-making?

Curating data

- Please describe the overall process of getting data into the system and then working with the data.

- Please tell me about any unexpected or problematic cases that made working with data particularly challenging.
- What made these cases unique or challenging?
- How did you resolve them or work towards a solution or viable outcome?

Accessing data

- Do you consider the system easy to access?
- Can you identify some challenges that pose as barriers to access?

- Who has access to data?
- How are decisions regarding access rights made?
- Can you tell me about any unnaceptable practices regarding accessing and sharing data?

Using data

- If you engage with the data with specific questions in mind, how do these questions emerge?
- What role does the data play in shaping the questions and analytical approach?

- Is the system amenable to exploratory or serendipitous modes of discovery?
- Please tell me about specific examples where you engaged with the data in this way.

- What features does the system have to view or export data?
- How easy is it to view, export or visualize data the data?
- Do you use the tools that are designed to export of visualize data, or do you prefer to use your own tooling?
- What are the reasons behind this preference?

Documentation

- How is the system documented?
- Who is responsible for creating documentation?
- Can you please tell me about a great example of documentation in your project?

- Overall, do you consider your project's documentation to be helpful?
- Why or why not?

- In your opinion, does the documentation accurately reflect the true nature of the documented data or work practices?
- Are specific things more accurately documented than others?
- Please tell me why you think some things are more accurately or less accurately documented.

- Can you recall any instances when documentation was updated?
- What prompted these updates?

- Do people ever get in touch to ask questions about specific aspects of the data or data curation procedures?
- What kinds of questions do they ask?
- What kinds of responses are given?

Relationships with Maelstrom

- Can you please concisely describe the role of Maelstrom as part of your project's overall initiative?

- What are the origins of your project's relationship with Maelstrom?
- How has this relationship changed over time?

- Does your project present any unique challenges or require any special attention?
- If so, please tell me about some unique cases or examples that demonstrate this unique relationship.

- Do you believe that Maelstrom is meeting your project's needs and enabling it to achieve its goals?
- In what ways is Maelstrom either satisfying or failing to meet your project's expectations or needs?
- How would you change the current system to better suit your project's needs more effectively?

- Do you engage with Maelstrom's other partners?
- If so, what is the nature of these relationships?

Ethics Protocol

Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Zachary Batist

2025-01-21

Project Title: Articulating epidemiological data harmonization initiatives as practical and collaborative experiences
Principal Investigator: Zachary Batist
Protocol Number: 25-01-057
Submitted:
Approved:

Recruitment and consent

Will this study involve recruitment of human study participants?

☒ Yes
☐ No

How are potential study participants identified and/or recruited to the study? Explain how potential participants are identified or introduced to the study, and who will recruit participants. Will the investigator/s require any special permissions or access to the target population e.g. clinic access, patient registries or records, mailing lists, community access?

Through consultation with key community stakeholders, the principal investigator will devise a list of prospective projects to serve as cases. The principal investigator will then write to the leaders of these projects inviting them to participate in the study. These invitations to project leaders will explain the project's purpose and scope, and will encourage the recipient to reply with any questions or concerns they may have. If they accept the invitation, the principal investigator will then work with project leaders to devise a list of individuals who may serve as interview candidates based on their roles in the project. The principal investigator will be clear with project leaders that they should not pressure those who work for them to participate in the study, and that individuals' participation should be treated as separate

from their regular duties; if project leaders cannot or will not abide by this condition, their project will be rejected as a prospective case. The principal investigator will then write to the recommended individuals to introduce the study and its objectives and to invite them to participate as research subjects. If these individuals express interest in participating in the study, the principal investigator will schedule a time to sit for an interview. Some interviews may be conducted remotely using internet-based video conferencing software, depending on participants' availability.

Describe the consent process. If alternate processes for seeking consent are planned (e.g. verbal, online, waiver), please provide a rationale and outline the procedure of obtaining and documenting consent and/or assent, where applicable.

Once individuals express their interest in participating, participants will provided with an informed consent document that outlines in more detail the goals of the study, the roles of the participant, how they will be recorded, how data pertaining to them will be retained, and the potential risks and benefits pertaining to their involvement. This document will also describe how participants' personally identifiable information will be managed and used. Participants will be asked to read and sign the document in order to obtain written informed consent. For interviews that will be held remotely using internet-based video conferencing software, participants will asked to send their signed informed consent documents in PDF format to the principal investigator. At the start of each interview the researcher will reiterate participants' rights and ask them to orally reaffirm their consent before proceeding.

Is there a relationship between the study participants and the person obtaining consent and/or the principal investigator/s?

☒ Yes
☐ No

If yes, please explain the nature of the relationship, and outline the steps that will be taken to avoid the perception of undue influence.

One project that serves as a case in this research is the Covid-19 Immunity Task Force (CITF), which the principal investigator currently serve as postdoctoral researcher. Some of the participants will therefore be his colleagues. The interviews will remain structured and limited in scope, and will not touch on matters relating to other aspects of their work. Moreover, prior to and throughout their involvement as research participants, frank and open discussion will be encouraged regarding collective expectations and to articulate the boundaries between participants' relationships with the principal investigator as colleagues and as research subjects.

The principal investigator will consult with David Buckeridge, who leads the CITF, as one key community stakeholder to help devise a shortlist of projects that may serve as prospective cases.

Risk-benefit assessment

Describe the foreseeable risks to study participants. What risks are attributable to the research, including cumulative risks? Which risks are participants normally exposed to in the course of their clinical care or in their daily activities as they relate to the research questions/objectives?

Participation in this study does not involve any physical, psychological or legal risks. However, the principal investigator will be asking participants to share detailed information about their work practices and work relationships, and public association with their responses may potentially disrupt or complicate their professional reputations. To mitigate against this potential harm, the principal investigator will give participants the option to render their responses confidential.

What procedures are in place to monitor and assess participant safety for the duration of the study?

Prior to each interview, and as part of the procedure for obtaining informed consent, participants will be asked about whether they want to render their responses confidential. Immediately after each interview, participants will be given an additional opportunity to reflect on their responses, and will be prompted to either confirm or alter their decision regarding whether or not to maintain confidentiality. Furthermore, for participants who have not requested that their responses be treated as confidential immediately before and after the interview, a follow-up email will be sent one week after the interview to reiterate the option to render their responses confidential.

Describe the potential benefits of the study for: (1) the study participants; (2) the population under investigation, and (3) the field of research.

This study contributes to the development of better epidemiological data-sharing infrastructures by articulating social, collaborative and discursive aspects of data harmonization, and how these factors relate to, overlap with or conflict with technical, institutional and epistemic factors. By explicitly framing data harmonization as a social and collaborative activity, we may devise more effective data-sharing infrastructures that better support the contextualization of data and enhance their value in contexts of data reuse. This work therefore poses new ways to document how epidemiologists mobilize distributed records in the constitution of synthetic knowledge and helps develop practical solutions that enable greater reflexivity. Additionally, this study may directly benefit participants by framing the experiences they address during interviews in ways that they might not have otherwise considered, thereby encouraging greater reflexivity in their own work.

Privacy and confidentiality

Please describe the measures in place for meeting confidentiality obligations. How is information and data safeguarded for the full cycle of the study: i.e. during its collection, use, dissemination, retention, and/or disposal?

The specific circumstances that frame each case are significant factors that will shape the findings, and the study will benefit from participants' consent to associate their identities with their interview responses. However, they may choose to render their interview responses confidential while maintaining their role a research participant. Participants may change their decision regarding whether or not to associate their identities with their interview responses up to one week after the interview, at which point the principal investigator will begin transcribing and analyzing the records pertaining to the interview. Participants will be reminded about this option immediately after the interview and one week following the interview via email.

The study engages with a relatively small community, and there is minimal social risk that others may be able to determine the identities of those whose research practices and professional relationships are being documented, even if their responses are rendered confidential. To address this issue, if any single participant from a case decides to render their responses confidential, the responses of all participants pertaining to that case will be rendered confidential as well, and the identify of the project that serves as the case will be obfuscated too.

In situations whereby a participant decides to render their responses confidential, or has their responses rendered confidential due to another member of their case deciding to do so, only the principal investigator will have access to records containing un-obfuscated information that may identify them. These un-obfuscated records, which may include audio and video records of interview sessions, as well as unedited transcripts and textual notes containing information that may reveal the participants' identities, will be kept in secure and encrypted media, and destroyed within five years of concluding the study, which provides sufficient time to revisit the data and produce additional research outputs. However, edited transcripts scrubbed of all information that may identify research participants may be kept, published and archived. If participants consent to maintaining association between their responses and their identities, un-obfuscated records and transcripts may be kept, published and archived.

The study is committed to adhering to fundamental data security practices, including those specified in McGill University's Cloud Directive which regulates the curation of sensitive research data. Physical records will be kept in a locked drawer in secure workspaces, either at McGill University's School of Public and Global Health or at the principal researcher's home office. Digital records will be stored on encrypted and password-protected drives and on secure servers approved or managed by McGill University under the Cloud Directive.

Recordings of remote interviews conducted using internet-based video conferencing software will be made using the software's built-in recording tools. Only video conferencing software approved by the Cloud Directive will be used. Participants will be instructed to disable their microphones or video cameras prior to initiating recording if they have opted to not be

recorded through these media. The researcher will record all media locally and refrain from using any cloud services to store or modify the records which the video conference software may provide.

If a contracted cloud/storage service provider or online survey tool is used, provide information on the service provider's security and privacy policy, location of its servers, data ownership, and what happens to the stored data after the contract is terminated. For more information, please consult the University's directive.

The study uses file-sharing software hosted by the Covid-19 Immunity Task Force at McGill University's School of Public and Global Health to backup all files maintained for this study. These backups will include files containing information that might reveal participants' identities. The software used to manage these backups is managed by McGill University and has been approved for storing sensitive research data by the Cloud Directive.

The study may use the secure GitLab instance hosted by the surveillance lab within the Clinical and Health Informatics Research Group at McGill University to store and track changes to sensitive research data. This software is managed by McGill University and has been approved for storing sensitive research data by the Cloud Directive.

The study maintains a website where the principal investigator shares documentation that supports the study and reflects on the work as it progresses. This is hosted using GitHub Pages and is backed up using Dropbox. No sensitive research data will pass through these services.

Recordings of remote interviews conducted using internet-based video conferencing software will be made using the software's built-in recording tools. Only video confering software approved by the Cloud Directive will be used. Participants will be instructed to disable their microphones or video cameras prior to initiating recording if they have opted to not be recorded through these media. The researcher will record all media locally and refrain from using any cloud services to store or modify the records which the video conference software may provide.

Please explain any reasonable and foreseeable disclosure requirements (e.g. disclosure to third parties such as government agencies or departments, community partners in research, personnel from an agency that monitors research, research sponsor, the REB/IRB, or regulatory agencies).

No disclosure requirements are foreseen.

If there are plans for retaining participant and/or study data for future use, please describe the context for its use, requirements for potentially re-contacting study participants and consent, and how the data will be stored and maintained for the long term.

Research data will be published in compliance with ethical standards for sharing open social science research data. Records that contain personally-identifying information pertaining to participants who have requested that their responses be rendered confidential and to those who

have had their responses rendered confidential due to another member of their case deciding to do so will not be published.

The database containing codings, memos and trends deriving from qualitative data analysis will be published only after being scrubbed of all personally-identifying information pertaining to participants who have requested that their responses be rendered confidential and to those who have had their responses rendered confidential due to another member of their case deciding to do so.

The principal investigator may follow up with the leaders of the data-sharing initiatives that serve as cases for this project to share the results with them and to present them with constructive feedback deriving from the study's findings. The principal investigator may also invite select participants to collaborate on a position paper advocating for reforms based on the project's findings.

Secondary use of data studies: if the study involves data linkage, please describe the data that will be linked and the likelihood that identifiable information will be created through the linkage.

This project does not rely on data deriving from other studies. The data may be reused in related work being undertaken under the same grant and by those who access the openly accessible data after they are published.

Managing conflicts of interest

Conflicts of interest do not imply wrong-doing. It is the responsibility of the investigator to determine if any conflicts apply to any person/s involved in the design and/or conduct of the research study or any member of their immediate family. Disclose all contracts and any conflicts of interest (real, perceived, or potential) relating to this research project. Conflict of interest may also arise with regard to the disclosure of personal health information.

- ☒ Not applicable. There are no conflicts of interest to disclose.
- ☐ Yes, there are conflicts of interest to disclose.

If yes, please describe the conflicts of interest (real, potential, and perceived), and the procedures for managing declared conflicts. Not applicable.

Participant Consent Form

Name of the Study: Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Sponsor: Canadian Institutes of Health Research / Public Health Agency of Canada 194216

Principal Researcher:
Zachary Batist
Postdoctoral Researcher
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health
McGill University
(438) 507-8685
zachary.batist@mcgill.ca

Faculty Supervisor:
David Buckeridge
Professor
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health
McGill University
(514) 934-1934 ext. 32991
david.buckeridge@mcgill.ca

Purpose: You are invited to participate in this study which investigates the social, technical, administrative and epistemic factors that scaffold data-sharing initiatives in epidemiological research. Data harmonization is one means of data-sharing that draws different sets of recorded observations into a unified formal schema, whose structure is driven by specific objectives and more general underlying suppositions and values. Although these schemas are arrived at through discussion, compromise and consensus-building, with an eye toward practical outcomes afforded by the data model, these socially mediated interactions are generally under-recognized as important factors contributing to data-sharing initiatives' success relative to their potential impact. The goal of this study is to survey how various factors are prioritized in harmonization activities, the rationales behind these decisions, and the relative efficacy of different approaches to data-sharing.

Study Procedures: If you choose to participate in this study, you will be asked to sit for an interview where you will be asked a series of questions concerning your experiences, attitudes and behaviours relating to data harmonization activities you have engaged in. With your consent, the interviews will be recorded using audio, video and textual media. Audio and video records will be transcribed, and all original records and transcripts may contribute to a corpus suitable for qualitative data analysis. You may decide to opt out of being recorded by any media type.

Voluntary Participation: Your participation in this study is voluntary. You may decline to answer any question or take part in any procedure. You may withdraw from the study at any time, for any reason. If you decide to withdraw from the study, information generated because of your participation will be destroyed, unless you give permission to retain those them. However, withdrawal is not possible after qualitative data analysis begins, since your interview responses will have already informed ongoing theory-building. Participation in this study is entirely independent of your regular work activities, and your responses will not be shared with your employer or other members of your research group. If the leader of the project that constitutes the case in which you are a member cannot or will not abide by this condition, the project will be rejected as a case and all information produced through interviews from participants who are members of the case will be destroyed. If you decide to render your responses confidential, the responses of all participants who are members of the same case will also be rendered confidential, and the identity of the project that serves as the case will be obfuscated as well.

Potential Risks: Although your participation in this study does not involve any physical, psychological or legal risks, I acknowledge the fact that you are part of a small community and that there is minimal social risk that others may be able to determine your identity, even if you decide to render your responses confidential. To mitigate this risk, if you decide to render your responses confidential, pseudonyms will be used instead of your real name in all publications and presentations of my findings; this will apply to all participants who are members of the same case, and the identity of the project that serves as the case will be obfuscated as well. This is meant to counteract the possibility of your identity being determined based

upon your association with the project and other related participants, which might be common knowledge. Additionally, interviews will be held in a private setting, and I welcome any suggestions that would enable you to feel more comfortable that are also conducive to the methods I employ to collect meaningful data.

Benefits: This study contributes to the development of better epidemiological data-sharing infrastructures by articulating social, collaborative and discursive aspects of data harmonization, and how these factors relate to, overlap with or conflict with technical, institutional and epistemic factors. By explicitly framing data harmonization as a social and collaborative activity, we may devise more effective data-sharing infrastructures that better support the contextualization of data and enhance their value in contexts of data reuse. This work therefore poses new ways to document how epidemiologists mobilize distributed records in the constitution of synthetic knowledge and helps develop practical solutions that enable greater reflexivity. Additionally, this study may directly benefit you by framing the experiences you address during interviews in ways that you might not have otherwise considered, thereby encouraging greater reflexivity in your own work.

Compensation: You will not be compensated for participating in this study.

Confidentiality: The specific circumstances that frame each case are significant factors that will shape the findings, and the study will benefit from your consent to associate your identity with your interview responses. However, you may choose to render your interview responses confidential as a research participant. You may change your decision regarding whether or not to associate your identity with your interview responses up to one week after the interview, at which point the researcher will begin transcribing and analyzing the records pertaining to your interview. You will be reminded about this option immediately after the interview and one week following the interview via email.

If you decide to render your responses confidential or have your responses rendered confidential due to another member of your case deciding to do so, only the principal researcher (Zachary Batist) will have access to records containing un-obfuscated information that may identify you. These un-obfuscated records, which may include audio and video records of interview sessions, as well as unedited transcripts and textual notes containing information that may reveal the participants' identities, will be kept in secure and encrypted media, and destroyed within five years of concluding the study, which provides sufficient time to revisit the data and produce additional research outputs. However, edited transcripts scrubbed of all information that may identify you may be kept, published and archived. If you consent to maintaining association between your responses and your identity, un-obfuscated records and transcripts may be kept, published and archived.

The study is committed to adhering to fundamental data security practices, including those specified in McGill University's Cloud Directive which regulates the curation of sensitive research data. Physical records will be kept in a locked drawer in secure workspaces, either at McGill University's School of Public and Global Health or at the principal researcher's home office. Digital records will be stored on encrypted and password-protected drives.

Dissemination of Results: The findings deriving from this study will be published in peer-reviewed journals, at conferences, and in media intended for lay-audiences (e.g. blogs, podcasts, social media). The findings may also serve as a source of evidence in policy briefs or recommendations to enhance the infrastructures and policy mandates that govern open data-sharing practices.

Questions: Please contact the principal researcher, Zachary Batist, if you have any questions about the study, either by phone at (438) 507-8685 or by email at zachary.batist@mcgill.ca.

If you have any ethical concerns or complaints about your participation in this study, and want to speak with someone not on the research team, please contact the Research Ethics Board Office, daniel.tesolin@mcgill.ca or 514-398-5410, citing REB file number 25-01-057.

you can contact the Office of Research Ethics at the University of Toronto at ethics.review@utoronto.ca or 416-946-3273.

Consent to Participate Statement:
Please sign below if you have read the above information and consent to participate in this study. Agreeing to participate in this study does not waive any of your rights or release the researchers from their responsibilities. To ensure the study is being conducted properly, authorized individuals, such as a member of the Research Ethics Board, may have access to your information. A copy of this consent form will be given to you and the researcher will keep a copy.

I agree that the researcher can audio-record our interview.
Yes ___ No ___

I agree that the researcher can video-record our interview.
Yes ___ No ___

I agree that the researcher can take written notes during our interview.
Yes ___ No ___

I agree maintain association between my interview responses and my identity.
Yes ___ No ___

Participant Name (please print): _____

Signature: _____

Date: _____

[date]

Dear [recipient's name],

I am Zachary Batist, a Postdoctoral Researcher at McGill University's Department of Epidemiology, Biostatistics and Occupational Health. As a scholar of scientific practice, my research is concerned with collaborative aspects of data work, records management and open science. More specifically, I'm interested in how scientists contribute to and govern collectively maintained information commons, and the technical, administrative and social structures that scaffold these efforts. To this end, I'm working with David Buckeridge on a study to articulate the motivations, values and challenges that inform epidemiological data harmonization initiatives through a series of interviews with people who lead and support these activities, and I'm writing to invite you to participate.

The study is centred around 3-5 cases, with each case representing a data harmonization initiative, such as [initiative's name]. If you decide to participate, [initiative's name] will serve as one of these cases, which will allow me to relate the specific motivations, challenges and values that shaped your work to similar factors experienced by other cases. To be clear, the goal is not to judge your performance in any way, but rather to ascertain what actions specific strategies entail, the circumstances in which each is adopted, the value that they bring, and the trade-offs involved. In other words, my goal is to reveal the diverse ways in which data harmonization occurs, and how different approaches impact the outcomes in different ways.

My intervention will take the form of a series of interviews with selected members of your team. During these interviews, I will ask participants about their professional backgrounds, their roles in [initiative's name], and the work they perform in fulfillment of [initiative's name]'s objectives. I aim to conduct interviews with 3-5 individuals who coordinate the project, foster productive relationships with collaborators, prepare data for harmonization, and draw analytic value from the integrated dataset. As the project leader, I would consult with you to determine who might serve as viable candidates to sit for an interview.

The study has been approved by McGill University's IRB (file #25-01-057), and I've prepared extensive documentation about the study's methods and procedures, which are available on a dedicated project website: https://zackbatist.info/CITF-Postdoc. I'd be happy to meet to discuss any questions or concerns you might have before deciding whether you want to participate.

In any case, I look forward to hearing back from you soon,

Zachary Batist, PhD
zachary.batist@mcgill.ca
Postdoctoral Researcher | CITF Databank
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health, McGill University

[date]

Dear [recipient's name],

I am Zachary Batist, a Postdoctoral Researcher at McGill University's Department of Epidemiology, Biostatistics and Occupational Health. As a scholar of scientific practice, my research is concerned with collaborative aspects of data work, records management and open science. More specifically, I'm interested in how scientists contribute to and govern collectively maintained information commons, and the technical, administrative and social structures that scaffold these efforts. To this end, I'm working with David Buckeridge on a study to articulate the motivations, values and challenges that inform epidemiological data harmonization initiatives through a series of interviews with people who lead and support these activities, and I'm writing to invite you to participate due to your involvement with [initiative's name].

The study is centred around interviews with participants from 3-5 cases, with each case representing a data harmonization initiative, such as [initiative's name]. Through a series of interviews with various members of the [initiative's name] team, I aim to identify and articulate the diverse ways in which data harmonization occurs, and how different approaches impact the outcomes in different ways. More specifically, I will ask participants about their professional backgrounds, their roles in [initiative's name], and the work they perform in fulfillment of [initiative's name]'s objectives. To be clear, the goal is not to judge your performance in any way, but rather to ascertain what actions specific strategies entail, the circumstances in which each is adopted, the value that they bring, and the trade-offs involved.

I received your contact information from [initiative's leader], who encouraged me to get in touch with you based on your role as [recipient's role]. This study is completely independent of your work at [initiative's name], and you are not obliged to participate.

The study has been approved by McGill University's IRB (file #25-01-057), and I've prepared extensive documentation about the study's methods and procedures, which are available on a dedicated project website: https://zackbatist.info/CITF-Postdoc. I'd be happy to meet to discuss any questions or concerns you might have before deciding whether you want to participate.

In any case, I look forward to hearing back from you soon,

Zachary Batist, PhD
zachary.batist@mcgill.ca
Postdoctoral Researcher | CITF Databank
Department of Epidemiology, Biostatistics and Occupational Health
School of Population and Global Health, McGill University

McGill | Faculty of Medicine and Health Sciences | Faculté de médecine et des sciences de la santé

Research Ethics (IRB)

Scientific / Peer Review Form

If your protocol has not been reviewed by a recognized scholarly or granting agency, please provide a scientific review from a recognized independent authority in your field or from your trainee's supervisory/advisory committee (see #8, below). **The completed scientific/peer review must be included with your ethics submission. Your submission will not be reviewed until it is complete.**

The person doing the scientific/peer review must have the expertise in the field and have no conflict of interest, so as to be able to give an **independent and unbiased** review.

The scientific/peer review must respond specifically to the questions below. The reviewer is requested to elaborate on any issues identified and to avoid a "yes/no" response:

Reviewer's name: David Buckeridge

Title: Professor, Department of Epidemiology, Biostatistics and Occupational Health, School of Population and Global Health, McGill University

Signature: 

Date: 2025-01-29

Title of Research Proposal and PI: Articulating epidemiological data harmonization initiatives as practical and collaborative experiences, Zachary Batist

1. What is the purpose of the study and how does the investigator propose to answer the research questions?

The study's objective is to articulate the social, technical, administrative and epistemic factors that scaffold efforts to harmonize epidemiological research data. Through qualitative analysis of a series of interviews with individuals who play key roles in data harmonization initiatives (including project leaders, coordinators and support workers), the study will investigate the values and objectives that inform these initiatives, as well as the challenges they experience and the strategies they implement to overcome them.

2. How do the proposed measures relate to the stated purpose and hypotheses?

The methodology is very well articulated and provides sufficient depth to explain the rationale behind the study's approach. The measures are well thought-out and are appropriate for achieving the study's aims.

3. How will the study achieve the stated goals? If the goals are achieved, how will the information contribute to the advancement of knowledge?

The study will implement qualitative data analysis methods following an abductive grounded theory approach to ascertain the values, actions and challenges elicited by research participants during a series of semi-structured interviews. This will enable the principal investigator to construct theories grounded in qualitative evidence that articulate and explain some challenges that are commonly experienced in data harmonization initiatives. This will contribute to constructive critique of the infrastructures, policies and protocols that are currently in place to support data harmonization work and will inform ongoing efforts to improve these systems.

4. Is the research design appropriate to the goals of the study?

- Will there be an intervention or treatment? What type of comparisons will be made?
- What procedures will be used to control variables?
- Are the risks associated with the procedures necessary and reasonable, and are they adequately assessed?
- If there are interventional, biochemical and/or physiological measures, how are these justified, and how is the level of risk documented?
- If study measures are to be used (i.e. questionnaires, instruments, tests) are they appropriate to the research questions, and do the concepts and measures correspond? Are they standard for this field of research, and how is the level of risk documented?

The methods outlined in the research protocol are appropriate means of achieving the objectives. They correspond with commonly accepted practices in the field of Science and Technology Studies to ascertain the values, objectives and strategies adopted by collective scientific endeavors.

The methods do not call for any kind of comparative treatment. As such, there is no need to control variables as would be common practice in clinical trials. An adequately detailed risk assessment is provided in the ethics protocol. No interventional, biochemical and/or physiological measures are to be implemented.

5. Is the study population clearly defined, and is it appropriate to the stated purpose?
    - Are inclusion and exclusion criteria presented?
    - With respect to the population being studied, how have any scientific or ethical challenges been explained? How will these challenges be handled?
    - What steps are taken to mitigate selection bias?

The study population is clearly defined and is appropriate to the stated purpose. Inclusion and exclusion criteria are clearly presented. The research protocol identifies some potential scientific challenges, namely with regards to framing and scope, and with ensuring that the sample is representative of a broader phenomenon of interest. The protocol clearly articulates the rationale behind how cases will be selected and the suitability of the case selection procedure to the study's overall methodology. Ethical challenges are well thought out and strike an acceptable balance between reducing risk of harm and enabling the study to achieve its objectives.

6. Are recruitment procedures specified? Describe any conflicts of interest that might affect the recruitment procedures.

Recruitment procedures are clearly specified. The ethics protocol identifies one potential conflict of interest whereby the principal investigator's supervisor leads an initiative that may serve as one prospective case. Data harmonization initiatives will be selected as based on the potential insight that may be obtained through their comparison, and not necessarily due to prior relationships with the principal investigator.

If this initiative is selected as a case, potential interviews may include the principal investigator's colleagues. This potential conflict will be mitigated by holding open discussion with the initiative's research team prior to and throughout their involvement in the study regarding their collective expectations and to articulate the boundaries between their relations as colleagues and as research subjects. This strategy for resolving the potential conflict is acceptable.

7. Is an explicit plan for data analysis provided?
    - How will the planned analysis respond to the research question?
    - Will this plan indicate exactly how the data to be collected will be utilized in statistical analysis?
    - Is there a justification provided for any data which will not be specifically utilized in statistical analysis?

- Is there a formal sample size calculation? How did the investigators arrive at the planned sample size, and how likely is it to provide statistical validity?

The plan for data collection and analysis is clearly articulated. Through qualitative data analysis following an abductive grounded theory approach, the study will devise theory grounded in qualitative evidence that articulates and explains challenges involved in data harmonization work.

The plan does highlight the limited role that statistical methods will play and how data may or may not be used to this end; basic summary statistics will be used to represent the distribution of codings across individual interviews or ranges of interviews, which will help to identify trends and associations as they pertain to their limited scopes. This will be used to support the principal investigator's theory-building but will not be used to infer generalizable causal relationships.

There is no formal sample size calculation. The sample size was arrived at through the principal investigator's prior experience implementing similar qualitative research projects, and recognition of the practical constraint of having to complete this project during the short period in which funding has been allotted. Since this study is rooted in case study methods, and the goal is to articulate the series of inter-woven factors that impact how epidemiological researchers coordinate and participate in data-sharing initiatives while explicitly accounting for and drawing from the unique and situational contexts that frame each case, the goal is not to define causal relationships or to derive findings that may be generalized across the whole field of epidemiology. As such, statistical representativeness is not a concern.

SECTION TO BE COMPLETED BY STUDENT'S/TRAINEE'S SUPERVISOR/SUPERVISORY COMMITTEE

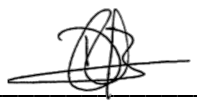
8. If this is a protocol submitted on behalf of a graduate student or other trainee, the student's supervisory committee should complete the scientific review by **providing responses to each of the questions above**, and the supervisor should submit the completed form, along with a signed copy of the following statement:

Oversight
I agree that I will provide oversight of this protocol by doing the following:

- Overseeing the design and conduct of the study
- Ensuring that the student/trainee is trained and competent to perform the required activities
- Reviewing the protocol application prior to submission to the REB
- Providing guidance in the protection of research subjects
- Assuring proper record keeping and reporting to the REB and other relevant bodies
- Working with students/trainees to identify modifications which may be needed, unanticipated problems or circumstances involving risks to participants or others

David Buckeridge, Professor
Name and title of Academic Supervisor or Supervisory Committee

2025-01-29
Date



Signature of Academic Supervisor

McGill | Faculty of Medicine and Health Sciences | Faculté de médecine et des sciences de la santé

IRB/Research Ethics Office
www.mcgill.ca/medresearch/ethics

Protocol Number: 25-01-057

Principal Investigator: Zachary Batist

Study title: Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

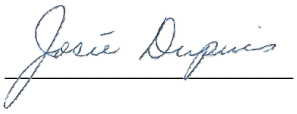
Departmental Chair / Director: Josée Dupuis

This signature confirms that:

1. The Department/School is aware of this application and agree with its submission.

2. The Department/School acknowledges the roles and responsibilities imparted in McGill University's Policy on the Ethical Conduct of Research Involving Human Participants and the Tri-Council Policy Statement *Ethical Conduct of Research Involving Humans* (TCPS2).

Date: January 30, 2025

Signature: 

PANEL ON RESEARCH ETHICS

Navigating the ethics of human research

TCPS 2: CORE 2022

Certificate of Completion

This document certifies that

Zachary Batist

successfully completed the Course on Research Ethics based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022)

Certificate # 0001391231

22 January, 2025