Leveraging expert knowledge and public data to inform health interventions: Novel practical approaches for building causal directed acyclic graphs

Stephanie Long

Department of Family Medicine,

McGill University, Montreal, Quebec, Canada

A thesis submitted to McGill University, Faculty of Medicine and Health Sciences in partial fulfilment of the requirements of the degree of Doctor of Philosophy (PhD) in Family Medicine and Primary Care.

Table of Contents

ABSTRACT

BACKGROUND:

Epidemiological and biostatistical methods applied in primary care research are evolving rapidly due to increasingly complex health data and limitations of traditional associational inference approaches. The COVID-19 pandemic accelerated this change, necessitating a critical re-evaluation of both conventional and modern methodologies. Integrating innovative approaches such as causal inference, adaptive trials, machine learning, and - most recently - large language models with traditional data analytical methods offers promising new possibilities for addressing complex research questions more efficiently and effectively. However, this integration presents practical challenges in implementation, as modern research paradigms increasingly leverage domain expert knowledge i.e., external information critical for informing structural and operational aspects of research. Within the modern causal inference archetype, directed acyclic graphs (DAGs) are central to encoding external information and are hence key in enabling effective knowledge integration for the practice of primary care research.

This doctoral research assessed current methodological shortcomings in integrating public data and expert knowledge in primary care research studies through a causal inference lens. Through a series of framework developments centred on DAGs, this dissertation addressed these challenges, demonstrating their feasibility and applicability in the context of chronic disease management and adaptive trials.

OBJECTIVES:

1. Evaluate epidemiological reporting standards using COVID-19 as a case study, with a focus on assessing the limitations of causal and actionable interpretations of data reported to the public.
2. Assess the potential of large language models in building directed acyclic graphs (DAGs) leveraging the vast corpus of public data for primary care research.
3. Establish a causal mapping approach of the HIV literature to identify frequently reported HIV-related individual-level outcomes with the goal of constructing a comprehensive DAG of HIV outcomes reported in the literature.

4. Develop and evaluate the feasibility of a novel approach for DAG development with domain experts.

METHODS:

Four studies were conducted using multiple methodological research approaches:

1. A longitudinal (real-time) critical appraisal of the causal utility of the Canadian COVID-19 data reporting of governmental bodies and news outlets between April 2020 and August 2021.
2. An empirical study of the utility of the large language model, GPT-3, an early pre-ChatGPT large language model, for building DAGs for primary care research.
3. A scoping review to develop a DAG describing relationships with HIV-related individual-level outcomes used in research studies conducted in high-income countries.
4. A feasibility study to develop and evaluate of an alternative approach to DAG development with domain experts, with a secondary goal of updating the DAG created in study 3.

RESULTS:

Study 1: Canadian COVID-19 data reporting exhibited varying case definitions, heterogeneous testing criteria, and lack of appropriate standardization. These findings highlight challenges in applying established epidemiological principles and their impact on public health policy.

Study 2: GPT-3's performance was promising, achieving greater than 50% accuracy on at least one of the tested settings (e.g., prompt, link verb, specificity of language). This demonstrated that GPT-3's accuracy in confirming edges in health-related DAGs depend on the language used in prompts used to describe relationships between variables e.g., "X is caused by Y" or "X is associated with Y". While LLMs show potential utility in extracting information from public data, combining this with expert knowledge and literature may offer a more efficient means to generate comprehensive DAGs.

Study 3: The scoping review found that physical health outcomes were the most frequently reported in HIV research studies, followed by social health, with limited focus on mental health.

Only 2.2% of included studies used surrogate outcomes, with CD4+ cell count being the most frequent, acting as a proxy for overall immune function or retention in care. The findings were used to create a DAG illustrating relationships amongst HIV-related outcomes. This initial DAG's predominant focus on physical and clinical outcomes highlighted a lack of known psychosocial and structural outcomes, illustrating the need for additional expert involvement when generating DAGs.

Study 4: Based on the findings of Study 3, an alternative DAG development approach with domain expert was proposed and iteratively developed based on domain expert feedback: (1) identify outcomes of interest, (2) elicit variables, (3) organize variables temporally, (4) consolidate domain expert data and create a DAG, and (5) review and finalize the DAG with domain experts. Seven DAG development sessions were conducted with seven domain experts to update the baseline DAG from Study 3, particularly focusing on engagement in care and ART adherence. Domain experts found the process stimulating, essential, and clear. They updated the DAG adding social and structural factors influencing the outcomes of interest – resulting in a more comprehensive DAG.

CONCLUSION:

This doctoral thesis highlights the evolving nature of knowledge synthesis approaches that aim to inform modern causal inference for health interventions. By critically appraising conventional data reporting practices, exploring the potential of machine learning in causal modeling, and developing a novel approach to incorporating domain expertise with DAGs, this work provides a series of methodological developments promoting the formal integration of expert knowledge and public data for causal model building. The findings underscore the importance of balancing technological advances with domain expertise, offering a pathway to more robust and contextually relevant primary care research.

RESUMÉ

CONTEXTE :

Les méthodes épidémiologiques et biostatistiques appliquées à la recherche en soins primaires évoluent rapidement en raison de la complexité croissante des données sanitaires et des limites des approches traditionnelles d'inférence associative. La pandémie de COVID-19 a accéléré ce changement, nécessitant une réévaluation critique des méthodologies conventionnelles et modernes. L'intégration d'approches innovantes telles que l'inférence causale, les essais adaptatifs, l'apprentissage automatique et, plus récemment, les grands modèles de langage avec les méthodes traditionnelles d'analyse des données offre de nouvelles possibilités prometteuses pour répondre à des questions de recherche complexes de manière plus efficace et efficiente. Cependant, cette intégration présente des défis pratiques dans la mise en œuvre, car les paradigmes de recherche modernes exploitent de plus en plus les connaissances des experts du domaine, c'est-à-dire des informations externes essentielles pour informer les aspects structurels et opérationnels de la recherche. Dans l'archétype moderne de l'inférence causale, les graphes acycliques dirigés (GAD) jouent un rôle central dans l'encodage des informations externes et sont donc essentiels pour permettre une intégration efficace des connaissances dans la pratique de la recherche sur les soins primaires.

Cette recherche doctorale a évalué les lacunes méthodologiques actuelles dans l'intégration des données publiques et des connaissances des experts dans les études de recherche sur les soins primaires à travers une lentille d'inférence causale. À travers une série de développements de cadres centrés sur les GAD, cette dissertation a abordé ces défis, démontrant leur faisabilité et leur applicabilité dans le contexte de la gestion des maladies chroniques et des essais adaptatifs.

OBJECTIFS :

1. Évaluer les normes de déclaration épidémiologique en utilisant COVID-19 comme étude de cas, en mettant l'accent sur l'évaluation des limites des interprétations causales et exploitables des données communiquées au public.
2. Évaluer le potentiel des grands modèles de langage dans la construction de graphes acycliques dirigés (GAD) en tirant parti du vaste corpus de données publiques pour la recherche sur les soins primaires.

3. Établir une approche de cartographie causale de la littérature sur le VIH afin d'identifier les résultats fréquemment rapportés par les patients en rapport avec le VIH, dans le but de construire un GAD complet des résultats du VIH rapportés dans la littérature.
4. Développer et évaluer la faisabilité d'une nouvelle approche pour le développement de GAD avec des experts du domaine.

MÉTHODES :

Quatre études ont été réalisées à l'aide d'approches méthodologiques multiples :

1. Évaluation critique longitudinale (en temps réel) de l'utilité causale des données canadiennes COVID-19 communiquées par les organismes gouvernementaux et les organes de presse entre avril 2020 et août 2021.
2. Une étude empirique de l'utilité du grand modèle linguistique GPT-3, un modèle linguistique antérieur à ChatGPT, pour la construction de GAD dans le cadre de la recherche sur les soins primaires.
3. Un examen approfondi pour développer un GAD décrivant les relations avec les résultats individuels liés au VIH utilisés dans les études de recherche menées dans les pays à revenu élevé.
4. Une étude de faisabilité pour développer et évaluer une approche alternative au développement de GAD avec des experts du domaine, avec un objectif secondaire de mise à jour du GAD créé dans l'étude 3.

RÉSULTATS :

Étude 1 : Les données canadiennes relatives à l'étude COVID-19 présentaient des définitions de cas variables, des critères de test hétérogènes et un manque de normalisation appropriée. Ces résultats mettent en évidence les difficultés d'application des principes épidémiologiques établis et leur impact sur la politique de santé publique.

Étude 2 : Les performances du GPT-3 étaient prometteuses, avec une précision supérieure à 50 % pour au moins l'un des paramètres testés (par exemple, l'invite, le verbe de liaison, la spécificité du langage). Cela a démontré que la précision du GPT-3 dans la confirmation des

arêtes dans les GAD liés à la santé dépend du langage utilisé dans les invites pour décrire les relations entre les variables, par exemple "X est causé par Y" ou "X est associé à Y". Bien que les LLM soient potentiellement utiles pour extraire des informations des données publiques, leur combinaison avec des connaissances d'experts et de la littérature peut constituer un moyen plus efficace de générer des GAD complets.

Étude 3 : L'examen exploratoire a révélé que les résultats en matière de santé physique étaient les plus fréquemment rapportés dans les études de recherche sur le VIH, suivis par la santé sociale, avec une attention limitée portée à la santé mentale. Seules 2,2 % des études incluses ont utilisé des résultats de substitution, la numération des cellules CD4+ étant la plus fréquente, en tant qu'indicateur de la fonction immunitaire globale ou de la rétention dans le système de soins. Les résultats ont été utilisés pour créer un GAD illustrant les relations entre les résultats liés au VIH. L'accent prédominant mis par ce premier GAD sur les résultats physiques et cliniques a mis en évidence le manque de résultats psychosociaux et structurels connus, illustrant la nécessité d'une implication supplémentaire des experts lors de la création des GAD.

Étude 4 : Sur la base des résultats de l'étude 3, une approche alternative de développement de DAG avec des experts du domaine a été proposée et développée de manière itérative sur la base des commentaires des experts du domaine : (1) identifier les résultats d'intérêt, (2) éliciter les variables, (3) organiser les variables dans le temps, (4) consolider les données des experts du domaine et créer un GAD, et (5) examiner et finaliser le GAD avec les experts du domaine. Sept sessions de développement du GAD ont été menées avec sept experts du domaine pour mettre à jour le GAD de base de l'étude 3, en se concentrant particulièrement sur l'engagement dans les soins et l'adhésion au traitement antirétroviral. Les experts ont trouvé le processus stimulant, essentiel et clair. Ils ont mis à jour le GAD en ajoutant des facteurs sociaux et structurels influençant les résultats d'intérêt, ce qui a permis d'obtenir un GAD plus complet.

CONCLUSION :

Cette thèse de doctorat met en évidence la nature évolutive des approches de synthèse des connaissances qui visent à informer l'inférence causale moderne pour les interventions de santé. En évaluant de manière critique les pratiques conventionnelles de communication des données, en explorant le potentiel de l'apprentissage automatique dans la modélisation causale, et en développant une nouvelle approche pour incorporer l'expertise du domaine avec les DAG, ce

travail fournit une série de développements méthodologiques promouvant l'intégration formelle de la connaissance des experts et des données publiques pour la construction de modèles causaux. Les résultats soulignent l'importance d'équilibrer les avancées technologiques avec l'expertise du domaine, offrant une voie vers une recherche en soins primaires plus robuste et contextuellement pertinente.

ACKNOWLEDGMENTS

This work would not have been possible without the support of many people. First and foremost, I owe my deepest appreciation to my primary supervisor, Dr. Tibor Schuster. Thank you for all the support and guidance, which have been instrumental in shaping my development as a researcher. Your passion and commitment to advancing knowledge and your dedication to excellence and rigour have consistently inspired me throughout this journey. I am so grateful for our many insightful discussions about this work and your timely, constructive encouragement during challenging times. It has been an absolute privilege working under your mentorship.

Thank you to my co-supervisor, Dr. Bertrand Lebouché, and my advisory committee, Drs. Alexandra de Pokomandy and Aude Motulsky, for your feedback and support. Your insights and constructive feedback have help shape the direction of this work. Thank you to Kim Engler for your valuable feedback on this work.

To my PhD colleagues, Lashanda Skerritt and Sarah Aboushawareb, this academic journey would not have been the same without you. Thank you for your unwavering support and understanding, you made the hard days easier and the easy days even better. I am so grateful for our many stimulating discussions over plates of delicious food. Thank you also to Sophia Siedlikowski, Nickoo Merati, Mary Henein, and Reem El Sherif for the support, laughs, and encouragement.

Thank you to my partner and collaborator, Alexandre Piché. I could not imagine going through this PhD journey without you by my side. Thank you for your love, your unwavering support, and being my light on dark days. To my beloved Shiba inu, Yuki, thank you for your stress-melting cuddles, you always knew how to make me feel better.

To my parents, Charlotte and Francis Long. This achievement is your achievement. Thank you for giving me the tools I needed for success. I am forever grateful for your endless support and love. A huge thank you to my dear sister, Jennifer Long for always being there for me and for all the adventures!

Acknowledgments for funding support

This doctoral research was funded by a Fonds de recherche du Québec – Santé (FRQS) Doctoral Training Award. This work was also supported by trainee scholarships from Graduate and Postdoctoral Studies at McGill University, the Department of Family Medicine at McGill University, and funds from my supervisors, Dr. Tibor Schuster and Dr. Bertrand Lebouché. I also received travel awards from the Canadian Institutes of Health Research (CIHR) and the Department of Family Medicine to attend and present at international conferences within Canada, the United States, and Europe.

LIST OF ABBREVIATIONS

HIV: *Human immunodeficiency virus*

PLWH: *People living with HIV*

ART: *antiretroviral therapy*

RCT: *randomized control trial*

DAG: *directed acyclic graph*

ML: *machine learning*

AI: *artificial intelligence*

LLM: *large language model*

LIST OF FIGURES

LIST OF TABLES

STATEMENT OF ORIGINALITY AND AUTHOR CONTRIBUTIONS

Contribution to original knowledge

This is a manuscript-based doctoral dissertation comprised of four manuscripts; two of which have been published, the two remaining are ready for submission to peer-reviewed journals. The work described in this dissertation presents original research and an original contribution to epidemiological and causal inferences methods applied in primary care research informing the development of health interventions.

In Manuscript 1 (Chapter 4), I evaluated Canadian epidemiological reporting during the COVID-19 pandemic longitudinally, specifically assessing its causal utility for informing health policy. The findings demonstrate that there are still challenges in applying established epidemiological principles to public health data, particularly in rapidly evolving situations such as the unprecedented COVID-19 pandemic. To my knowledge, this was one of the first studies examining COVID-19 data reporting in Canada.

Manuscript 2 (Chapter 5) was an empirical investigation of the utility of an early large language model, GPT-3, a precursor of ChatGPT, in constructing directed acyclic graphs (DAGs) for primary care research. This work was initiated, published, and presented prior to the release of ChatGPT at the end of November 2022. As such, it was one of the first papers exploring the use of large language models in causal modelling.

Manuscript 3 (Chapter 6) is a scoping review that explored the types of HIV-related individual outcomes reported in HIV studies, while also focusing on the use of surrogate outcomes in this context. This manuscript developed a directed acyclic graph illustrating the causal relationships among frequently used HIV outcomes, which informed Manuscript 4 (Chapter 7). This review developed the DAG illustrating the relationships amongst commonly reported HIV outcomes, provided an overview of surrogate outcome use and an exploration of prominent HIV-related individual outcomes used. From the findings, it became clear that research in this domain still primarily focused on physical or clinical aspects of health, though more social and mental health outcomes were becoming more prominently used.

In Manuscript 4 (Chapter 7), I developed and assessed the feasibility of a novel approach to DAG development with domain experts. In this study, the DAG constructed in Manuscript 3

(Chapter 6) was updated with domain experts, resulting in the creation of a more comprehensive DAG acknowledging social- and mental-health related and structural aspects contributing to HIV care. Further, this study expands on limited, but existing literature on involving domain expertise in DAG development, by presenting a practical approach cognizant of time commitments and cognitive burden. To my knowledge, this is one of very few studies describing an approach to involving domain experts in DAG construction.

Contribution of authors

As a doctoral candidate and the first author of all the chapters and manuscripts included in this doctoral dissertation, I was responsible for the conception of the thesis, research design, analysis, interpretation, presentation of results, and writing of each manuscript in this thesis. The direction of this research work was guided by my primary supervisor, Dr. Tibor Schuster, who provided extensive guidance and support in the conceptualization of this thesis through many engaging discussions.

A list of all four manuscripts with specific author contributions is provided below:

Manuscript 1: Limitations of Canadian COVID-19 data reporting to the general public

Stephanie Long; David Loutfi; Jay Kaufman; Tibor Schuster

Long, S., Loutfi, D., Kaufman, J. S., & Schuster, T. (2022). Limitations of Canadian COVID-19 data reporting to the general public. *Journal of Public Health Policy, 43*(1), 203–221. https://doi.org/10.1057/s41271-022-00337-x

SL and TS conceptualized the study and performed data extraction. SL conducted the statistical analysis, data visualization, and led the interpretation of the findings with input from TS, JK, and DL. SL drafted the manuscript. JK provided epidemiological expertise and guidance. All authors revised the manuscript and approved the final version for publication.

Manuscript 2: Can large language models build causal diagrams?

Stephanie Long; Tibor Schuster; Alexandre Piché

Long S, Piché A, Schuster T. (2023) *Can large language models build causal graphs?* NeurIPS 2022 Workshop on Causal Machine Learning for Real-World Impact (CML4Impact, 2022), New Orleans, USA.

SL and AP conceptualized the study design, analysis, and interpretation of the data. SL developed the DAGS on which the LLM were evaluated upon and drafted the manuscript. SL and AP analyzed the results and interpreted the findings. TS and AP contributed to the drafting, revision, and final approval of the manuscript.

Manuscript 3: HIV-related individual-level outcomes commonly reported in research studies conducted in high-income countries: A scoping review. To be submitted.

Stephanie Long, Guowei Zhong, Kim Engler, Bertrand Lebouche, Tibor Schuster

SL conceptualized the study, study design including reviewing the search strategy with McGill librarian, Genevieve Gore, and implemented the search strategy. SL and GZ screened all titles and abstracts; SL screened all full-text articles, with GZ screening 10%. SL extracted data from all included articles, and GZ extracted 10%. SL conducted the quantitative and qualitative data analysis and drafted the manuscript. TS, GZ, KE, and BL revised the manuscript.

Manuscript 4: Leveraging expert knowledge for the co-creation of causal diagrams in inform clinical trial planning: A feasibility study in the context of HIV

Stephanie Long, Kim Engler, Bertrand Lebouche, Tibor Schuster

SL conceptualized this study with input from TS, BL, and KE. SL recruited study participants. SL and TS facilitated the interviews and data collection sessions. SL conducted the content analysis of the interview data and consolidated the causal diagram data. SL interpreted the findings with input from TS and KE. SL drafted the manuscript. TS and KE revised the manuscript.

Ethics approval

This study was approved by the Research Ethics Board of the Research Institute of the McGill University Health Network in Montreal, Canada. Please see Appendix A for ethics approval.

1. CHAPTER 1: INTRODUCTION

"The charm of history and its enigmatic lesson consist in the fact that, from age to age, nothing changes and yet everything is completely different." – Aldous Huxley

Evolving research methods in epidemiology and primary care

Quantitative research designs and analytical approaches applied in primary care research studies have undergone significant transformations in recent decades, driven by technological advancements, shifting research paradigms, and evolving public health challenges. The growing complexity of health data and the limitations of traditional associational inference approaches have necessitated a re-evaluation of established methodologies [1]. This evolution has been particularly apparent in epidemiology, a field adjacent to primary care research that gained unprecedented prominence during the COVID-19 pandemic.

This global health crisis presented a unique opportunity to observe how health research would be conducted and how epidemiological principles would be applied and used to inform policy under highly dynamic challenging conditions [2].

After nearly one year into the pandemic (February 2021), close to 3000 COVID-19 related randomized control trials (RCTs) were registered in the COVID-evidence database [3, 4]. However, many of these RCTs were small-scale and investigated highly similar interventions [4], revealing systematic inefficiencies in research coordination. While these research initiatives were necessary, the lack of collaboration and coordination highlighted a significant gap in the current research landscape. The situation underscored the need for more flexible and practical research approaches, which could have potentially offered greater efficiency in resource allocation and generation of results.

Innovative approaches, such as adaptive multi-arm trials give promise to address complex research questions more effectively and efficiently. There is a growing need for flexible, adaptive, and pragmatic methodologies that can respond to rapidly changing health landscapes while maintaining scientific rigor [1].

The role of modern causal inference

Over the last 40 years, advances in causal inference have led to a fundamental paradigm shift in how to formulate and rigorously answer cause-and-effect research questions outside of experimental settings. Until recently, applying causal inference beyond RCTs seemed inconceivable due to the inherent confounding in observational data and the absence of a formal framework for articulating causal research questions in statistical terms [5]. For example, a common question in health research, "*how effective is a given treatment X in preventing a disease Y?*" was impossible to state in traditional mathematical terms. This was because established analytical approaches, such as regression modeling, implied a symmetrical relationship between variables, and there was no explicit way to indicate whether X caused Y or vice versa [6]. This led to a dichotomy in how cause-and-effect research has continued to be conducted and reported (using exclusively *associational* notions) and, in contrary, how research findings were (and *still* are) being used to inform policies and interventions.

Causal inference is the study of how real or conceptual actions, interventions, or treatments affect outcomes of interest [7]. It uses notation that explicitly recognizes counterfactual events and variables, which enables identification of the causal effect of an exposure X on an outcome Y. Counterfactuals are hypothetical scenarios that did not actually occur but are constructed to enable causal inference by comparing potential outcomes under different conditions [8]. For example, the expressions Y_x, $Y(x)$, Y^x or $Y_{do\,=\,x}$, all represent an outcome variable Y (e.g., incident COVID-19 in the following year) under the counterfactual scenario that intervention X (e.g., COVID-19 vaccination status) had been set, for everyone in the population, to level x [9]. Considering a second exposure level x' allows definition of a causal contrast using counterfactual statistical notation such as $E[Y_x] - E[Y_{x'}]$. This quantifies the difference in the expected population outcome Y under two counterfactual scenarios: setting exposure status for everyone to X=x (i.e., everyone received the vaccine) versus setting exposure status for everyone to X=x' (i.e., no one received the vaccine). Thus, being able to define this causal contrast in which everyone in a population was vaccinated vs. everyone was not vaccinated, one can determine the actual causal effect of the exposure (e.g., COVID-19 vaccination status) on the outcome (e.g., incident COVID-19 in the following year).

Important examples of causal inference include mediation analysis [10], inverse probability weighting to address time-dependent confounding [11], and methods to handle selection biases and common missing data issues in observational research studies [12].

Graphical approaches to causal inference

Another important distinction of causal inference to associational inference methods is the central role of *causal diagrams,* such as directed acyclic graphs (DAGs). These graphical models, pioneered by Pearl (2009) [8], systematically encode contextual knowledge about observable and unobservable variables, representing their structural causal relationships and potential confounding pathways. Causal inference pioneer Judea Pearl describes the nodes (variables) in a causal diagram as a "society of listening variables" [5]. The term "listening" emphasizes the defining property of directed and acyclic relationships between the variables. This asymmetrical nature, where variable A listening to variable B, does not imply the reverse, underpins the concept of DAGs [13, 14].

DAGs allow researchers to visually represent and analyze complex causal relationships, providing a powerful complement to statistical methods. Moreover, they enable the verification of identifiability—determining whether an (average) causal effect can be recovered from measured data and an appropriate model, assuming the DAG accurately depicts the true data generating process [15].

Complex data and related limitations of causal inference approaches

With the growing complexity of health and medical data being routinely collected, research databases are reaching dimensions that limit the possibility of manual data handling and careful crafting of statistical inference models [16]. This has led to increased interest in machine learning approaches which offer promise in handling large-scale complex datasets [17].

Paradoxically, while there is profound understanding of RCTs and their ability to confirm the utility of interventions and policies, the excitement surrounding machine learning has been overshadowed by its current limitations in confirming the effectiveness of interventions [18]. *Machine learning* refers to computational techniques that enable algorithms to automatically extract patterns, insights, and build predictive models from large amounts of data through iterative learning processes [19]. For a given task, an algorithm is given a set of training

examples in the form of "inputs" (e.g., data features) and "outputs" (e.g., labels or scores). The algorithm then attempts to learn a function that maps the input variables to the output. For example, when predicting the viral load of a person living with HIV (e.g., output), an algorithm is given CD4+ cell count and ART adherence (e.g., inputs). The algorithm then takes all the input features and creates a series of functions aimed at predicting the outcome i.e., viral load. Selection of the best available function is actually an *Empirical Risk Minimization* problem in that the goal is to select the function that minimizes the discrepancy between the actual observed outcome and the predicted value by the function [20].

Machine learning in primary care research: capabilities, limitations, and challenges

While machine learning approaches excel at pattern recognition and prediction tasks, they face challenges when applied to causal inference problems. Firstly, although the algorithmic and computational components of machine learning approaches can typically be precisely described using statistical, mathematical, and/or programming notations, many of such algorithms are considered 'black boxes'. This is because the computationally complex nature in which they make predictions is very difficult for humans to interpret or because they are proprietary [21]. Secondly, machine learning algorithms typically do not make explicit assumptions considering prior knowledge or regarding the sequence of input variables, thus limiting their ability to make causal inferences. Instead they consider the joint distribution of variables with the outcome to build prediction rules [21]. This lack of structural assumptions can result in algorithms that draw on spurious (i.e., non-causal) associations projected by the data, rendering outcome predictions and variable importance assessments invalid. Consequently, machine learning algorithms may identify reverse causation, as temporality is often not explicitly considered.

Unlike machine learning, causal inference enables explicit consideration of the causal dependencies between observed and unobserved variables. The insufficiency of encoding structural knowledge in machine learning algorithms can lead to largely inaccurate results and misleading conclusions [22], especially when applied to datasets that are not representative of the target population e.g., datasets lacking gender and ethnic/racial diversity. This is particularly problematic when the findings of these machine learning algorithms are then used to make policy decisions or for resource allocation.

The use of machine learning in answering clinical questions has increased, but reporting, interpreting, and evaluating the validity of machine learning produced clinical findings can be challenging due to the limited familiarity among researchers, peer-reviewers, and readers in some clinical disciplines [23]. To assist the clinical audience, critical appraisal tools have been developed, such as ROBUST-ML, a quality appraisal checklist for machine learning studies for clinicians [24] and Faes et al.,'s [25] clinician's guide to artificial intelligence, which provides key points to consider when critically appraising machine learning applications in clinical research. Stevens et al., (2020) has also published recommendations for reporting machine learning analyses in clinical research [26].

Despite these efforts, there remains a lack of consensus on the guidelines and quality appraisal tools that help safeguard the development and performance assessment of machine learning algorithms used in practice. The available tools vary in scope and focus, highlighting the need for more comprehensive and standardized approaches to evaluate and reporting machine learning applications in primary care research.

Large language models: a rapidly evolving frontier in AI

As the limitations of machine learning approaches in causal inference are considered, it is crucial to explore emerging technologies such as large language models (LLMs). They represent a cutting-edge development in artificial intelligence (AI) that, while still part of the broader machine learning family, possess unique characteristics that could potentially address some of the causal inference challenges. Unlike conventional machine learning algorithms, LLMs are trained on vast amounts of textual data, including scientific literature, which inadvertently incorporates a wealth of domain knowledge, and potentially causal relationships described in natural language. This implicit encoding of external knowledge allows LLMs to potentially capture and reason about causal structures in ways that other machine learning models cannot.

LLMs differ from conventional machine learning models in terms of versatility and scope. While traditional models are often designed for specific tasks, LLMs demonstrate capabilities across a diverse range of applications such as answering knowledge base questions, creative writing, generating code, and performing classification or generation [27]. Their potential in causal inference tasks, such as building directed acyclic graphs, is particularly intriguing. By leveraging

their ability to process vast amounts of text data, LLMs could potentially assist in identifying causal relationships from scientific literature, helping researchers construct DAGs more efficiently.

However, LLMs also have limitations. They are sensitive to prompts [28, 29] and a tendency to hallucinate i.e., producing text that appears factual but is actually false or unsupported [30]. This propensity for generating plausible sounding but incorrect information poses challenges for their use in scientific research, where accuracy and reliability are imperative. Moreover, LLMs still inherit some of the fundamental limitations of machine learning approaches and require careful consideration in their application to causal inference tasks.

The potential of domain expert engagement to address limitations of machine learning

While LLM and other AI tools offer exciting opportunities for data analysis and causal inference, they also highlight the irreplaceable value of human expertise in guiding and validating AI-driven research.

Machine learning models make predictions by identifying patterns in large amounts of data, which is why the quality of data on which they are trained is of utmost importance. When machine learning models are trained on datasets that underrepresent certain groups i.e., people of colour, marginalized and vulnerable populations, while overrepresenting others, inferences made from such models may be invalid for, and hence, inadvertently discriminatory against these populations. Furthermore, inherent population selection mechanisms can induce spurious associations due to the phenomenon of collider-stratification [31].

One way to mitigate structural data and modeling deficiencies is to involve key domain experts in the development of algorithms to signal the absence or overrepresentation of certain data features i.e., race/ethnicity or gender and to identify potentially detrimental selection mechanisms [32, 33]. Their involvement can help ensure that the models are more representative, fair, and accurate across diverse populations.

Leveraging machine learning in modern adaptive trial designs

Despite the shortcomings of traditional RCTs and machine learning, evidence-based medicine can still effectively leverage their utility if rigorously designed and applied. While machine learning methods are superior at pattern recognition and classification tasks such as diagnostics, i.e., answering "what is?" questions, queries regarding the preventive or curative utility of interventions, i.e., answering "what if?" questions, can only be addressed using appropriate causal inference frameworks including (but not limited to) RCTs [34].

From an innovation and regulatory point of view, the development of novel interventions, particularly treatments or therapeutic agents, may outpace traditional pipelines for assessing efficacy or effectiveness, delaying critical access to interventions with high utility. For example, through leveraging on existing vaccine development knowledge, Moderna's mRNA-1273 COVID-19 vaccine was designed in two days [35], while efficacy trials took nearly one year to reach confirmatory status [36]. Despite being a stark example, the developmental pipeline for this vaccine was already expedited due to the urgency of demand given the coronavirus pandemic. In the United States, it takes an average of 12 years [37] and 1.5 to 2 billion USD to bring a drug from pre-clinical testing to market approval [38]. Despite the large investments of time and money that go into drug development and discovery, the ratio between drugs gaining regulatory approval and R&D spending each year has been steadily declining [38, 39].

Emerging *adaptive trial designs* are a promising solution to address the current limitations of confirmatory trials. These innovative trial designs are finally entering a renaissance epoch, empowered by advancements in information technology [40], endorsed by research groups worldwide, and gaining attention in high-ranked international journals. In a recent reflection, the head of the Melbourne School of Population and Global Health, Tony Blakely stated "advances in causal inference methods and the emergence of big, complex longitudinal data as well as data science, will profit from incorporating methods such as machine learning into epidemiological causal inference" [17]. By leveraging the strengths of these fields and adapting the emerging developments to the context of clinical trials, there is promise that some key challenges within existing trial designs can be effectively addressed including costs associated with inefficient static designs aspects, lack of interim efficacy information on interventions [17], patient cohort

selection and recruitment [7, 18], endpoint/outcome selection [19], and evaluation of multiple and upcoming interventions [20].

Adaptive trial designs: integrating causal inference and machine learning

The intersection of machine learning, causal inference, and innovative trial designs presents a promising frontier in health research. While machine learning excels at pattern recognition and handling large datasets, causal inference provides the theoretical framework for understanding cause-and-effect relationships. Adaptive trial designs, in turn, offer a flexible and efficient approach to confirmatory analysis in settings where (randomized) assignment of interventions to individuals is ethically and operationally permissible.

Adaptive platform trials are a novel type of adaptive clinical trial that allows the simultaneous and perpetual evaluation of multiple interventions against a common control group, with (new) interventions allowed to enter or leave the trial based on a pre-defined decision algorithm [41, 42]. Due to their flexibility, *adaptive platform trials* address many of the aforementioned issues with RCTs and enable a more time-efficient and seamless evaluation of collected data, allowing for adaptation of key design aspects (e.g., allocation ratios into trial arms, sample size, criteria leading to the termination of the trial, or an ineffective trial arm).

Integrating machine learning and causal inference in adaptive platform trials may enhance various aspects of the research process, including outcome selection, predictive modeling, causal discovery, data-driven adaptations, and identification of heterogeneous treatment effects. This integration addresses key challenges in health research, such as balancing speed and rigor, handling complex data, establishing causality, and addressing ethical considerations.

Adaptive platform trials frequently adopt a more pragmatic approach of implementation than standard clinical trials which enables the collection of evidence that is closer to the real world. In fact, the origin of adaptive platform trials stems from population-wide studies in rare diseases as well as epidemics of highly lethal infectious diseases [43]. Both are situations where rapid assessment of multiple competing treatment strategies is critical, making standard trial designs unfeasible or inefficient.

The fundamental role of appropriate outcome selection in platform trials

Integrating modern adaptive (platform) trial design, causal inference, and machine learning is highly promising approach to answering research questions more effectively, efficiently, and ethically. One of the key challenges that spans across these three methodological domains and is hence a fundamental decision point in any research endeavour, is how to best select the surrogate outcome measures. In clinical trial settings, there are often multiple ways to measure an individual's health status and/or response to treatment. In practice, however, it is not always clear which outcomes are most sensitive, reliable, and informative for the purpose of answering the research question under study [44]. Additionally, there is a delicate balance between clinical relevance, feasibility, and costs of certain outcome measures. Traditionally, outcome measures are selected by experts i.e., clinicians or trialists, however, in complex data settings and with increasing access to data, relying solely on experts may result in missed opportunities for identifying clinically relevant but also statistically robust and efficient outcome measures. In the context of platform trials, surrogate outcomes, i.e., measures that can reliably act as an early indicator for mid- or long-term outcomes, play a particularly central role [42]. This is due to the adaptive randomization procedures embedded in such trials which use surrogate measures to identify the least effective trial arms that are eventually dropped from the trial [45].

1.1 Overarching Research Objective

Although much progress has been made in biomedical, clinical and health research, the uptake *of important methodological advancements from the three emerging fields* of machine learning, causal inference, and adaptive trial designs is a bottleneck [17]. These interrelated themes underscore the challenges with integrating expert knowledge and public data. Stakeholders such as clinicians, trialists, and even the regulatory authorities lack the expertise and skills to apply such rapidly evolving methods and are thus missing out on more effective and efficient ways to answer pertinent research questions.

The overarching aim of this doctoral research is to assess, through a causal inference lens, current methodological shortcomings in integrating public data and expert knowledge in primary care research studies. Through a series of framework developments related to DAGs, these shortcomings were addressed, and the feasibility and acceptability of the proposed frameworks was demonstrated in the context of chronic disease management and adaptive trials.

Chapter 1 - Introduction

This research explores ways to integrate public data and expert knowledge into primary care research methodologies. It examines frameworks using directed acyclic graphs and causal inference, aiming to contribute to research approaches that can address both 'what is' and 'what if' questions. The goal is to support the development of more informative and practical methods for addressing complex health challenges in primary care settings.

2. CHAPTER 2: LITERATURE REVIEW

2.1 Premise

This dissertation is situated at the intersection of three emerging methodological domains: causal inference, machine learning, and adaptive trials. This literature review provides an overview of these topics. While not exhaustive, it offers sufficient detail to understand the methods applied in this thesis.

2.2 Role of randomized controlled trials vs observational studies

According to the levels of evidence originally described in a report by the Canadian Task Force on the Periodic Health Examination in 1979 [46], randomized controlled trials (RCTs) are the highest form of evidence followed by observational studies and expert opinions. These levels of evidence were further elaborated upon by Sackett [47] in a 1986 article on levels of evidence for antithrombotic agents, where RCTs with low Type I and Type II error were described as the highest level of evidence. Both hierarchies ranked the evidence according to the probability of bias, and thus, placed RCTs at the top and case reports/series and expert opinions at the bottom. At that time, RCTs were ranked the highest form of evidence because they are designed to be unbiased and present less risk of systematic error [48].

A hallmark of an RCT is the fixed random assignment of participants to a control group or one or more treatment groups. This methodological approach systematically mitigates two critical sources of bias: selection bias and confounding. By randomly distributing participants across groups, RCTs create statistically comparable cohorts with balanced baseline characteristics, ensuring that any observed differences in outcomes can be more confidently attributed to the intervention rather than pre-existing variations. By doing this, RCTs create a near-perfect counterfactual scenario, enabling the identification of a causal contrast by constructing a synthetic comparison that closely approximates what would have happened to the treatment group had they not received the intervention. This exchangeability between treatment groups is what allows, under further regularity conditions, for the estimation of average causal treatment effects [49].

The hierarchy of evidence has been further expanded upon and elaborated by numerous groups, including the Oxford Centre for Evidence-Based Medicine [50] which ranks RCTs as a high

level of evidence, but ranks systematic reviews of RCTs even higher, at the top of the hierarchy. It has been argued that if a single RCT provides good evidence, then the best evidence would be a systematic review with meta-analysis because it synthesizes all the relevant evidence and provides more reliable evidence than a single study [51]. Regardless, this placement has been challenged as heterogeneity (clinical, methodological, or statistical) is an inherent limitation of meta-analyses and systematic reviews that can never be fully eliminated, only managed [51].

Randomized controlled trials have many strengths but are not insusceptible to flaws. To lessen the risk of bias, they have very strict inclusion and exclusion criteria, which influences the generalizability of findings to other populations. Additionally, larger sample sizes are required to generate enough power to identify a true effect [52]. Thus, despite being lower in the hierarchy of evidence, observational studies e.g., cohort studies and case-control studies, are still valuable and important study designs, when conducting an RCT is unethical or infeasible e.g., randomizing participants to environmental hazards or disease states. RCTs tend to evaluate interventions under controlled clinical conditions among specific populations, while observational studies *observe* effects in "real-world" settings. Since observational studies typically have higher external validity (or generalizability) than RCTs, due to their inclusion of a patient sample that is representative of the average patient population [53]. Additionally, they may provide evidence of effectiveness in the general population and better understanding of care and outcomes in populations underrepresented in RCTs [54] such as women [55, 56], the elderly [56, 57], and ethnic minorities [55, 56, 58].

Observational studies have clear value, but the characteristics of their design limits their ability to control for unmeasured or unknown confounders and establish causation between the outcome and exposures.

2.3 Adaptive trial designs

Adaptive trial designs enable more flexible trial conduct through leveraging *interim results* to modify the characteristics of a trial (e.g., randomization ratio, number of treatment groups, number and frequency of interim analyses, and the patient population under study) in accordance with pre-specified decision rules [59, 60]. Compared to traditional trials i.e., RCTs, adaptive trials are often more efficient, informative, and ethical as they are more time- and resource-

efficient and due to interim analyses, *can* increase the probability that trial participants will be assigned to the best performing treatment group [59, 60].

Despite the advantages of adaptive trial designs, randomized controlled trials are still the long-standing 'gold-standard' for comparing different interventions. RCTs, however, can be costly, resource-heavy, inefficient, and do not provide information regarding the efficacy of a new intervention until the trial is completed. Awareness of the shortcomings of RCTs has led to growing interest in novel and more flexible trial designs such as adaptive trial designs.

Table 2-1: Table 2 1: Comparing RCTs and adaptive trials [41, 60, 61]

	Randomized controlled trials	**Adaptive trials**
Scope	Evaluating efficacy of a single intervention in a homogenous population	Evaluating multiple interventions in a heterogeneous population explicitly assuming treatment effects may be heterogeneous
Use of surrogate outcome	Enable faster and more cost-effective trial design by using intermediate outcomes that predict primary endpoints with shorter follow-up periods.	Interim analysis to update key trial characteristics.
Duration	Finite	Potentially long-term, new interventions may be added
Number of groups	Pre-defined number of treatment groups, typically limited	Multiple treatment groups, with the number of groups and specific treatments evaluated changing over time
Stopping rules	The trial may be stopped early due to success, futility, or harm.	Individual treatment groups can be added or dropped during the conduct of the trial based on efficacy or futility of treatments
Group allocation strategy	Fixed randomization	Response-adaptive randomization
Sponsor funding	Supported by a single federal or industrial sponsor	Since they are master protocols, they may be supported by multiple federal or industrial sponsors

Unlike conventional RCTs where data analysis only begins at trial completion, adaptive trial designs allow for interim analysis as data accumulates in order to adapt key design aspects such as sample size, eligibility criteria, medication dosage, allocation proportions to study arms,

addition or elimination of treatment arms, rules to change from one study phase to another, and early stopping due to superiority or futility [41, 61] (Table 2-1).

These adaptive trial designs also fall under the broad category of *master protocols*, which are one overarching study designed to address multiple research questions, which may include multiple subpopulations, interventions, or target diseases [62, 63]. Three types of study designs fall under this category: basket trials, umbrella trials, and platform trials, the latter of which will be considered in this thesis. Briefly, basket trials evaluate a single therapeutic agent in multiple patient populations in several parallel studies [63]. Umbrella trials evaluate multiple therapeutic agents in a single disease in parallel treatment arms [63]. Platform trials will be discussed in greater detail in the following section.

Adaptive trial designs come with many advantages including increased efficiency, broader scope in assessing competing interventions, flexible objectives, and increased stakeholder enthusiasm and involvement [63, 64]. Since these types of trials are master protocols that can evaluate multiple interventions or multiple subpopulations with a focus not only on detecting large efficacy signals or large treatment effects, smaller sample sizes can be used [64]. Master protocols also allow investigators to more rapidly initiate new studies by adding onto existing substudies [63].

Since master protocols can address multiple research questions simultaneously, they often use a master budget with a common infrastructure to share costs with other investigators or sponsors [63, 64].

2.4 Adaptive platform trials

Adaptive platform trials are a subclass of adaptive trials that allow new intervention arms to be added in the course of the trial while potentially discontinuing trial regimens that show inferior performance [64]. Key aspects of *platform trials* include interim analysis and response-adaptive randomization. Platform trials are similar to umbrella trials in that more than one therapeutic agent is studied for a single disease. Unlike umbrella trials, platform trials allow treatments to enter or exit the trial based on a decision algorithm applied on data collected throughout the trial

i.e., interim analysis [62]. Because platform trials are able to evaluate multiple treatments across multiple subpopulations, they are highly useful to examine the efficacy of combination treatments or directly compare competing treatments [65].

2.4.1 Interim analyses and trial monitoring

One of the main differences between a platform trial and a traditional RCT is the ability to analyse outcome data throughout the trial and use these interim results to adapt key trial characteristics during the trial i.e., interim analyses. Though traditional RCTs can also pre-specify interim analyses in their protocols, they are typically planned to review the efficacy and safety of the intervention and not to adjust key trial characteristics as in adaptive trials [66, 67].

Interim analyses in platform trials are motivated by ethical and resource considerations as the interim information obtained is used to adapt allocation probabilities of participants to treatment arms [68]. If emerging evidence on the efficacy, effectiveness, or safety of a treatment arm suggests inferiority compared to other treatment arms, these allocation probabilities can reach zero, leading to an exclusion of the respective treatment arm from the trial.

Data monitoring and interim analyses plans must be pre-specified in protocols and should include the number of evaluations, the time interval between evaluations, decision rules, and outcomes assessed at each timepoint (the focus of this thesis). When using conventional hypothesis testing approaches, there are several statistical concerns that can arise with interim analyses in adaptive platform trials. For instance, the number of interim evaluations is important because the probability of false hypothesis rejections increases with the number of hypotheses tests being performed, especially without proper statistical adjustments [42]. Additionally, in platform trials that enable early removal of poorly performing treatment arms, without adequate statistical adjustment, there is an elevated risk of dropping actually effective arms.

Timing of the interim evaluations is also an important consideration, especially when the first evaluation will be conducted. Smaller datasets are prone to random error; thus, it is important not to start interim evaluations too early in a trial [45, 61]. Like most master protocols, platform trials require an adequate “burn-in” period, in which enough data have been collected to allow

for sufficient precision to detect the presence or absence of relevant effect [42]. Unlike most traditional RCTs, adaptive platform trials employ Bayesian inference methods or reinforcement learning approaches for interim analysis purposes. These approaches are not prone to Type I error, as decisions are based on posterior probability distributions and how they match pre-defined clinical parameter margins of interest: in contrast to statistical testing, the likelihood of falsely rejecting (or accepting) a certain range of parameter values does not increase with the number of assessments being done.

2.4.2 Decision rules

Statistical decision rules are an important aspect of interim analyses, which must be established prior to the initiation of a trial. Decision rules are criteria used to determine termination of treatment arms or adapt the allocation probabilities of participants to treatment arms i.e., inform response-adaptive randomization. Additional decision rules include pre-established quantitative criteria for re-estimating the sample size, adapting patient eligibility, or selecting new arms to inform dose estimations [61]. Selection of appropriate decision rules is central in platform trials to minimize risk of biased and inefficient decision-making during interim evaluations [42].

2.4.3 Response-adaptive randomization

Response-adaptive randomization is a group allocation procedure that utilizes data accrued during the ongoing trial to adapt the randomization probabilities such that a higher proportion of participants are allocated to the treatment arm with the most favourable interim results as the trial progresses [69]. For example, during interim analysis of a three-arm trial comparing two interventions (A, B) against a control group (C), it may be determined that intervention A is more efficacious than intervention B and C, thus, the allocation ratios is adjusted to 2:1:1 for Intervention A: Intervention B: Control C from its initial ratio of 1:1:1 [45, 68]. This group allocation procedure differs from the fixed randomization approach traditionally used in RCTs, which typically assigns individuals to treatment groups in equal and fixed proportions e.g., ratio 1:1, treatment : control.

Advantages of implementing response-adaptive randomization include the reduction of potentially deleterious clinical outcomes observed during the trial and a smaller overall sample size without substantial loss of statistical precision [45]. Though seemingly favourable, response-adaptive randomization does have some disadvantages that present certain statistical challenges.

Temporal trends in the prognostic characteristics of the patient population during trial enrollment may systematically bias the results of the trial. For example, if in the beginning of the trial, participants are randomly assigned equally to the experimental and control arms, but later in the trial, a much greater proportion of participants are randomly assigned to treatment arms vs. control arms, then an improving prognostic pool of patients being randomly assigned in the trial will translate into a bias in favour of the treatment arms. Due to the potential of such bias, response-adaptive randomization is not recommended for long-term trials.

2.4.4 Surrogate outcome selection

Traditional clinical trials typically employ a primary outcome that directly addresses the main research question. In traditional clinical trials, this outcome is used for the final analysis. Adaptive platform trials, however, one *can* select different outcomes for the interim analyses and the final analysis to adapt key trial characteristics during trial conduct [42].

Surrogate outcomes are often used in settings where the primary outcome has limitations. These limitations may include a long delay before observation, it is an invasive measurement, the cost of its measurement is prohibitively expensive [45, 70]. In such circumstances, a surrogate outcome that is easier, more efficient, and less invasive to measure may be used for the interim evaluations [71, 72]. The use of surrogate outcomes in both traditional and adaptive trials is based on the assumption that changes in the interim outcome are associated with changes the primary outcome [71].

The use of a surrogate outcome can therefore be an efficient way of evaluating interventions, but its utility is highly dependent on the choice of outcome [73]. One way to assess the appropriateness of a surrogate outcome is to determine if it is strongly correlated with a clinically meaningful endpoint e.g., overall survival [74]. In HIV care, for example, lymphocyte TCD4+ cell count is often used as a surrogate measure for disease progression [72, 75].

The selection of surrogate outcome requires careful consideration and evaluation on a case-by-case basis. Surrogate outcomes effective in one clinical context may often not be applicable in another, even if they appear similar. For example, evidence suggests that progress-free survival

is highly correlated with overall survival in patients with chronic lymphocytic leukemia [76], this relationship does not hold for other tumour types such as metastatic breast cancer [77]. Some researchers argue that the use of a surrogate outcome is justifiable if it has been evaluated in a trial pilot study [78].

Selecting an interim outcome not strongly correlated with the primary outcome can lead to erroneous findings that may result in "effective" treatment arms being dropped from the trial or "ineffective" treatment arms being selected [42]. Thus, the choice of interim outcome is critical in platform trials, where it has important implications on response-driven adaptive randomization.

Aside from statistical considerations, there is limited literature on how to select appropriate surrogate outcomes in a platform trial. Platform trials present additional complexities, requiring both primary outcomes and statistically associated, relevant surrogate outcomes. Thus, a careful examination of the available data, literature, and domain expert insights are needed for optimal outcome selection in platform trials.

2.5 Causal Inference

Advances in causal inference have had important implications in empirical research as most research questions asked in health and medical research are not statistical, but causal in nature. Examples of such research questions include: *What is the efficacy of a given drug in a given population*? *What is the effect of a given intervention on a given outcome? What factors explain a given outcome?* Common amongst these research questions is the desire to uncover the cause-effect relationships amongst a set of variables i.e., treatments, interventions, and outcomes. Such *causal* questions cannot be answered without knowledge of the data-generating processes, directly from the data itself, or from the distributions that govern said data [8]. Prior to the formal establishment of causal theory, such *causal* questions could only be answered with statistical notation with specific assumptions regarding how the data were collected [9].

Causal inference pioneer, Judea Pearl defines *causal inference* as a method that takes three inputs and produces answers for two types of causal questions [6].

Inputs:

1. What we *wish* to know: A sequence of actions that will lead me to consequence, etc.
2. What we do *already* know: Directed Acyclic Graphs (how are the variables related?), types of relationships/functions (linear, log linear) etc., willing to defend on scientific grounds.
3. What type of *data* do we have available? Experimental, observational, etc. under what conditions were the data collected, population.

Types of causal questions:

1. Effects of pending *interventions*
2. Effects of *undoing past events* i.e., counterfactuals: something happened and what if some event in the past did not occur, what would take place? I have been fired because I missed a meeting, what if my plane had arrived on time?

There is an important distinction to be drawn between statistical models used for associational analysis and causal inference: Models for associational inference evaluate the relationship between two variables over a population i.e., defining a relationship in terms of a joint distribution between two variables [9, 79]. Language used to describe associational concepts include: correlation, regression, likelihood, risk ratio, odds ratio, conditionalization [9]. Causal relationships, on the other hand, cannot be discerned from solely the joint distribution [9] as they reflect probabilities under changing conditions e.g., changes induced by an intervention or treatment [8]. Causal concepts are reflected in language such as randomization, effect, confounding, intervention, spurious correlation [9].

In order for the (existing) statistical notation to be used to make causal inferences, new notation was required to express causal assumptions and claims, for example, counterfactual events [9]. For example, probability calculus does not have expressions to disentangle statistical dependence from causal dependence, thus, there is no way to convey “symptoms do not cause diseases”, merely that if a symptom of a disease encountered, it is likely the disease will also be encountered [8]. This illustrates the importance of a well-defined and unambiguous notation for expressing causal assumptions, so judgments regarding the validity and plausibility of such assumptions can be made.

Causal expressions in statistical literature are not new and have roots in the *potential-outcome* notation [80-82]. Such causal expressions can be easily identified through the subscripts, superscripts, or parenthetical expressions which represent counterfactual events and variables, for example, $Y_x(u)$, $Y_{x,}$ $Y(x)$, Y^x or $Y^{do\,=\,x}$. These causal expressions can be interpreted as the value of outcome *Y* under the counterfactual scenario that intervention or exposure *X* had been set (for everyone in the population) to level *x*. This notation allows the formal definition of causal contrasts, quantifying the difference in the expected population outcome *Y* under two counterfactual scenarios: exposure status set for the entire population to $X=x$ vs. setting the exposure status for the entire population to $X=x'$. Another important distinction between associational inference and causal inference is the use of structured graphs called directed acyclic graphs (*DAGs)* which encode the structural relationships between variables of interest e.g., outcome, exposures, and unmeasured variables. **Table 2-2** provides a comparison of associational and causal inference.

Table 2-2: Comparing Associational Inference with Causal Inference

	Associational Inference	**Causal Inference**
Research Questions	**What is?** What is the relationship between an outcome and an exposure?	**What if? (Interventions)** What is the efficacy of Drug A in a population of cancer patients? **Why? (Counterfactuals)** Would the patient had died if they had not received heart surgery?
Notation	**P(Y|A)** Probability of Y given A	**P(Y|A=a)** Probability of Y given A if A was set to a

2.5.1 Directed acyclic graphs

Causal models are typically accompanied by graphical representations i.e., *Directed Acyclic Graphs (DAGs)* which succinctly illustrate the qualitative assumptions made by the models, not captured by conventional statistical models [13, 14]. In epidemiological research, *DAGs* serve a variety of purposes including: (1) representing the causal relationships amongst variables [14, 83, 84]; (2) identifying the potential confounding variables which need to be controlled for in order

to estimate causal effects [14, 83, 85, 86]; and more recently (3) as a means of classifying the types of causal relationships that may give rise to selection bias [87].

A DAG is composed of variables (*nodes*), both measured and unmeasured, and their connections displayed via line segments (*directed edges*) [13, 87]. The absence of an edge between variables indicates the absence of a causal relationship between the variables. If the *edge* has an arrowhead, *variable* at the tail is the *parent node* and the variable at the arrowhead is the *child node* [13]. An *edge* is any line (with an arrowhead or not) that connects two variables [84]. **Figure 2-1a** illustrates the most basic DAG, with X representing the *parent node* (or in this case, cause of Y) and Y being the *child node* (and in this case, outcome). Two variables in a graph are *adjacent* if they are directly connected with an *edge* [14]; in **Figure 2-1b**, Z and V are *adjacent*, but V and Y are not.

A *path* between two variables, *X* and *Y*, is a series of adjacent *edges* connecting *X* and *Y*. A *directed path* is path in which each *child* in the sequence is the *parent* of the subsequent *node* [13]. A *backdoor path* is a non-causal path e.g., alternative path between two variables, in **Figure 2-1b**, X → Y has the backdoor path X → Z → Y.

Therefore, a DAG is causal if: (1) the arrows between variables can be interpreted as *direct causal effects*, and (2) all common causes of any pair of variables are present [87]. The causal effects are 'direct' relative to certain degrees of abstraction in that the DAG does not include any variables that may mediate the effect [13]. As the name suggests, *Directed Acyclic Graphs* are acyclic because a variable cannot be the cause of itself, either directly or indirectly through another variable i.e., there are no feedback loops; as illustrated by each DAG in **Figure 2-1** [87]. Additionally, in DAGs, causal pathways are represented with *directed paths* from the starting variable to the final variable; thus, a variable is the cause of its *descendants* and an *effect* of its *ancestors* [13].

In addition to illustrate the causal relationships between variables, DAGs can also encode the causal determinants of statistical associations [87]. In causal DAGs, the association between the exposure and outcome can be produced by three causal structures [83]:

1. *Cause and effect*: If an exposure X causes outcome Y, then they are associated (**Figure 2-1a**).
2. *Common causes*: If there is a common cause of two variables, in general, they will be associated even if one is not the cause of the other. In **Figure 2-1c**, V is the common cause of X and Y, thus, they are associated even though X does not cause Y.
3. *Common effects*: An exposure X and outcome Y will be conditionally associated if they have a common effect S, and the association measure is computed within levels of S (**Figure 2-1d**). In other words, selection bias due to conditioning on a common effect of the exposure and outcome i.e., collider stratification bias.

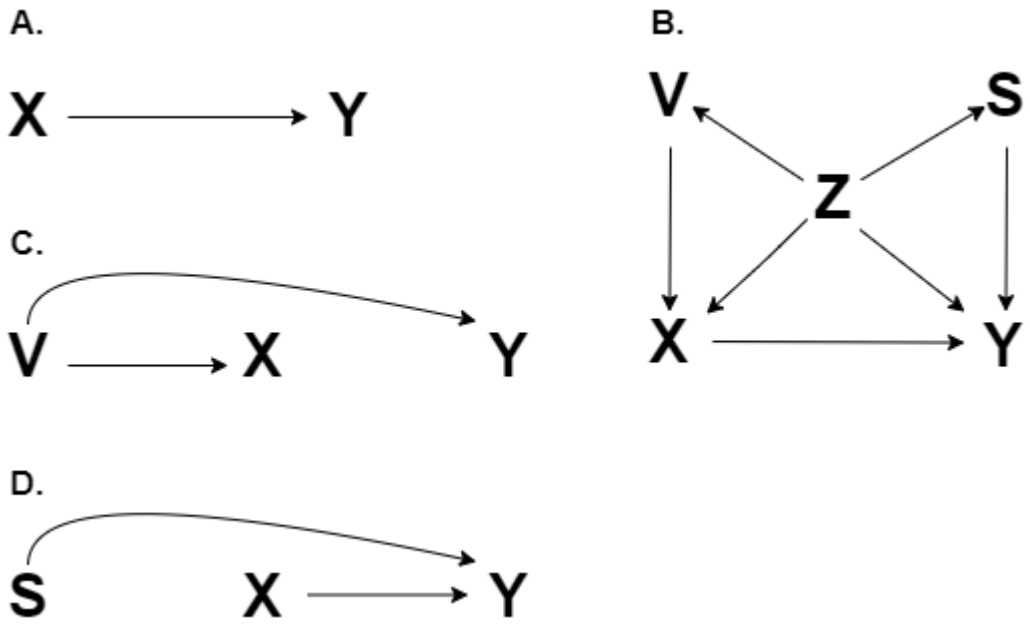


Figure 2-1: Four causal directed acyclic graph examples details in the body of text.

2.6 Machine Learning

The section provides a brief introduction to machine learning, describing how it generally makes predictions. While not exhaustive, it offers sufficient information to understand the methods described in this thesis.

Finding meaningful patterns in data has been a fundamental and longstanding problem throughout scientific history. *Machine learning*, a subset of pattern recognition and artificial

intelligence, uses computer algorithms to detect regularities in data and *classify* these regularities into meaningful (or not) categories [88]. As an automated data analysis approach, machine learning enables the analysis of much larger quantities of data more efficiently. However, it is not without limitations which will be discussed in later in 2.6.2 Machine Learning Fallacies.

Consider an example where a machine learning algorithm must classify a chest X-ray as cancerous or not. The algorithm would be given a portion of the dataset called a *training set,* which is a set of *N* X-ray images used to tune the model parameters [88]. The training set would include the *labels (*or *target vectors)* indicating whether each X-ray is cancerous or non-cancerous, allowing the *algorithm* to *learn*. The algorithm would then create a set of functions that take the *input vectors* (in this case, X-rays) and generate an *output vector* i.e., prediction of whether it is cancerous or not. The goal of the algorithm is to select the most accurate function for prediction (or categorization), which is essentially an *Empirical Risk Minimization* problem [20].

Once trained, the model can be used to determine the identity of new *input* chest X-rays from a different portion of the original dataset called the *test set*. In practical applications, another goal of the algorithm is *generalization* i.e., the ability to identify new inputs not found in the *training set* [88].

This application exemplifies a *supervised learning* problem, where the training data are comprised of examples of the *input vectors* with their corresponding *target vector* i.e., its label [88]. Problems like this X-ray diagnostic example are called *classification* problems. *Supervised learning* problems will be the primary focus of this thesis. Tasks with continuous variable outputs are called *regression* problems; for example, predicting an individual's weight given their height, age, and gender.

Other pattern recognition tasks involve *unsupervised learning,* where the *training set* includes *input vectors X* without corresponding *target vectors* [88]. Here, the goal is not to *classify* or *regress*, but to discover similar examples within the data known as *clustering*, to determine the distribution of data within the input space (*density estimation*), or to reduce data from high-dimensional space to two or three dimensions for the purpose of *visualization* [88].

There is a final class of machine learning called *reinforcement learning,* which differs significantly from *supervised* and *unsupervised learning*. In this approach, the algorithm is not told what actions to take but must discover which actions yield the greatest rewards through trial and error [89].

Recent advancements in machine learning have led to the development of deep learning techniques, which use artificial neural networks with multiple layers to learn complex patterns in data [90]. These methods have shown remarkable success in various fields, including image and speech recognition, natural language processing, and within healthcare, disease diagnosis and drug discovery [90].

2.6.1 Large language models

One of the most significant advancements in deep learning has been within the field of natural language processing (NLP), specifically with the development of large language models (LLMs). These models are designed to understand and generate human-like text, among other tasks. LLMs such as OpenAI's GPT-4 [91], Meta's Llama-2 [92], and Anthropic's Claude [93] are trained on vast corpora of textual data, often comprising hundreds of terabytes, and contain hundreds of billions of parameters. These parameters are essentially the model's learned weights that control how it processes, understands, and generates language [94, 95]. Temperature is one example of a parameter that controls an LLM's output, influencing whether it is more random or ordered and deterministic [96].

The fundamental architecture of most modern LLMs is based on the Transformer model, introduced by Vaswani et al., in 2015 [97]. A transformer is a type of neural network architecture used for processing sequential data such as text. This architecture uses a mechanism called "attention" to weigh the importance of different words in a sentence with processing language, allowing the model to focus on the most relevant input data, instead of treating all input data equally. This enables the model to improve accuracy and efficiency.

The training process of LLMs involves two main stages: pre-training and fine-tuning. During pre-training, the model learns general language understanding on a large corpus of text data. Specifically, in the text corpus, a fraction of tokens (i.e., words, characters, subwords) are masked and the model is required to predict the masked tokens. This process, known as self-

supervised learning, allows the model to capture complex patterns and relationships in language without the need for manually labeled data (as in supervised learning settings) [98].

After pre-training, LLMs can be fine-tuned on specific tasks or domains, which adapts their general language understanding to more specialized applications. This adaptability has led to impressive performance across a wide range of NLP tasks such as translation, extraction of text data, and answering medical questions [99]. Despite their remarkable performance in generating human-like text, LLMs are prone to hallucinations i.e., producing plausible sounding text that is actually false [30]. Additionally, LLMs are susceptible to the limitations inherent in general machine learning algorithms as well (and described below) and careful consideration should be applied when using for causal inference tasks, such as drawing DAGs.

2.6.3 Machine Learning Fallacies

Machine learning algorithms make predictions by "learning" information from the data on which they are trained. As previously mentioned, unlike statistical modelling, machine learning algorithms make no assumptions regarding the inherent data structures or dependencies between measured or unmeasured variables. Lack of assumptions paired with non-representative data can lead algorithms to make biased and inaccurate predictions on marginalized populations. This illustrates why the quality of data upon which algorithms are trained is of dire importance, especially when the results are used to inform policy making or resource allocation [100].

Machine learning algorithms are typically trained on labeled data e.g., for facial recognition tools, algorithms are trained on datasets including a series of photographs of human faces with an associated label e.g., "woman". Researchers recently found that machine learning algorithms trained with biased data i.e., inaccurate or unrepresentative data, can lead to algorithmic discrimination [22, 101, 102]. Bolukbasi et al., (2016) determined that a popular word embedding space for natural language processing, Word2Vec, encoded societal gender biases. The authors used Word2Vec to train an analogy generator that fills in the missing word in an analogy, for example, "man is to king, as woman is to X" [101]. They discovered that the word embeddings in Word2Vec had implicit sexism encoded whereby an analogy was completed as "man is to 'computer programmer', as woman is to 'homemaker'" conforming to the stereotype that programming is associated with men, and women are associated with homemaking [101]. The biases in Word2Vec are likely to propagate in any system that uses them. More recent

studies have quantified gender and ethnic stereotypes in word embeddings [103], and researchers have raised concerns about the potential dangers of large language models perpetuating harmful biases [104]. The biases in these models are likely to propagate in any system that uses them.

Additionally, many widely used datasets that machine learning algorithms are trained upon do not equally represent all race and ethnicities within a society, which results in inaccurate predictions on those populations. For example, a widely used algorithm applied on 200 million patients in US hospitals to predict which patients would more likely require extra medical care "assigned the same level of risk to Black patients that [were] sicker than White patients" and consequently, it was estimated by the authors "that this racial bias reduces the number of Black patients identified for extra care by more than half" [22]. The authors concluded that "bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients" [22]. This example also highlights issues that arise with the inappropriate selection of a proxy outcome.

Another poignant example is an analysis of two commercial datasets of unfiltered faces, IARPA Janus Benchmark-A (IJB-A) and Adience were found to be "overwhelming composed of lighter-skinned subjects," 79.6% and 86.2% respectively [22]. When machine learning algorithms are trained on gender- and race-imbalanced datasets and the findings are used to inform policy decisions or medical decisions, there can be serious discriminatory consequences. For example, machine learning algorithms trained on gender-imbalanced datasets perform worse at reading chest X-rays of the underrepresented gender [105] and researchers have expressed concern that skin-cancer detection algorithms trained on predominantly fair-skinned individuals will underperform on darker-skinned individuals [106]. Usually, researchers or the public do not have full access to these algorithms as the companies that produce them are protective of the source code due to high development costs and the competitive nature of the field. The public generally must take the word of developers that their proprietary algorithm performs as stated.

2.6.3 Fairness in Machine Learning

The examples outlined in the previous section (2.6.2 Machine learning fallacies) illustrate the potential harms and discriminatory biases that can result from machine learning models. The irony lies in that automated data analysis by machine learning has been praised for its efficiency

in time-consuming processes, ability to improve accuracy and performance of tasks, and due to the lack of human input, its ability to remain neutral and free of human biases. This line of thinking was coined the 'neutrality fallacy,' [107] which is the misconception that machine learning will not perpetuate the trends in data which are (unintentionally) often encoded with human biases i.e., provide a more objective treatment of individuals.

As use of machine learning continues to proliferate across fields, particularly in fields that have ethical and legal implications like healthcare, it is important to address the potential for discrimination against certain subpopulations or based on protected attributes (e.g., gender, race, sexuality, religion) [108]. Consequently, there is growing interest in designing algorithms that make *fair* predictions and mitigate the harmful effects of algorithmic discrimination. There has also been a focus in developing frameworks and tools that other researchers can use in their application of responsible and fair machine learning [109]. This leads to two important questions, "what does it mean for an algorithm to be *fair*?" and "how can we quantify *fairness* in machine learning?"

Extensive discussions on the definitions of "fairness" and "discrimination" have been ongoing in the social sciences community for decades [110, 111]. Similar debates have been occurring in the field of computer science, particularly surrounding individual [112] vs. group fairness [113] and the quantification of discrimination via the development of scores [114-117]. These parallel debates highlight a difference in how "fairness" is understood in decision-making, whereby there are different interpretations in the extent to which characteristics of an individual beyond their control should be included in decisions about them. Generally speaking, in decision-making, *fairness* is the "absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics" [118]. Thus, an "unfair" algorithm can be understood as one whose decisions are skewed towards a particular group of people, like the examples presented in the previous section (2.6.2 Machine Learning Fallacies).

Most of the proposed definitions of fairness in machine learning are observational i.e., they depend on the joint distribution of the predictor $\hat{Y}$, protected attributes of an individual A, observed features (i.e., covariates) X, relevant latent unobserved variables U, and the outcome to be predicted Y [108]. Protected attributes A represent variables that must not be discriminated against in the formal sense of the different mathematical definitions of fairness [119]. Predictor $\hat{Y}$

is a random variable that depends on *A, X, U*, and is produced by the machine learning algorithm as its prediction of outcome *Y* [119].

2.7 Clinical Context: HIV infection

Incidence rates of HIV infections and associated mortality have been declining over the past few decades due to the improved effectiveness of antiretroviral therapy (ART). The number of people living with HIV (PLHV), however, remains high, with an estimated of 39 million infected worldwide [120]. As HIV shifts from a being a terminal disease to a chronic, manageable condition, we need to re-evaluate what health outcomes are most relevant and appropriate for use in adaptive platform trials. Viral load and CD4+ cell count have been the preferred indicators of HIV treatment success and are frequently reported primary outcomes in the trial literature [121]. They do not however give the full picture of health, as defined by the World Health Organization as "a state of complete *physical*, *mental*, and *social* well-being, and not merely the absence of disease" [122]. There are many other outcomes that can describe an individual's state of health. The HIV care continuum [123] (to be discussed in further detail below) outlines five stages towards successful HIV care, each representing a relevant outcome to be evaluated.

Given that HIV is a complex chronic condition affected by a constellation of factors with many associated health outcomes, the choice of outcome in an adaptive platform trial may not be so straight forward. Considerations must be made to ensure the selected outcome is representative, fair, and measurable. These characteristics make HIV management an appropriate clinical context to develop a framework for outcome selection in adaptive platform trials.

2.7.1 HIV Outcomes

Since the 1980s, numerous landmark studies [124-127] have led to the development of antiretroviral therapies that now allow those with HIV to have near normal lifespans [128]. Effective treatments are able to achieve undetectable plasma HIV RNA levels (copies/ml) [129], which has transformed HIV into a chronically manageable disease [130]. However, HIV requires lifelong follow-up, self-management, and antiretroviral adherence to maintain an undetectable viral load and avoid transmitting the virus [131]. Since taking ART is central to HIV care, many widely used HIV outcomes are treatment-related:

- **Viral load suppression**: viral load below the detection threshold using viral assays [132]. Viral suppression is an indicator of treatment success and reduced transmission potential [132].
- **CD4 cell count**: CD4 cell count is a laboratory test that measures the number of CD4 T-cells in a sample of blood. The normal range is between 500 to 1500 cells/mm^3 [133], people living with HIV have CD4 cell counts below the normal range, usually <500 cells/mm^3. CD4 cell count is an important laboratory indicator of immune function and a strong predictor of HIV progression [134].
- **ART adherence**: "the extent to which a person's behaviour – taking [ART] medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a health care provider" [135].

There are other HIV outcomes related to the provision of care, best described through the "*HIV treatment cascade and care continuum*" framework developed in 2013 [136]. This framework describes HIV care as a dynamic and bidirectional progression of five main steps: (1) diagnosis, (2) linkage to care, (3) retention in care, (4) adherence to ART, culminating in (5) viral suppression (Figure 3) [123, 136].

The abovementioned HIV outcomes are influenced by a variety of sociodemographic factors including age, sex, marital status education level, annual household income, having young children, and mental illness [137-141].

Viral suppression: A study of sociodemographic factors of people living with HIV found that men, individuals aged 30-49 years, as well as those with employment, annual incomes above $10,000 USD, and higher levels of educational attainment had greater odds of viral suppression compared to women, individuals aged 18-29 years, the unemployed, and those with incomes below $10,000 [137]. Additionally, individuals who were married/living together or never married had lower odds of viral suppression compared to those with other relationship statuses [137].

ART adherence: A 2018 qualitative literature synthesis unveiled six interrelated barriers to ART adherence: (1) cognitive and emotional aspects (*affect, beliefs, acceptance, motivation, knowledge*), (2) lifestyle factors (*life demands and organizational issues, substance use*), (3)

social and material context (*social interaction, support and relationships; HIV stigma and concealment; material and structural challenges*), (4) characteristics of ART (*side effects, instructions, physical features*), (5) health experience and state (*body monitoring, comorbidity, manifestations of HIV disease and general health*), and (6) healthcare services and system (*patient-provider relationship, HIV clinic and healthcare system issues, pharmacy issues, health insurance*) [142].

2.8 Knowledge Gaps

A review of the existing literature on adaptive trial designs, causal inference, and novel machine learning approaches such as large language models highlight some important knowledge gaps. The application of epidemiological and biostatistical methods in primary care have recently undergone rapid changes due to increasingly complex health data and limitations of traditional associational inference. Implementation and integration of different types of knowledge e.g., public data from governmental reporting, literature, and domain expertise remains a challenge in practical research settings. Specifically:

> **Gap #1:** How did advances in causal inference and epidemiology influence data reporting and use during the COVID-19 pandemic?
>
> **Gap #2:** It is unknown whether large language models can build directed acyclic graphs in the medical context.
>
> **Gap #3:** With the successful transition of HIV from a terminal disease to a chronic manageable condition, it is unclear what types of health outcomes are most relevant and reported in current HIV studies.
>
> **Gap #4:** Aside from some limited materials on building DAGs, there lacks practical approaches to integrating domain expertise in the process. There is a lack of consensus on how to integrate diverse knowledge sources into causal modelling.

3. CHAPTER 3: STUDY OBJECTIVES

3.1 Research Objectives

The **research objectives** were to:

1. Evaluate epidemiological reporting standards using COVID-19 as a case study, with a focus on assessing the limitations of causal and actionable interpretations of data reported to the public.
2. Assess the potential of large language models in building directed acyclic graphs (DAGs) leveraging the vast corpus of public data for primary care research.
3. Establish a causal mapping approach of the HIV literature to identify frequently reported HIV-related individual-level outcomes with the goal of constructing a comprehensive DAG of HIV outcomes reported in the literature.
4. Develop and evaluate the feasibility of a novel approach for DAG development with domain experts.

4. CHAPTER 4: LIMITATIONS OF CANADIAN COVID-19 DATA REPORTING TO THE GENERAL PUBLIC (MANUSCRIPT 1)

"You keep using that word. I do not think it means what you think it means."

– Inigo Montoya, The Princess Bride

4.1 Preamble

Epidemiology has been defined as "the study of determinants, occurrence and distribution of health and disease in a defined population" [1]. It primarily focuses on describing patterns of disease occurrence through *associational inference*. These analyses evaluate relationships between two variables in a population i.e., joint distribution between two variables [2]. Such analyses typically express these relationships using effect measures such as relative risk, odds ratios, or hazard ratios or through quantifying levels of co-occurrences applying undirected correlation or associational languages [3]. While these measures may be sensitive to and hence reflective of underlying average causal effects, they are also prone to distortion due to unaccounted confounding and selection mechanisms.

Population-wide epidemiological data serves multiple critical functions: guiding the planning and evaluation of strategies to prevent illness and serving as a reference for the management of patients in whom the disease has already developed [4]. However, a disconnect often exists between the associational nature of traditional epidemiology and its application in policymaking, which is inherently causal in intent. Unlike associational inferences, causal relationships cannot be derived solely from joint distributions, as they reflect probabilities under changing conditions such as those induced by an intervention, treatment, or policy [4].

Surprisingly, public health bodies have continued using primarily associational data to inform policy decisions aimed at impacting public health. This practice highlights a crucial challenge in translating epidemiological evidence into effective interventions. The utility and effectiveness of public health policy depends on strong epidemiological evidence and methodological rigor that permits causal conclusions [5]. Ideally, policy decisions would be based on causal inferences, but

in reality, they frequently must be made based on the best available evidence, which often consists of well-established associations from observational studies.

In March 2020, quarantined in my apartment, I watched the COVID-19 pandemic unfold through breaking news broadcasts. I observed inconsistent reporting from news outlets and governmental agencies, unclear and varying definitions of COVID-19 cases across provinces, and comparisons being made across incomparable geographic regions without proper denominators. Additionally, it quickly became apparent that low-income workers and, more often, people of colour were disproportionately impacted by the pandemic; however, no race/ethnicity data was being reported for COVID-19 cases. These observations inspired this manuscript and my desire to contribute as a researcher in solving apparent issues related to how public health data had been reported and potentially misused for decision-making.

Given the importance of valid epidemiological data in informing public health policy, my first manuscript examined this topic within the context of the COVID-19 pandemic. Recognizing and observing firsthand the potential gaps between epidemiological data reported to the public and its application in public health policy decisions, I conducted a longitudinal critical appraisal of the COVID-19 epidemiological data reporting from governmental and news sources from April 2020 to August 2021. This chapter, which was published as manuscript in the Journal of Public Health Policy has been cited 7 times, provides insights into the challenges of translating epidemiological findings into effective public health interventions during a global health crisis.

Long, S., Loutfi, D., Kaufman, J. S., & Schuster, T. (2022). Limitations of Canadian COVID-19 data reporting to the general public. *Journal of Public Health Policy, 43*(1), 203–221. https://doi.org/10.1057/s41271-022-00337-x

Chapter 4 – Manuscript 1

4.2 Title Page

Limitations of Canadian COVID-19 data reporting to the general public

Stephanie Long PhD candidate[a], David Loutfi PhD[a], Jay S Kaufman PhD[b], Tibor Schuster PhD[a]

[a]Department of Family Medicine, McGill University, 5858 Chemin de la Cote-des-Neiges, Suite 300, Montreal, Quebec, Canada, H3Z 1Z1

[b]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 2001 McGill College Avenue, Montreal, Quebec, H3A 1Y7

Keywords: Coronavirus · COVID-19 · Epidemiological reporting standards · Disease reporting

Corresponding author:

Tibor Schuster PhD

Tibor.Schuster@mcgill.ca

4.3 Abstract

Canadian coronavirus (COVID-19) case statistics reported by governmental bodies and news outlets are central to inform the public and to guide health policy. We searched Canadian governmental and news outlets websites to determine how COVID-19 case statistics were reported to the general public, whether they were reported with appropriate denominators, data sources, and accounted for age, sex, and race or ethnicity. Canadian COVID-19 data reporting practices were found to have limited utility due to varying case definitions, heterogeneous and dynamic testing criteria, lack of appropriate standardization accounting for dynamics, sizes, and characteristics of the populations being tested. Population-wide representative COVID-19 testing should be implemented to enable accurate estimation of the scale and dynamics of the epidemiological situation. Comprehensive COVID-19 data on underrepresented and marginalized populations should be collected and reported in an effort to develop equitable health policies.

4.4 Key Message

1. Current COVID-19 case statistics reported to the public by Canadian news outlets and governmental websites do not abide by epidemiological reporting standards and show important data gaps such as lack of COVID-19 case data on race and ethnicity.
2. Population-wide representative COVID-19 testing should be implemented to allow for accurate monitoring of the scale and dynamics of the COVID-19 epidemic in Canada.
3. As currently used indicators for monitoring the pandemic (e.g., COVID-19 case statistics) are surrogate measures prone to large imprecision, more focus should be given to resource-centric measures such as required hospitalizations and occupation of intensive care unit beds in relation to known capacities.

4.5 Introduction

The coronavirus (COVID-19) pandemic presents unprecedented challenges. Widespread cases of COVID-19 and related preventive measures recommended by health authorities and implemented by governments have negatively impacted national economies and societal life. As this unparalleled situation evolves in Canada, provincial and territorial governments continue to affect people's day-to-day lives, and to an unforeseeable extent, the national economy. Thus, a critical appraisal of the foundation of decision and policy making is essential: the epidemiological COVID-19 data collected and reported to the public. Several authors have already commented on challenges related to accurate COVID-19 disease surveillance and modelling of the epidemic situation [6-8].

Minimum requirements for robust and practically relevant inference from population disease data have long been established in the epidemiological literature:

- **Consistent case definition:** consistent *case definition* and unambiguous application of a clinically meaningful diagnostic criterion of the disease in the target population [9-11],
- **Large and representative samples:** use of sufficiently large, *representative* and repeated samples to monitor the prevalence, incidence, and spatio-temporal spread of the disease [7, 12, 13], and
- **Use of appropriate denominators:** appropriate *standardization* and *representation* of disease cases to enable an unbiased evaluation of the epidemic over time and across geographical regions or sub-populations [14, 15].

During the ongoing COVID-19 pandemic, we have observed with utmost concern that none of these criteria have been met for epidemiological data routinely presented in public communications by official news outlets and government bodies. Researchers raised similar concerns about the quality of reporting about the 2014 Ebola outbreak epidemic in West Africa [11]. A recent systemic analysis of 69 Ebola epidemic reports found that only 70% included case definitions and 84% included proportions of patient outcomes such as hospitalizations, mortality, and ICU admittance [12]. These findings draw attention to a serious weakness: that use of appropriate epidemiological standards continues to be a challenge in disease reporting. This analysis [12] included only articles published in scientific journals, not reporting by news outlets

or governmental bodies. Some articles [11, 13, 14] did assess news outlet and social media reporting of the Ebola epidemic, but primarily looking at the impact of sensational reporting, with less focus on epidemiological reporting standards. Regardless, the central message remained – the way in which infectious disease data are reported can greatly influence the public perception of risk associated with a disease.

We sought to appraise the COVID-19 data reporting of the Canadian government (federal, provincial, and territorial) and major Canadian news outlets over time. We demonstrate in several examples why these routinely reported COVID-19 data are of limited utility for informing public health policy. We also illustrate why reporting and comparing absolute case counts alone (such as confirmed cases without appropriate scaling or denominators to the population being tested or eligible for testing, or both) is not only sub-optimal but may misguide public health policy.

4.6 Overview of epidemiological reporting guidelines

4.6.1 The importance of reporting case statistics with appropriate denominators (not only absolute case counts)

Proportions and rates are fundamental concepts in descriptive statistics and epidemiology, because putting observations (such as case counts) in relation to time and populations- at- risk enables a fair comparative assessment of the relevance and dynamics of the problem [2]. Epidemiologists commonly use two indices to describe the presence and emergence of a disease: prevalence and incidence. The former describes the relative frequency (for example, proportion) of a condition in a population at a single point in time or during a specific period of time [15-20]. The latter captures the rate of emerging cases in a specific population, typically reported as the number of new cases per total observation time, often per 100,000 person-years [15-20].

In the definitions of both indices, ‘population’ refers to a distinct group of individuals who are, in general, ‘at risk’ of developing the condition of interest, the population-at-risk [18-20]. In practice, however, the population-at-risk may take into account individuals who actually have zero probability of acquiring the condition under study. For instance, when estimating the prevalence or incidence of shingles in a population, the population-at-risk is typically defined as ‘all adults’, disregarding that a shingles infection is predicated upon a previous varicella-zoster

virus infection (chicken pox) potentially miscounting individuals who have not had a previous infection. Although such imprecisions may affect the overall accuracy of the estimated indices, we can assume the consistency of these errors over time can often be assumed, allowing for informative monitoring of changes of the population disease burden over time.

4.6.2 The curse of dynamically changing invisible denominators

In the context of an epidemic or pandemic disease surveillance, the population-at-risk is not only defined by potentially susceptible individuals, but more strictly by individuals-at-risk who actually have the opportunity to undergo diagnostic testing [19]. In settings where a large proportion of infected individuals remains asymptomatic throughout the course of the disease and testing is primarily available to symptomatic individuals or selected subpopulations (such as close contacts of individuals who tested positive or health professionals potentially exposed to infected individuals) – neither the numerator (case count) nor denominator (population-at-risk to be diagnosed) are truly informative measures. They do not reflect the actual population quantities of interest, rendering prevalence and incidence estimates invalid [17]. In situations where specific sub-populations (such as health workers at a particular site) undergo routine testing, estimates may be useful for monitoring the epidemic situation in this sub-population.

Some might argue that counting confirmed cases in these selected populations suffices to approximate the infamous “[epidemic] curve [to be flattened]”. This approach is, however, problematic for three reasons. First, the number of positive test results strictly depends on the availability of testing and the number of tests conducted in a specific region and population at a given time or time period. These capacities are largely time-dynamic and selection criteria for testing change over time, often in response to emerging evidence on local outbreaks or potential mass exposure to infection. Such events inevitably lead to dynamic changes in the population-at-risk with access to diagnostic tests; hence they lead to unpredictable variations of cases expected over time.

4.6.3 Importance of large and representative samples

Performing diagnostic testing predominately in symptomatic individuals and non-representative subpopulations does not allow for estimation of the prevalence of currently infectious individuals who pose immediate risk to others. Nor does it allow for estimation of the proportion of infected

individuals who show few or no symptoms, an important index for understanding the utility of symptom (self-) screening or monitoring implemented as one of Canada's COVID-19 pandemic response criteria [21]. Several types of COVID-19 tests are available across Canada. Canadian health authorities employ molecular polymerase chain reaction (PCR) tests to detect the presence of the COVID-19 viral DNA via nose swab, throat swab, or saliva sample [22]. For rapid screening of COVID-19 cases, Canada uses point-of-care tests such as rapid antigen tests. Administration of the latter is more rapid, but the results are less accurate than PCR tests in detecting exposure to COVID-19. Though point-of-care tests may be more accurate for detection of transmissible disease [23]. To identify a previous COVID-19 infection, Canadian health authorities use antibody (serology) tests, but these offer limited diagnostic value [22]. Eligibility for PCR COVID-19 testing varies across Canada's geographical regions (provinces and territories) and typically has depended on the presence of symptoms. Availability of the types of COVID-19 tests also varies across regions; some provincial or territorial websites do not explicitly state the type of COVID-19 tests used. As of 5 August 2021, only 7 of 13 provincial and territorial websites explicitly stated the types of COVID-19 testing available. News reports frequently do not differentiate among types of tests. If symptoms that qualify individuals to undergo diagnostic testing are not disease specific and are associated with other conditions prevalent in the population e.g., influenza, seasonal variations in the manifestation of these alternate conditions are also important for determining which populations to test.

4.7 Methods

We are guided by the World Health Organization's definition of health, as "the state of complete *physical*, *mental*, and *social well-being*" [16]. Thus, an assessment of the health-based criteria (such as age, gender) and the disparities in the social determinants of health is essential to a full and meaningful discussion about health, particularly in this unprecedented pandemic.

We sought to appraise the COVID-19 reporting of official governmental websites of the provincial and territorial health institutions and the top 15 Canadian news outlets according to a large international media outlet database (www.allyoucanread.com) in Canada over time from 2020-2021: on 28 April, 2 June, 29 June, 15 September 2020; 15 January, and 19 August 2021. We chose to appraise the news outlets reporting of COVID-19 because, according to framing theory, the way the media *frames* (or reports) an issue can influence individuals' perceptions of

it, and affect their attitudes or behaviours [17]. As observed during other infectious disease reporting such as with Ebola [18], information reported by news outlets may influence actions (or inactions) of individuals in society to protect public health.

Relevant surveillance documents formalize the complexity of case diagnostics and reporting, but none provides guidance for appraisal, though most do promote use of epidemiological indices. A 2001 guideline from the United States Centers for Disease Control and Prevention (CDC) on surveillance [19] provides general recommendations on quality assurance of data reported from a surveillance system. These recommendations confirm the importance of epidemiological principles and of random sampling to verify reported data [20].

A recent literature review identified methodological quality assessment tools available for primary and secondary medical studies, including research pertaining to epidemiological questions such as prevalence [21]. Among the 27 tools described [21], two appeared to be relevant to the objectives of this study: the Joanna Briggs Institute's Checklist for Prevalence Studies (from the University of Adelaide's Faculty of Medical Sciences in South Australia) [22] and the United States Agency for Healthcare Research and Quality (AHRQ) methodology checklist for prevalence study quality [23]. Both checklists proved relevant and covered established criteria for disease reporting described in the epidemiological literature. We deemed neither of the two checklists as sufficiently comprehensive to serve as a standalone tool for appraising the quality of data acquisition, reporting, and interpretation in the current context of the COVID-19 pandemic. Thus, we based our appraisal on standard criteria for rigorous conduct and reporting of epidemiological research from the established literature, including consistent and standardized case definitions, use of appropriate denominators to report case statistics, and large and representative samples for testing.

Two members of our research team (S.L., T.S.) reviewed the provincial and territorial websites and the news outlets. The list of COVID-19 data items they extracted appears in Table 4-1. We created graphs presented in this article in R programming software [24] using the *ggplot2* package [25].

Table 4-1: Data Extraction Form

Governmental websites	News outlets
• COVID-19 case definition • COVID-19 symptoms* • COVID-19 testing eligibility criteria • Use of denominators when reporting COVID-19 case statistics (such as per 100,000) • Whether sources were provided for their COVID-19 case statistics • Relevant population characteristics such as age, sex or gender, and race or ethnicity.	• Presence of dedicated COVID-19 tracker • Use of denominators when reporting COVID-19 case statistics • Whether sources were provided for their reported COVID-19 case statistics • Relevant population characteristics such as age, sex or gender, and race or ethnicity.
*Symptom data was extracted verbatim from provincial and territorial websites, with no changes made to how each reported symptoms or grouped similar symptoms.	

4.8 Results

4.8.1 COVID-19 case definitions, use of denominators, and data sources

Figure 4-1 displays the types of COVID-19 case definitions we found on each provincial or territorial website. **Figure 4-2** presents the epidemiological reporting standards (case definitions, use of denominators to report COVID-19 case counts, and data sources) of the 10 provinces and for the 15 news outlets at each extraction point.

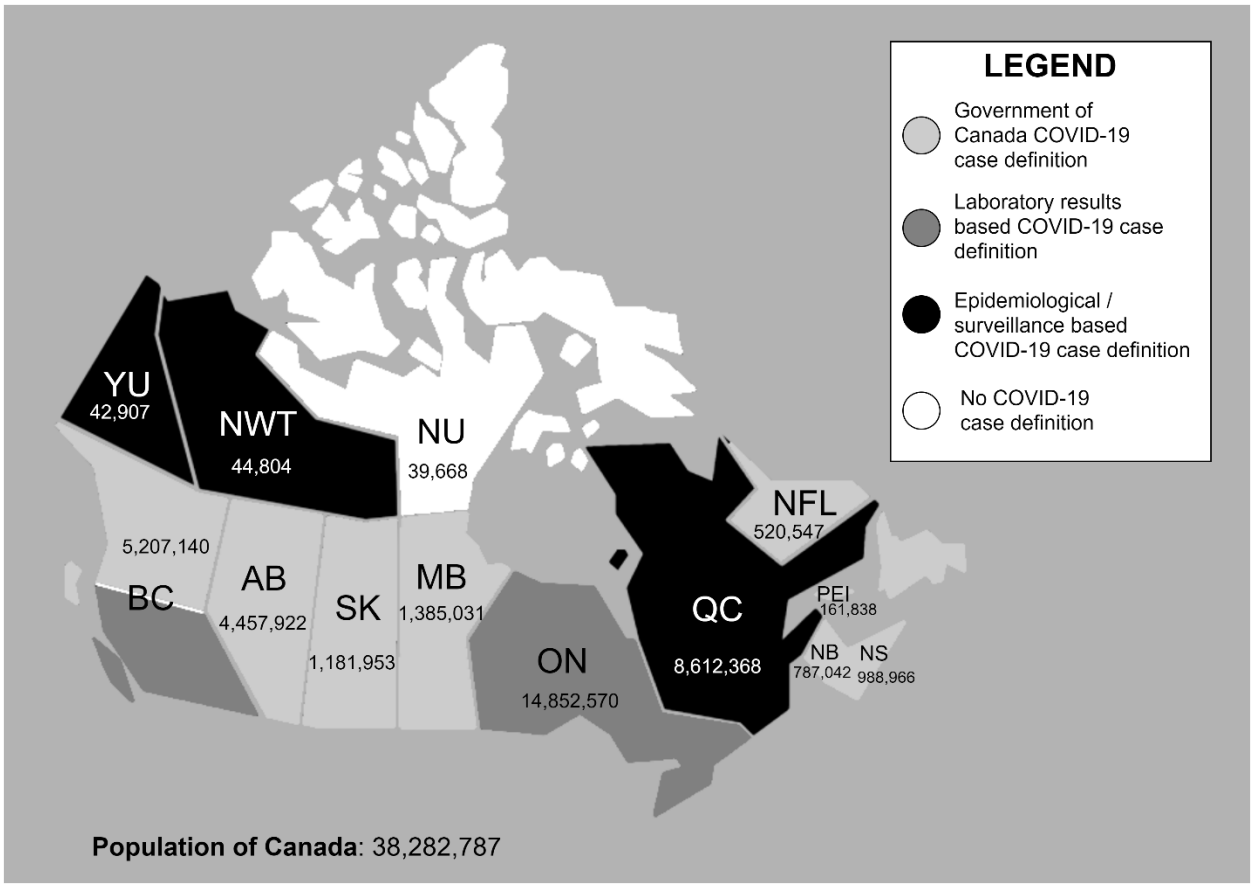


Figure 4-1: COVID-19 Case Definitions used Across Canada. Population sizes of each province and territory have been included (as of 23 July 23, 2021) [81].

At the first timepoint (28 April 2020), only 6 of 13 provinces or territories included COVID-19 *case definitions* on their respective governmental websites. All provincial and territorial websites reported COVID-19 case numbers as *absolute values without a denominator* such as population size, per 100,000, or number of tested individuals. None reported a source of their data. By 2 June 2020, all 10 provinces and 2 of 3 territories displayed *case definitions*; Nunavut was the exception. British Columbia provided the source of its data, an easily accessible one. By 29 June 2020, Alberta, Saskatchewan, and Ontario joined British Columbia in providing data *sources*. By 15 September 2020, *case definitions* and *sources* showed no additional changes. But Newfoundland began to report certain COVID-19 case statistics with denominators.

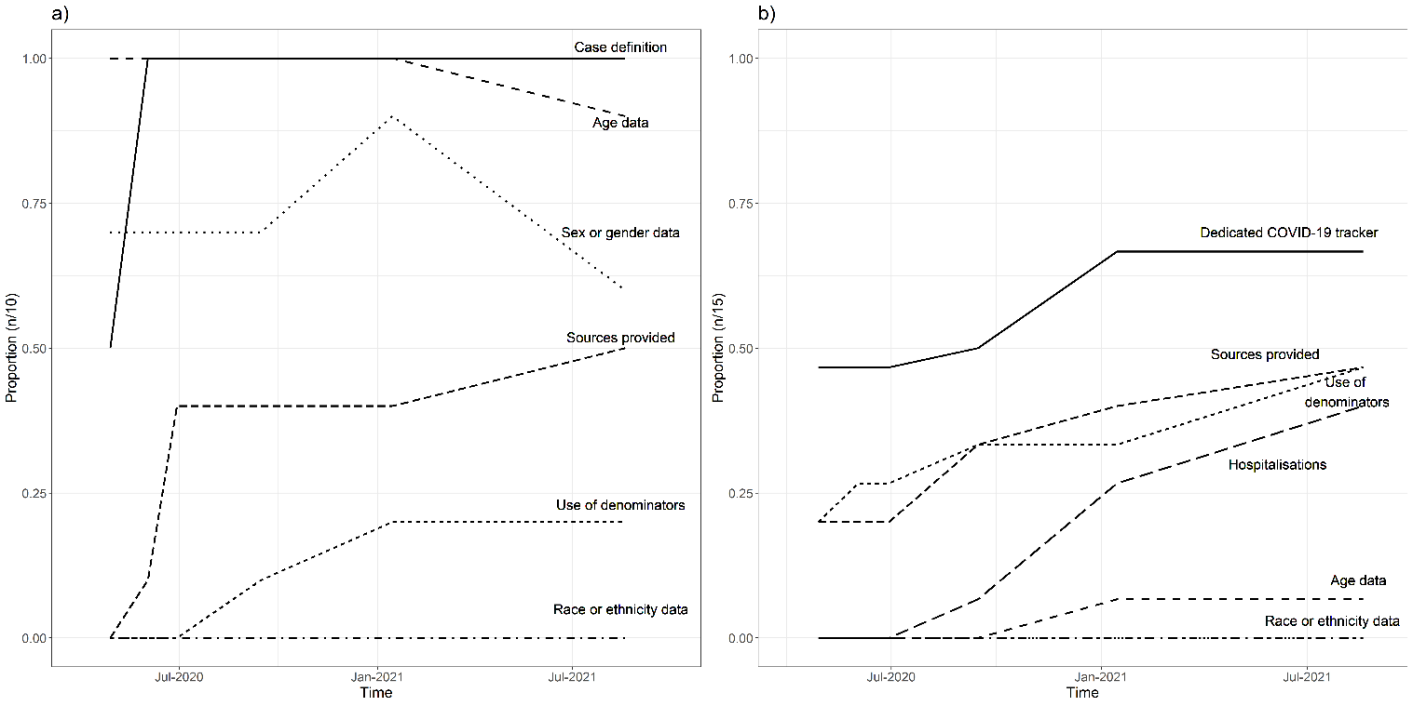


Figure 4-2: Epidemiological reporting standards of a) Canadian Provinces and b) Canadian News Outlets

By 15 January 2021, case definitions showed no changes, and Quebec [26] started providing *sources* for their data. In addition, Newfoundland [27] and Ontario [28] began reporting COVID-19 case statistics with *denominators*. By the final extraction point (19 August 2021), only two additional major changes occurred: Nova Scotia [29] reported COVID-19 case statistics with denominators and provided sources to their data, and Newfoundland [27] ceased use of denominators.

Among the 15 news outlets, two-thirds reported absolute *case counts without applying any denominators*. Few news outlets reported any data sources or provide links directly to those data. Others indicated use of "Government Sources".

4.8.2 COVID-19 symptomatic versus asymptomatic testing

By 28 April 2020, all 13 provinces and territories recommended COVID-19 testing for individuals who had recently travelled out of the country and had reason to believe they had been exposed to COVID-19, or for those experiencing symptoms of COVID-19 (Table 4-2).

COVID-19 testing criteria of all 13 provinces and territories remained unchanged at the second extraction point (2 June 2020). By 29 June 2020, the official governmental websites of Alberta, Saskatchewan, and Manitoba explicitly stated eligibility for certain asymptomatic individuals to undergo COVID-19 testing. Alberta provided the broadest testing, allowing any individual to be tested whether or not that person had any symptoms [30]. Manitoba's COVID-19 testing guidelines allowed testing of asymptomatic individuals, or patients who visited an emergency department, or those admitted into acute care or long-term care facilities [31]. Saskatchewan offered COVID-19 asymptomatic testing only to immunocompromised individuals [32].

By 15 September 2020, most provinces and territories promoted testing of symptomatic individuals on their websites. Alberta and Saskatchewan remained the only two provinces to explicitly encourage testing of asymptomatic individuals. Saskatchewan broadened its testing capacity to allow anyone to receive COVID-19 testing. Manitoba "… developed several options for testing, including introducing voluntary asymptomatic testing for clients in a number of health-care settings and for truck drivers travelling outside of Manitoba to further monitor the presence of COVID-19 in the province" [33].

By 15 January 2021, most provincial and territorial websites still predominately promoted testing of symptomatic individuals. There were, however, some caveats for testing of asymptomatic individuals. These included: individuals having had close contact with a COVID-19 positive person [34-37], individuals requested to test by public health authorities [34, 35], or people who received an exposure notification via the Canadian COVID Alert app [34, 35]. Ontario also identified certain groups as eligible for asymptomatic testing: workers of long-term care facilities, homeless shelters, or other shelters; farmers; Indigenous people; individuals requiring a COVID-19 test prior to surgery; international students who had completed a 14-day quarantine; and individuals who received a positive result from a COVID-19 antigen test [35]. Only two provinces offered asymptomatic testing to any individual: Saskatchewan [32] and Nova Scotia [37]. Previously Alberta [38] and Manitoba [33] had offered asymptomatic testing, then paused by 15 January 2021.

By 19 August 2021, most provincial and territorial websites still predominately promoted testing of symptomatic individuals with the caveats noted above. At that time, Saskatchewan [32],

Quebec [34], New Brunswick [39], Nova Scotia [37], and the Northwest Territories [40] allowed asymptomatic testing.

4.8.3 COVID-19 case data by age, sex, and racial or ethnic minority status

Table 4-2 displays the types of data reported by provinces as of 19 August 2021. As of 15 July 2020, all 10 provinces reported age-stratified data on COVID-19 cases but only Quebec reported age-standardized mortality data [41]. Seven out of ten provinces (British Columbia [42], Alberta [43], Manitoba [44], Quebec [41], Ontario [45], Nova Scotia [29], and Prince Edward Island [46]) provided sex-stratified data. New Brunswick did provide sex-stratified data, but only for COVID-19 tests; it was not clear whether this represented a breakdown of positive test results or of all testing performed in general [47]. None of the territories (Northwest Territories, Yukon Territory, or Nunavut) reported any age- or sex-stratified data on COVID-19 cases. No province or territory reported any COVID-19 data stratified by race or ethnicity. By 15 September 2020, Quebec [41] no longer reported sex-stratified data. By 15 January 2021, reporting remained the same except Quebec [26] began to release sex-stratified data. As of 19 August 2021, reporting remained mostly unchanged, except Quebec [41] and Prince Edward Island [46] no longer released sex-stratified data, and New Brunswick [47] no longer released sex-stratified and age-stratified data.

Figure 4-2 demonstrates that by the final extraction time point (19 August 2021), fewer than half of the news outlets reported COVID-19 case statistics with a denominator and provided sources of data. (One we assessed, Huffington Post Canada, ceased operation as of 9 March 2021.)

Table 4-2: COVID-19 symptoms or testing criteria (or both) and data reported (as of 19 August 2021)

Symptoms or Data reported (or both)	BC	AB	SK	MB	ON	QC	NB	NS	PEI	NFL	YU	NWT	NU
Sex	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Race or ethnicity													
Asymptomatic testing?		~~✓~~[a]	✓[b]	~~✓~~[a]		✓[e]	✓[e]	✓[d]				✓[e]	
Fever	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
New or worsening cough	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Respiratory difficulties or shortness of breath, or both	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sore throat	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Runny, stuffy, or congested nose	~~✓~~[2.1]	✓	✓	✓	✓	~~✓~~[1.1,2.2]	✓	✓	✓	✓	✓	✓	
Sudden loss of smell without nasal congestion	✓	✓[1]	✓	✓	✓	✓	✓		✓	✓[1.1]	✓[1]	✓	
Muscle aches	✓	✓[1]	✓	✓		✓[1.2]	~~✓~~[2.1]	~~✓~~[2]	✓	✓	✓	✓	
Loss of taste	✓	✓[A,1.1]	✓	✓	✓	✓	✓		✓	✓[1.1]	✓[1]	✓	
Extreme fatigue	✓	✓[1]	✓	~~✓~~[2.1]	✓	✓[1.1]	~~✓~~[2.1]	~~✓~~[2]	✓	✓[1.1]	✓		
Diarrhea	✓	✓[A, 1.1]	✓			✓[1.1]	✓	~~✓~~[2]	~~✓~~[2.2]	✓[1]	✓	✓	
Nausea or vomiting	✓	✓[A, 1.1]	✓	✓	✓[1.2]	✓[A]			~~✓~~[2.2]	✓[1.1]	✓	✓	
Loss of appetite	✓		✓	✓[1]	✓[A]	✓[1.2]				✓[1]	✓[1.2]	✓	
Headache	✓	✓[1]	✓	✓		✓[1.2]	✓[1.2]	✓		✓[1.1]	✓[1.2]	✓	
Chills	✓	✓[1]	✓	✓[1.2]	✓[1]			✓	✓	✓	✓[1.2]		
Pink eye	~~✓~~[2.1]	✓[1]	✓	✓[1.1]	✓								
Dizziness or light-headedness or confusion, or a combination of these	~~✓~~[2.1]		✓		✓[1]					~~✓~~[2.1]			
Red, purple, or blueish lesions on extremities (toes, feet, fingers)	~~✓~~[2.1]						✓[A]	~~✓~~[2]		~~✓~~[1, 2.2]			
Hoarse voice				✓				~~✓~~[2]					
Digestive issues		~~✓~~[1, 2.2]			✓								
Painful or difficulty swallowing		✓			✓[1.1]					~~✓~~[2.1]			
Falling down often					✓								
Pneumonia requiring ventilator													✓
Abdominal pain	~~✓~~[2.1]												
Chest pain					✓								
Skin rash of unknown cause				✓							~~✓~~[1.1,2.2]		
Barking cough, making a whistling noise					✓[1.1]								
Stomach aches						✓[1.1]							
Poor feeding				✓[A,1.2]									

BC British Columbia, *AB* Alberta, *SK* Saskatchewan, *MB* Manitoba, *ON* Ontario, *QC* Quebec, *NB* New Brunswick, *NS* Nova Scotia, *PEI* Prince Edward Island, *NFL* Newfoundland and Labrador, *YU* Yukon Territories, *NWT* Northwest Territories, *NU* Nunavut

[A]Only in children or infants

[1]New symptoms added between third (29 June 2020) and fourth (15 September 2020) extraction points

[1.1]New symptoms added between fourth (15 September 2020) and fifth (15 January 2021) extraction points

[1.2]New symptoms added between fifth (15 January 2021) and final (19 August 2021) extraction points

[2]Removed from symptoms list between third (29 June 2020) and fourth (15 September 2020) extraction points

[2.1]Removed from symptoms list between fourth (15 September 2020) and fifth (15 January 2021) extraction points

[2.2]Removed from symptoms list between fifth (15 January 2021) and final (19 August 2021) extraction points

[a]As of 8 June 2020. Asymptomatic testing no longer available, identified between fourth (15 September 2020) and fifth (15 January 2021) extraction points. As of 29 July 2021, testing is only available to symptomatic Albertans, those linked to a known outbreak (symptomatic or not), those travelling, or requiring a Point-of-Care test through their workplace

[b]As of 29 June 2020

[c]As of 19 August 2021 asymptomatic testing no longer available

[d]Became available between fourth (15 September 2020) and fifth (15 January 2021) extraction points

[e]Became available between fifth (15 January 2021) and final (19 August 2021) extraction points

4.9 Discussion

The purpose of this article was to identify and compare the COVID-19 case definitions of all Canadian provinces and territories, to illustrate that COVID-19 case data routinely published and disseminated to the general public is not representative of the respective target populations, and to explain why the reporting and comparing of absolute case counts alone may misguide public health policy. We believe it is important to appraise and understand variations in governmental and news outlet reporting of COVID-19 to the public because reporting may be unintentionally biased which may result in neglect of key public health guidelines.

4.9.1 Relevance of reporting appropriate denominators

Our assessment of COVID-19 case reporting (as of 19 August 2021) revealed that 11 of 13 provincial and territorial websites and 8 of 15 news outlets reported COVID-19 case counts only as absolute numbers. Two exceptions, Nova Scotia [29] and Ontario [28], reported case counts in reference to denominators, COVID-19 cases per 100,000 population. At the penultimate extraction point, Newfoundland [27] had reported case statistics with denominators, but by the final extraction point, no longer did so. The reporting of *only* absolute COVID-19 case counts prevents accurate comparison of the disease spread across geographic regions. Fixed denominators such as population size, however, have only limited utility when assessing the spread of disease over time, as eligibility criteria for testing and testing capacities vary widely, even within one region. Although *population-at-risk* is the ideal denominator, positivity rate, the proportion of all tests performed that are *actually* positive (in a given period of time) [48], is another meaningful measure. Low positivity rates indicate low viral prevalence and adequate surveillance capacity; high positivity rates reflect high viral prevalence or testing strategies focused primarily on symptomatic individuals, or both. Despite its prevalence in news reporting, positivity rates may be biased due to differences in test-seeking or care-seeking behaviour of individuals [6], asymptomatic cases of COVID-19 [49], changes in testing capacities, and imperfect test sensitivity [50]. Obtaining accurate estimates of the burden of the disease is crucial to informing the public health response [7]. Despite this fact, news outlets and governmental bodies remain inclined to compare the disease prevalence and incidence across cities, regions, and countries (Figure 4-3). For example, news outlets repeatedly called Montreal the “epicentre of the pandemic” [51, 52] in Canada, as did public health officials [53, 54] based on its high

absolute number of COVID-19 cases. Even Prime Minister Justin Trudeau [55] expressed concern for Montreal residents. None point outed that Laval, a city north of Montreal and the third largest city in Quebec following Montreal and Quebec City, experienced similar proportions of positive COVID-19 cases and death rates of COVID-19 as Montreal (Figure 4-3).

Table 4-3: Examples of news outlet reporting of COVID-19 pandemic

Date	**Headline & Quote**	**Issue**	
		Reporting absolute counts of COVID-19 cases	Comparing across geographical regions
3 April 2020	"'Montreal is the epicentre of the pandemic,' public health director says "We are starting to come into the ascending slope of the epidemic," with 480 new cases for a total of 2,642."" – *Montreal Gazette*[a]	✓	
22 April 2020	"With 9,856 cases, Montreal region remains Canada's COVID-19 epicentre." - *Montreal Gazette*[b]	✓	✓
11 May 2020	"Trudeau fears COVD-19 deaths will spike in Montreal, Canada's virus epicentre, as Legault reopens Quebec." - *The National Post*[c]		✓
15 May 2020	"'Society failed': Legault visits Montreal as Quebec becomes the world's seventh deadliest COVID-19 epicentre." - *The National Post*[d]		✓
21 May2020	"Nurses recount 'hell' in Laval, Canada's new COVID-19 epicentre, and ask what they can withstand." – *CTV News*[e]		✓
13 January 2021	"Montreal 'once again the epicentre' of COVID-19 crisis as city adds hundreds of hospital beds." – *Global News*[f]		✓
26 February 2021	"B.C. the Florida of Canada? Epidemiologist says his comparison was meant as a warning" – *CTV News*[g]		✓

24 March 2021	"How Regina's COVID-19 cases compare to other Canadian, American cities" – *CTV News*[h]	✓	✓
15 May 2021	"Why does Manitoba have nearly twice as many COVID-19 deaths as Saskatchewan?" – *CBC News*[i]		✓
5 August 2021	"Ontario reports 213 new COVID-19 cases; 14 more deaths with 12 due to data clean-up." – *CP24*[j]	✓	

[a]'Montreal is the epicentre of the pandemic,' public health director says. 2020 [cited 8 June 2020]. Available from: https://montrealgazette.com/news/local-news/montreal-hit-by-rapid-rise-in-number-of-covid-19-cases

[b]Lalonde M. With 9,856 cases, Montreal region remains Canada's COVID-19 epicentre. Montreal Gazette. 22 April 2020

[c]The Canadian Press, National Post Staff. Trudeau fears COVID-19 deaths will spike in Montreal, Canada's virus epicentre, as Legault reopens Quebec. National Post. 11 May 2020

[d]The Canadian Press. 'Society failed': Legault visits Montreal as Quebec becomes the world's seventh deadliest COVID-19 epicentre. The National Post. 15 May 2020

[e]Greig K, Ross S. Nurses recount 'hell' in Laval, Canada's new COVID-19 epicentre, and ask what they can withstand. CTV News. 23 May 2020

[f]Laframboise K. Montreal 'once again the epicentre' of COVID-19 crisis as city adds hundreds of hospital beds 2021 [updated 13 January 2021. Available from: https://globalnews.ca/news/7574222/montreal-coronavirus-update-january-2021/

[g]Holliday I. B.C. the Florida of Canada? Epidemiologist says his comparison was meant as a warning. CTV News. 26 February 2021

[h]Soloman M. How Regina's COVID-19 cases compare to other Canadian, American cities. CTV News. 2021

[i]MacLean C. Why does Manitoba have nearly twice as many COVID-19 deaths as Saskatchewan? CBC News. 2021

[j]Wilson K. Ontario reports 213 new COVID-19 cases; 14 more deaths with 12 due to data clean-up. CP24. 5 August 2021

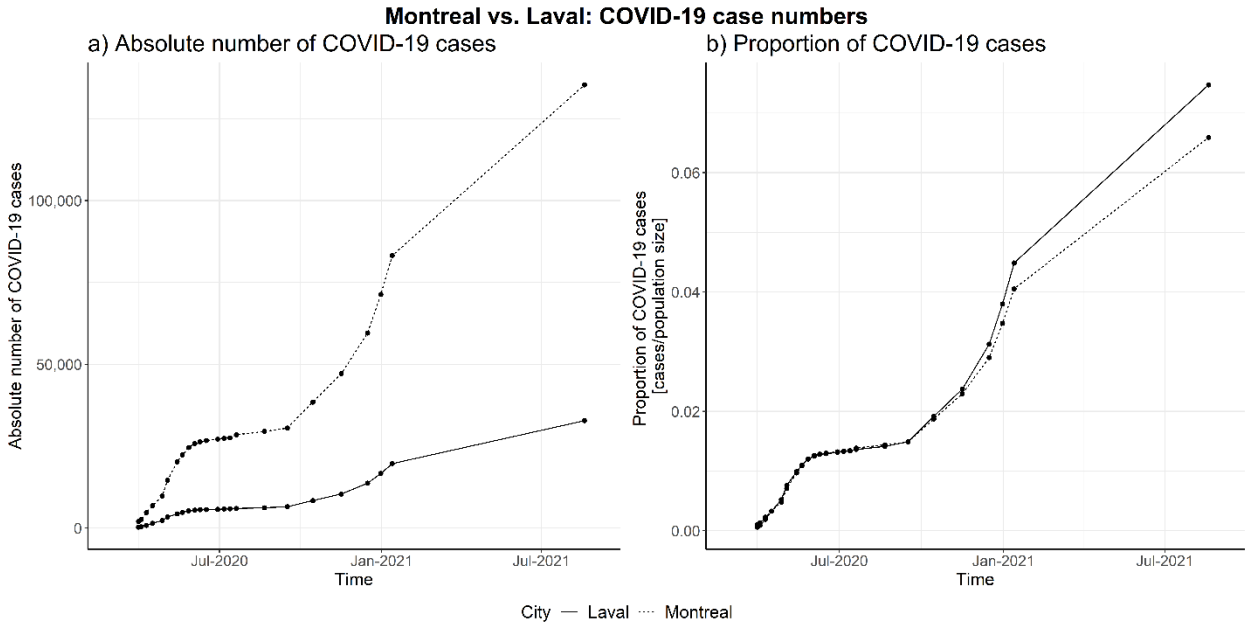


Figure 4-3 Comparing COVID-19 case statistics in Montreal and Laval (Quebec, Canada).

Data source: Government of Quebec (as of 19 August 2021): a) Absolute case counts and b) proportions (cases/population size). Note: These curves do not reflect changes in testing capacities or selection of individuals being testing over time

4.9.2 Symptom-based testing predominates

We found that most provincial and territorial websites recommended COVID-19 testing primarily to individuals experiencing symptoms of COVID-19. And some provinces (British Columbia and Ontario) explicitly discouraged testing of individuals without symptoms: "if you don't have any symptoms, testing is not recommended even if you are a contact" [56] and "[Public Health Ontario] does not currently recommend routine testing of asymptomatic persons for COVID-19." [57]. By 15 September 2020, only the websites of Alberta [30], Saskatchewan [32], and Manitoba [58] explicitly noted availability of COVID-19 testing to asymptomatic individuals or certain priority groups. By 15 January 2021, however, Alberta and Manitoba had paused their asymptomatic testing. By 19 August 2021, more provinces (including Quebec, Nova Scotia, and New Brunswick) allowed for asymptomatic testing.

Despite the content of postings for the public on provincial and territorial websites, in practice, COVID-19 testing may be more widely available. For instance, Public Health Ontario

recommended that healthcare providers "should continue to use their discretion to make decisions on which individuals to test [for COVID-19]" [57]. Additionally, although not always stated explicitly on the websites, some provinces may have expanded COVID-19 testing for priority groups such as healthcare workers, first responders, teachers, immunocompromised individuals, patients who had been admitted to acute care hospitals, among others. COVID-19 case statistics that rely on symptomatic testing may grossly underestimate the true extent of spread of the epidemic. Thus, these findings highlight the need for large-scale representative testing to enable accurate estimation of the disease's scale and dynamics.

Despite the importance of large-scale representative testing, it has taken nearly 1.5 years after the initial lockdowns in Canada for health authorities to implement it. On 6 July 2021, Statistics Canada and the COVID-19 Immunity Task Force released preprint data on COVID-19 seroprevalence of >10,000 Canadians tested between November 2020 and April 2021 [59]. Researchers found that between those dates, 2.6% of Canadians had COVID-19 antibodies, another 1% had the antibodies due to vaccinations. (COVID-19 vaccines were not widely available during the survey period) [59].

4.9.3 Lack of data on racial and ethnic minorities

Our analysis of COVID-19 reporting by governmental websites also uncovered absence of any reporting on race or ethnicity. After nearly 1.5 years since the initial lockdowns in Canada, no province or territory reported any COVID-19 data on race or ethnicity. This failure prevents Canadian public health authorities from understanding how COVID-19 impacts these groups. This is particularly problematic given growing evidence that COVID-19 disproportionately affects racial and ethnic minorities. According to the COVID Racial Data Tracker Project, a collaboration between *The Atlantic* and Boston University aimed at gathering race and ethnicity data on COVID-19 in the United States, "nationwide, Black people are dying at 1.5 times the rate of White people" [60]. Other racial and ethnic minority groups are also adversely affected; Indigenous and Latinos experience mortality rates of 138 and 121 deaths per 100,000 respectively compared to 98 for White Americans (as of 26 January 2021) [60].

These disparities may be attributed to inequities in the social determinants of health such as access to healthcare, socioeconomic conditions (including poverty and the stress that accompanies it), housing, and occupation [61]. An additional explanation that must not be

ignored is systemic racism, a term used to convey "racism [that] is embedded in the policies [and practices] of public and private institutions" [62]. Systemic racism can exist even if no one in the institution is racist, but historically architects of the system and structure of the institution built these in a way that favours certain groups over others. Racial and ethnic minorities are more likely to be low-income, frontline workers (healthcare workers, caretakers, delivery drivers, among others), and live in housing and multi-generational homes [63, 64] under "conditions ripe for [the] spread of coronavirus" [65].

A Statistics Canada report found that neighbourhoods in Quebec, Ontario, and British Columbia with the highest proportions of visible minority residents (>25%) had an age-standardized COVID-19 mortality rate per 100,000 population at least two times that of neighbourhoods with less than 1% visible minority residents [66]. These results align with those reported earlier by CBC News in Montreal [63] and Toronto [64]. The data gaps in race and ethnicity led to numerous calls to collect COVID-19 data on race and ethnicity from committees and community groups in Montreal [67, 68], Toronto [69], Vancouver [70], and Nova Scotia [71]. These data would allow us to better understand changes in the COVID-19 pandemic and identify the most vulnerable at-risk groups.

4.9.4 Limitations

This study's limitations stem from its design. As only two members of the research team reviewed governmental and news outlet websites, we may have missed some data. We attempted to mitigate this by reviewing the sources at least twice at each extraction point and by consulting the Internet Archive (www.archive.org). Another limitation is the lack of consistent time intervals between the data extraction points. We found, however, that the data reporting methods of the governmental bodies and news outlets did not evolve as rapidly as the COVID-19 pandemic itself.

4.10 Conclusion

Accurate monitoring of the course of the COVID-19 epidemic is critical for determining which population-wide measures are necessary (or unnecessary) to prevent spread of the disease or subsequent 'waves' of the pandemic, or both. With the currently implemented measures on testing for, and reporting of COVID-19 cases, it is statistically difficult to arrive at a valid

numerical projection of reality. As travel restrictions are still in place and household members likely accessible to those gathering data, (repeated) random sampling of the population to be tested would enable a more accurate and precise estimate of the status and development of the epidemic. Random (household) sampling enables an accurate representation of the population and its manifold characteristics when determining the epidemic situation in a population. It may be surprising to the reader, but even with a size n=500 random samples (in a specific neighborhood, for example), an estimated proportion (the prevalence of COVID-19 positive persons) would yield a half-width of less than 5% for the respectively associated 95% confidence interval. That is, random sampling and symptom-independent antibody testing would enable us to learn about the percentage of the population still at risk of acquiring a COVID-19 infection or who have already been exposed to the virus (seropositive) [72, 73]. This would help citizens and policy makers understand the actual scale of the ongoing epidemic and provide invaluable guidance on which preventive measures are most effective, thus necessary, and which are not, now, or soon. We also need greater focus on monitoring clinically relevant case trajectories such as people affected by COVID-19 requiring hospitalization or intensive care (rather than counts of positive tests). Understanding these case statistics in relation to available healthcare capacities is imperative as they are the key indicators of the direct impact of the ongoing pandemic on the health system and people's lives.

4.11 References

1. Brachman P. Epidemiology In: Baron S, editor. Medical Microbiology. Galveston (TX): University of Texas Medical Branch at Galveston

Copyright © 1996, The University of Texas Medical Branch at Galveston.; 1996.

2. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.

3. Pearl J. Statistics and causal inference: A review. Sociedad de Estadistica e Investigacion Operativa. 2003;12(2):281-345.

4. Coggon D, Rose GA, Barker DJP. Epidemiology for the uninitiated. 5th ed ed. London: BMJ Books; 2003.

5. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(Volume 34, 2013):61-75.

6. Vespignani A, Tian H, Dye C, Lloyd-Smith JO, Eggo RM, Shrestha M, et al. Modelling COVID-19. Nat Rev Phys. 2020:1-3.

7. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions. Am J Public Health. 2020;110(7):949-51.

8. Yiannakoulias N, Slavik CE, Sturrock SL, Darlington JC. Open government data, uncertainty and coronavirus: An infodemiological case study. Soc Sci Med. 2020;265:113549.

9. Norms and standards in epidemiology: case definitions Epidemiol Bull. 1999;20(1):12-3.

10. Wharton M, Chorba T, Vogt R, Morse D, Buehler J. Case definitions for public health surveillance. MMWR Recommendations and Reports. 1990;39:1-43.

11. Coggon D, Martyn C, Palmer K, Evanoff B. Assessing case definitions in the absence of a diagnostic gold standard. International Journal of Epidemiology. 2005;34:949-52.

12. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. BJPsych bulletin 2016;40(2):57-60.

13. Fraser GE. The estimation of disease frequency using a population sample. International Journal of Epidemiology. 1978;7(3):277-84.

14. Rose GA, Barker DJP. Epideiology for the uninitiated: Comparing rates. British Medical Journal. 1978;2:1282-3.

15. Broeck J, Brestoff JR, Kaulfuss C. Statistical Estimation. Epidemiology: Principles and Practical Guidelines. Heidelberg: Springer; 2013. p. 417.

16. World Health Organization. Constitution of the World Health Organization. 54th ed: World Health Organization; 2006.

17. Scheufele DA. Framing as a Theory of Media Effects. Journal of Communication. 1999;49(1):103-22.

18. Humphries B, Radice M, Lauzier S. Comparing "insider" and "outsider" news coverage of the 2014 Ebola outbreak. Can J Public Health. 2017;108(4):e381-e7.

19. U.S. Department of Health and Human Services. Updated guidelines for evaluating public health surveillance systems: Recommendations from the Guidelines Working Group. Atlanta, GA: Centers for Disease Control and Prevention (CDC); 2001. Contract No.: RR-13.

20. Klevens RM, FLeming PL, Neal JJ. Mode of Transmission Validation Study Group. Is there really a heterosexual AIDS epidemic in the United States? Findings from a multisite validation study. Am J Epidemiol. 1999;149:75-84.

21. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res. 2020;7(1):7.

22. The Joanna Briggs Institute. Checklist for Prevalence Studies. The Joanna Briggs Institute; 2017.

23. Rostom A, Dubé C, Cranney A, Saloojee N, Sy R, Garritty C, et al. Celiac DIsease. Agency for Healthcare Research and Quality; 2004.

24. R Core Team. R: A language and environment for statistical computing. Vienna, Austraia: R Foundation for Statistical Computing; 2019.

25. Wickham H. ggplot2: Elegant Graphics for Data Analytics. New York, NY: Springer-Verlag; 2016.

26. Government of Quebec. Coronavirus disease (COVID-19) in Québec 2020 [Available from: https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/.

27. Government of Newfoundland. COVID-19 Home: Pandemic Update 2020 [Available from: https://covid-19-newfoundland-and-labrador-gnl.hub.arcgis.com/.

28. Government of Ontario. All Ontario: Case numbers and spread 2021 [Available from: https://covid-19.ontario.ca/data.

29. Government of Nova Scotia. COVID-19: case data in Nova Scotia - Find information on COVID-19 cases in Nova Scotia. Data includes age range, gender and location (by NSHA zone). Data is reported daily. 2020 [updated July 15, 2020. Available from: https://novascotia.ca/coronavirus/data/.

30. Alberta Health Services. Novel coronavirus (COVID-19): COVID-19 Testing Available for Everyone 2020 [updated June 30, 2020. Available from: https://www.albertahealthservices.ca/topics/Page16944.aspx.

31. Shared health Manitoba. COVID-10 Asymptomatic Surveillance: Information for Providers. 2020 June 12, 2020.

32. Government of Sasketchewan. Testing Information 2020 [Available from: https://www.saskatchewan.ca/government/health-care-administration-and-provider-resources/treatment-procedures-and-guidelines/emerging-public-health-issues/2019-novel-coronavirus/testing-information.

33. Government of Manitoba. COVID-19 Testing 2020 [Available from: https://manitoba.ca/covid19/updates/testing.html.

34. Government of Quebec. Testing for COVID-19 2021 [Available from: https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/testing-for-covid-19/.

35. Government of Ontario. COVID-19 test and testing location information 2021 [updated January 6, 2021. Available from: https://covid-19.ontario.ca/covid-19-test-and-testing-location-information#where-and-when-to-get-tested.

36. BC Centre for Disease Control. Testing information 2020 [Available from: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/testing.

37. Government of Nova Scotia. Coronavirus (COVID-19): symptoms and testing 2021 [Available from: https://novascotia.ca/coronavirus/symptoms-and-testing/#who-can-be-tested.

38. Government of Alberta. Symptoms and testing 2021 [Available from: https://www.alberta.ca/covid-19-testing-in-alberta.aspx.

39. Government of New Brunswick. Coronavirus disease (COVID-19): If you think you have symptoms 2020 [Available from: https://www2.gnb.ca/content/gnb/en/corporate/promo/covid-19.html.

40. Government of Northwest Territories. Getting tested for COVID-19 2020 [updated May 25, 2020. Available from: https://www.gov.nt.ca/covid-19/en/services/getting-tested-covid-19.

41. Gouvernment du Quebec. Situation of the coronavirus (COVID-19) in Québec 2020 [Available from: https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/situation-coronavirus-in-quebec/#c51839.

42. BC Centre for Disease Control. British Columbia COVID-19 Dashboard 2020 [updated July 14, 2020. Available from: https://experience.arcgis.com/experience/a6f23959a8b14bfa989e3cda29297ded.

43. Government of Alberta. COVID-19 Alberta statistics: Interactive aggregate data on COVID-19 cases in Alberta 2020 [Available from: https://www.alberta.ca/stats/covid-19-alberta-statistics.htm.

44. Government of Manitoba. Manitoba COVID-19 2020 [updated July 14, 2020. Available from: https://experience.arcgis.com/experience/f55693e56018406ebbd08b3492e99771.

45. Government of Ontario. How Ontario is responding to COVID-19: Learn about coronavirus (COVID-19) cases in Ontario and how the province is keeping people safe. 2020 [updated July 14, 2020. Available from: https://www.ontario.ca/page/how-ontario-is-responding-covid-19#section-0.

46. Government of Prince Edward Island. PEI COVID-19 Testing Data 2020 [updated July 15, 2020. Available from: https://www.princeedwardisland.ca/en/information/health-and-wellness/pei-covid-19-testing-data.

47. Government of New Brunswick. New Brunswick COVID-19 Dashboard 2020 [updated July 15, 2020. Available from: https://experience.arcgis.com/experience/8eeb9a2052d641c996dba5de8f25a8aa.

48. Mercer TR, Salit M. Testing at scale during the COVID-19 pandemic. Nat Rev Genet. 2021;22(7):415-26.

49. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. N Engl J Med. 2020;382(12):1178-9.

50. Wu SL, Mertens AN, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. Nat Commun. 2020;11(1):4507.

51. Laframboise K. Montreal 'once again the epicentre' of COVID-19 crisis as city adds hundreds of hospital beds 2021 [updated January 13, 2021. Available from: https://globalnews.ca/news/7574222/montreal-coronavirus-update-january-2021/.

52. The Canadian Press. 'Still worrisome': Montreal remains COVID-19 epicentre of Quebec, with hospitals at full capacity 2021 [updated January 22, 2021. Available from: https://montreal.ctvnews.ca/still-worrisome-montreal-remains-covid-19-epicentre-of-quebec-with-hospitals-at-full-capacity-1.5278603.

53. 'Montreal is the epicentre of the pandemic,' public health director says [Internet]. 2020 [cited June 8, 2020]. Available from: https://montrealgazette.com/news/local-news/montreal-hit-by-rapid-rise-in-number-of-covid-19-cases.

54. Maratta A. COVID-19: Quebec reports 169 new cases as province steps up testing capacity 2020 [updated July 26, 2020. Available from: https://globalnews.ca/news/7219152/quebec-covid-19-coronavirus-july-26/.

55. The Canadian Press, National Post Staff. Trudeau fears COVID-19 deaths will spike in Montreal, Canada's virus epicentre, as Legault reopens Quebec. National Post. 2020 May 11, 2020.

56. BC Centre for Disease Control. Testing information 2020 [Available from: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/testing.

57. Public Health Ontario. COVID-19 Laboratory Testing in Ontario 2020 [Available from: https://www.publichealthontario.ca/en/diseases-and-conditions/infectious-diseases/respiratory-diseases/novel-coronavirus/lab-testing-ontario.

58. Government of Manitoba. Cases and Risk of COVID-19 in Manitoba 2020 [Available from: https://www.gov.mb.ca/covid19/updates/data.html.

59. Study reveals children and youth had highest rates of SARS-CoV-2 infection in Canada before third wave [press release]. Montreal, QC2021.

60. The COVID Tracking Project. The COVID Racial Data Tracker 2020 [Available from: https://covidtracking.com/race.

61. Centers for Disease Control and Prevention. Health Equity Considerations and Racial and Ethnic Minority Groups 2020 [updated July 24, 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html.

62. Harmon A, Mandavilli A, Maheshwari S, Kantor J. From Cosmetics to NASCAR, Calls for Racial Justice Are Spreading. The New York Times. 2020 June 13, 2020.

63. Rocha R, Shingler B, Montpetit J. Montreal's poorest and most racially diverse neighbourhoods hit hardest by COVID-19, data analysis shows 2020 [updated June 11, 2020. Available from: https://www.cbc.ca/news/canada/montreal/race-covid-19-montreal-data-census-1.5607123.

64. CBC News. Lower income people, new immigrants at higher COVID-19 risk in Toronto, data suggests 2020 [updated May 12, 2020. Available from: https://www.cbc.ca/news/canada/toronto/low-income-immigrants-covid-19-infection-1.5566384.

65. Olson I, Mignacca F. Montréal-Nord responds to call for help as COVID-19 cases climb in the borough Montreal2020 [updated April 30, 2020. Available from: https://www.cbc.ca/news/canada/montreal/montr%C3%A9al-nord-covid-19-highest-rate-1.5548712.

66. Subedi R, Greenburg L, Turcotte M. COVID-19 mortality rates in Canada's ethno-cultural neighbourhoods. Statistics Canada,; 2020.

67. Carpenter P. Montreal groups demand the collection of race-based data during COVID-19 testing 2020 [updated May 6, 2020. Available from: https://globalnews.ca/news/6913506/montreal-groups-race-based-data-coronavirus-testing/.

68. Mignacca F. Quebec is still not publishing race-based data about COVID-19. These community groups aim to fill the void: CBC News; 2020 [updated August 19, 2020. Available from: https://www.cbc.ca/news/canada/montreal/community-groups-launch-national-covid-19-race-database-1.5691937.

69. CBC News. Toronto will start tracking race-based COVID-19 data, even if province won't Toronto2020 [updated April 22, 2020. Available from: https://www.cbc.ca/news/canada/toronto/toronto-covid-19-race-based-data-1.5540937.

70. Watson B. Race-based COVID-19 data collection should be mandatory, says City of Vancouver committee Vancouver2020 [updated June 9, 2020. Available from: https://www.cbc.ca/news/canada/british-columbia/city-committee-race-data-covid19-1.5604442.

71. Field A, Quon A. Canadian officials urged to collect race-based health data during COVID-19. Global News. 2020 July 15, 2020.

72. Baum JCA, Rowley R. We need random testing to gather data on COVID-19. The Toronto Star. 2020.

73. Simpson E, Harris T. A National Random Testing is Essential to Determine the True Infection and Mortality Rate of Coronavirus: Frontier Centre for Public Policy; 2020 [Available from: https://fcpp.org/2020/04/25/a-national-random-testing-is-essential-to-determine-the-true-infection-and-mortality-rate-of-coronavirus/

5. CHAPTER 5: CAN LARGE LANGUAGE MODELS BUILD CAUSAL GRAPHS? (MANUSCRIPT 2)

"The limits of my language mean the limits of my world" – Ludwig Wittgenstein

5.1 Preamble

Findings from my first manuscript demonstrated that Canadian COVID-19 data reporting on governmental and news outlet websites fell short of proper epidemiological standards e.g., varying case definitions, inappropriate or absent denominators [1]. This shortcoming raised concerns regarding the data's use in informing health policy, which inherently involves causal considerations [2]. Consequently, this data provided limited utility for guiding policy development, as relying on non-representative or incomplete data may introduce systematic biases and potentially harmful outcomes.

These limitations highlighted the need for more routine implementation of causal inference modelling strategies in public health epidemiology [2]. Directed acyclic graphs (DAGs) are the central tool, as their creation is typically the first step in causal modelling. DAGs are powerful tools for representing causal relationships, and their accurate construction is vital for sound causal inference and modeling [3]. Domain expertise remains the most valuable tool for creating DAGs; however, the ever-increasing amount and complexity of emerging health data would necessitate continual updating of these DAGs. Leveraging machine learning tools to automate this process could enable a more efficient approach to this step of causal modelling.

Therefore, this chapter investigated whether LLMs, a class of machine learning algorithms, could assist in constructing DAGs. LLMs are trained on a vast corpus of textual data and differ from other trained algorithms in their unique ability to interpret and extract knowledge from text data encoding human conversations and written text. Additionally, it is entirely possible that a domain expert may have omitted some nodes or edges that actually exist in the data. While other machine learning models that primarily identify patterns in structured data, LLMs are designed

to understand and generate human-like text. This capability allows them to potentially capture the nuanced reasoning and domain expertise embedded in human language, bringing us closer to the requirement of external expertise in building DAGs.

The technological breakthrough that enabled the remarkable performance of today's LLMs is the transformer [4], a type of neural network architecture that can focus its attention on the most relevant and important parts of the input sequence. This was a significant advancement from earlier natural language processing feature extraction methods such as term-frequency inverse document frequencies (TD-IDF, *numerical statistic which reflects the importance of words in a specific document*) and bag-of-words (*an unordered collection of words and their frequencies*), which do not take into account the sequence of words and thus are incapable of tasks like machine translation [5, 6]. Subsequent early language models used recurrent neural networks, which are neutral networks that do consider the temporal sequence of input data [7], a very important aspect of text data.

The potential of LLMs in causal inference lies in their ability to construct DAGs not just from data points, but by leveraging human-generated text and other literature. This approach may bridge the gap between purely data-driven models and expert-crafted DAGs, potentially leading to more comprehensive and efficient causal modelling. Assuming the veracity of the information in their training data, LLMs may offer an opportunity to move beyond pattern recognition towards a form of causal learning that more closely mimics human reasoning.

When this work was initiated, LLMs had not yet reached the prominence of ChatGPT, which was months away from being released. LLMs could perform some natural language processing tasks like classification and sentiment analysis but lacked the advanced capability of contemporary chatbots as we now know them (e.g., OpenAI's ChatGPT [8], Anthropic's Claude [9], and Meta's Llama [10]).

This chapter was published and presented at the Conference on Neural Information Processing Systems (NeurIPS) 2022 Workshop on Causal Machine Learning for Real-World Impact (CML4Impact, 2022) in New Orleans in November 2022. NeurIPS is the top-tier high impact conference for research in machine learning (ML) and artificial intelligence (AI). ML/AI conferences like NeurIPS, differ from those in medicine and health science in that abstract submissions require full and completed manuscripts (8-pages maximum), not 300-word abstracts

[11]. The submissions undergo peer review from at least 3 reviewers from the field with expertise in the domain. For the full conference and associated workshops, NeurIPS 2022 received 10,411 abstract submissions, 25.7% (n=2671) were accepted; of those, 23.9% were poster presentations and 1.8% were oral presentations [12]. In 2022, there were 15,530 hybrid registrations to the conference, 9,560 attended in-person [13]. The number of submissions to NeurIPS has increased substantially in recent years, from 1,420 total in 2013 to 12,345 in 2023 [12]; illustrating the growing prominence of this field and topic of study.

Beyond attendance numbers (875 attendees in 2022), I was unable to find similar statistics for North American Primary Care Research Group (NAPCRG) conference, which faculty and students of Department of Family Medicine at McGill University attend each year. In response to an email inquiring about acceptance and submission rates, a representative of NAPCRG informed me, “Our acceptance rates are based on the number of sessions we can accommodate at the conference venue and that varies each year. There is not a consistent or average rate we aim for, so I don’t have any data I can share.”

At the time of publication, this manuscript was one of the very few exploring LLM use in causal modelling. Notably, this chapter was presented just days before the release of ChatGPT, positioning it at the forefront of this emerging field. It has since between cited 37 times.

Long S, Piché A, Schuster T. (2023) *Can large language models build causal graphs?* NeurIPS 2022 Workshop on Causal Machine Learning for Real-World Impact (CML4Impact, 2022), New Orleans, USA.

Long, S., Schuster, T., & Piché, A. (2024). Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

5.2 Title Page

Can large language models build causal diagrams?

Stephanie Long PhD candidate[a], Tibor Schuster PhD[a], Alexandre Piché PhD candidate[b, c]

[a]Department of Family Medicine, McGill University, 5858 Chemin de la Cote-des-Neiges, Suite 300, Montreal, Quebec, Canada, H3Z 1Z1

[b]Mila – Quebec AI Institute, Université de Montréal, 6666 Rue Saint-Urbain, Montreal, Quebec, Canada, H2S 3H1

[c]Service Now Research, 6650 Rue Saint-Urbaine, Suite 500, Montreal, Quebec, H2S 3G9

Corresponding author:

Stephanie Long

Stephanie.Long@mail.mcgill.ca

5.3 Abstract

Building causal graphs can be a laborious process. To ensure all relevant causal pathways have been captured, researchers often have to discuss with clinicians and experts while also reviewing extensive relevant medical literature. By encoding common and medical knowledge, large language models (LLMs) represent an opportunity to ease this process by automatically scoring edges (i.e., connections between two variables) in potential graphs. LLMs however have been shown to be brittle to the choice of probing words, context, and prompts that the user employs. In this work, we evaluate if LLMs can be a useful tool in complementing causal graph development.

5.4 Introduction

Advances in causal inference have important implications in empirical research as most research questions asked in the health and medical context are not associational, but causal in nature. Examples of such research questions include: What is the efficacy of a given drug in a given population? What is the expected effect of a given intervention on a specific outcome? Common amongst these research questions is the desire to uncover the cause-and-effect relationships amongst a set of variables i.e., treatments, interventions, and outcomes. Such causal questions cannot be answered from (observed) data alone or from the distributions that govern said data [14]. In addition, external knowledge is needed to understand the underlying data-generating mechanisms to enable the setup of an appropriate 'inference engine'.

Causal diagrams play a central role in causal inference because they encode contextual knowledge of the observable and unobservable variables, and their causal dependencies. Causal inference pioneer Judea Pearl refers to the nodes in a causal diagram as a "society of listening variables" [15]. The term "listening" stresses the defining property of directed and acyclic relationships between the variables, i.e., listening being asymmetrical, variable A listening to variable B, does not imply variable B listening to variable A, motivating the commonly adapted nomenclature of Directed Acyclic Graphs (DAGs) [16, 17].

The first step when aiming to address causal questions using data is to draw a causal diagram e.g., a causal DAG. However, with the growing complexity and depth of health and medical knowledge being generated and increasing availability of new research articles daily, research

databases are reaching dimensions that limit the possibility of parsing through the enormity of evidence needed to craft comprehensive DAGs [18]. Though expert opinion is the most valuable tool for drawing DAGs, experts do not always generate perfect DAGs, sometimes missing important confounding pathways [19]. Additionally, obtaining the opinions of numerous experts is costly both in time and resources. Thus, the ongoing developments of Large Language Models (LLM) may offer promise to help overcome some of these challenges by leveraging existing text data that may express causal sentiments (e.g., "X causes Y").

This research aims to answer the question, "Can large language models help researchers build causal diagrams in the medical context using existing text data?" Here we will conduct experiments to determine under what conditions (e.g., prompt engineering, use of alternative language) GPT-3 [20] is able to provide accurate answers regarding the relationship between variables in a medical context and what are its limitations in doing so.

The main contributions of this paper are:

- Determining whether GPT-3 can signal the presence or absence of an edge between two variables in a directed acyclic graph from the medical context.
- Evaluating whether the use of certain language in prompts or linking verbs improves the classification accuracy of GPT-3.
- Exploring the limitations of GPT-3 in understanding the causal relationships between variables in the medical context.

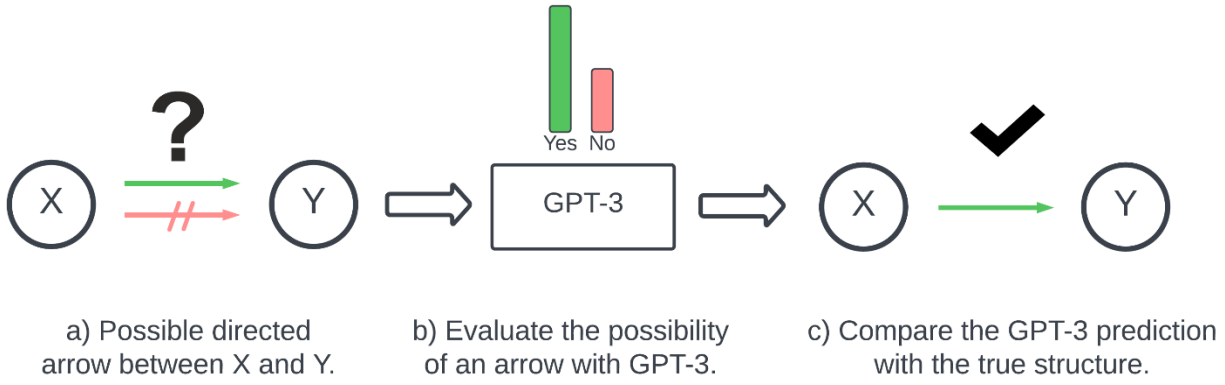


Figure 5-1: Overview of the evaluation

Figure 5-1 To predict the structure of a given causal graph, for every ordered variable pair, we scored two statements using GPT-3, where the first statement implied the presence of an arrow and the second implied the absence of an arrow. GTP-3 was accurate if the correct statement had a higher accuracy score than the incorrect statement. For example, GPT-3 would be accurate if the statement implying the presence (or absence) of an arrow had a higher accuracy than the incorrect statement and the arrow was present (or absent) in the true DAG.

5.5 Background

5.5.1 Large language models

Large language models capture non-trivial relationships and knowledge about the datasets they have been trained upon. This knowledge has the possibility to unlock numerous applications in healthcare such as summarizing research papers, assessing patient risks from subjective symptoms, and diagnosing patients from clinical notes. Although LLMs perform well on general natural language processing (NLP) tasks, its performance has been shown to be sensitive to its prompt [21, 22]. The advent of prompt-based learning introduced a possible solution to context sensitive text, by querying LLMs with a prompt that uses in-domain examples or task descriptions [23]. For example, chain-of-thought prompts such as "Let's take this step by step" have been shown to trigger multi-step reasoning in solving arithmetic problems [24]. Such prompts have also been shown to significantly improve performance in reasoning about medical questions [25].

Large language models are also sensitive to the type of text data they are trained on. For instance, GPT-3 [20] was trained on the corpus of text information on the internet. As one can imagine, the entirety of the internet would include a range of text data from lay and casual use of language on social media to more formal language in news articles. These differences in writing styles may influence the frequency of the use of causal language describing non-causal relationships. For instance, an individual writing a social media post may use the word 'cause' more lightly than medical researchers in medical journals.

5.5.2 Causal diagram overview

Causal models are typically accompanied by graphical representations i.e., Directed Acyclic Graphs (DAGs) which are acyclic graphs that succinctly illustrate the qualitative assumptions made by the models, not captured by conventional statistical models or machine learning algorithms [17, 26].

In epidemiological research, DAGs have a variety of purposes including: (1) representing the causal relationships amongst variables [3, 17, 26]; (2) identifying the potential confounding variables which need to be controlled for in order to estimate causal effects [3, 16, 27, 28]; and more recently (3) as a means of classifying the types of causal relationships that may give rise to selection bias [29].

A DAG is composed of variables (nodes), both measured and unmeasured, and their connections are displayed via line segments (directed edges) [17, 29]. The absence of an arrow between variables indicates the lack of a direct relationship between the variables. If the edge has an arrowhead, the variable at the tail is the parent node and the variable at the arrowhead is the child node [16]. An edge or arc is any line (with an arrowhead or not) that connects two variables [26]. The main characteristics of DAGs are that they are: (1) directed i.e., the edge has a defined direction (arrowhead), and (2) acyclic i.e., lack of cycles or loops within the graph.

A DAG is causal if: (1) the arrows between variables can be interpreted as direct causal effects, and (2) all common causes of any pair of variables are present [29]. The causal effects are 'direct' relative to certain degrees of abstraction in that the DAG does not include any variables that may mediate the effect [16]. As the name suggests, DAGs are acyclic because a variable cannot be the cause of itself, either directly or indirectly through another variable i.e., there are no feedback loops; as illustrated by each DAG in Figure 5-2 [29]. Additionally, in DAGs, causal

pathways are represented with directed paths from the starting variable to the final variable; thus, a variable is the cause of its descendants and an effect of its ancestors [16].

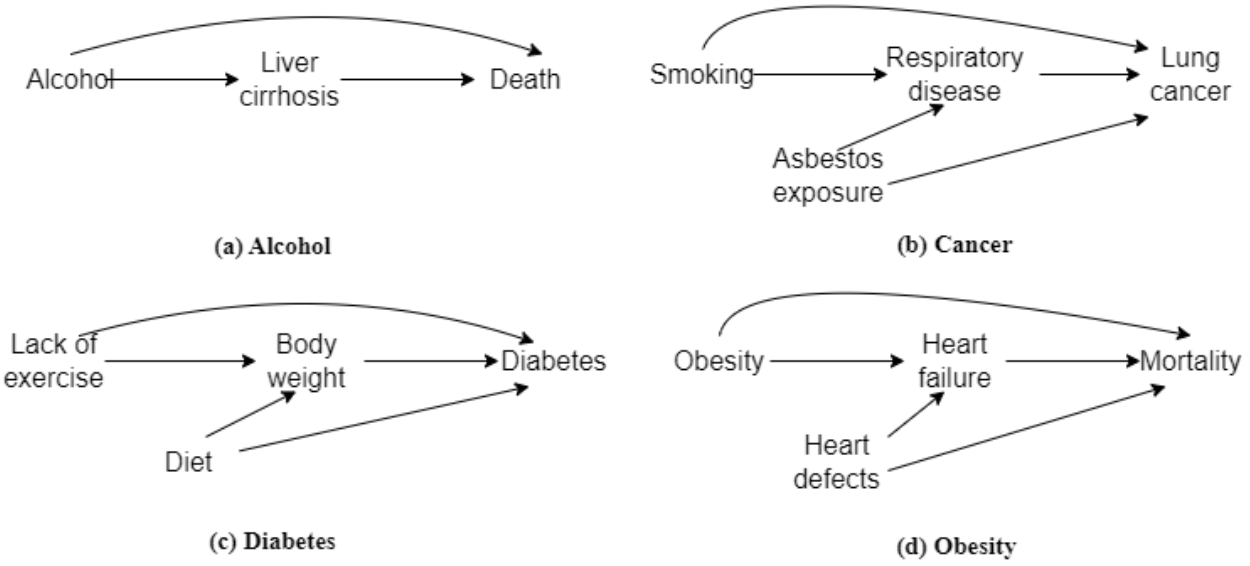


Figure 5-2: Ground Truth Directed Acyclic Graphs.

Figure 5-2 Four DAGs illustrating well-known exposure-outcome effects in the medical literature. DAG (A) represents the simplest DAG evaluated by GPT-3. DAGs (B-D) represent more complex structures involving a collider variable (node with two arrows pointing into it e.g., 'respiratory disease', 'body weight', and 'heart failure') with a common cause with the outcome.

5.6 Experiments

5.6.1 Experimental details

To empirically assess the potential effectiveness of LLMs in building DAGs, we used four DAGs representing well-known exposure-outcome relationships in the medical literature (Figure 5-2) as the Ground Truth. These DAGs are varied in complexity, amount of variables, and reflect different medical contexts. For a DAG of N variables, there are $\binom{N}{2}$ possible edges between two variables, and there are twice this amount of possible arrows since the arrows are directed. For example, a DAG of 4 variables has $2 \times \binom{4}{2} = 12$ possible arrows.

For each DAG, we looped through every ordered variable pair, and asked GPT-3 to score two statements per pair: (1) one implying the presence of a directed edge from variable 1 to variable

2, and (2) one implying absence of a directed edge from variable 1 to variable 2. The presence or absence of an edge between two variables is a binary decision (Yes / No), thus, we defined the prediction as accurate, if GPT-3 scored the correct statement higher than the incorrect one. We reported the accuracy or the proportion of correct predictions of our model.

5.6.2 Results

Q1: Does using prompt engineering lead to more accurate answers?

We investigated if the prediction accuracy of GPT-3 could be improved by prompting the statements with a reference to a medical authority. For example,

"According to X, var1 increases the risk of developing var2",

instead of

"Var1 increases the risk of developing var2" (baseline),

where X is an individual or entity with medical authority or expertise, e.g., medical doctors, medical studies, or "Big Pharma". These prompts were chosen as they vary in their credibility with the public. We found that in 2 cases (Diabetes and Obesity DAGs) prompt engineering did not help and baseline (no prompting individual or authority) outperformed all other prompts. While in 2 other cases, the "According to medical doctors," prompting significantly improved the accuracy of GPT-3. Interestingly, conditioning on "According to Big Pharma," decreases the accuracy of 3 of the 4 DAGs compared to the baseline. Furthermore, prompting the model on medical studies or medical doctors resulted in different results for half the DAGs. See Table 5-1 for all result.

Table 5-1: Prompt engineering: The medical authority used to prompt the statement

DAG name	Prompt	Accuracy
Alcohol	Baseline	0.33
	Big Pharma	0.50
	Medical doctors	**0.83**
	Medical studies	0.67
Cancer	Baseline	0.75
	Big Pharma	0.58
	Medical doctors	**1.00**
	Medical studies	**1.00**
Diabetes	**Baseline**	**0.67**
	Big Pharma	0.50
	Medical doctors	0.33
	Medical studies	0.42
Obesity	**Baseline**	**0.75**
	Big Pharma	0.58
	Medical doctors	**0.75**
	Medical studies	**0.75**

Q2: Does the verb used to denote the relationship between the variables have an impact on accuracy?

For instance, "Variable 1 X Variable 2" where X represents the verb (or phrase) that denotes the relationship between the variables, e.g., "causes" or "increases the risk".

Our results demonstrated that while no verb consistently improved classification accuracy, the choice of verb linking the two variables of interest influenced accuracy. 'Increases risk' had the highest accuracy for three of the four DAGs. Though it did not achieve the highest accuracy in the Alcohol DAG. Overall, the use of 'cause' yielded decent results for all DAGs. Results are reported in Table 5-2.

Table 5-2: Linking verb: The verb or phrase used to link the two variables of interest.

DAG name	Linking Verb	Accuracy
Alcohol	Cause	0.33
	Increases likelihood	**0.50**
	Increases risk	0.33
Cancer	Cause	0.58
	Increases likelihood	0.58
	Increases risk	**0.75**
Diabetes	Cause	0.58
	Increases likelihood	0.42
	Increases risk	**0.67**
Obesity	Cause	0.58
	Increases likelihood	0.42
	Increases risk	**0.75**

Q3: Does specificity in language improve accuracy?

We investigated if making our statements more specific or descriptive improved GPT-3's accuracy. Unsurprisingly, rephrasing the "alcohol" variable to "excessive alcohol consumption" increased the accuracy of GPT-3 on the Alcohol DAG. However, being more specific about the number of cigarettes being smoked and using a clinical term to qualify obesity resulted in worse accuracy for the Cancer and Obesity DAGs. Overall, in this analysis, more specific statements did not increase the accuracy and often resulted in worse accuracy for different linking verbs. Results are reported in Table 5-3.

Table 5-3: Specificity: More extensive descriptions of variables/concepts.

DAG name	Variable Name	Linking Verb	Accuracy
Alcohol	Alcohol	Cause Increases risk	0.33 0.50
	Excessive alcohol consumption	Cause **Increases risk**	0.33 **0.67**
Cancer	**Cigarette smoking**	Cause **Increases risk**	0.58 **0.67**
	Smoking 100 cigarettes a day	Cause Increases risk	0.50 0.58
Obesity	**Obesity**	Cause **Increases risk**	0.58 **0.67**
	Excessive fat accumulation	Cause Increases Risk	0.58 0.58

5.7 Discussion

In this work, we explored if LLMs could be used to complement and speed up the workflow of researchers by automatically scoring edges in potential DAGs. For the relatively simple and well-studied DAGs that we tested GPT-3 on, the results were overall encouraging as the performance reached much higher than 50% accuracy (random guessing) on all DAGs for at least one of the tested settings (e.g., prompt or linking verb). In this analysis, we found that GPT-3's accuracy performance was influenced by different prompts and linking verbs between variables of interest.

To the best of our knowledge, this is the first study to examine using LLM for causal diagram development in the medical context. Though there is growing interest, to date, there are few studies exploring the utility of LLM in causal diagram development. A recent study by Willig et al., (2022) [30] compared the performance of three query LLMs in making causal graph predictions in a general context. There also has been some interesting works applying causal inference in the LLM context. For instance, Vig et al., (2020) [31] investigated gender bias present in LLM using causal mediation analysis. Feder et al., (2021) [32] released a preprint of a consolidated exploration of causal inference situated in NLP. These works suggest more focus is being devoted to researching how causal inference can be applied to LLMs and NLP.

Furthermore, there has been some research investigating LLM's ability to answer and reason with medical text data. Several recent studies [25, 33] showed promising results on LLMs ability to answer medical exam questions. Others [21, 22] have shown that context-specific LLMs such as BioBert are able to outperform GPT-3 in medical domain NLP tasks.

Limitations: This study has some limitations. First, it must be acknowledged that the updating of LLMs, themselves as well as the data they are trained upon, lags behind the availability of new medical literature, and, thus may not be useful for informing the building of DAGs for novel diseases. Additionally, GPT-3 was trained upon the corpus of text data uploaded to the internet. The language used on the broader internet is likely more casual with the use of causal language than the medical academic literature [34]. Lastly, the way in which we probed GPT-3's ability to draw an edge between variables assumes that the causal connections between variables would be well-established in the corpus of text data.

Future work: Future work aims to use a medical language context-specific LLM such as web-GPT with PubMed or BioBert [35] to signal the presence or absence of edges in DAGs using medical terminology. Additionally, since our preliminary evaluations only examined the presence/absence of arrows and their direction, upcoming projects will be focused on controlling for acyclicity amongst variables, another important characteristic of DAGs.

5.8 Conclusion

Our results illustrate that GPT-3's level of accuracy in confirming an edge connecting two variables in a DAG depends on the language used to describe the relationship. Presently, expert opinion is the most valuable tool for constructing DAGs; however, like LLMs, experts are not exempt from making errors resulting in imperfect or erroneous DAGs via omission of important confounder variables [19]. These imperfections highlight that the use of LLMs to build DAGs should be, at present, only conducted with expert verification. We see LLMs providing utility in extracting common knowledge from medical text which when paired with expert knowledge may present a more efficient means to generate comprehensive DAGs. Large Language Models represent an exciting opportunity to extract common knowledge from the medical literature to complement and speed up DAG creation, but further research must be done to address the limitations reported above.

5.9 References:

1. Long S, Loutfi D, Kaufman JS, Schuster T. Limitations of Canadian COVID-19 data reporting to the general public. J Public Health Policy. 2022;43(2):203-21.
2. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(Volume 34, 2013):61-75.
3. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-710.
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
5. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation. 2004;60(5):503-20.
6. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P. Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists. Radiographics. 2021;41(5):1420-6.
7. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016.
8. OpenAI. GPT-3.5. 2021.
9. Anthropic. Introducing Claude 2023 [updated March 14, 2023. Available from: https://www.anthropic.com/news/introducing-claude.
10. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
11. NeurIPS. Call for papers 2022 [Available from: https://neurips.cc/Conferences/2022/CallForPapers.
12. Paper Copilot. NeurIPS Statistics 2024 [Available from: https://papercopilot.com/statistics/neurips-statistics/.
13. NeurIPS. 36th Annual Conference of Neural Information Processing Systems (NeurIPS): First Hybrid Program 2022 Fact Sheet 2022 [Available from: https://media.neurips.cc/Conferences/NeurIPS2022/NeurIPS_2022_Fact_Sheet.pdf.
14. Pearl J. Causal inference in statistics: An overview. Statistics Surveys. 2009;3(none).
15. Pearl J. The Eight Pillars of Causal Wisdom. Los Angeles, California: UCLA; 2017 April 24, 2017.
16. Greenland S, Pearl J. Causal Diagrams. Encyclopedia of Epidemiology2006.
17. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
18. Raghupathi W, Raghupathi V. Big data analytics in healthcare- promise and potential. Health Information Science and Systems. 2014;2(3):1-10.
19. Oates CJ, Kasza J, Simpson JA, Forbes AB. Repair of Partly Misspecified Causal Diagrams. Epidemiology. 2017;28(4):548-52.
20. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
21. Moradi M, Blagec K, Haberl F, Samwald M. Gpt-3 models are poor few-shot learners in the biomedical domain. arXiv preprint arXiv:210902555. 2021.
22. Gutierrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about gpt-3 in-context learning for biomedical ie? think again. arXiv preprint arXiv:220308410. 2022.
23. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys. 2023;55(9):1-35.
24. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Advances in neural information processing systems. 2022;35:22199-213.

25. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5(3).
26. Greenland S, Brumback B. An overview of relations among causal modelling methods. International Journal of Epidemiology. 2002;31:1030-7.
27. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12(3):313-20.
28. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequesite for confounding evaluation: An application to birth defects epidemiology. American Journal of Epidemiology. 2002;155(2):176-87.
29. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.
30. Willig M, Zečević M, Dhami DS, Kersting K. Can foundation models talk causality? arXiv preprint arXiv:220610591. 2022.
31. Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, SInger Y, et al. Investigating gender bias in language models using causal mediation analysis. Advances in Neural Information Processing Systems. 2020.
32. Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Transactions of the Association for Computational Linguistics. 2022;10:1138-58.
33. Guo Q, Cao S, Yi Z. A medical question answering system using large language models and knowledge graphs. International Journal of Intelligent Systems. 2022;37(11):8548-64.
34. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA, et al. Causal and Associational Language in Observational Health Research: A Systematic Evaluation. Am J Epidemiol. 2022;191(12):2084-97.
35. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-4

6. CHAPTER 6: HIV-RELATED INDIVIDUAL-LEVEL OUTCOMES COMMONLY REPORTED IN RESEARCH STUDIES CONDUCTED IN HIGH-INCOME COUNTRIES: A SCOPING REVIEW (MANUSCRIPT 3)

"A correlate does not a surrogate make." [1]

6.1 Preamble

Chapter 4 (Manuscript 1) of this thesis revealed shortcomings in COVID-19 epidemiological data reported by governmental bodies and news outlets [2]. The data often failed to meet epidemiological standards, lacking appropriate denominators, clear definitions of cases and populations of interest, and inclusivity of gender and race/ethnicity. Despite these limitations, public health officials continued to use this data to inform public health policy, raising concerns about using associational data for decision-making during the pandemic [3]. This observation revealed the need for more routine causal modelling methods to be implemented in public health epidemiology. The amount and complexity of public health data suggest the need for automated approaches that enable a more efficient development of causal modelling.

In response to this challenge, Chapter 5 (Manuscript 2) explored the potential of LLMs in constructing DAGs describing known exposure-outcome relationships in the medical literature. The aim was to investigate whether the LLM, GPT-3, could make use of the vast corpus of internet data their trained upon, to extract information regarding causal relationships between variables and aid in constructing DAGs. GPT-3's in determining the presence or absence of an edge between pairs of variables was dependent on the language used in prompts, the verb used to denote the relationship between variables (e.g., cause, increases likelihood, increases risk), and the specificity of the variables examined [4]. Furthermore, GPT-3 performed better on DAGs that described certain medical exposure-outcome pairs, for instance, DAGs describing cigarette smoking and lung cancer had higher accuracies in all tested settings than DAGs illustrating the relationship between alcohol and liver cirrhosis. It is possible that this performance may be due to lack of text data consistently describing certain relationships between variables in the training

data. This performance may be improved through model pre-training on the scientific literature describing actual research findings. This limitation, while revealing current constraints of LLM in causal modeling, encouraged a return to the basics: synthesis of domain-specific scientific literature.

Building on these insights, the current chapter presents a scoping review of the HIV literature. This review aims to address the gap identified in Chapter 5 and provides an evidence-based synthesis of HIV-related patient outcomes were reported in studies conducted in high-income countries. By broadly searching and synthesizing HIV-related research, this reviews aimed to create a comprehensive DAG of HIV-related individual outcomes. The focus on surrogate and primary outcomes that could be applied in adaptive platform trials serves as a bridge between the theoretical work on causal modeling and practical applications in clinical research. The findings from this review informed subsequent research reported in manuscript 4 of this thesis.

Chapter 6 – Manuscript 3

6.2 Title page

HIV-related patient outcomes commonly reported in research studies conducted in high-income countries: A scoping review

Stephanie Long[1], Guowei Zhong[1], Kim Engler[2,3], Bertrand Lebouche[1-5],Tibor Schuster[1]

[1]Department of Family Medicine, McGill University, Montreal, Canada

[2]Research Institute of the McGill University Health Network, McGill University, Montreal, Canada

[3]Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada

[4]Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, Quebec, Canada

[5]Chronic Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, Quebec, Canada

Key words: HIV, outcomes, outcome measures, patient-reported outcomes

6.3 Abstract

Introduction:

Antiretroviral therapy has transformed HIV into a chronic, manageable condition with near-normal life expectancies, shifting research focus from traditional outcomes like mortality and survival to more comprehensive and holistic health outcomes such as those described in the HIV care continuum. This shift necessitates a re-evaluation of the outcomes that are relevant and appropriate for use in research studies, particularly adaptive trials. These trials efficiently evaluate multiple interventions by adjusting key trial characteristics based on interim results, relying on surrogate outcomes, which are less invasive, cheaper, and faster to measure than primary endpoints, thus enabling timely trial adaptations.

Methods:

We conducted a scoping review to identify HIV-related outcomes used in high-income countries, particularly aiming to uncover surrogate-primary outcome pairs and to develop a comprehensive directed acyclic graph illustrating relationships amongst these outcomes. We searched four databases for quantitative or mixed methods studies evaluating HIV care provided to adults living with HIV during the era of highly active combination therapy (2006 – February 8, 2024). Data were analyzed using framework analysis and descriptive statistics, and a DAG was constructed.

Results:

Of 9443 unique articles identified, 681 met inclusion criteria after full-text review. Most studies (83.4%) were conducted in North America, with 88.8% being observational studies, and very few randomized control trials. Outcomes identified predominantly assessed physical health, followed by social health, with limited focus on mental health. Only 2.2% of included studies used surrogate outcomes, primarily CD4+ count (35.7%) used as a surrogate marker of overall immune function or for outcomes along the HIV care continuum. The most commonly used primary outcomes were HIV viral suppression (n=91, 12.8%), retention in care (n=60, 8.5%), and ART adherence (n=54, 7.6%).

Conclusion:

This review reveals a shift in focus towards holistic health outcomes in HIV research in high-income countries, though mental health remains underrepresented. Despite increased attention to diverse populations, men continue to be the predominant population in HIV research. There is a pressing need for more inclusive and comprehensive health evaluation in HIV care research.

6.4 Introduction

Numerous landmark studies [5-8] have led to the development of antiretroviral therapies (ART), dramatically reducing HIV-associated morbidity and mortality. Despite this progress, HIV remains a global health challenge with an estimated 39 million people living with HIV (PLWH) worldwide [9]. Effective ART regimens have transformed HIV infection into a manageable chronic condition [10, 11], allowing PLWH to have near-normal lifespans and achieve undetectable viral loads, rendering the virus non-transmissible (U=U) [12].

However, HIV management requires lifelong follow-up, self-care, and strict ART adherence to maintain an undetectable viral load and prevent virus transmission [13]. As HIV transitions from a being a terminal disease to a chronic, manageable condition, there is a need to re-evaluate which health outcomes are most relevant and appropriate for use in clinical trials, particularly adaptive multi-arm multi-stage trials such as platform trials which enable adaptive randomization based on interim results [14].

Adaptive randomization based on interim results depends on selection of an appropriate interim (a.k.a. surrogate) outcome. Surrogate outcomes are intermediate measures of primary endpoints that offer an efficient way of evaluating interventions, especially when there is a long delay before the primary outcome can be observed, if it involves invasive measurements or the cost of measurement is prohibitive. The choice of surrogate outcome is crucial as it influences the adaptive trials' ability to efficiently evaluate interventions and make accurate decisions about treatment arms [14].

Over the last decades, viral load and CD4+ cell count have been the preferred indicators of HIV treatment success and disease progression and are frequently reported primary outcomes in the trial literature [15]. They are also often used as surrogates for immune system response and disease progress [16]. These outcomes, however, do not give the complete picture of health, as defined by the World Health Organization as "a state of complete *physical*, *mental*, and *social* well-being, and not merely the absence of disease" [17]. There are many other outcomes that can describe an individual's state of health. The HIV care continuum [18] outlines five stages towards successful HIV care, each representing a relevant outcome to be evaluated.

Given that HIV is a complex chronic condition affected by a constellation of factors with many associated health outcomes, the choice of adequate study endpoints and associated surrogate

outcome measures in adaptive platform trials poses a new challenge for investigators. Considerations must be made to ensure the selected outcomes are relevant and accurately measurable.

We conducted a scoping review of HIV-related outcomes reported in research studies conducted in high-income countries. Our aim was to identify and map the range of outcomes currently used in HIV research, with particular attention to surrogate-primary outcome pairs. This review may help inform the selection of appropriate outcomes for future adaptive trials in HIV research, ensuring that they capture the full spectrum of health and well-being for PLWH in the era of effective ART.

The primary objective of this scoping review was to comprehensively synthesize the HIV literature to identify HIV-related patient outcomes reported in HIV research studies with a focus on surrogate and primary outcomes that could be used in adaptive platform trials. The secondary objective was to develop a comprehensive directed acyclic graph (DAG) depicting the pathways amongst the HIV outcomes from the findings of review. This DAG will undergo subsequent verification with domain experts in another study.

6.5 Methods

Applying the five-step scoping review methodology of Arksey and O'Malley [19], we searched the HIV literature to identify HIV-related outcomes reported in studies conducted in high-income countries (according to the World Bank Classification [20]). We focused particularly on identifying surrogate outcome and primary endpoint pairs. Our findings are reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews (PRISMA-ScR) checklist [21].

6.5.1 Research question

This scoping review was guided by the research question, "*What individual-level health outcomes (primary and surrogate) are being reported in HIV research studies conducted in high-income countries?*"

6.5.2 Search strategy

In collaboration with an academic librarian, a search strategy was developed and executed to identify relevant literature using controlled vocabularies (e.g., MeSH terms) and keywords (See

Appendix A). The search strategy was adapted and implemented in four major databases: EMBASE (Ovid), MEDLINE (Ovid), CINAHL, and Scopus. We limited our search to studies published between 2006 and February 8, 2024 (date of search execution). We chose 2006 as the starting point because it marks the advent of combination antiretroviral therapy for HIV, ensuring that the identified outcomes would be highly relevant to current practice. Given that the purpose of this review was to identify HIV-related outcomes used in care settings in high-income countries, we determined that citation tracking of key articles was not necessary or relevant to our objectives.

6.5.3 Eligibility criteria

Included articles: 1) were quantitative or mixed methods studies, 2) conducted in a high-income country (according to the World Bank classification [20]); 3) had adults 18 years and above living with HIV only (no co-infections) as the study population of interest; 4) evaluated individual-level HIV-related outcomes; and 5) were in English or French. We excluded drug trials focusing primarily on toxicity or safety outcomes, as they are more concerned with assessing the drug itself rather than overall patient health. Additionally, we omitted literature or systematic reviews, meta-analyses, feasibility studies, case reports, study protocols and conference abstracts.

Two authors (S.L., G.Z.) independently screened all titles and abstracts applying the eligibility criteria using Rayyan (http://rayyan.qcri.org) [22]. A third reviewer (T.S.) resolved discrepancies. Cohen's Kappa was used as a measure of inter-rater reliability.

Articles included for full-text review were exported to EndNote X9.3.3 [23]. Full-text screening was performed by a single reviewer (S.L.), with a second reviewer (G.Z.) verifying 10% of full-text articles for inclusion.

6.5.4 Data extraction

As the purpose of this scoping review was to identify HIV-related primary and surrogate outcomes that could be used in an adaptive trial and construct a DAG illustrating relationship amongst these outcomes, we extracted a comprehensive range of data types (e.g., bibliometric, causal, statistical, sample population, and participatory data) that may be useful in the planning such a trial. Table 6-1 displays the extracted data.

6.5.4.1 Study outcomes:

We extracted the primary and surrogate outcomes reported in each article from the methods section, where study outcomes and statistical analysis are typically described. Surrogate outcomes were identified as any outcome that was described in the included articles as a "surrogate", "proxy for", "indicator of", "intermediate of" the primary outcome. For each study, we recorded whether a surrogate outcome was used and for which primary outcome it served as a surrogate.

Each primary and surrogate outcome was categorized as per the WHO aspect of health it assessed - *physical*, *mental*, or *social* [17]. Further, we noted whether each outcome was a clinical measurement or self-report (or patient-reported) measure.

6.5.4.2 Causal:

This data was extracted to address the secondary objective of constructing a DAG. Further, to discern whether surrogates must be 'causally' linked and not solely 'associational', we extracted the study claims (i.e., a statement made in the discussion section regarding study findings, typically the first sentence) and recommendations (e.g., 'change practice', 'intervene now', 'further research needed'). This allowed us to determine whether the study design was congruent with the language used in the article. For example, unlike observational studies, RCTs have the capacity to establish causality regarding the relationship between the exposure and the outcome.

Table 6-1: Data extraction form

Category	Type of data
Bibliometric	- Year of publication - Continent - Study design (observational, randomized control trial) - Methods (descriptive or inferential statistics, machine learning, causal inference)

Outcomes	- Primary outcome
	  o Patient-reported outcome (yes/no)
	  o Clinical measure (yes/no)
	  o World Health Organization [17] outcome category (physical, mental, social)
	- Surrogate outcome
	  o Used surrogate outcome (yes/no)
	  o Name of surrogate outcome

Causal	- Claims made (associational vs. causal) - Recommendations ('change practice', 'further research needed', 'intervene now')
Sample Population	- Participants (inclusion/exclusion criteria) - Explicitly recorded race/ethnicity (yes/no) - Outcomes reported by relevant subgroups (yes/no)

6.5.5 Data synthesis and analysis

The data synthesis focused on bibliometric description and framework analysis [24]. The results are reported according to the PRISMA extension for Scoping Reviews (PRISMA-ScR) [25].

Quantitative analysis

Descriptive statistics were used to summarize the nature and distribution of the studies (e.g., study design, year of publication, study population). All variables were categorical or binary and were expressed as frequencies and percentages.

To report frequency counts of outcomes used, we tabulated primary outcomes and calculated their absolute usage. We computed frequency by dividing each outcome's count by the total number of outcomes used. Since each study has at least one primary outcome, t

We presented the top 21 reported primary outcomes in a mosaic plot, where the size of each tile represents the frequency of reporting. We elected to report only the top 21 outcomes to enhance readability of the figure, as outcomes infrequently reported were very difficult to visualize.

Qualitative framework analysis

Data analysis was guided by framework analysis, a pragmatic analysis approach developed by Ritchie and Spencer for systematically analyzing large qualitative datasets [24]. This method allows for the organization and reduction of data into a matrix format, facilitating the identification of themes and relationships. For analytical purposes, we categorized HIV-related outcomes using the World Health Organization's definition of 'health' i.e., "a state of complete *physical*, *mental*, and *social* well-being, and not merely the absence of disease" [17]. We further classified these outcomes according to the taxonomy of outcomes in medical research [26], which evaluates seven main domains: death, physiological/clinical, life impact, resource use, and adverse events. This approach enabled us to systematically analyze the complex landscape of HIV outcomes in high-income countries, exploring both their existence and interrelationships, particularly their potential use as surrogate outcomes in adaptive trials.

Directed acyclic graph (DAG)

We constructed a DAG to visually represent the causal relationships amongst the identified primary and surrogate outcomes. DAGs are graphical tools that illustrate the qualitative assumptions made by causal models, which are not captured by conventional statistical models [27, 28]. They are useful for summarizing and organizing knowledge from research and experts and can assist with trial planning and data analysis.

To construct the DAG, we followed the 'Evidence synthesis for constructing DAGs' (ESC-DAGs) protocol [29]. The protocol consists of four steps: (i) mapping (*identifying primary and surrogate outcomes from the studies*), (ii) translation (*identifying the relationships between outcomes*); (iii) integration I (*synthesizing a DAG illustrating the relevant relationships*); and (iv) integration II (*grouping similar variables within the DAG*) [29]. This approach allows for the synthesis of knowledge from previously published studies into a unifying DAG.

6.6 Results

6.6.1 Search Results:

Figure 6-1 shows the PRISMA-ScR diagram of the article screening and reasons for exclusion. The search strategy yielded 13044 articles, with 9443 unique records after removal of duplicates. Two reviewers independently reviewed all titles and abstracts, achieving a 90.2% agreement

rate. Title and abstract screening excluded 8147 records, leaving 1296 full-text articles to review. Full-text review excluded 612 articles, resulting in 681 full-text articles meeting the inclusion criteria for synthesis.

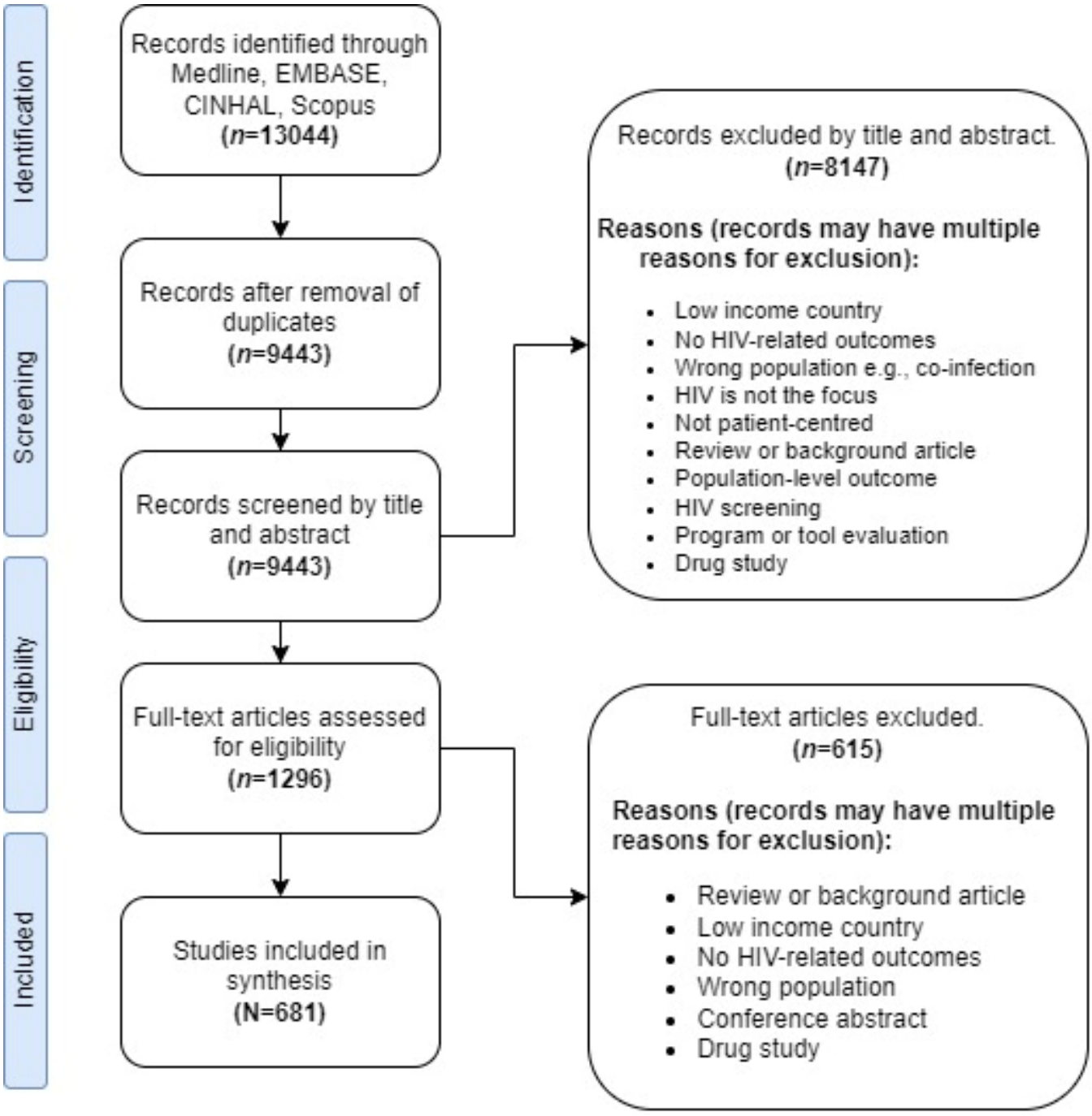


Figure 6-1: PRISMA-ScR diagram

6.6.2 Characteristics of included studies:

Table 6-2 summarizes the bibliometric characteristics of included articles. Included studies were published between 2006 and February 8, 2024, across 168 journals. The top three journals were

AIDS and Behaviour (n=87, 12.8%), Journal of Acquired Immune Deficiency Syndromes (n=69, 10.0%), or AIDS Care (n=66, 9.5%) (see Appendix B for details). Most studies were from North America (n=568, 83.4%) and Europe (n=74, 10.9%). The majority were observational (n=612, 89.9%), with few RCTs (n=69, 10.1%). 97.7% of included studies used descriptive/inferential statistical analysis, 0.2% (n=15) studies applied causal inference, and 0.01% used machine learning. Participants were primarily those living with HIV: adults (65.0%, n=442), more specifically including women (3.4%, n=23), and drug users (2.6%, n=18). Most studies (89.3%) recorded the race/ethnicity of their study participants, but less than half reported the outcomes by these subgroups.

Table 6-2: Bibliometric characteristics of studies included in this review (N=681)

Characteristic	**No. of publications (%)**
Continent:	
North America	568 (83.4)
Europe	74 (10.9)
Asia	23 (3.5)
Oceania	8 (1.2)
International	8 (1.2)
Study Design:	
Observational	612 (89.9)
Randomized control trial	69 (10.1)
Methods:	
Descriptive / Inferential Statistics	665 (97.7)
Causal Inference	15 (2.2)
Machine Learning	1 (0.1)
Number of studies including these participants**:	
Female	602 (88.3)
Male	613 (90.0)
Transgender	154 (22.6)
Participants, People living with HIV (PLWH):	
Adults	442 (65.0)
Women	23 (3.4)
Substance users	18 (2.6)
Men who have sex with men (MSM)	17 (2.5)
Released from prison	16 (2.3)
Incarcerated adults	15 (2.2)
Veterans	12 (1.8)
Transgender women of colour	10 (1.5)
Young adults	10 (1.5)
Black MSM	9 (1.3)
Newly diagnosed	7 (1.0)
Transgender women	7 (1.0)

Pregnant women	6 (0.9)
Men	6 (0.9)
Homeless or unstably housed	6 (0.9)
Transgender women and MSM	5 (0.7)
ART naïve	4 (0.6)
Older adults	4 (0.6)
Outpatients	3 (0.4)
Postpartum women	3 (0.4)
Active-duty military personnel	3 (0.4)
Heavy drinkers	3 (0.4)
Black men	3 (0.4)
Black adults	2 (0.3)
Women of colour	2 (0.3)
Women sex workers	2 (0.3)
Serodiscordant male couples	2 (0.3)
Not in care	2 (0.3)
Starting ART	2 (0.3)
Women with perinatally acquired HIV	2 (0.3)
Women or transgender women	2 (0.3)
In care	1 (0.1)
Hispanic adults	1 (0.1)
With mental illness	1 (0.1)
With mental illness and substance abuse	1 (0.1)
Black Transgender women	1 (0.1)
Black Women	1 (0.1)
Immigrants from Africa	1 (0.1)
Marginalized populations*	1 (0.1)
Native Americans	1 (0.1)
Adults marginally engaged in care	1 (0.1)
Adults on parole not engaged in care	1 (0.1)
Adults with desire to switch ART	1 (0.1)
With suboptimal ART adherence	1 (0.1)
With treatment failure	1 (0.1)
With heterosexually acquired HIV	1 (0.1)
Latino adults	1 (0.1)
Treatment-experienced adults	1 (0.1)
Virally unsuppressed adults	1 (0.1)
Women out of care	1 (0.1)
Youths transitioning to adult care	1 (0.1)
African American/Black and Latino adults	1 (0.1)
Black sexual minority	1 (0.1)
MSM drug users	1 (0.1)
Newly out-of-care	1 (0.1)
Admitted to ICU	1 (0.1)
Nursing home residents	1 (0.1)
Sexual minority older	1 (0.1)
Virally suppressed	1 (0.1)
Who changed ART	1 (0.1)

With complex psychosocial needs	1 (0.1)
Sexual and gender minority youth	1 (0.1)
Treatment-naïve and obese or overweight	1 (0.1)
Women at HIV menopause clinic	1 (0.1)
Recorded race/ethnicity of participants:	
Yes	592 (86.3)
No	91 (13.6)
Reported outcomes according to subgroups:	
Yes	333 (48.5)
No	352 (51.5)
*hard-to-reach individuals selected from an outreach service [30] **8 of the included articles only reported the total number of participants and/or did not provide a breakdown of the sex of their participants [31-40]. Thus, the proportions are calculated from a denominator that reflects this limitation (N=673).	

6.6.3 Primary Outcomes:

6.6.3.1 Mosaic plot of primary outcomes reported

Figure 6-2 visualizes the top 21 reported primary outcomes identified in this review in a mosaic plot, organized by WHO aspect of health and type of medical outcome [26]. The size of each tile corresponds to the frequency of reporting of each primary outcome, highlighting viral suppression, retention in care and ART adherence as the most common. Physical health outcomes dominate, followed by social aspects, with mental health significantly underutilized.

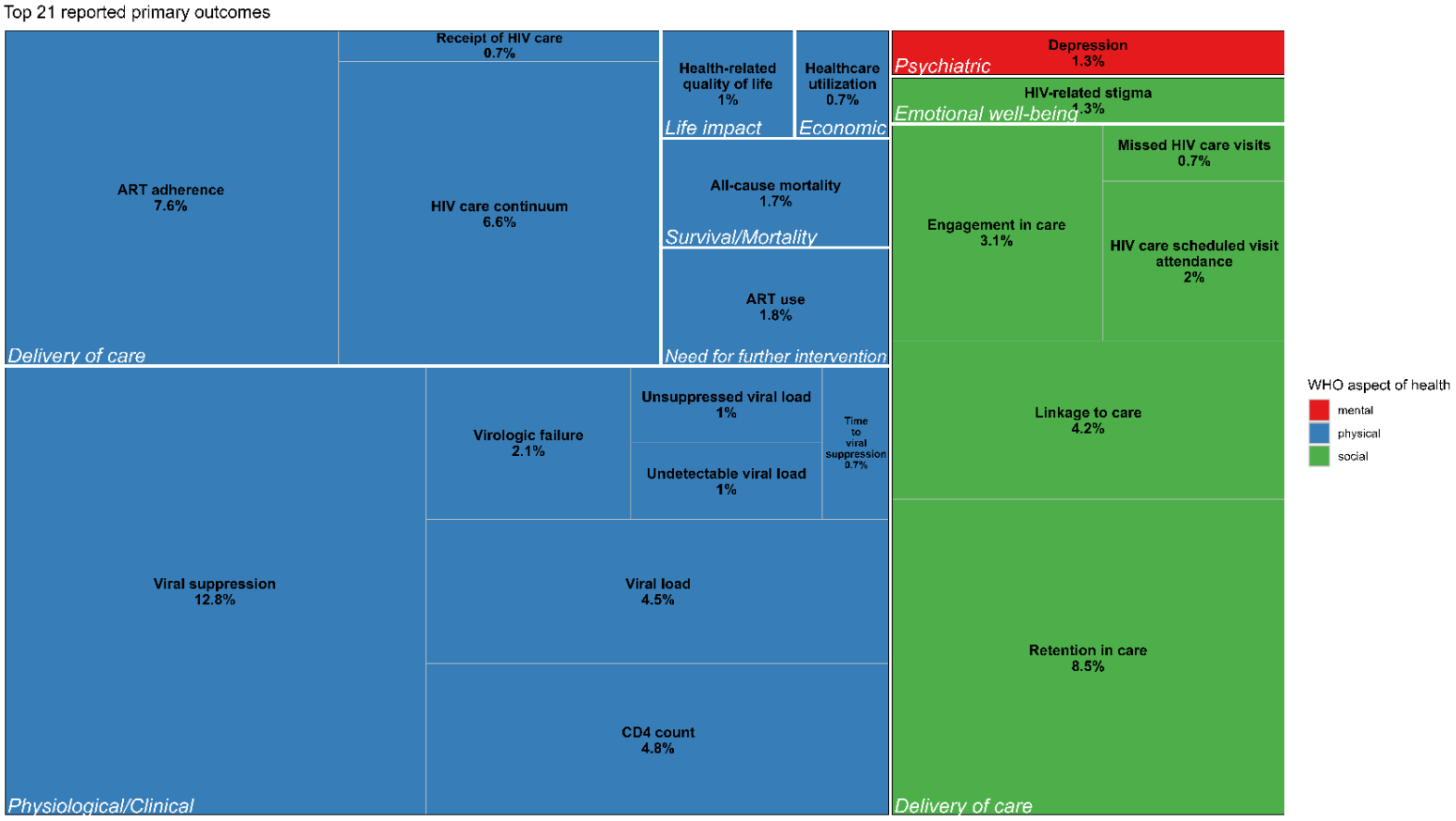


Figure 6-2: Mosaic plot of top 21 reported primary outcomes

Table 6-3 presents all of the identified primary outcomes (n=135). 62% (n=84) of the identified outcomes assessed *physical* health, particularly physiological/clinical domains (immune system outcomes) and life impact (delivery of care, quality of care) outcomes. *Social* aspects of health, accounted for 32% (n=43), primarily assessing life impact factors such as emotional well-being and delivery of care. Seven outcomes related to mental health were reported, but only three had be used in more than one study: depression, recent opioid misuse, and self-rated health.

The most studied primary outcomes were viral suppression (n=91, 12.8%), retention in care (n=64, 9.0%), ART adherence (n=56, 7.9%), and CD4+ count (n=40, 5.6%). All assessed physical aspects of health, except 'retention in care' a social component of health.

Table 6-3: All primary outcomes (N total outcomes =135) identified in this review. We report number of studies that used the outcome as well as the proportion from the total number of outcomes used (n, %). Outcomes in blue were only reported in one study (n=1, 0.14%).

	Physical **N=854**	**Mental** **N=8**	**Social** **N=43**
Death (n=3)	- All-cause mortality (13, 1.8) - Time to all-cause mortality (2, 0.3) - Time to AIDS-defining illness or death		
Physiological / clinical (n=45)	***Immune system outcomes (n=24)*** - Viral suppression (98, 13.8) - CD4+ cell count (40, 5.6) - Viral load (39, 5.5) - Virologic failure (15, 2.1) - Unsuppressed viral load (7, 1.0) - Time to viral suppression (5, 0.7) - AIDS-defining illness (5, 0.6) - AIDS progression (4, 0.6) - HIV symptoms (4, 0.6) - Viremia (4, 0.6) - Detectable viral load (4, 0.5) - Viral rebound (4, 0.4) - Virologic success (2, 0.3) - Immunological and virological response (2, 0.3) - Weight change (2, 0.3) - Advanced HIV disease - Change in CD4 or CD8 count - Viral failure or viral suppression - Time from diagnosis to viral suppression - Time to CD4CD8 ratio >1 and time to CD4 count >900 cells per ul - Timing of HIV rebound - Time spent living with viral load >1500 copies/ml - Time from ART initiation to viral suppression - Time to CD4 count >350 cells/uL ***Pregnancy, puerperium, and perinatal outcomes: (n=1)*** - Perinatal transmission	***Psychiatric outcomes: (n=2)*** - Depression (10, 1.4) - Recent opioid misuse (2, 0.3)	***N=19*** ***Social functioning: (n=2)*** - No show to psychiatric appointment - Social support satisfaction ***Delivery of care: (n=7)*** - Disengagement in care - In care - Initiation of care - Re-engagement in care - Time from HIV diagnosis until entry to HIV care - Timeliness of HIV diagnosis - Time to linkage to care ***Patient satisfaction: (n=5)*** - Patient's trust in HIV care provider - Patient activation - Patient ratings of clinician communication - Clinician and patient communication behaviours - Medical mistrust ***Personal circumstances: (n=5)*** - Homelessness - Housing stability - Undocumented status - Incarceration status - Medicaid enrollment
Life impact (n=63)	***Delivery of care: (n=12)*** - ART adherence (56, 7.9) - Receipt of HIV care (5, 0.7) - ART receipt (4, 0.6)	***N=6*** ***Perceived health status: (n=1)*** - Self-rated health (3, 0.4)	***Emotional well-being: (n=22)*** - Stigma (10, 1.5) ***Social functioning:*** - Unsafe sexual behaviour (2, 0.3)

	- Late-stage HIV diagnosis (4, 0.6) - ART prescription (3, 0.4) - Delayed presentation to care - Early presentation to HIV care - Entry into HIV care - Late presentation for HIV care - Days between referral and initial appointment date - Pre-ART CD4 testing - Treatment satisfaction ***Quality of life: (n=2)*** - Health-related quality of life (7, 1.0) - Quality of life (3, 0.4) ***Physical functioning: (n=8)*** - Number of alcohol drinks per week (2, 0.3) - Change in systolic and diastolic BP - Activities of daily living changes - Age at diagnosis of comorbidity - Change in comorbidity burden - Early menopause - Change in BMI - Time to weight or BMI increase ***Health behaviour and management: (n=11)*** - Complementary and alternative medicine use - Not in care - Not taking ART - Non-engagement in ART - Participation in opioid agonist therapy - Cessation of substance use - Reduction in alcohol use - Attendance at primary care and specialty outpatient clinic - Entry into opioid agonist therapy - Reduction to low-risk or no alcohol use - Unhealthy alcohol screening at primary care visit	***Social functioning:*** - Loneliness ***Delivery of care:*** - Care transition readiness ***Cognitive functioning***: - Cognitive function - Health literacy and numeracy ***Emotional well-being:*** - Positive affect	- HIV disclosure status (2, 0.3) ***Delivery of care:*** - Retention in care (64, 9.0) - HIV care continuum (47, 6.9) - Linkage to care (33, 4.6) - Engagement in care (24, 3.4) - HIV care visit attendance (14, 2.1) - Missed HIV care visits (6, 0.7) - Missed ART doses (2, 0.3) - Access to care (4, 0.6) - Failure to be retained in care (3, 0.4) - Gaps in care (3, 0.4) - HIV visit adherence (3, 0.4) - Patient satisfaction with HIV care (3, 0.4) - Lost to follow-up (3, 0.4) - ART interruption (2, 0.3) - ART side effects (2, 0.3) - ART initiation (2, 0.3) - Relinkage to care (2, 0.3) - Re-engagement in care (2, 0.3) - Time from HIV diagnosis to first ART dispensation (2, 0.3)
Resource use (n=24)	***Economic: (n=5)*** - Healthcare utilization (5, 0.7) - Healthcare costs (4, 0.6) - HIV-related healthcare costs (2, 0.3)		***Societal/carer burden: (n=2)*** - Advance care planning - Unmet needs

	- Qualified health plan enrollment - Quality adjusted life years ***Hospital: (n=5)*** - Emergency department use (3, 0.4) - Emergency department visits and hospitalizations (3, 0.4) - Hospital readmission - Preventability of hospital readmissions - Time to first hospitalization ***Need for further intervention: (n=12)*** - ART use (13, 1.8) - ART readiness - ART regime switch - Number of pharmacy visits - Poly-drug use - Early refills of ART at 12 months - Heavily-treatment experienced - Time to ART use - Time to first ART switch - Frailty - Potential drug-drug interactions with ART - Receipt of any opioid prescription		
Adverse events (n=1)	- Severe clinical event (2, 0.3)		

6.6.4 Surrogate outcomes:

Table 6-4 shows the surrogate outcomes and their corresponding primary outcome. Only 2.1% of included studies used surrogate outcomes; 93.3% of which were observational studies. CD4+ count was the most common, accounting for 37.5% (n=6) of the surrogate outcomes used, often as a surrogate for outcomes along the HIV care continuum i.e., retention in care, engagement in care as well as presentation to HIV care.

Table 6-4: Surrogate outcomes identified: This table shows the primary and surrogate outcome pairs reported in the identified studies, whereby the [number] refers to their citation.

		Surrogate for:									
		Early presentation to care	**Late presentation to care**	**Linkage to care**	**Retention in care**	**Engagement in care**	**HIV care visits**	**ART adherence**	**Disease progression**	**Advanced HIV disease**	**Social & medical instability**
Outcome used:	**All-cause mortality**								[41]		
	AIDS-related mortality								[41]		
	CD4+ count	[42]	[43]			[34, 44]				[43]	
	CD4%					[44]					
	Viral load			[45]	[46]	[34, 44]	[47]				
	Viral suppression							[45]*			
	Undetectable viral load							[48]*			
	Emergency department use										[49]
	HIV lab tests				[50]		[51]				

*To achieve viral suppression or undetectable viral load, one must be adherent to their ART [52].

6.6.6 Trends across study design and participant type:

Table 6-5 presents the most common characteristics across study designs. Observational studies primarily used viral suppression (13.4%, n=82), retention in care (9.2%, n=56), and the HIV care continuum (7.4%, n=44) as primary outcomes. RCTs mostly reported ART adherence (18.6%, n=13), viral suppression (12.9%, n=9) and viral load (10.0%, n=7). Adults living with HIV were the predominant focal population in both observational studies (66.2%) an RCTs (54.3%). Only 2.3% of observational studies, and 1.4% of RCTs assessed surrogate outcomes.

Table 6-4: Top 3 characteristics of included studies across study design: Observational vs RCT

	Observational Study (N=612) (n, %)	**RCTs (N=70) (n, %)**
Continent of publication	1. North America (306, 50) 2. Europe (23, 3.8) 3. Asia (5, 0.8)	1. North America (59, 84.3) 2. Europe (8, 11.4) 3. International (3, 4.3)
Participants	1. Adults (405, 66.2) 2. Women (21, 3.4) 3. Drug users (15, 2.5) 3. Men who have sex with men (15, 2.5)	1. Adults (38, 54,3) 2. Black men who have sex with men (4, 5.7) 3. Prisoners (4, 5.7) 4. Released from prison (4, 5.7)
Primary outcome	1. Viral suppression (82, 13.4) 2. Retention in care (56, 9.2) 3. HIV care continuum (44, 7.2)	1. ART adherence (13, 18.6) 2. Viral suppression (9, 12.9) 3. Viral load (7, 10.0)
Use of surrogate outcome	• Yes (14, 2.3) • No (598, 97.7)	• Yes (1, 1.4) • No (69, 98.6)

6.6.8 Directed acyclic graph

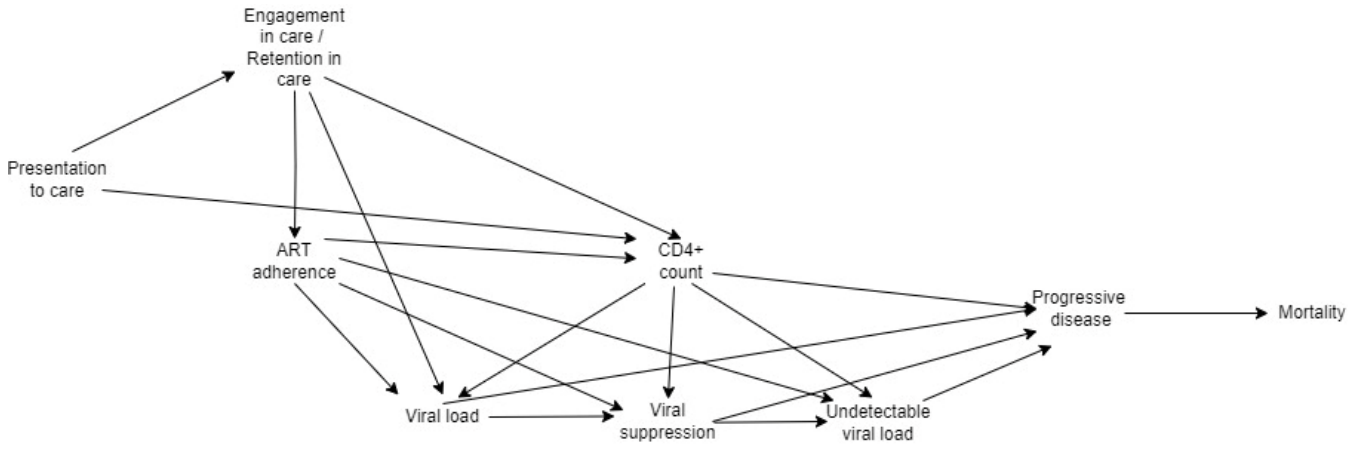


Figure 6-3 Directed acyclic graph constructed from the surrogate-primary outcome pairs identified in this review.

We constructed a DAG depicting the causal relationships amongst HIV-related individual outcomes used in research studies conducted in high-income countries, based on the surrogate and primary outcome pairs identified (Figure 6-3). CD4+ cell count was the most frequently used surrogate outcome identified in this synthesis, acting as a surrogate indicator for time to HIV care presentation (late or early) [42, 43], engagement with care [34, 44], and progression to advanced HIV disease [43]. Viral load was the next most frequently used surrogate outcome, serving as a surrogate for engagement in care [34, 44], linkage to care [45], and ART adherence [48]. Both viral suppression [45] and undetectable viral load [48] were used as surrogate outcomes for ART adherence; however, achievement of both is contingent on being adherent to ART [53]. Laboratory tests, viral load, and CD4+ count [44, 46, 51] were used as surrogates for retention or engagement in care. Retention/engagement in care refers to an individual's continued engagement in their HIV health care [54].

6.7 Discussion

This scoping review provides a comprehensive overview of the study outcomes used to evaluate health in HIV care in high-income countries published between 2006 to present. Among the 681 included studies, we identified 140 unique primary outcomes, 56 of which were used in at least two articles. A majority (58%) of these 56 primary outcomes focused on the physical aspects of health, particularly the physiological/clinical components (e.g., immune system outcomes or survival/mortality) or life impact factors such as delivery of care or quality of life. With an increasing emphasis on the HIV care continuum, social aspects of health (e.g., retention in care, engagement in care, etc.,) were also frequently assessed. However, there is a notable lack of focus on mental health related outcomes within the HIV literature, with depression and substance use being the only measures identified.

6.7.1 Surrogate outcomes in adaptive trials

Our review found that only 15 studies (2.2%) used some form of surrogate outcome, with CD4+ count accounting for 37.5%. CD4+ count served as a proxy for advanced HIV disease [43], (late or early) presentation to care [42, 43], and engagement with care [34, 44]. Viral load was the next most frequently used surrogate outcome, acting as a surrogate for engagement with care [34, 44], retention in care [46], routine HIV care visits [47], linkage to care [45], and ART adherence [48].

While both CD4+ count and viral load have long been used as markers of immune function, disease severity, mortality, and ART treatment response respectively, their use as surrogates for engagement in care marks a significant shift. This change emphasizes patient involvement in ART management, highlighting the crucial role of patient engagement in sustaining treatment success [55]. This transition may be beneficial, as a review of HIV/AIDS trials revealed that the effect of treatment on CD4+ count did not consistently predict the effects of treatment on disease progression or time to mortality [56].

Our current understanding of the relationship between CD4+ count and viral load and engagement in care is largely associational. This limited insight invites further examination of these biomarkers as surrogates for engage in care and other HIV care continuum outcomes. While some studies in our review used these measures as surrogates of engagement in care, suggesting that changes in these outcomes precede changes in engagement in care, other research suggests the opposite causal direction. Studies have shown that poor engagement is associated with worsened CD4+ count and viral load [73, 74], hinting at bidirectional influences. This complexity emphasizes the importance of careful selection of surrogate outcome measures for adaptive trials. It also highlights the value of using causal diagrams and involving stakeholders to clarify causal pathways and potential feedback loops. Such careful consideration is crucial when designing adaptive trials to ensure that surrogate outcomes serve as reliable predictors of the primary endpoint, moving beyond association to capture meaningful causal relationships.

6.7.2 Constructed DAG:

We synthesized the literature to create a DAG illustrating the causal relationships amongst commonly used primary and surrogate outcomes in HIV research studies. CD4+ cell count emerged as the most frequently reported surrogate outcome, serving as a proxy for retention and engagement in care as well as presentation to care.

For decades, CD4+ cell count measurement has been used as a crucial indicator of presentation to care, HIV disease progression, and AIDS-related mortality [57]. It also is used to predict the risk of specific opportunistic infections and guide decisions about prophylaxis [58]. Typically, one of the first clinical tests performed on newly diagnosed individuals, CD4+ cell count is

measured prior to the initiation of ART to establish the need for prophylaxis against opportunistic infections [58] as well as to stage and monitor disease progression. Lower CD4+ cell counts indicate compromised immune function, greater likelihood of opportunistic infections, and heightened risk for mortality [57].

Viral load was the second most prominently used surrogate outcome identified in this review. Viral load is an important indicator of ART response [59] and a crucial predictor of HIV transmission risk. Similar to CD4+ count, viral load measurements prior to ART initiation and the subsequent decline after initiation provide crucial information about the likelihood of disease progression and treatment response [60]. Viral suppression and undetectable viral load are defined as confirmed viral load levels at the lower limit of detection (200 copies of HIV / mL blood) or below the level of detection, respectively [61].

According to treatment guidelines, viral load and CD4+ counts should be monitored every 3 – 12 months depending on the individual's progression through treatment [16]; thus, routine laboratory measures have been used to reflect an individual's retention/engagement in care.

This DAG contained many of the known and very frequently reported outcomes in HIV research; however, there was an absence of other known influences such as social and mental health factors as well as structural influences such as socioeconomic status or the healthcare system itself. This highlights the value of domain expertise in creating more comprehensive DAGs.

6.7.3 Breadth of HIV outcomes captured

As ART has become more accessible and effective, HIV has transitioned from a fatal disease to a chronic, manageable condition. This evolution has necessitated a re-evaluation of what outcomes are most relevant in measuring and determining the health of individuals living with HIV today. While traditional outcomes such as survival and all-cause mortality remain important, they are no longer the primary focus. In the included studies, all-cause mortality was a primary outcome in 13 cases, but survival was not identified as a primary outcome.

With HIV now considered a chronic condition, the assessment of primary outcomes has shifted towards a more holistic approach considering aspects of health beyond just the physical. However, there is a clear gap in the assessment of mental health, with depression, opioid misuse, and self-rated health being the only mental health outcomes evaluated, appearing in 10, 2, and 2

of the included studies, respectively. The most frequently evaluated primary outcomes in this synthesis were viral suppression (cited in 91 studies), retention in care (64 studies), ART adherence (56 studies), and the HIV care continuum (47 studies). These outcomes assess two of three of the WHO aspects of health: physical and social. Despite this broader focus, mental health remains significantly underrepresented.

Mental health disorders remain a significant source of mortality and morbidity in the general population [62]. They are also one of the most common comorbidities amongst PLWH, including depression, anxiety, and severe mental illness (e.g., substance use disorders, psychoses, schizophrenia, and bipolar disorder) [63]. Some studies indicate that PLWH experience higher rates of mental illness compared to the general population [64-67]. A North American cohort study of 122,896 PLWH between 2008 and 2018 found that 55% were diagnosed with at least one of the following mental health disorders: depressive disorder, anxiety disorder, bipolar disorder, and schizophrenia [68]. Such psychiatric comorbidities have been linked to poorer health outcomes across the HIV care continuum, with the relationship between ART adherence and mental health being the most studied. A meta-analysis of 95 independent samples found depression to be significantly associated with ART non-adherence [69].

6.7.4 Underrepresented populations: Gender and race

Despite a rise in the inclusion of women and transgender individuals as participants, men remain the predominant group of focus in HIV studies. Although the majority of studies included in this review focused on PLWH in general, men significantly outnumbered women and transgender individuals in these studies. Out of 435 studies with PLWH as the focal population, only 17 reported a higher number of women participants compared to men.

By the end of 2020, in Canada, 75.4% of PLWH were men and 24.6% were women [70]. While the proportion of new infections among typically high-prevalence groups such as gay, bisexual, and men who have sex with men (MSM) has been declining, the incidence of HIV has slightly increased in people who inject drugs, women, and Indigenous peoples [70]. Following PLWH, the next most studied groups in this review were women living with HIV (23 studies, 3.4%) and drug users (18 studies, 2.6%). This indicates a shift in the epidemiological landscape, with increased attention on more diverse populations.

Women and transgender individuals were not the only groups underrepresented in the reviewed studies. We identified 64 distinct participant groups including demographic groups (e.g., MSM, women of colour, Black men, etc.,), health and treatment status (e.g., ART naïve, virally unsuppressed, etc.,), behavioural and lifestyle factors (e.g., with mental illness, drug users) and specific health contexts (e.g., youth transitioning to adult care, older adults, pregnant women). Of these 64 groups, 14 represented racial/ethnic minorities, with the most prevalent being transgender women of colour (10 studies, 1.5%), Black MSM (9 studies, 1.3%), and Black men (3, 0.4%). Additionally, while 86.3% of the included studies collected information on the race/ethnicity of their participants, less than half reported outcomes according to these subgroups. This demonstrates a gradual improvement in the inclusion of minorities but highlights a significant gap in the diversity of populations studied and the reporting of outcomes, which may affect the generalizability and applicability of research findings across different communities. Moreover, minorities whether defined by their gender, sexuality, or race/ethnicity identity, may be affected by socio-economic determinants of health e.g., gender, income, education, access to health services, racism, stigma impacting their HIV health outcomes. This emphasizes the need for more comprehensive research approaches that not only include diverse populations, but also analyze and report data specific to these groups. By doing so, researchers can better understand the complex interplay of socio-economic determinants and health outcomes, ultimately contributing to more effective and tailored interventions that address the unique challenges faced by these underrepresented communities in the treatment of HIV.

6.7.5 Limitations

This review had some limitations. First, the broad research question guiding this scoping review aimed to identify HIV-related outcomes reported in research studies conducted in high-income countries, resulting in the inclusion of a vast array of studies. Despite the extensive number of included studies, an overrepresentation of certain populations (e.g., men) and underrepresentation of others (e.g., women, transgender individuals, racial/ethnic minorities) was evident. This imbalance was also evident in the outcomes identified, with a predominant focus on physical and, to a lesser extent, social health outcomes, while mental health outcomes were notably sparse.

Moreover, the review was restricted to only quantitative or mixed methods studies, encompassing both observational and experimental designs. This inclusivity may have introduced a variability in study designs, potentially impacting the comparability of the outcomes reported. Additionally, by limiting the scope to studies conducted in high-income countries, valuable research from other settings may have been inadvertently excluded, potentially overlooking insights relevant to a broader context.

This review was also restricted to articles to those published from 2006 onwards, marking the period when combination ART became widely available. We observed an interesting trend showing an increase in the number of articles meeting our eligibility criteria i.e., quantitative or mixed methods studies conducted in high-income countries with PLWH (see Appendix C). While HIV has transitioned from a terminal disease to a chronic, manageable condition, this alone cannot fully explain the increase in published studies. A more likely explanation is the overall landscape of published research. Of the 681 included studies, published from 168 unique journals, 48.3% were secondary analyses i.e., a re-analysis of another study's data, use of data not collected for research purposes (electronic medical records, surveillance data, insurance claims, etc.,). This trend may reflect a dilution of information from original, primary research studies.

Despite these limitations, this scoping review provides a comprehensive overview of the outcomes and populations studied in the HIV literature from the advent of combination ART to present, highlighting critical trends and gap in current research.

6.8 Conclusion

The transition of HIV from an acute disease with inevitable mortality to a chronic condition managed by millions of people worldwide has transformed the outcomes used to assess health and treatment success. Our scoping review included a wide array of studies which revealed a shift from conventional clinical outcomes such as mortality and survival to more comprehensive measures of health like stigma or the HIV care continuum. However, immunological measures such as CD4+ cell count, viral load, and viral suppression remain the predominant health outcomes reported. Notably, mental health outcomes continue to be underrepresented in HIV care. Although there is a growing inclusion of underrepresented groups such as women, transgender individuals, and racial/ethnic minorities, men still predominantly populate HIV

research. There is pressing need to enhance the inclusivity of diverse populations and to broaden the scope of health outcomes to encompass more holistic aspects of health such as mental health.

6.9 References

1. Fleming TR, DeMets DL. Surrogate End Points in Clinical Trials: Are We Being Misled? Annals of Internal Medicine. 1996;125(7):605-13.
2. Long S, Loutfi D, Kaufman JS, Schuster T. Limitations of Canadian COVID-19 data reporting to the general public. J Public Health Policy. 2022;43(2):203-21.
3. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(Volume 34, 2013):61-75.
4. Long S, Schuster T, Piché A, Research S. Can large language models build causal graphs? arXiv preprint arXiv:230305279. 2023.
5. Fischl MA, Richman DD, Greico MH, Gottlieb MS, Volberding PA, Laskin OL, et al. The efficacy of azidothymidine AZT in the treatment of patients with AIDS and AIDS-related complex. N Engl J Med. 1987;317(4):185-92.
6. The Strategies for Management of Antiretroviral Therapy (SMART) Study Group. CD4+ count-guided interruption of antiretroviral treatment. N Engl J Med. 2006;355(20):2283-97.
7. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC, et al. Effect of Early vs. Deferred Antiretroviral therapy for HIV on Survival. N Engl J Med. 2009;360(18):1815-26.
8. Group ISS, Lundgren JD, Babiker AG, Gordin F, Emery S, Grund B, et al. Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection. N Engl J Med. 2015;373(9):795-807.
9. World Health Organization. Global Health Observatory (GHO) data 2020 [Available from: https://www.who.int/gho/hiv/en/.
10. Kobin A, Sheth N. Levels of adherence required for virologic suppression among newer antiretroviral medications. The Annals of Pharmacotherapy. 2011;45:371-80.
11. Iacob SA, Iacob DG, Jugulete G. Improving the Adherence to Antiretroviral Therapy, a Difficult but Essential Task for a Successful HIV Treatment-Clinical Points of View and Practical Considerations. Front Pharmacol. 2017;8:831.
12. Okoli C, Van de Velde N, Richman B, Allan B, Castellanos E, Young B, et al. Undetectable equals untransmittable (U = U): awareness and associations with health outcomes among people living with HIV in 25 countries. Sex Transm Infect. 2021;97(1):18-26.
13. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, van Lunzen J, et al. Sexual Activity Without Condoms and Risk of HIV Transmission in Serodifferent Couples When the HIV-Positive Partner Is Using Suppressive Antiretroviral Therapy. JAMA. 2016;316(2):171-81.
14. Berry S, Connor J, Lewis R. The Platform Trial: An efficient strategy for evaluating multiple treatments. JAMA. 2015;313(16):1619-20.
15. O'Brien N, Chi YL, Krause KR. Measuring Health Outcomes in HIV: Time to Bring in the Patient Experience. Ann Glob Health. 2021;87(1):2.
16. Clinical Info HIV.Gov. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV 2022 [updated September 21, 2022. Available from: https://clinicalinfo.hiv.gov/en/guidelines/hiv-clinical-guidelines-adult-and-adolescent-arv/plasma-hiv-1-rna-cd4-monitoring.
17. World Health Organization. Constitution of the World Health Organization. Geneva, Switzerland: World Health Organization; 2006.
18. Kay ES, Batey DS, Mugavero MJ. The HIV treatment cascade and care continuum: updates, goals, and recommendations for the future. AIDS Res Ther. 2016;13:35.

19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Scoping studies: towards a methodological framework. 2005;8(1):19-32.
20. World Bank. New country classifications by income level: 2019-2020 2019 [cited 2020 March 21]. Available from: https://blogs.worldbank.org/opendata/new-country-classifications-income-level-2019-2020.
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467-73.
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan — a web and mobile app for systematic reviews. Systematic Reviews. 2016;5(210).
23. The EndNote Team. EndNote. EndNote X8.0.2 ed. Philadelphia, PA: Clarivate Analytics; 2013.
24. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: A. B, Burgess RG, editors. Analyzing qualitative data. London, UK: Routledge; 1994.
25. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26.
26. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol. 2018;96:84-92.
27. Greenland S, Pearl J. Causal Diagrams. Encyclopedia of Epidemiology2006.
28. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
29. Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. Int J Epidemiol. 2020;49(1):322-9.
30. Cunningham CO, Sohler NL, Wong MD, Relf M, Cunningham WE, Drainoni M, et al. Utilization of health care services in hard-to-reach marginalized HIV-infected individuals. AIDS Patient Care & STDs. 2007;21(3):177-86.
31. Grace C, Kutzko D, Alston WK, Ramundo M, Polish L, Osler T. The Vermont model for rural HIV care delivery: eleven years of outcome data comparing urban and rural clinics. Journal of Rural Health. 2010;26(2):113-9.
32. Rebeiro PF, Cesar C, Shepherd BE, De Boni RB, Cortes CP, Rodriguez F, et al. Assessing the HIV Care Continuum in Latin America: progress in clinical retention, cART use and viral suppression. J Int AIDS Soc. 2016;19(1):20636.
33. Rebeiro PF, McPherson TD, Goggins KM, Turner M, Bebawy SS, Rogers WB, et al. Health Literacy and Demographic Disparities in HIV Care Continuum Outcomes. AIDS & Behavior. 2018;22(8):2604-14.
34. Williams EC, McGinnis KA, Edelman EJ, Matson TE, Gordon AJ, Marshall BDL, et al. Level of Alcohol Use Associated with HIV Care Continuum Targets in a National U.S. Sample of Persons Living with HIV Receiving Healthcare. AIDS & Behavior. 2019;23(1):140-51.
35. Farag E, Bozicevic I, Tartour AI, Nasreldin H, Daghfal J, Himatt S, et al. HIV case reporting and HIV treatment outcomes in Qatar. Frontiers in public health. 2023;11:1234585.
36. Lopez C, Moreland A, Goodrum N, Davies F, Meissner E, Danielson C. Association of mental health symptoms on HIV care outcomes and retention in treatment. Gen Hosp Psychiatry. 2023;82:41-6.

37. Norwood J, Kheshti A, Shepherd B, Rebeiro P, Ahonkhai A, Kelly S, et al. The Impact of COVID-19 on the HIV Care Continuum in a Large Urban Southern Clinic. AIDS and Behavior. 2022;26(8):2825-9.
38. Sanders R, Dombrowski J, Hajat A, Buskin S, Erly S. Associations between adverse childhood experiences, viral suppression, and quality of life among persons living with HIV in Washington state. AIDS care. 2024:1-9.
39. Shah S, Reist B, Sawyer J, Chiao A, Hodge S, Jones C, et al. Evaluating Evidence-Informed Behavioral Health Models to Improve HIV Health Outcomes: Quantitative Findings from the Ryan White HIV/AIDS Program Special Projects of National Significance Black Men Who Have Sex with Men Initiative. AIDS Patient Care and STDs. 2022;36:S3-S20.
40. Turner C, Trujillo D, Le V, Wilson E, Arayasirikul S. Event-Level Association Between Daily Alcohol Use and Same-Day Nonadherence to Antiretroviral Therapy Among Young Men Who Have Sex With Men and Trans Women Living With HIV: Intensive Longitudinal Study. JMIR mHealth and uHealth. 2020;8(10):e22733.
41. Mehta SH, Lucas G, Astemborski J, Kirk GD, Vlahov D, Galai N. Early immunologic and virologic responses to highly active antiretroviral therapy and subsequent disease progression among HIV-infected injection drug users. AIDS Care. 2007;19(5):637-45.
42. Mugavero MJ, Pence BW, Whetten K, Leserman J, Swartz M, Stangl D, et al. Childhood abuse and initial presentation for HIV care: an opportunity for early intervention. AIDS Care. 2007;19(9):1083-7.
43. Oliva J, Diez M, Galindo S, Cevallos C, Izquierdo A, Cereijo J, et al. Predictors of advanced disease and late presentation in new HIV diagnoses reported to the surveillance system in Spain. Gac Sanit. 2013;28(2):116-22.
44. Hemmy Asamsama O, Squires L, Tessema A, Rae E, Hall K, Williams R, et al. HIV Nurse Navigation: Charting the Course to Improve Engagement in Care and HIV Virologic Suppression. Journal of the International Association of Providers of AIDS Care. 2017;16(6):603-7.
45. Westergaard RP, Hochstatter KR, Andrews PN, Kahn D, Schumann CL, Winzenried AE, et al. Effect of Patient Navigation on Transitions of HIV Care After Release from Prison: A Retrospective Cohort Study. AIDS & Behavior. 2019;23(9):2549-57.
46. Hemmige V, Flash CA, Carter J, Giordano TP, Zerai T. Single tablet HIV regimens facilitate virologic suppression and retention in care among treatment naïve patients. AIDS Care. 2018;30(8):1017-24.
47. Loeliger KB, Meyer JP, Desai MM, Ciarleglio MM, Gallagher C, Altice FL. Retention in HIV care during the 3 years following release from incarceration: A cohort study. PLoS Medicine. 2018;15(10):1-28.
48. Fischetti B, Sorbera M, Michael R, Njeim N. Evaluation of rates of virologic suppression in HIV-positive patients with varying numbers of comorbidities. American Journal of Health-System Pharmacy. 2022;79(2):72-7.
49. Boyd AT, Song DL, Meyer JP, Altice FL. Emergency department use among HIV-infected released jail detainees. Journal of Urban Health. 2015;92(1):108-35.
50. Rebeiro P, Althoff KN, Buchacz K, Gill J, Horberg M, Krentz H, et al. Retention among North American HIV-infected persons in clinical care, 2000-2008. J Acquir Immune Defic Syndr. 2013;62(3):356-62.

51. Keller SC, Yehia BR, Eberhart MG, Brady KA. Accuracy of definitions for linkage to care in persons living with HIV. Journal of Acquired Immune Deficiency Syndromes. 2013;63(5):622-30.
52. Byrd KK, Hou JG, Hazen R, Kirkham H, Suzuki S, Clay PG, et al. Antiretroviral Adherence Level Necessary for HIV Viral Suppression Using Real-World Data. J Acquir Immune Defic Syndr. 2019;82(3):245-51.
53. Ray M, Logan R, Sterne JA, Hernández-Díaz S, Robins JM, Sabin C, et al. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. Aids. 2010;24(1):123-37.
54. World Health Organization. HIV service delivery Geneva2024 [Available from: https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/treatment/service-delivery-adherence-retention.
55. Mugavero MJ, Davila JA, Nevin CR, Giordano TP. From access to engagement: measuring retention in outpatient HIV clinical care. AIDS Patient Care & STDs. 2010;24(10):607-13.
56. Fleming TR. Surrogate markers in aids and cancer trials. Statistics in Medicine. 1994;13(13-14):1423-35.
57. Battistini Garcia SA, Guzman N. Acquired Immune Deficiency Syndrome CD4+ Count Treasure Island: StatPearls Publishing 2023 [updated August 14, 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513289/#:~:text=The%20CD4%20count%20normal%20range,for%20the%20diagnosis%20of%20AIDS.
58. Graham NM, Park LP, Piantadosi S, Phair JP, Mellors J, Fahey JL, et al. Prognostic value of combined response markers among human immunodeficiency virus-infected persons: possible aid in the decision to change zidovudine monotherapy. Clin Infect Dis. 1995;20(2):352-62.
59. H. I. V. Surrogate Marker Collaborative Group. Human Immunodeficiency Virus Type 1 RNA Level and CD4 Count as Prognostic Markers and Surrogate End Points: A Meta-Analysis. AIDS RESEARCH AND HUMAN RETROVIRUSES. 2000;16:1123-34.
60. Murray JS, Elashoff MR, Iacono-Connors LC, Cvetkovich TA, Struble KA. The use of plasma HIV RNA as a study endpoint in efficacy trials of antiretroviral drugs. Aids. 1999;13(7):797-804.
61. World Health Organization. Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach. World Health Organization,; 2016.
62. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry. 2015;72(4):334-41.
63. De Francesco D, Sabin CA, Reiss P. Multimorbidity patterns in people with HIV. Curr Opin HIV AIDS. 2020;15(2):110-7.
64. Cook JA, Burke-Miller JK, Steigman PJ, Schwartz RM, Hessol NA, Milam J, et al. Prevalence, Comorbidity, and Correlates of Psychiatric and Substance Use Disorders and Associations with HIV Risk Behaviors in a Multisite Cohort of Women Living with HIV. AIDS Behav. 2018;22(10):3141-54.
65. Rubin LH, Maki PM. HIV, Depression, and Cognitive Impairment in the Era of Effective Antiretroviral Therapy. Curr HIV/AIDS Rep. 2019;16(1):82-95.
66. Nanni MG, Caruso R, Mitchell AJ, Meggiolaro E, Grassi L. Depression in HIV infected patients: a review. Curr Psychiatry Rep. 2015;17(1):530.

67. Gooden TE, Gardner M, Wang J, Chandan JS, Beane A, Haniffa R, et al. The risk of mental illness in people living with HIV in the UK: a propensity score-matched cohort study. The Lancet HIV. 2022;9(3):e172-e81.
68. Lang R, B. H, J. Z, K. M, J. L, P. Z, et al. The prevalence of mental health disorders in people with HIV and the effects on the HIV care continuum. Aids. 2023;37(2):259-69.
69. Gonzalez JS, Batchelder AW, Psaros C, Safren SA. Depression and HIV/AIDS treatment nonadherence: a review and meta-analysis. J Acquir Immune Defic Syndr. 2011;58(2):181-7.
70. Public Health Agency of Canada. Estimates of HIV incidence, prevalence and Canada's progress on meeting the 90-90-90 HIV targets, 2020 Ottawa2022 [Available from: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/estimates-hiv-incidence-prevalence-canada-meeting-90-90-90-targets-2020.html#a3.

6.11 Appendix A: PubMed Search Strategy

(hiv[ti] OR "hiv infections"[mesh])

AND

(hiv care[tiab] OR hiv management[tiab] OR care continuum[tiab] OR delivery of care[tiab] OR delivery of healthcare[tiab] OR delivery of health care[tiab] OR health care services[tiab] OR "Delivery of Health Care"[Mesh:noexp] OR "Patient Care"[Mesh:noexp] OR "Delivery of Health Care, Integrated"[Mesh:noexp])

AND

(outcome*[tiab] OR "outcome assessment (health care)"[mesh])

6.12 Appendix B: Journals of all articles included in this review

Journals (N=168)	No. of publications (%)
AIDS and Behavior	88 (12.8)
Journal of Acquired Immune Deficiency Syndromes	69 (10.1)
AIDS Care	65 (9.5)
AIDS Patient Care & STDs	50 (7.3)
AIDS	33 (4.8)
PLoS ONE	29 (4.2)
Clinical Infectious Diseases	27 (3.9)
Open Forum Infectious Diseases	15 (2.3)
HIV Medicine	14 (2)
Journal of the International AIDS Society	9 (1.3)
International Journal of STD and AIDS	9 (1.3)
BMC Public Health	7 (1)
AIDS Research and Human Retroviruses	7 (1)
Journal of the International Association of Providers of AIDS Care	7 (1)
Drug & Alcohol Dependence	6 (0.9)
HIV Research and Clinical Practice	5 (0.7)
International Journal of Environmental Research and Public Health	5 (0.7)
Journal of General Internal Medicine	5 (0.7)
AIDS Research and Therapy	5 (0.7)
American Journal of Public Health	4 (0.6)
Frontiers in Public Health	4 (0.6)
Journal of Infectious Diseases	4 (0.6)
Journal of Medical Internet Research	4 (0.6)
Journal of Rural Health	4 (0.6)
Journal of Antimicrobial Chemotherapy	4 (0.6)
International Journal of Antimicrobial Agents	3 (0.4)
JAMA Internal Medicine	3 (0.4)
Annals of Internal Medicine	3 (0.4)
Health Psychology	3 (0.4)
International Journal of Prisoner Health	3 (0.4)
Journal of Behavioral Medicine	3 (0.4)
Journal of HIV/AIDS & Social Services	3 (0.4)
Journal of Public Health Management & Practice	3 (0.4)
Journal of the Association of Nurses in AIDS Care	3 (0.4)
Sexual Health	3 (0.4)
Sexually Transmitted Diseases	3 (0.4)
The Lancet HIV	3 (0.4)
Journal of Immigrant and Minority Health	3 (0.4)

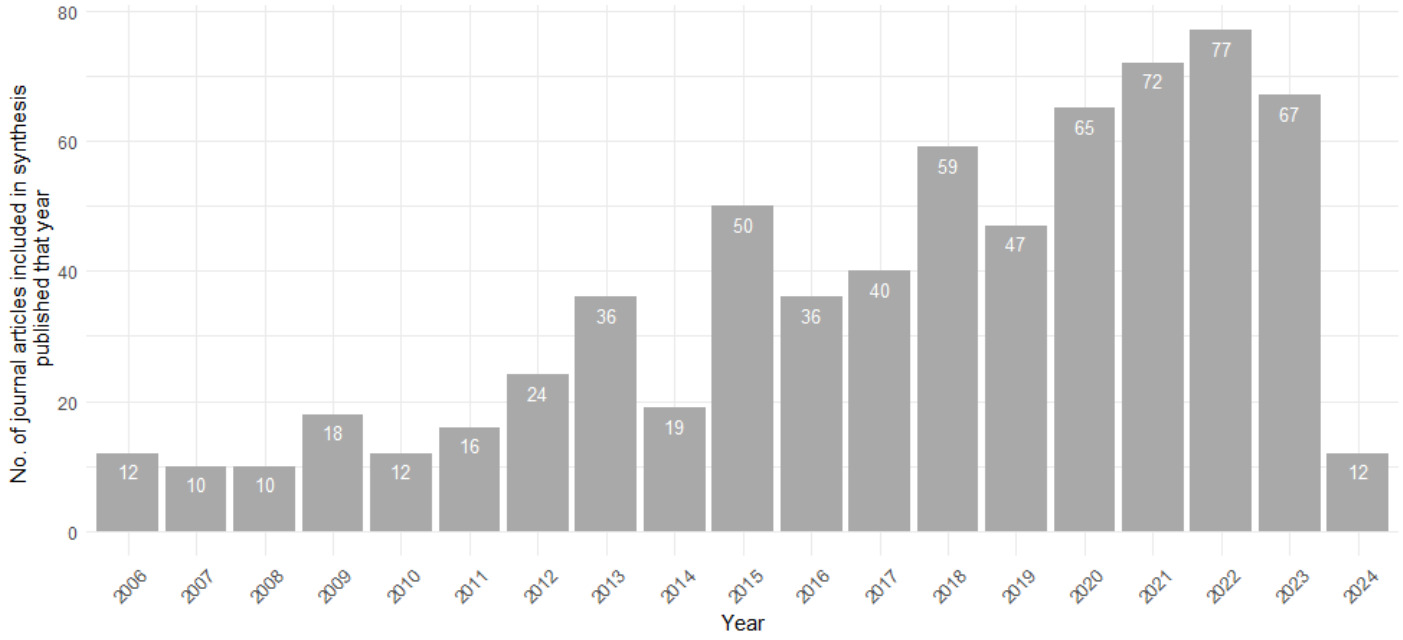
Addiction Science & Clinical Practice	2 (0.3)
BMC Health Services Research	2 (0.3)
Canadian Journal of Infectious Diseases and Medical Microbiology	2 (0.3)
ClinicoEconomics and Outcomes Research	2 (0.3)
HIV Clinical Trials	2 (0.3)
International Journal of Behavioral Medicine	2 (0.3)
International Journal of Drug Policy	2 (0.3)
International Journal of Infectious Diseases	2 (0.3)
JAMA	2 (0.3)
Journal of AIDS and Clinical Research	2 (0.3)
Journal of Community Health	2 (0.3)
Journal of Consulting & Clinical Psychology	2 (0.3)
Journal of Managed Care Pharmacy	2 (0.3)
Journal of Pain and Symptom Management	2 (0.3)
Journal of Pharmacy Practice	2 (0.3)
Journal of Urban Health	2 (0.3)
Medical Care	2 (0.3)
Patient Education & Counseling	2 (0.3)
Psychiatric Services	2 (0.3)
Public Health Reports	2 (0.3)
Social Work Research	2 (0.3)
Southern Medical Journal	2 (0.3)
General Hospital Psychiatry	2 (0.3)
Infectious Diseases Now	2 (0.3)
JMIR mHealth and uHealth	2 (0.3)
Journal of Microbiology, Immunology and Infection	2 (0.3)
Patient Preference and Adherence	2 (0.3)
Sexually Transmitted Infections	2 (0.3)
Therapeutic Advances in Infectious Disease	2 (0.3)
BMC Infectious Diseases	2 (0.3)
Drug and Alcohol Dependence	2 (0.3)
Journal of NeuroVirology	2 (0.3)
Journal of the American College of Clinical Pharmacy	2 (0.3)
Academic Emergency Medicine	1 (0.1)
AIDS Research & Therapy	1 (0.1)
Annals of the Academy of Medicine, Singapore	1 (0.1)
Archives of Internal Medicine	1 (0.1)
Asian Nursing Research	1 (0.1)
BMC Infectious Disease	1 (0.1)
Canadian Family Physician	1 (0.1)
Care Management	1 (0.1)
Clinical Journal of Pain	1 (0.1)
EClinicalMedicine	2 (0.1)

European Journal of Health Economics	1 (0.1)
Family Practice	1 (0.1)
Gaceta Sanitaria	1 (0.1)
Health and Justice	1 (0.1)
Health Policy	1 (0.1)
Infectious Diseases in Obstetrics & Gynecology	1 (0.1)
Internal Medicine Journal	1 (0.1)
International Journal of Medical Informatics	1 (0.1)
JAMA Psychiatry	1 (0.1)
Journal of Addiction Medicine	1 (0.1)
Journal of Alternative & Complementary Medicine	1 (0.1)
Journal of Clinical Psychology in Medical Settings	1 (0.1)
Journal of Correctional Health Care	1 (0.1)
Journal of Medical Virology	1 (0.1)
Journal of Substance Abuse Treatment	1 (0.1)
Journal of the American Board of Family Medicine	1 (0.1)
Journal of the International Association of Physicians in AIDS Care	1 (0.1)
Journal of the National Medical Association	1 (0.1)
Lancet Infectious Diseases	1 (0.1)
Medical Decision Making	1 (0.1)
New Zealand Medical Journal	1 (0.1)
Open AIDS Journal	1 (0.1)
Pediatric Infectious Disease Journal	1 (0.1)
PeerJ	1 (0.1)
Rehabilitation Counseling Bulletin	1 (0.1)
Rural & Remote Health	1 (0.1)
The Journal of Mental Health Policy & Economics	1 (0.1)
Value in Health Regional Issues	1 (0.1)
Violence and Gender	1 (0.1)
Women's Health Issues	1 (0.1)
Journal of the Association of Medical Microbiology and Infectious Disease Canada	1 (0.1)
Addiction	1 (0.1)
Addiction Science and Clinical Practice	1 (0.1)
AIDs Care	1 (0.1)
Alcoholism: Clinical and Experimental Research	1 (0.1)
American Journal of Obstetrics and Gynecology MFM	1 (0.1)
Behavioral Medicine	1 (0.1)
Enfermedades Infecciosas y Microbiologia Clinica	1 (0.1)
Frontiers in Immunology	1 (0.1)
Future Virology	1 (0.1)
Health Education AND Behavior	1 (0.1)
Health Equity	1 (0.1)

Hispanic Health Care International	1 (0.1)
JMIR Formative Research	1 (0.1)
Journal of Managed Care and Specialty Pharmacy	1 (0.1)
Journal of Public Health Management and Practice	1 (0.1)
Journal of Substance Use and Addiction Treatment	1 (0.1)
Journal of the American Medical Informatics Association	1 (0.1)
Lancet Regional Health - Americas	1 (0.1)
mHealth	1 (0.1)
Military Medicine	1 (0.1)
PharmacoEconomics	1 (0.1)
Porto Biomedical Journal	1 (0.1)
Preventive Medicine	1 (0.1)
Psychology, Health & Medicine	1 (0.1)
Research on Social Work Practice	1 (0.1)
Substance Use and Misuse	1 (0.1)
Viruses	1 (0.1)
Age and Ageing	1 (0.1)
Alcoholism, Clinical and Experimental Research	1 (0.1)
American Journal of Health-System Pharmacy	1 (0.1)
American Journal of Managed Care	1 (0.1)
BMJ Open	1 (0.1)
Critical Care	1 (0.1)
Drug Design, Development and Therapy	1 (0.1)
Enfermedades infecciosas y microbiologia clinica (English ed.)	1 (0.1)
Epidemiology and Infection	1 (0.1)
Frontiers in Sociology	1 (0.1)
Healthcare	1 (0.1)
Infectious Disease Reports	1 (0.1)
JMIR Medical Informatics	1 (0.1)
JMIR Mhealth Uhealth	1 (0.1)
Journal of Adolescent Health	1 (0.1)
Journal of Advanced Nursing	1 (0.1)
Journal of Infection and Public Health	1 (0.1)
Journal of interpersonal violence	1 (0.1)
Journal of Investigative Medicine	1 (0.1)
Journal of Microbiology Immunology and Infection	1 (0.1)
Journal of Public Health and Emergency	1 (0.1)
Journal of the American Geriatrics Society	1 (0.1)
Medicine	1 (0.1)
Patient Education and Counseling	1 (0.1)
PLOS Medicine	1 (0.1)
Population Medicine	1 (0.1)
Public Health Nutrition	1 (0.1)

Public Health Rep	1 (0.1)
The Lancet	1 (0.1)

6.13 Appendix C: Frequency of publications in articles included in this review: 2006 – 2024



CHAPTER 7: LEVERAGING EXPERT KNOWLEDGE FOR THE CO-CREATION OF CAUSAL DIAGRAMS IN INFORM CLINICAL TRIAL PLANNING: A FEASIBILITY STUDY IN THE CONTEXT OF HIV (MANUSCRIPT 4)

"Data do not understand causes and effects; humans do." – Judea Pearl

7.1 Preamble

Chapter 4 (Manuscript 1) demonstrated there were challenges in implementing established epidemiological principles to public health data. Calling for more routine implementation of causal modelling in public health epidemiology [1]. With the growing amount of data, we are reaching dimensions that limit the manual handling and creation of DAGs. To address these challenges, Chapter 5 (Manuscript 2) examined the utility of the LLM, GPT-3 in identifying the direction and/or presence of an edge between two variables. Prediction accuracy in edge direction between variables varied by medical context, language used in prompts e.g., more specific or use of certain terms [2]. These findings suggest improvements may be made with pre-training on actual scientific literature [3].

As such, Chapter 6 (Manuscript 3) was naturally a return to the scientific literature. I conducted a scoping review of HIV-related observational studies and RCTs conducted in high-income countries between 2006 to 2024. This review resulted in the creation of a DAG illustrating the causal relationships among commonly reported HIV-related individual level outcomes. The outcome measures (primary and surrogate) identified in this review revealed a predominant focus on physical and clinical outcomes and limited use of surrogate outcomes. Social health outcomes were the next most reported type of outcomes, with very limited focus on mental health outcomes. From a face validity perspective, this DAG, though aiming to be a comprehensive overview of the HIV literature, had missing known social, mental, and structural factors influencing the existing variables in this DAG.

Thus, Chapter 7 (Manuscript 4) is the final study addressing the aims of this dissertation to assess methodological shortcomings in integrating public data and expert knowledge into DAG

construction. To complete this comprehensive assessment of the different types of data contributing to the development of DAGs, this final chapter incorporates domain expertise.

Informed by the DAG created from Chapter 6 (Manuscript 3), this chapter developed and assessed the feasibility of an alternative approach to integrating domain expertise in DAG construction. From firsthand experience of creating DAGs and helping teach graduate students how to draw DAGs in an introductory epidemiology and biostatistics course as a teaching assistant, I know drawing a DAG can be an overwhelming and laborious task - a reality rarely mentioned in the literature.

At their most basic, a *directed acyclic graph* is composed of two elements: variables (nodes; e.g., representing treatments, exposures, health outcomes, or patient characteristics) and arrows between nodes (directed edges; which depicted known or suspected causal relationships), which cannot introduce cycles (*acyclic)* i.e., a node cannot be the cause of itself [4]. Though some guidance does exist [5, 6], there is little acknowledgment about the practicality of the task, particularly in research scenarios with many interrelated variables, which may lead to the creation of a DAG that is difficult to effectively manage and visualize [7].

Involving domain experts also presents other practical challenges due to scheduling conflicts and their limited ability to commit substantial time to research tasks. Typically, reasons for declining to participate in research are practical in nature, with potential participants citing inability to get time off work or unwillingness to travel [8]. Thus, a central goal of this chapter was to develop a practical approach to DAG development with domain experts mindful of time constraints, but still capable of developing comprehensive DAGs.

Chapter 7 – Manuscript 4

7.2 Title Page

Leveraging expert knowledge for the co-creation of causal diagrams to inform clinical trial planning: A feasibility study in the context of HIV research

Stephanie Long[1], Kim Engler[2,3], Bertrand Lebouché[1-5], Tibor Schuster[1]

[1]Department of Family Medicine, McGill University, Montreal, Canada

[2]Research Institute of the McGill University Health Network, McGill University, Montreal, Canada

[3]Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada

[4]Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, Quebec, Canada

[5]Chronic Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, Quebec, Canada

7.3 Abstract

Background:

Selecting appropriate surrogate outcomes, intermediate measures used as substitutes of the outcome of interest, is crucial in adaptive trials. Surrogate outcomes must reliably predict primary endpoints and be sensitive to intervention effects to have utility. Directed acyclic graphs (DAGs) offer a rigorous method for representing causal relationships between variables, providing potential utility in surrogate outcome selection. Despite their widespread use in epidemiology, DAGs are rarely used health research, and practical guidance on creating DAGs with domain experts is limited. Traditional approaches, which involve identifying and connecting all possible nodes with edges, are often time-consuming and result in cluttered DAGs. We developed and assessed the feasibility of an alternative approach to DAG development with domain experts, informed by a baseline DAG created from the findings of a scoping review on HIV-related patient outcomes.

Objectives:

1. To assess the feasibility of an alternative approach to DAG development with domain experts within the HIV field.
2. To update a baseline DAG of HIV-related patient outcomes with these experts.

Methods:

For this feasibility study, participants were recruited via convenience sampling and included researchers, clinicians, and patients with expertise in HIV affiliated with the McGill University Health Centre (Montreal, Canada). The DAG development approach involved individual virtual sessions where domain experts were asked to modify a baseline DAG derived from a scoping review. Participants were instructed to add, remove, or reposition nodes and edges, placing them temporally where they believed it had the most impact: not in care, transition to care, in care. Semi-structured interviews with domain experts were conducted to gather feedback on the process's feasibility and acceptability.

Results:

Seven sessions with seven domain experts, averaging 43.5 minutes, were conducted between February and June 2024. After the initial pilot test revealed the original approach was overly

complicated, we refined the method to focus on two outcomes of interest: engagement in care and medication adherence. We maintained the temporal organization and added a step to elicit variables by categorizing them on based their impact (help or hinder) on the outcome of interest. The revised approach comprised five steps, with domain experts involved in steps 1-4: (1) introduction, (2) elicit variables, (3) organize variables temporally, (4) completion of DAG session, (5) consolidate domain expert input and create the DAG. In interviews, domain experts reported a positive experience, appreciating the structured yet flexible space for reflection and knowledge sharing. They suggested improvements, including considering patients' perspectives on instructions and repeating definitions for clarity. The final updated DAG incorporated more social and structural factors in HIV care than the initial scoping review, extending beyond physical and clinical outcomes.

Conclusion:

This proposed DAG development approach offers a structured, efficient method for knowledge elicitation in applied health research. It presents a more feasible and practical approach to conventional approaches, considering study recruitment restraints and time commitments. This approach is particularly appropriate for settings where the potential DAG has many interrelated variables and shows promise for the selection of surrogate outcomes in adaptive trial designs by simplifying the process of including domain experts. Domain experts positively received the method, appreciating its structured yet flexible nature. The resultant DAG became more comprehensive, with domain experts incorporating social and structural factors in HIV care that were not captured in the scoping review. This approach enhances the feasibility of creating DAGs with domain experts in complex health research settings, potentially improving the quality and applicability of graphical causal models in fields such as HIV research.

7.4 Introduction

The selection of appropriate outcome measures in clinical trials remains a fundamental challenge that must take ethical, scientific, feasibility, and economic aspects into account. Surrogate outcomes, which are biomarkers or intermediate outcomes that are used as substitutes for the primary outcome of interest, are particularly desirable in settings where the outcome of interest takes a long time to observe or is prohibitively expensive to measure [9, 10]. Modern clinical trial designs such as *adaptive trials* [11], leverage outcome-driven randomization schemes (i.e., adaptive randomization) based on interim analyses. The utility of these analyses relies on the specification of appropriate surrogate outcomes, in addition to a primary study endpoint.

To ensure a surrogate outcome is a valid proxy for the primary study endpoint, it must fulfill two fundamental criteria. First, the effect of the intervention (or exposure) on the surrogate outcome is strongly correlated to the primary endpoint [12, 13]. Second, the surrogate outcome must fully capture the net effect of the intervention (or treatment) on the primary endpoint i.e., must be a mediator of the exposure (intervention) effect on the primary outcome [12] (Figure 7-2). While the first criterion is relatively easy to verify, the second is more challenging and thus is often unmet for a variety of reasons. The second criterion is a strong but necessary criterium, as a strong correlation alone is often insufficient, as illustrated in Figure 7-1, where the correlations can be spurious.

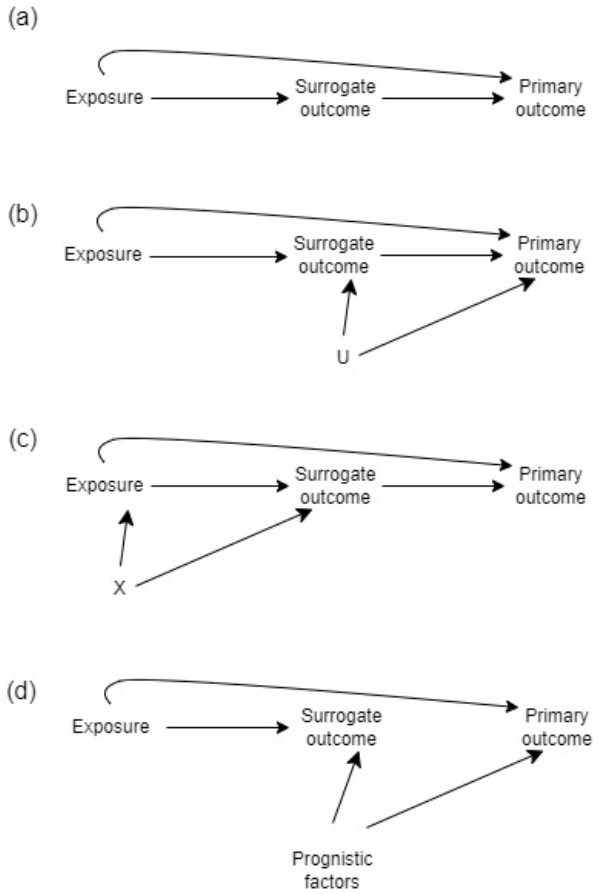


Figure 0-1: Four directed acyclic graphs describing potential surrogate outcome scenarios.

A particularly concerning phenomenon in the use of surrogate outcomes is the "surrogate paradox". This occurs when a treatment shows a positive effect on the surrogate outcome, which is in turn positively associated with the outcome of

interest; yet the treatment's effect on the primary outcome is null or even negative [10]. This paradox highlights the potential for misleading conclusions when relying solely on surrogate outcomes, even when they appear to satisfy both fundamental criteria. The surrogate paradox can arise due to complex causal pathways, unmeasured confounding, or treatment effect heterogeneity [14]. This phenomenon was tragically observed in trials evaluating the effect of drugs on ventricular arrythmia, used as a surrogate for mortality in cardiac patients [15, 16]. While the drugs effectively suppressed ventricular arrthymia (surrogate outcome), they unexpectedly increased mortality (primary outcome) [15, 17]. The possibility of the surrogate paradox underscores the critical importance of careful selection of surrogate outcomes.

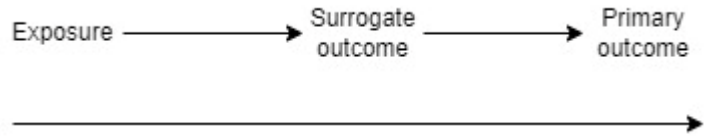


Figure 0-2: Ideal surrogate outcome scenario. The exposure acts on the causal pathway of the true primary outcome, fully mediated by the surrogate outcome. Thus, the exposure and primary outcome are conditionally independent based on the surrogate outcome.

Figures 7-1 and 7-2 illustrate various scenarios for surrogate outcomes, each with implications for surrogate utility. Figure 7-1a shows partial mediation, where the surrogate captures only part of the exposure's effect on the primary outcome; while this surrogate may have some utility, the partial mediation reduces it, as changes in the exposure could affect the primary outcome through pathways not captured by the surrogate. Figure 7-1b introduces an unmeasured confounder of the surrogate and primary outcomes, potentially creating spurious associations between the surrogate and primary outcome, biasing the surrogate's predictive capacity. Alternatively, there could be a known confounder X of the exposure and surrogate outcomes, as in Figure 7-1c but due to the direct effects of the exposure on the primary outcome, this surrogate may retain utility if the confounder is controlled for. Figure 7-1d depicts a scenario where prognostic factors for the surrogate and primary outcomes can create an apparent association between the surrogate and outcome despite their independence, undermining the surrogate's utility. Figure 7-2 illustrates the ideal scenario, where the surrogate fully mediates the exposure's effect on the primary outcome. These scenarios underscore the critical importance of careful

consideration of causal structures when selecting and interpreting surrogate outcomes in clinical trials.

Evidence synthesis of the literature can provide insight on the causal pathways among disease, intervention, primary endpoints, and surrogate outcomes. However, surrogate outcomes in one clinical context may not be applicable in another, even if they are similar. For example, while evidence suggests that progress-free survival is highly correlated with overall survival in patients with chronic lymphocytic leukemia [18], this relationship does not hold for other tumour types, such as metastatic breast cancer [19]. Furthermore, a surrogate widely used as a marker for a clinical outcome is not necessarily valid or predictive. A review found that CD4+ count, a commonly used indicator of disease progression and time to mortality in HIV/AIDS, did not consistently mirror the effect of treatment on these outcomes [20].

One approach to mitigating the selection of inappropriate surrogate outcomes not on the causal pathway to the primary outcome is to co-create causal diagrams such as directed acyclic graphs (DAGs) with domain experts. Although DAGs have been widely used in epidemiology, their use in applied health research is rare [7]. While instructions exist on how to construct a DAG in general and from synthesizing the literature [4, 6, 21-23], there is limited guidance for constructing DAGs with domain experts. Progress is being made in this area, as evidenced by Rodrigues et al., [24] who recently published findings from a DAG building workshop with stakeholders. Their workshop, involving 20 domain experts over 2.5 hours included four phases: brainstorming, refinement, exposition, and reconciliation. The authors faced challenges related to conveying complex concepts to participants with no prior knowledge of DAGs, the software used workshop facilitation, and time constraints [24]. Generally, as the number of nodes and edges increases, drawing a DAG becomes more complicated and more cognitively demanding. For example, a DAG containing only 10 possible variables (nodes), requires verification of 45 potential directed node connections.

Study objectives

Acknowledging these challenges, the primary objective of this study was to assess the feasibility and acceptability of an approach to construct DAGs with domain experts. The secondary objective was to update a baseline DAG created from a scoping review of the HIV literature,

with domain experts using this alternative approach. For readers unfamiliar with DAGs, more detailed information about them can be found in Supplemental Appendix A.

7.5 Methods

7.5.1 Design

This study was a multi-method feasibility study of a novel DAG development approach with domain experts.

7.5.2 Original DAG development approach

DAG development was planned to proceed through virtual individual sessions with each expert (on Zoom), following these four steps:

1) **Introduction**: A brief Microsoft Power Point presentation on DAGs and their associated nomenclature, with time allotted for the domain expert to ask questions and verify their understanding of DAGs.
2) **Baseline DAG Presentation:** The domain expert is then shown the baseline DAG, which they will be asked to update in the following step. In particular, domain experts were asked to focus on engagement and retention in care as well as antiretroviral therapy (ART) adherence, to identify factors that could serve as early indicators of these outcomes.
3) **DAG modification (temporal organization):** The domain expert is tasked with updating the baseline DAG based on their experiential expertise by adding, removing, or reorganizing variables and the connections between them. When adding nodes, domain experts are asked to place them temporally, where they were believed to have the greatest impact i.e., when an individual is *not yet in care*, *transitioning to care*, and *in care*.
4) **Conclusion:** The session ends when the domain expert confirms that their DAG is complete, with no further modifications needed.

Throughout the process, a set of guiding questions was used to facilitate discussion. The Microsoft PowerPoint presentation served as a visual aid for both the educational component and DAG development process.

7.5.3 Ethics

This study received ethics approval from the McGill University Health Centre (MUHC) Research Ethics Board.

7.5.4 Participants

Eligible participants were adults with knowledge and expertise in HIV such as individuals living with HIV, and HIV care providers and researchers, herein referred to as *domain experts*. The domain experts were recruited via convenience sample from the McGill University Health Centre, in Montreal, Quebec (Canada) and our professional network.

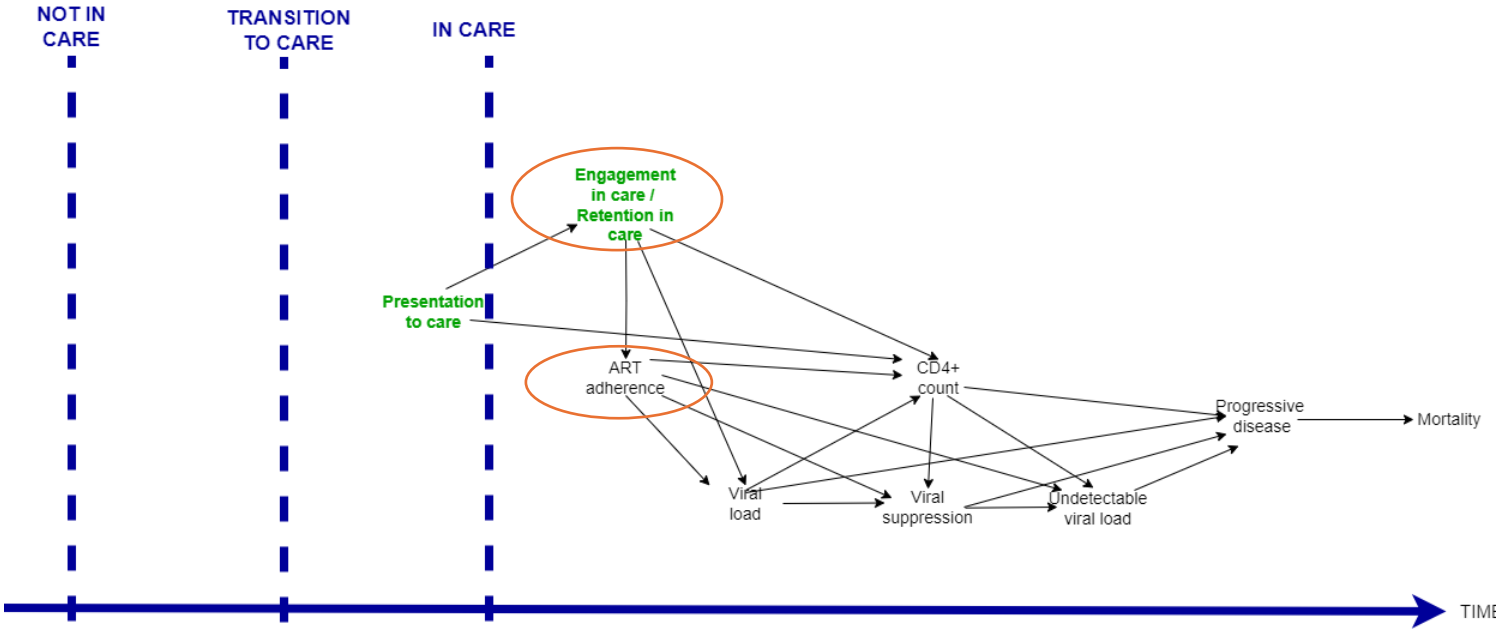


Figure 0-3: Baseline DAG to be updated with domain experts. Timeline indicating temporal organization along care continuum. Variables (nodes) are colour-coded according to WHO aspects of health: physical (black), social (green), and mental (red). Circled variables are the outcomes of interest for this study.

7.5.5 Outcomes: Feasibility and Acceptability

The feasibility of the approach was based on the number of potential participants screened; recruitment rate (number of participants recruited per month), retention rate (number of participants remaining in the study / total number of participants recruited), as well as the proportion of participants completing the full protocol with reasons for non-completion [25].

Additionally, we recorded the length of each causal mapping session and compared the actual duration to the predicted time of 45-60 minutes.

We also explored the feasibility and acceptability via semi-structured interviews with open-ended questions immediately after each session. The interview guide contained 11 questions and sub-questions: (1) describe experience of process, (2) feasibility of the approach (i.e., length of sessions, clarity of instructions), (3) acceptability of the approach (i.e., challenges experienced, likes and dislikes), (4) researcher only questions (i.e., *whether researchers would use this approach in their own research*), and (5) final remarks (Full interview guide can be found in Supplemental Appendix B).

7.5.6 Analyses

For our primary objective, quantitative feasibility outcomes were represented by descriptive statistics. For qualitative data, interviews were audio recorded, transcribed verbatim, and de-identified to protect participant confidentiality. We then analyzed this data using content analysis [26]. A single reviewer (SL) independently analyzed the interview transcripts using an inductive approach, systematically developing new meaning units as they arose. These meaning units were then condensed into codes, and subsequently grouped into broader, interpretive categories that captured the essential aspects of feasibility of the tested approach.

For our secondary objective to update an existing DAG (Figure 7-3), we consolidated and illustrated the variables identified by the domain experts. We illustrated these using heatmaps that maintained the categorizations (e.g., *help vs. hinder; not in care, transition to care, in care*) whilst reflecting frequency counts.

7.6 Results

7.6.1 Participants

Seven domain experts participated in this study between February and June 2024: four PhD-level researchers, a PhD candidate, a MSc student, and an MD-PhD (Table 7-1).

Table 0-1: Characteristics of domain experts (N=7)

	Gender	Education level	Expertise
1	F	Doctorate	Qualitative research, HIV care
2	M	Doctorate	Qualitative research, anthropology, HIV care
3	F	Doctorate, MD	Quantitative research, pregnancy intentions of women living with HIV, pediatric neurology
4	M	Doctorate	Statistics, HIV care
5	M	Doctorate	Mixed methods research, migrant health in HIV
6	M	MSc student	HIV-related expertise in community health, vulnerable populations
7	M	PhD candidate	Sexual dysfunction among men who have sex with men with HIV

7.6.2 Pilot test 1

During the very first session, following the original protocol, the domain expert quickly added a large number of nodes and edges, which overwhelmed the limited screen space available during Zoom sessions, creating a very cluttered and difficult to read DAG. The session was discontinued after 1.5 hours as the screen was too overcrowded to add more components.

The domain expert's feedback indicated that there was a lot of information to keep track of, making the task mentally taxing. They described it as "*a bit of a colossal task… for the participant who puts pressure on themselves, but also for the person leading the sessions, because keeping track of all of the ideas is…quite complex.*" Regarding the visual clutter, they also noted that, "*if you have to write everything up on the screen…it doesn't seem very feasible.*" (Domain Expert 1).

The complexity of the task was further highlighted by feedback on the information presentation itself: "*a lot of jargon…like 'nodes'…there's a lot of important information in your presentation*" (Domain Expert 1). This feedback prompted a need to adapt the approach to make it more practical.

7.6.3 Approach modifications following pilot test 1

As a result of this first test, we substantially reduced the amount of content presented at the introductory presentation, focusing more on introducing the clinical context and clearly defining

the outcomes of interest, and significantly reduced the amount of DAG-related content. This was intended to reduce the information, and the time required to conduct a session as well as to facilitate recruitment of domain experts. The revised approach was simplified, no longer updating the baseline DAG directly, temporally organizing added variables. We maintained the temporal organization, but added a preceding step, asking domain experts to identify variables according to their impact (help vs. hinder). Following this step, the process went forward as originally planned.

To mitigate the cognitive load and time commitment of domain experts and streamline the DAG development process, we revised the construction of DAGs with domain experts. Table 7-2 shows a comparison of the original and revised approach, highlighting differences.

Table 0-2: Comparison of original vs. revised approach

	Original process	**Revised process**
Number of steps	4	5
Presentation: # of slides	32	20
Topics covered	- DAGs and associated nomenclature - Baseline DAG inclusive of clinical context and outcomes of interest	- Clinical context - Outcomes of interest
Starting point of process	Baseline DAG derived from scoping review	Outcome of interest(s)
Steps:	1. Introduction 2. Baseline DAG presentation 3. DAG modification (temporal organization) 4. Conclusion [Remaining step is for researchers only] 5. Consolidate individual domain expert DAGs into final DAG	1. Introduction 2. Elicit variables (help & hinder) 3. Temporal organization of variables 4. Conclusion [Remaining step is for researcher only] 5. Consolidate domain experts' variables into final DAG
Output from domain experts	Updated DAG incorporating domain expert input	Lists of variables associated with the outcome of interest organized in two ways: impact (help or hinder) and temporal organization (not in care, transition to care, in care).
Length of time of each session	75 minutes	45 – 60 minutes

7.6.4 Pilot tests 2 – 7: Final approach to DAG Development with domain experts

For pilot tests 2 to 7, DAG development sessions lasted 60 minutes and involved one or two facilitators. The revised approach comprised of the following steps:

1) **Introduction**: We begin with a brief Microsoft PowerPoint presentation defining the specific outcomes of interest in which we want the domain experts to build a DAG around. This step focuses the domain expert's efforts on a well-defined objective and context, reducing the cognitive burden of considering an overly broad scope.
2) **Elicitation of variables:** In this step, domain experts were asked to identify variables or factors that either *help* or *hinder* the specified outcome(s) of interest. This categorization provides a framework for which the domain expert can use to brainstorm and systematically consider relevant factors without needing to think in terms of causal relationships.
3) **Temporal organization of variables:** The purpose of this step was to guide domain experts to organize the variables or factors they identified in step 2 across three temporal stages: not yet in care (*not yet receiving medical attention, treatment, or support for a health condition*), transition to care (*movement towards receiving medical attention, treatment, and support for a health condition*), and in care (*currently receiving medical attention, treatment, and support for a health condition*). These timepoints were selected as the aim was to identify early indicators of poor downstream outcomes. In situations where the domain expert believed a variable could have an effect at multiple timepoints, variables were placed according to *when* they believed it would have the greatest impact on the outcome of interest.

   This temporal structuring implicitly suggests causal relationships without requiring domain experts to explicitly draw edges between nodes. By organizing variables temporally, implied edges naturally emerge based on the chronological progression from one stage to the next. This method leverages an individual's intuitive understanding of time and sequence, reducing the cognitive effort needed to establish direct causal links between variables [27]. This approach aligns with the Bradford Hill criteria of causality, *temporality*, in which the cause must precede the effect, thereby reinforcing the logical flow of causal relationships.

4) **Conclusion***:* The session ends when the domain expert confirms that their DAG is complete, with no further modifications needed.

Remaining step is for the researcher only:

5) **Consolidation of input from domain experts and create DAG**: The purpose of this penultimate step was for the research team to independently consolidate the variables identified by the domain experts into a final DAG, expanding upon the baseline DAG (Figure 5). Thus, we aimed to find literature that supported the domain experts claims. Additionally, to reduce redundancy, we combined multiple similar nodes. We were guided by the protocol of Ferguson et al., (2022) [6]. The protocol consists of four steps: (i) mapping (*identifying primary and surrogate outcomes from the studies*), (ii) translation (*identifying the relationships between outcomes*); (iii) integration I (*synthesizing a DAG illustrating the relevant relationships*); and (iv) integration II (*grouping similar variables within the DAG*) [29].

7.6.5 Quantitative feasibility indicators:

7.6.5.1 Recruitment and retention:

- 7 potential participants were screened.
- 1.4 participants were recruited per month
- The retention rate was 100% (7 participants remained in the study out of 7 recruited)

7.6.5.2 Completion indicators:

- 6 out of 7 participants completed the entire session.
- Reasons for not completing session: pilot test 1 unveiled significant feasibility issues with original approach.
- The average length of time per DAG development session was of 43.5 ± 13.3 minutes (range: 36 – 60 minutes).
- Actual time of sessions did not exceed predicted time of DAG development sessions of 45 – 60 minutes.

7.6.6 Qualitative assessment of feasibility and acceptability:

Domain experts responded to open-ended questions regarding what they liked and disliked about the DAG development approach, and potential improvements in terms of the implementation of

the approach and its use. These semi-structured interviews were an average of 14.1 ± 6.2 minutes (range: 4-22 minutes).

7.6.6.1 Positive experiences and perceived benefits of the alternative DAG development approach

Domain experts generally found the alternative DAG development approach to be a positive experience. The approach provided a structured yet flexible space for reflection and knowledge sharing, while the open-endedness of the questions and discussions facilitated recall and brainstorming.

One domain expert highlighted the approach's ability to foster brainstorming and idea elaboration:

"I thought it was really interesting. I thought the structure was one where there was…a lot of space for brainstorming and…elaborating on different ideas. And I felt like the interaction that we had also helped me…think more or about different examples…you know, clarify ideas." – (Domain Expert 3)

Another domain expert appreciated the dual-set approach, emphasizing its role in reinforcing ideas and encouraging comprehensive thinking:

"And by utilizing…a dual set approach, where we first indicate both the facilitators and barriers, I think that's important that you addressed it twice. Because it created some sort of reinforcement. And I think I came up with more concepts that I would have initially, if it was a regular semi-structured interview. It gave me an opportunity to revisit a lot of the key outcomes and indicators that impact patients and the providers." (Domain Expert 6)

The approach was also seen as potentially beneficial for patients beyond its research purposes. As one domain expert observed: "*...so even for a patient… [it's] good to talk about those aspects and to put them in [context] just like we did*." (Domain expert 4)

7.5.6.2 Feasibility:

Length appropriateness and flexibility:

Domain experts generally found the session length appropriate, while also recognizing the potential need for flexibility. One domain expert described the sessions as "*short but good,*" while another noted that the ideal length might vary depending on the interviewee: "*Some people*

may have more to share, others not so much, that can impact the timing" (Domain Expert 6). This sentiment was echoed by another, who commented, "*I thought it was really good, like a good amount of time. Yeah, I think I had a good amount of time to…think about ideas.*" (Domain Expert 3). These responses suggest that while the current session length of 45-60 minutes was generally suitable, there may be value in maintaining some flexibility to accommodate different participants' experiences.

Clarity of instructions and suggestions for improvement:
Domain experts generally found the instructions clear and comprehensive. One domain expert appreciated the approach, stating, "*pretty comprehensive, I appreciate that. You provided a quick overview, first, in the form of a PowerPoint, just to know what the scope was,*" (Domain Expert 6). While another felt the instructions were very clear from a researcher's perspective, they expressed curiosity about how patients might perceive them. Echoing this sentiment, another domain expert suggested providing the definitions of outcomes throughout the session as a reminder because, "*if you have referred that in the very beginning of the presentation, perhaps at this point the person doesn't remember*" (Domain Expert 4). These comments indicate overall satisfaction with the clarity of instructions while also offering constructive suggestions for potential improvements, particularly for accommodating different types of domain expert participants.

7.6.6.3 Acceptability:

Challenges and potential improvements in the DAG development process:
Domain experts identified a few challenges in the DAG development process, primarily related to recall and time constraints. One domain expert noted difficulty in recollection, stating, "*the recall, like it takes a minute to think back to all the different studies that I did*" (Domain Expert 5). While generally satisfied with session length, some experts, felt that with additional time might they "*might have come up with more ideas*" (Domain Expert 3). To address these challenges, another domain expert suggested sharing the topics of interest prior to the sessions, explaining:

> *"It can give us some time to contemplate on what has worked for us what hasn't. and they might produce something richer… it might also…create a bottleneck effect almost where you're only thinking within those parameters. So…they both have their pros and cons" (Domain Expert 6).*

This suggestion highlights the potential trade-off between preparation and spontaneity in the process.

Another challenge emerged regarding the experts' perspectives. One domain expert expressed difficulty in navigating their role, stating: "[I struggled with my] *position, like my positionality in this, like I'm a researcher, I do have an expertise, but maybe not, you know, an expertise…like from a patient's experience*" (Domain Expert 3). This comment underscores the complexity of balancing different types of expertise within the DAG development process.

Positive aspects of the DAG development approach: structure, guidance, accessibility

Domain experts highlighted several positive aspects of the DAG development approach, particularly appreciating its structured yet flexible format. Several enjoyed the "*very open questions,*" (Domain Expert 2), while another valued the interviewer's guidance: "*I appreciated you giving me prompts to…elicit some more information from me*" (Domain Expert 5). Another domain expert echoed this sentiment, stating "*I also liked that I didn't feel complete left to…make sense of…all the categorization on my own like there was opportunities to kind of ask [the interviewer…for a clarification*" (Domain Expert 3).

The iterative nature of the sessions was another appreciated feature. A domain expert noted, "*I think the structure is perfect, because we can come back to it as well.*" (Domain Expert 6).

The interviewer's approach was crucial in creating an enjoyable experience as emphasized by a domain expert:

> *"So it's important to be able to… make the meeting something funny…So instead of being the cold academic person asking that question, blah, blah, you are laughing, you're gentle. So putting some emotional and personal aspects in the meeting can help a lot. And to make the meeting. Interesting, pleasant and enjoyable." (Domain Expert 4)*

Accessibility and convenience were enhanced through the use of virtual sessions via Zoom, which one domain expert found "*very convenient, very convenient for me*" (Domain Expert 4) and "*makes it accessible*" (Domain Expert 4).

The practicality and implementation-focused nature of the approach were highlighted, as one domain expert explained:

"I think it's practical. It's practical in the sense that it allows us to look at the main outcomes that are associated with the treatment process. And I think it's more tailored towards... It's more about implementation in my mind... knowledge generation, we know what exists already, there's many tools to evaluate the criteria of which we're looking for. But at this phase, I think we're past the knowledge synthesis aspect, or trying to create almost like a mind map as to what are the key areas that we need to address. And I think this will also highlight the gaps." *(Domain Expert 5)*

Another domain expert further emphasized the approach's practicality in addressing cognitive challenges in implementation:

"I think your approach is really addressing a really practical problem that happens when you try to do that with participants. Because it…can be quite…cognitive draining, to ask participants to …go through the exercise of doing all this brainstorming and then drawing all these arrows and think…it's quite a lot" (Domain Expert 3).

This feedback suggests that the approach successfully balanced structure with flexibility, provided necessary guidance, and addressed some practical challenges in DAG development. It also highlights its efficiency in managing the cognitive load on participants during the complex task of DAG development.

7.6.6.4 Prior experience with DAGs and related methods:

Most domain experts had limited to no prior experience with DAGs, with exposure "*mostly…in academic environments*" (Domain Expert 5). Two reported experience with related mapping techniques such as "*fuzzy cognitive mapping as part of my PhD work. And in that, I had a bit of a similar idea of…trying to map concepts using participants understanding of causal relationships*" (Domain Expert 3) and "*cognitive mapping…[but] is not the same though*" (Domain Expert 7).

This section highlights the novelty of DAG use for most of these domain experts, while also acknowledging some familiarity with related conceptual mapping techniques.

7.6.6.5 Potential future applications in research:

Domain experts expressed varying levels of interest in applying this approach to their future research. Some saw potential benefits, with one referencing a desire to "*go further than statistical associations*" (Domain expert 4).

Another expressed interest in using the approach based on past challenges in their own work\:

"At some point…it started to be like…you had to either sacrifice, kind of shortening the sessions, and then only focusing on the key important concepts, or doing…hours and hours long of mapping, which…you know…was a problem that I definitely experienced in my own work" (Domain Expert 3).

Another domain expert saw potential utility in addressing complex research questions: "*I would use in future research because I think… we're having a huge difficulty with seeing how external or context-specific factors are impacting patient outcomes*" (Domain Expert 6).

However, not all domain experts saw immediate applications, with one noting that "*none of my research questions really would need this kind of methodology or method*" (Domain Expert 5).

This illustrates the diverse perspectives on the approach's potential future applications, ranging from enthusiastic interest to more reserved consideration.

7.6.7 DAG development:

7.6.7.1 Variables elicited

Figures 7-4 and 7-5 are heatmaps depicting the frequency of variables identified by domain experts as helping or hindering the outcomes of interest: engagement in care (Figure 7-4) and ART adherence (Figure 4). Darker green colour indicates greater frequency of identification by domain experts.

Engagement in care (Figure 7-4):

Healthcare coverage and stigma were the most frequently identified variables, both helping and hindering engagement in care. Holistic care, social support system, and proximity to care were

often cited as helping factors. Healthcare mistrust, stigma from the community, and other life priorities were identified as hindering engagement in care.

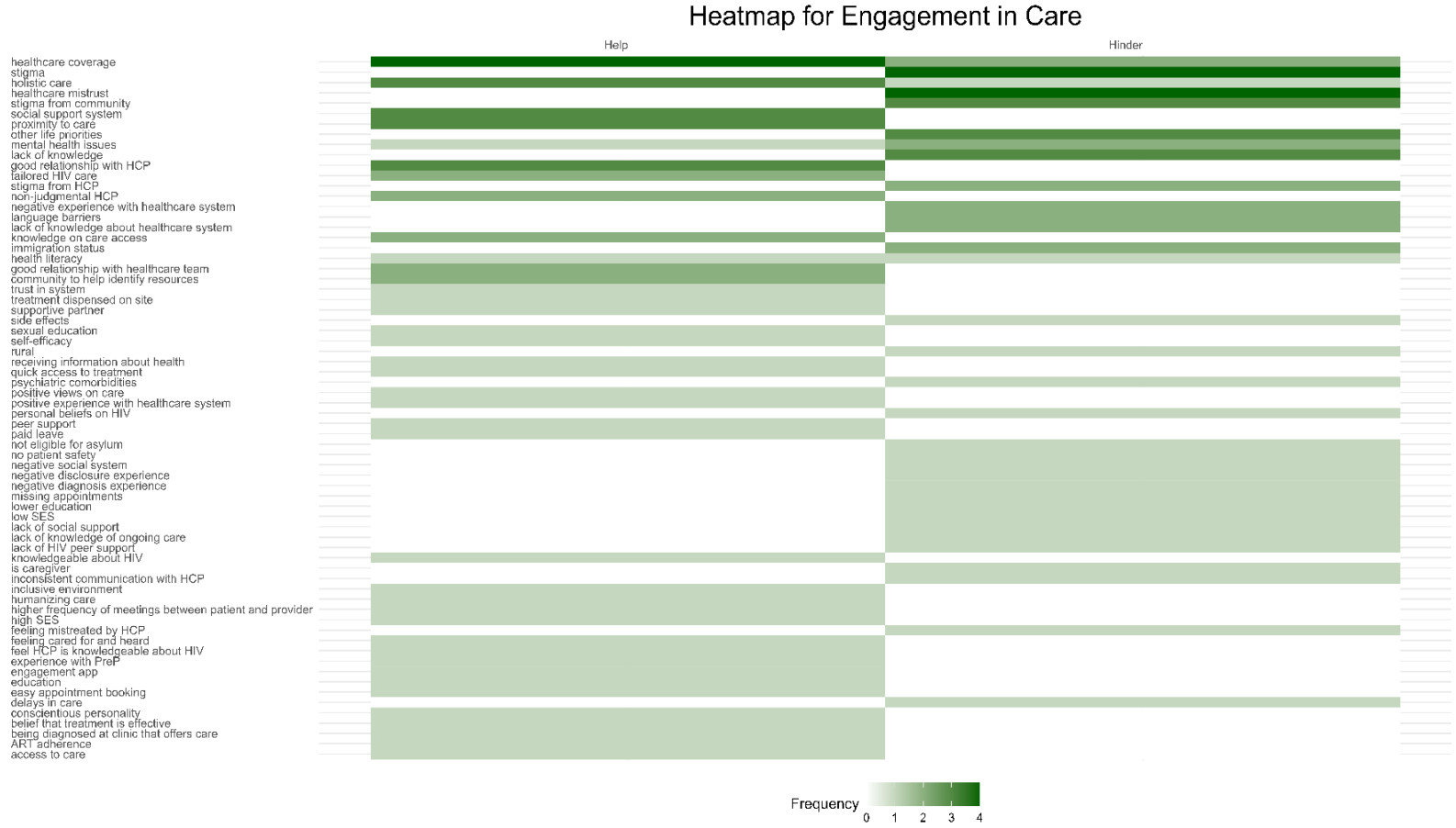


Figure 0-4: Heatmap of variables helping and hindering engagement in care

ART adherence (Figure 7-5):

Healthcare coverage, side effects, stigma, and being engaged in care and treatment were the most frequently identified variables both helping and hindering ART adherence. Factors such as treatment regime, supportive partner, and social support system were frequently identified as helping adherence. Comorbidities, mental health issues, and drug use were often noted as hindering adherence.

In both outcomes, healthcare coverage stands out as an important factor, being the most frequently mentioned variable for both outcomes. Stigma also appears as a significant factor across both outcomes, highlighting its pervasive impact on HIV care.

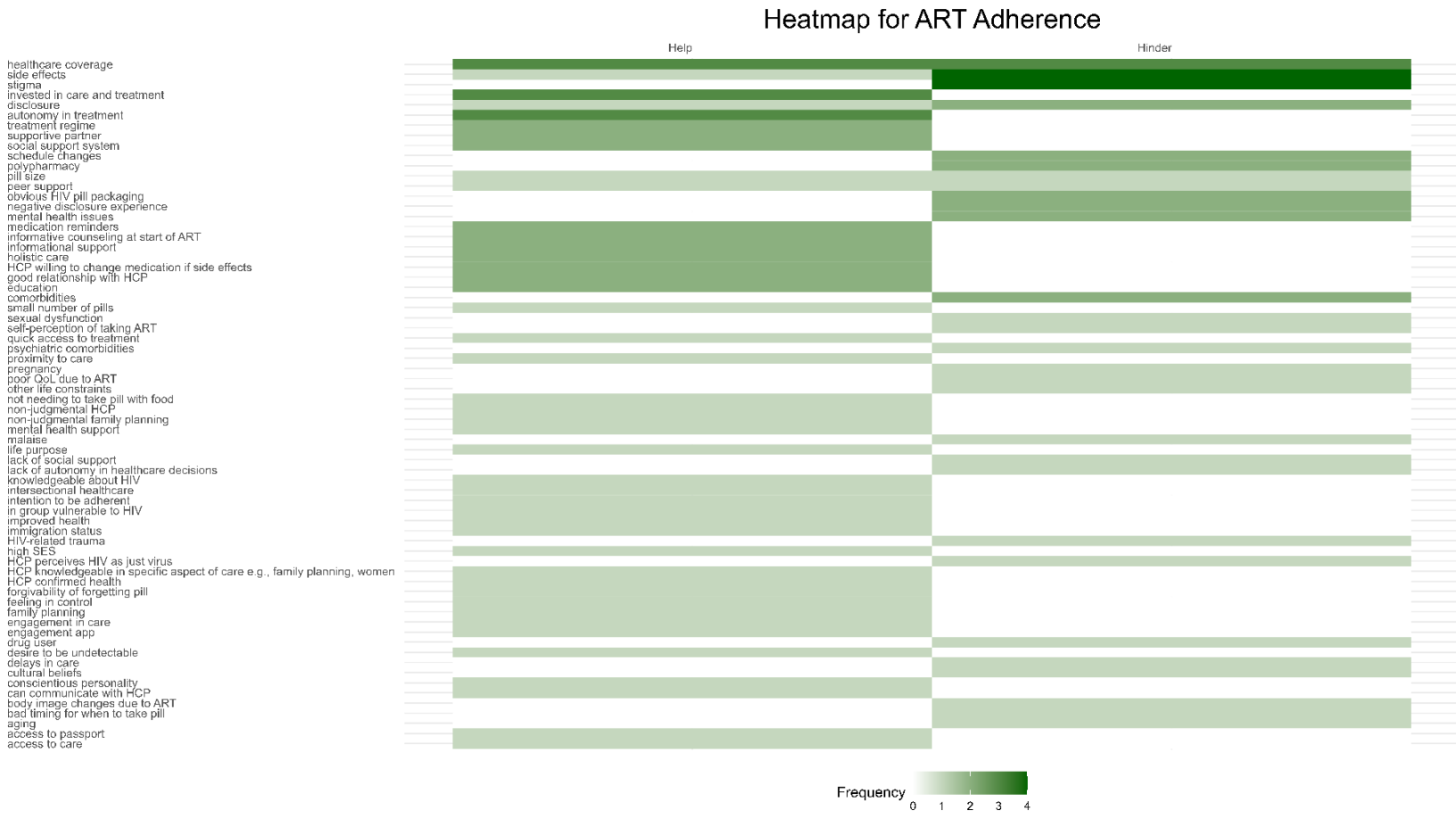


Figure 0-5: Heatmap of variables helping and hindering ART adherence

7.6.7.2 Temporally organized variables:

Figures 7-6 and 7-7 are heatmaps illustrating the frequency of variables organized temporally as 'not in care', 'transition to care', and 'being in care' for engagement in care (Figure 7-6) and ART adherence (Figure 7-7). Darker blue colour indicates higher frequency of identification by domain experts.

Engagement in care (Figure 7-6):

- 'Not in care': healthcare coverage and stigma were most frequently identified.
- 'Transition': healthcare coverage remained the most prominent variable.
- 'In care': holistic care emerged as the most frequently identified variable, followed by social support system.

ART adherence (Figure 7-7):

- 'Not in care': stigma and healthcare coverage were more frequently identified variables.
- 'Transition': Healthcare coverage was overwhelming the most prominent variable.
- 'In care': side effects and disclosure were the most frequently cited factors.

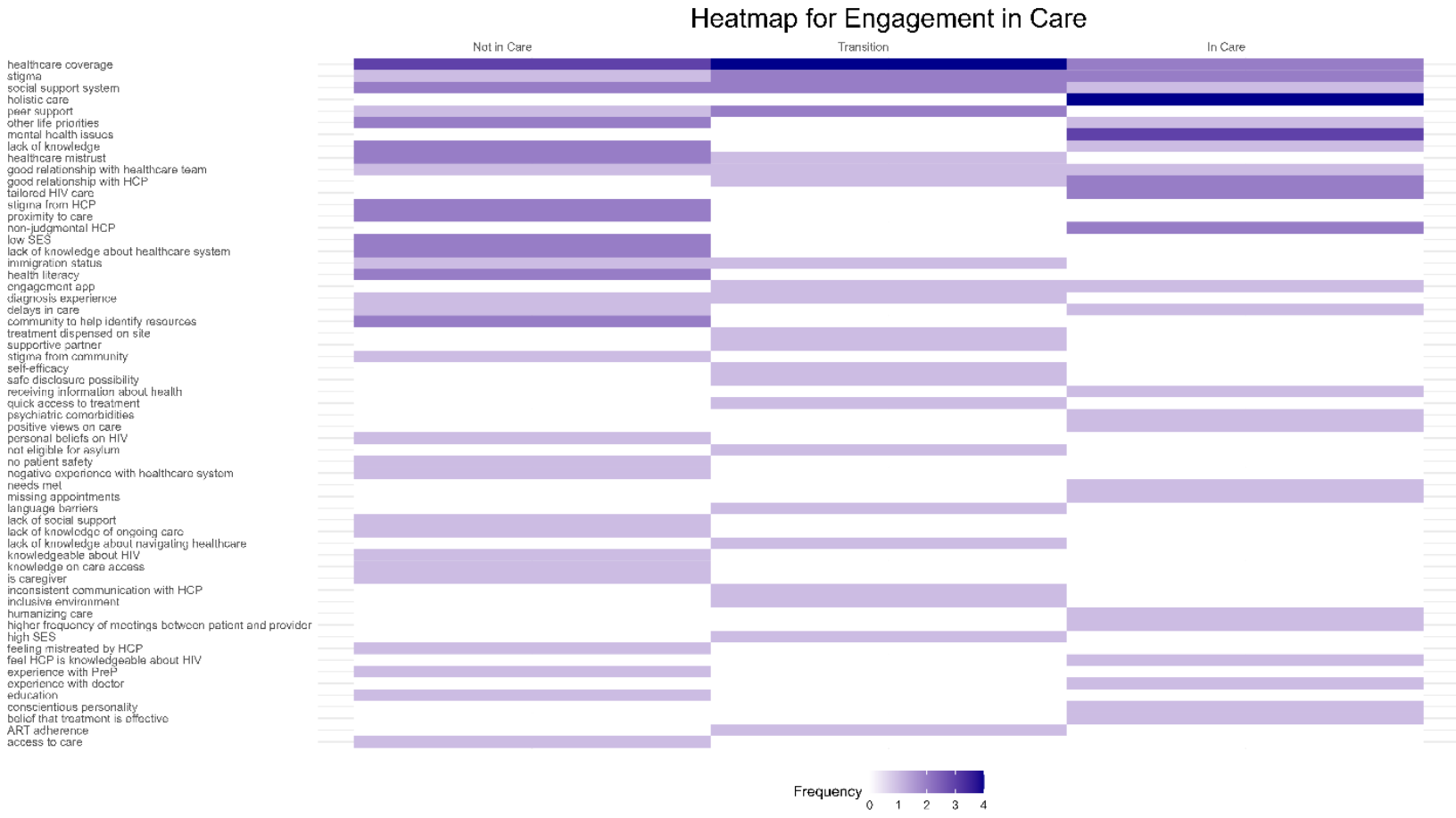


Figure 0-6: Heatmap representing frequency of variables impacting engagement in care from diagnosis to being in care

Across both outcomes, healthcare coverage consistently appeared as a critical factor, especially in the 'not in care' and 'transition' stages. Stigma also played a significant role, particularly in the early stages of care.

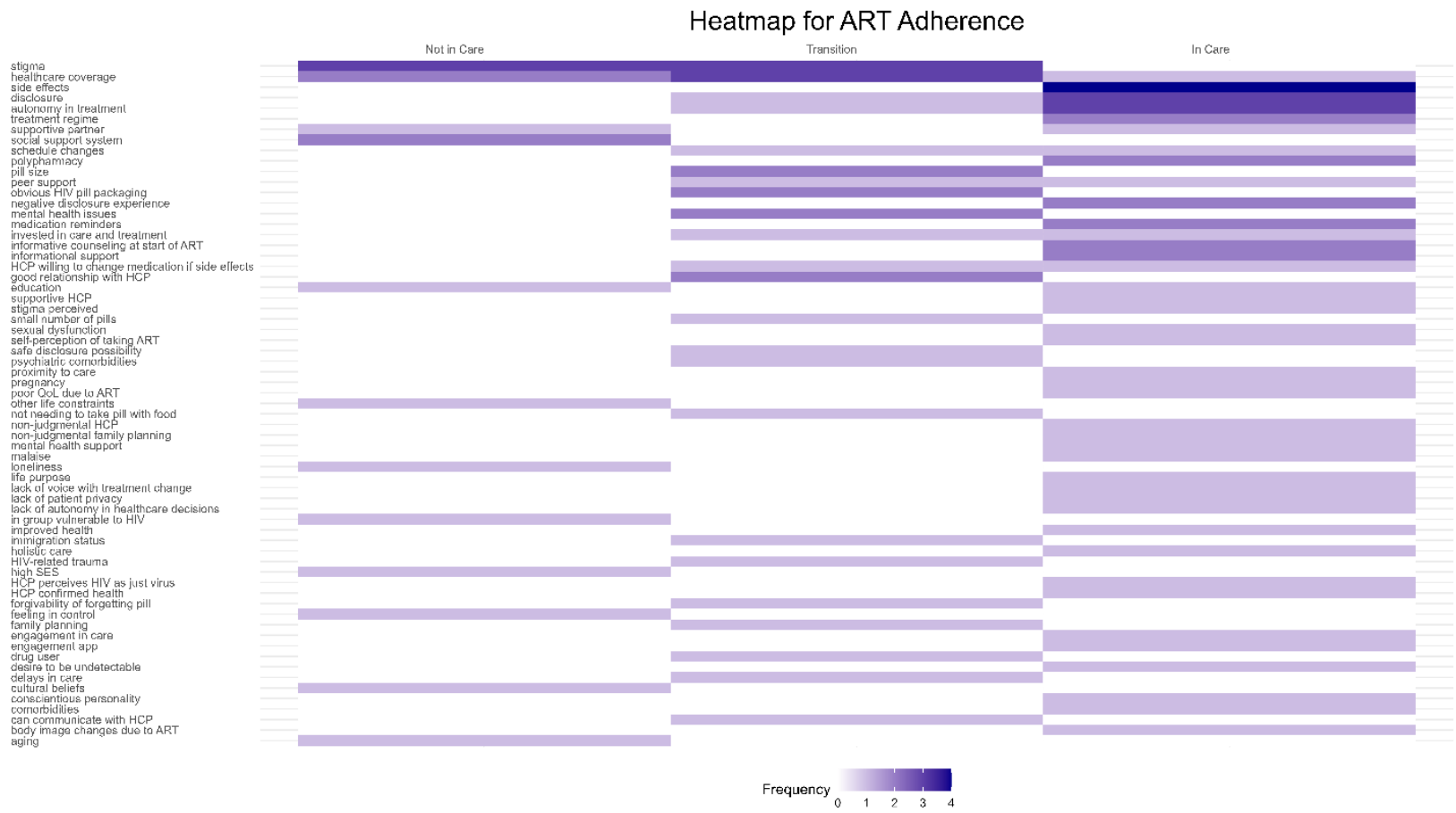


Figure 0-7: Heatmap representing frequency of variables impacting engagement in care from diagnosis to being in care

7.6.8 Updated DAG

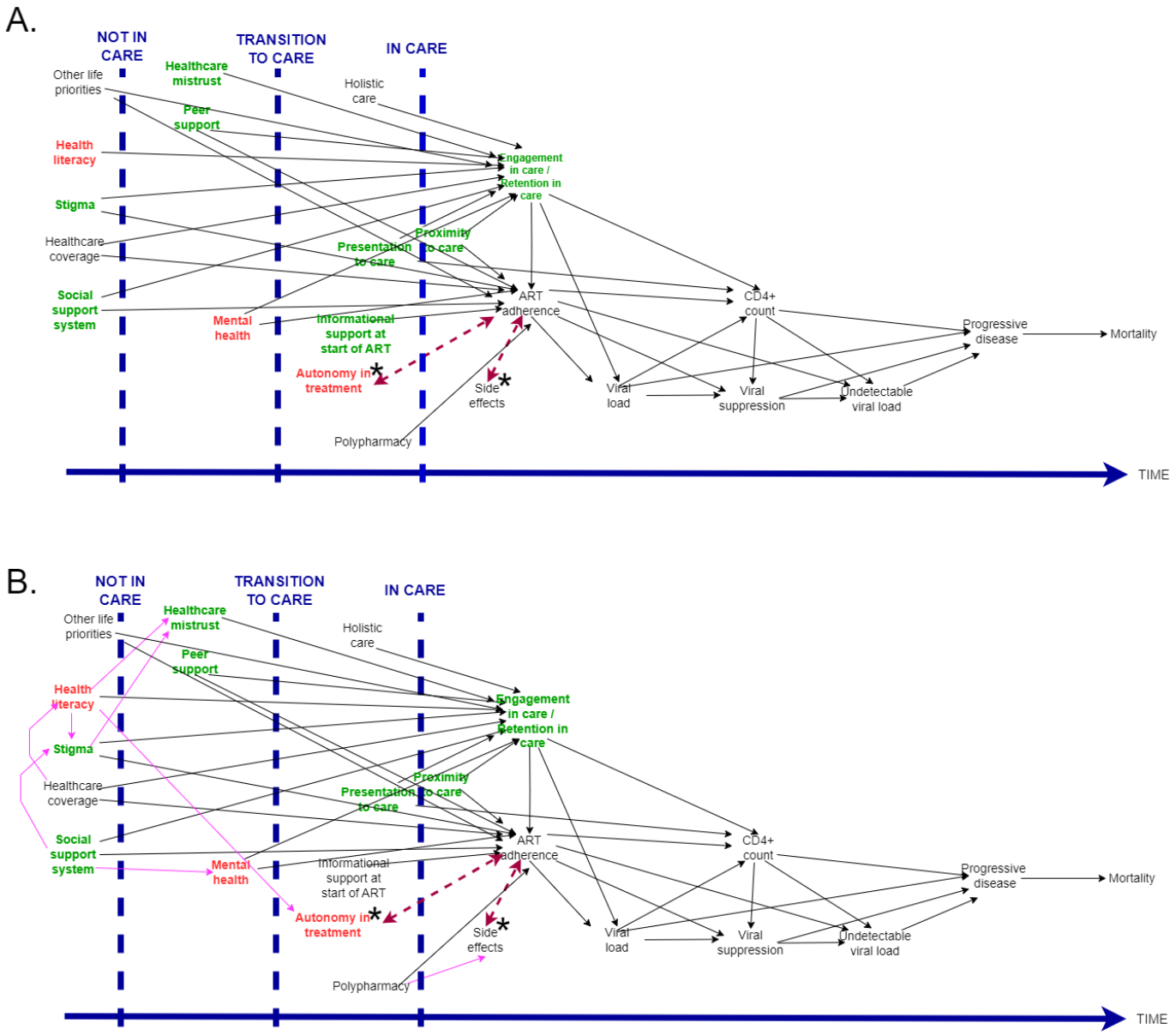


Figure 0-8: Both A and B depict HIV-related outcomes across three temporal stages: 'not in care', 'transition to care' and 'in care', as indicated by the timeline arrow at the bottom.

A) Initial draft of the DAG created with domain experts after the elicitation of variables then temporally organized. Asterisk and the dashed arrows indicates bidirectional relationships, e.g., side effects result from ART adherence, but may subsequently affect ART adherence. Similarly, greater autonomy in treatment may increase engagement and adherence, while adherence may lead to healthcare providers granting more treatment autonomy.*

B) Final DAG created by the research team during the consolidation phase. Light pink arrows represent edges added by the research team based on literature review. Variables (nodes) are colour-coded according to WHO aspects of health: physical (black), social (green), and mental (red).

7.7 Discussion

Our study initially aimed to assess the feasibility of updating a baseline DAG, derived from a scoping review of HIV literature, with domain experts. During pilot testing, we experienced some challenges, including visual clutter, cognitive burden, and time constraints for domain experts. In response, we iteratively developed and assessed the feasibility of an approach incorporating early domain expert feedback aimed at reducing task complexity and time commitment. This involved streamlining the introductory content, focusing primarily on the clinical context, while minimizing DAG-specific terminology.

Generally, domain experts felt positively about the proposed approach. Citing the structured format, use of open-ended questions, clarity of instructions, and virtual format as positive aspects of the approach. By simplifying the process, it allowed for more efficient and focused knowledge elicitation, as indicated by domain expert feedback. The stepwise approach, particularly the categorization of variables into those that "help" and "hinder", provided a structured yet intuitive framework that encouraged and supported brainstorming. Our approach aimed to overcome some of the issues of complexity and time constraints reported in previous studies, such as Rodrigues et al., [24]. While both approaches were conducted online, our one-on-one interviews with individual domain experts enabled completion within one hour, taking an average of 43.5 minutes; compared to their workshop took 2.5 hours. Our approach also required less resources, needing one facilitator compared to three. Additionally, we opted for one-on-one interviews with domain experts in an attempt to facilitate easier recruitment of domain experts, as involving multiple domain experts in a single session can be difficult to schedule.

A key innovation of this approach is the temporal organization of variables. By aligning with the Bradford Hill criteria of temporality [28], it inherently supports the logical flow of causal relationships. This organization helped domain experts to think systematically about the progression of HIV care, from diagnosis through to being in care with treatment. It enabled the highlighting of important factors, predating linkage to care, with important downstream consequences, effectively identifying early indicators of poor outcomes downstream.

The focus on identifying early indicators makes this approach particularly valuable to adaptive trial designs, where selection of appropriate early surrogate outcomes is crucial. By identifying variables appearing early in the care continuum that are causally linked to later downstream

outcomes, researchers may be able to better understand the relevant causal pathways amongst the exposure, surrogate outcome, and primary endpoint, facilitating selection of appropriate surrogate outcomes. This is especially relevant for trials employing adaptive randomization, where early indicators of efficacy are imperative [29].

The baseline DAG, informed by a scoping review of the HIV literature, demonstrated a focus on physical health outcomes, followed by social health, with limited attention to mental health (Figure 4). Interestingly, domain experts' contributions reversed this trend, emphasizing social aspects of health such as stigma and social support systems, systemic factors such as healthcare coverage and holistic care provision, and mental health outcomes such as mental health issues and support and autonomy in treatment. This resulted in a more comprehensive, holistic view of health in the updated DAG.

7.7.1 Limitations

The primary limitation of this study is the small sample size and use of convenience sampling. We included only seven domain experts– recruited from our professional and clinical network. This may raise concerns about the generalizability of our findings. However, the study's primary purpose was to assess the feasibility of a novel DAG development approach, not to evaluate the effect of an intervention.

We ceased recruitment after the seventh domain expert due to time constraints, however, we observed that new information was producing minimal changes to the DAG [30] and we were receiving no new domain expert feedback on the DAG development approach. While this suggests we captured a range of perspectives, a larger more diverse sample may have yielded additional insights. As the secondary study objective was to update a baseline DAG of HIV-related patient outcomes, we focused recruitment on those with expertise in this content. Thus, it is possible that the limited feedback on the approach may be explained in part by the limited knowledge of causal inference and DAGs by the domain experts in our sample. Only one of the interviewed domain experts had extensive experience with causal mapping, while others at most had some exposure through coursework. Re-evaluating the approach with domain experts with more experience with causal approaches may prompt more useful feedback to further refine the approach.

The underrepresentation of patients and healthcare providers in our sample is notable limitation. A more balanced representation of patients and other healthcare providers could have provided a more comprehensive view of HIV care. One domain expert noted the challenge of considering her expertise in relation to a patient's perspective. Additionally, since participants were recruited from our professional network, there may be some social desirability bias at play, potentially leading to overly positive feedback on the approach.

Future studies could address these limitations by including a larger, more diverse sample of domain experts, particularly patients. Additionally, comparing this approach with traditional DAG development methods to validate its effectiveness and efficiency.

7.8 Conclusion

Our study evaluated the feasibility of proposed approach DAG development with domain experts. Domain expert feedback was positive, citing the structured flexible approach which supported brainstorming and knowledge elicitation. This feedback suggests that the approach may have addressed some of the challenges encountered during pilot testing, including visual clutter, cognitive burden, and time constraints for domain experts. This novel approach simplifies the process while still effectively capturing causal relationships in complex healthcare scenarios. By streamlining the process, we have made it more manageable for domain experts in a shorter period time, and potentially easier to recruit domain experts for future studies.

Our approach attempts to make DAG development with domain experts more practical via:

1. A stepwise approach that encourages brainstorming through categorization of variables into those that "help" and "hinder".
2. Temporal organization of variables, aligning with the Bradford Hill criteria of temporality, supporting the logical flow of causal relationships and identifies early indicators of downstream outcomes.
3. One-on-one semi-structured interviews that enable completion within an hour, requiring fewer resources compared to other workshops.

In the context of HIV care, this method has highlighted the interconnectedness of clinical, social, and structural factors impacting patient outcomes. Notably, domain experts emphasized social

and mental health aspects, as well as systemic structural factors, resulting in a more holistic view of health than initially derived from the scoping review.

Furthermore, this bridges the gap between theoretical causal modeling and practical knowledge elicitation from domain experts, offering a valuable tool for researchers and clinicians. Providing relevance to adaptive trial designs, facilitating the selection of appropriate surrogate outcomes of primary endpoints.

Overall, this novel approach to DAG development offers a structured and efficient method for knowledge elicitation from domain experts in applied health research. By simplifying the process, this approach provides a feasible method of drawing DAGs with domain experts in complex health settings. It is a particularly relevant approach in scenarios with many interrelated variables.

7.9 References

1. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(Volume 34, 2013):61-75.
2. Long S, Schuster T, Piché A, Research S. Can large language models build causal graphs? arXiv preprint arXiv:230305279. 2023.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
4. VanderWeele TJ, Hernan MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiology. 2008;19(5):720-8.
5. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol. 2022;142:264-7.
6. Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. Int J Epidemiol. 2020;49(1):322-9.
7. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021;50(2):620-32.
8. Newington L, Metcalfe A. Factors influencing recruitment to research: qualitative study of the experiences and perceptions of research teams. BMC medical research methodology. 2014;14:1-11.
9. Christensen R, Ciani O, Manyara AM, Taylor RS. Surrogate endpoints: a key concept in clinical epidemiology. J Clin Epidemiol. 2024;167:111242.
10. Vanderweele TJ. Surrogate measures and consistent surrogates. Biometrics. 2013;69(3):561-9.
11. Berry S, Connor J, Lewis R. The Platform Trial: An efficient strategy for evaluating multiple treatments. JAMA. 2015;313(16):1619-20.
12. DeMets DL, Psaty BM, Fleming TR. When Can Intermediate Outcomes Be Used as Surrogate Outcomes? JAMA. 2020;323(12):1184-5.
13. Park JJH, Harari O, Dron L, Lester RT, Thorlund K, Mills EJ. An overview of platform trials with a checklist for clinical readers. J Clin Epidemiol. 2020;125:1-8.
14. Parast L, Cai T, Tian L. Using a surrogate with heterogeneous utility to test for a treatment effect. Stat Med. 2023;42(1):68-88.
15. Moore TJ. Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster: Simon & Schuster; 1995.
16. Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and Morbidity in Patients Receiving Encainide, Flecainide, or Placebo. New England Journal of Medicine. 1991;324(12):781-8.
17. Fleming TR, DeMets DL. Surrogate End Points in Clinical Trials: Are We Being Misled? Annals of Internal Medicine. 1996;125(7):605-13.
18. Beauchemin C, Johnston JB, Lapierre ME, Aissa F, Lachaine J. Relationship between progression-free survival and overall survival in chronic lymphocytic leukemia: a literature-based analysis. Curr Oncol. 2015;22(3):e148-56.
19. Cortazar P, Zhang JJ, Sridhara R, Justice RL, Pazdur R. Relationship between OS and PFS in metastatic breast cancer (MBC): Review of FDA submission data. Journal of Clinical Oncology. 2011;29(15_suppl):1035-.

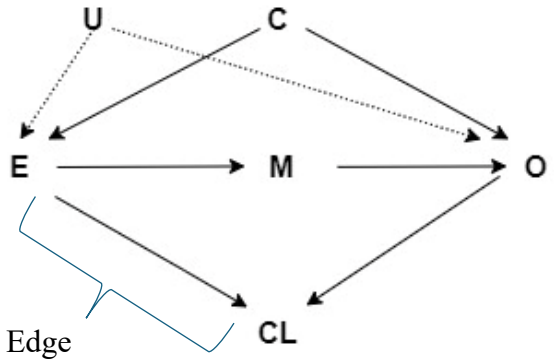
20. Fleming TR. Surrogate markers in aids and cancer trials. Statistics in Medicine. 1994;13(13-14):1423-35.
21. VanderWeele TJ, Robins JM. Signed directed acyclic graphs for causal inference. J R Stat Soc Series B Stat Methodol. 2010;72(1):111-27.
22. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. Am J Epidemiol. 2007;166(9):1096-104.
23. VanderWeele TJ, Robins JM. Four types of effect modification: A classification based on directed acyclic graphs. Epidemiology. 2007;18:561-8.
24. Rodrigues D, Kreif N, Lawrence-Jones A, Barahona M, Mayer E. Reflection on modern methods: constructing directed acyclic graphs (DAGs) with domain experts for health services research. International Journal of Epidemiology. 2022;51(4):1339-48.
25. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for Designing and Evaluating Feasibility Pilot Studies. Med Care. 2022;60(1):95-103.
26. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med. 2017;7(3):93-9.
27. Ellison GTH. Might Temporal Logic Improve the Specification of Directed Acyclic Graphs (DAGs)? Journal of Statistics and Data Science Education. 2021;29(2):202-13.
28. Hill AB. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? Proc R Soc Med. 1965;58(5):295-300.
29. Park JJ, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. Clin Epidemiol. 2018;10:343-51.
30. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Social Science & Medicine. 2022;292:114523.
31. Greenland S, Pearl J. Causal Diagrams. Encyclopedia of Epidemiology2006.
32. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
33. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-710.
34. Greenland S, Brumback B. An overview of relations among causal modelling methods. International Journal of Epidemiology. 2002;31:1030-7.
35. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12(3):313-20.
36. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequesite for confounding evaluation: An application to birth defects epidemiology. American Journal of Epidemiology. 2002;155(2):176-87.
37. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.

7.10 Supplemental Appendix A: Overview of Directed Acyclic Graphs (DAGs)

Directed Acyclic Graphs (DAGs) are graphical models that illustrate the qualitative assumptions made by causal models, not captured by conventional statistical models [13, 14]. In epidemiological research, *DAGs* have a variety of purposes: (1) representing the causal relationships amongst variables [14, 83, 84]; (2) identifying the potential confounding variables which need to be controlled for in order to estimate causal effects [14, 83, 85, 86]; and more recently (3) as a means of classifying the types of causal relationships that may give rise to selection bias [87].

A DAG is composed of variables (nodes), both measured and unmeasured, and their connections displayed via arrows (edges) [13, 87]. The absence of an arrow between variables indicates the lack of a causal relationship between the variables. If the edge has an arrowhead, the variable at the tail is the parent node and the variable at the arrowhead is the child [13]. An *edge* is any line (with an arrowhead or not) that connects two variables [84]. As the name suggests, *Directed Acyclic Graphs* are *acyclic* because a variable cannot be the cause of itself, either directly or indirectly through another variable i.e., there are no feedback loops [87]. Additionally, in DAGs, causal pathways are represented with *directed paths* from the starting variable to the final variable; thus, a variable is the cause of its *descendants* and an *effect* of its *ancestors* [13]. See Box 1 for more details.

Box 1: Terminology of DAG



Types of variables that can be found in a DAG.

1. Exposure/Treatment/Intervention (E): main cause.
2. Outcome (O): main effect.
3. Mediator (M): caused by E, and subsequently causes O.
4. Confounder (C): common cause of E and O
5. Collider (CL): common effect of any two variables.
6. Unmeasured confounder (U): unmeasured common cause of E and O.

7.11 Supplemental Appendix B: Interview guide

Outcome/Topic assessed	**Question**
General introduction	How would you describe your **experience** taking part in this causal mapping activity?
Feasibility	What did you think of the overall **length** of the causal mapping sessions? What did you think of the **instructions** provided on what you needed to do to create a causal diagram? - How could they be improved? - What aspects of the instructions would you change? If any.
Acceptability	What **challenges**, positive and negative, did you face while creating the causal diagram (with or without ChatGPT)? What did you **like** about using causal mapping? What did you **dislike** about using causal mapping?
For researchers only	What was your experience with causal mapping prior to participating in this study? Would you use causal mapping in your research? Why or why not?
Final remarks	Any further comments?

7. CHAPTER 8: DISCUSSION

8.1 Overview

This dissertation explored the challenges and opportunities of integrating public data and expert knowledge in primary care research through a causal inference-informed approach, focusing particularly on the creation of DAGs to inform health interventions. To accomplish this aim, I conducted a series of four studies that identified shortcomings in implementation and presented practical solutions for use in primary care research. This work identified the gaps and limitations of different methods of developing causal models e.g., DAGs and presents a series of approaches to integrating diverse data sources and methods. Taken together, the findings emphasize the importance of a practical approach to integrating public data and domain expertise in causal modelling for primary care research.

8.2 Summary of key dissertation findings

Key dissertation findings:

1. Canadian COVID-19 data reported to the general public by government and news websites demonstrated limited utility for informing health policy due to varying case definitions, heterogeneous and dynamic testing criteria, lack appropriate standardization accounting for dynamics, sizes, and characteristics of populations being tested.
2. Accuracy of the large language model, Generative Pre-trained Transformer-3 (GPT-3) in confirming the presence and direction of an edge between two variables in a medical-context directed acyclic graph (DAG) varies by prompt language, verb describing relationship between variables, and specificity of variable description.
3. This scoping review was designed to broadly map the HIV literature to develop a highly comprehensive DAG illustrating causal relationships among HIV-related outcomes. However, the resulting DAG did not pass face validity, missing known social, mental, and structural influences.
4. Domain experts may add nodes or edges to a DAG that were not identified from a literature synthesis, resulting in a more comprehensive DAG. Their involvement in this process is highly valuable but is facilitated by a practical and straightforward approach cognizant of time constraints and cognitive burdens. Recognizing that building a DAG at

once was not a feasible approach, this process was designed to be iterative and more manageable, allowing experts to contribute their knowledge in stages.

8.3 Main results

The overall aim of this thesis was to assess the current methodological shortcomings in integrating public data and expert knowledge in primary care research. Below, I discuss the main findings related to my four specific objectives.

8.3.1 Objective 1: to evaluate epidemiological reporting standards using COVID-19 as a case study, with a focus on assessing the limitations of causal and actionable interpretations of reported data.

The first objective was to evaluate whether Canadian COVID-19 data abided by established epidemiological reporting standards, focusing on the limitations of causal and actionable interpretations of this reported data. This objective was addressed in Chapter 4 (Manuscript 1) with a longitudinal critical appraisal of the Canadian COVID-19 data reported by governmental agencies and news outlets. This study examined the governmental and news outlet COVID-19 data reporting between April 2020 and August 2021 to examine whether they were reported with appropriate denominators, data sources, and accounted for age, sex, and ethnicity.

This study found that Canadian COVID-19 data reporting exhibited varying case definitions, heterogeneous testing criteria, and lacked or used inappropriate denominators for standardization [143]. Although most provinces and territories reported data on sex and age, none reported any statistics on race or ethnicity from the beginning to the end of the observation period. These findings highlight a gap of implementing long known and well-established epidemiological principles in actual practice. Additionally, they show continued gaps and challenges in collecting sufficient and complete data on population characteristics such as race and ethnicity.

Before, during, and after the COVID-19, racial and ethnic disparities in health care have existed, which have been extensively documented over decades [144, 145]. Numerous experts and studies have emphasized that addressing racial and ethnic disparities in health outcomes and healthcare access is crucial for overall quality improvement in the healthcare system [146-148]. In fact, this manuscript (Chapter 4) was cited heavily by research groups calling for change of such practice [149]. Canadian health agencies such as the Canadian Institute for Health Information (CIHI) agree, per their website:

"In Canada, differences across population subgroups (or inequalities) are significant for a range of health care indicators and are generally persisting or worsening over time. Health systems with a commitment to health equity recognize that measurement matters. The collection, measurement and reporting of socio-demographic data enables health systems to identify inequalities in care across populations, inform meaningful strategies and monitor progress in improving care for all patients" [150].

In fact in May 2020 in support of more race-based data collection and health reporting, CIHI proposed an interim race data collection approach, SPARK study (Screening for Poverty And Related social determinants and intervening to improve Knowledge of and links to resources) to facilitate collection of higher quality race-based data [151].

Which makes it all the more surprising and disappointing that during the COVID-19 pandemic, no race and ethnicity data was reported on any of the governmental websites of any province or territory between April 2020 and August 2021 [143]. Widespread, reliable, and consistent data on racial and ethnic characteristics within a population is essential for identifying, understanding, and addressing health disparities [144]. Such information is crucial for identifying the scope and nature of disparities, directing targeted quality improvement efforts, and tracking progress over time.

Despite public perception that the federal government and private sector collect large amounts of data, racial and ethnic data in the healthcare system is limited [144]. As recent as June 2023, the Canadian Medical Association Journal (CMAJ) published a commentary by Pinto et al., calling for the collection of data on race and Indigenous identity across Canadian jurisdictions [152]. Health policies are often developed using data from research studies, surveillance surveys, and other sources that provide insights into health outcomes and trends.

When decision-making on health policies is based on data that lacks information on race or ethnicity or is not representative of the target population, systematic biases or spurious associations could arise. This deficiency in data can lead to several issues: first, health policies may not account for significant differences in health outcomes and risks among various racial or ethnic groups, potentially perpetuating existing health disparities [144]. For example, COVID-19 disproportionately affected visible minorities [153], and this lack of data may have led to interventions not being appropriately targeted. Additionally, the lack of race or ethnicity data can result in misleading conclusions about the effectiveness of health interventions or policy, as the

results may not reflect various in impact across different groups. This can lead to inequitable resource allocation, where some populations receive insufficient attention or support [154].

8.3.2 Objective 2: to assess the potential of large language models in building directed acyclic graphs leveraging the vast corpus of public data for primary care research.

The second objective was to evaluate whether LLMs could help researchers build DAGs in a medical context. This objective was addressed in Chapter 5 (Manuscript 2) with an empirical investigation of the utility of GPT-3 in confirming the presence and direction of edges between two variables in DAGs illustrating well-known exposure-outcome effects in the medical literature. The accuracy of GPT-3 in confirming edges was dependent on the language used in the prompts, e.g., the medical authority used to prompt the statements, the linkage verb denoting the relationship between the two variables, and the specificity of language describing the variables [155]. This finding brings both promise in LLMs' ability to complement DAG construction and concerns about reproducibility.

This study found that GPT-3 was sensitive to prompts [155]. Even when a LLM is queried with the exact same prompt, it may provide different responses. This is due to the *temperature* of the LLM, which is a parameter that influences the model's output – making it more random or more focused and deterministic [96]. Temperature can be set to any value between 0 and 1; whereby the lowest temperature 0 will always produce the same output for a given prompt and a temperature of 1 will deliver inconsistent results. Thus, depending on its setting, the response of a LLM to a query can vary, which brings concerns for the reproducibility of its findings. Additionally, open-source LLMs such as GPT-3 and others are constantly undergoing updates and improvements, thus, even within given periods of time, it is entirely possible that not the exact same version of the model is being queried. For example, the version of the LLM (GPT-3) we conducted the study on is no longer available as it has been replaced by GPT-4 and GPT-4o. Though complete reproducibility may not be possible, for transparency purposes, it is recommended to record the time, date, and specific model of the LLM when queried.

Since this manuscript was published and examined GPT-3 in November 2022, six more LLMs have been released: OpenAI's GPT-4 (March 2023) [156] , Anthropic's Claude (March 2023) [93], Google Deepmind's Gemini (December 2023) [157], Mistral's Mistral 72B (September 2023) [158], Meta's LLaMA 2 (July 2023) [92], and Cohere's Command R (April 2024) [159].

With each new release, LLMs are becoming better and better at generating text and images. As their performance improves, LLMs are also becoming cheaper to use, for example, a series of system-wide optimizations of ChatGPT have reduced its cost by 90% since its initial release in December 2022 [160]. Thus, it is possible that some of the concerns raised in this chapter may be eventually resolved through improvements and refinements to LLMs.

Though LLMs have made remarkable strides in the past year and have become capable of non-trivial tasks such as engaging in human-like conversation, generating amazing images based on prompts (DALL-E) [161], and even creating websites or helping doing one's taxes (GPT-4o) [162]. Despite this progress, as of this evening, August 10, 2024, at 12:49 AM, when GPT-40 was given the following prompt:

"I would like you to draw a directed acyclic graph. I will give you a health context with exposure, outcome, and variables. Please build it informed by literature.

The health context is HIV care and management.

Exposure: being in HIV care

Outcome: antiretroviral therapy adherence.

Variables: engagement in care, viral load, CD4+ count, SES, proximity to care, language spoken"

Figure 8-1 displays the output provided by ChatGPT-4o, which does not actually illustrate a DAG as there are cycles, and many of the node labels are illegible and nonsensical. To test whether this strange output was limited to the HIV context or not, I also queried asking for a DAG with the following criteria: exposure = cigarette smoking, outcome = lung cancer; variables = asthma, age, family history of lung cancer, SES. Figure 8-2 shows the output, which was similar, but slightly better, as the variables were legible unlike in Figure 8-1.

The likely explanation of ChatGPT-4o's difficulty to produce correctly spelled text in images is because the underlying technology behind image and text generation are different. DALL-E, the image generator model behind ChatGPT is a text-to-image diffusion model developed by OpenAI [161]. It reconstructs images based on what it has seen in training and typically text in an image takes up a very small portion of the total image. Consequently, image generators often learn patterns that obscure or distort these pixels [163].

Thes examples illustrate that there is still progress to made by LLMs, in this case, GPT-4o (using DALL-E for image generation) and although they are quite adept at other generative tasks, they are not yet ready to directly draw DAGs in this context.

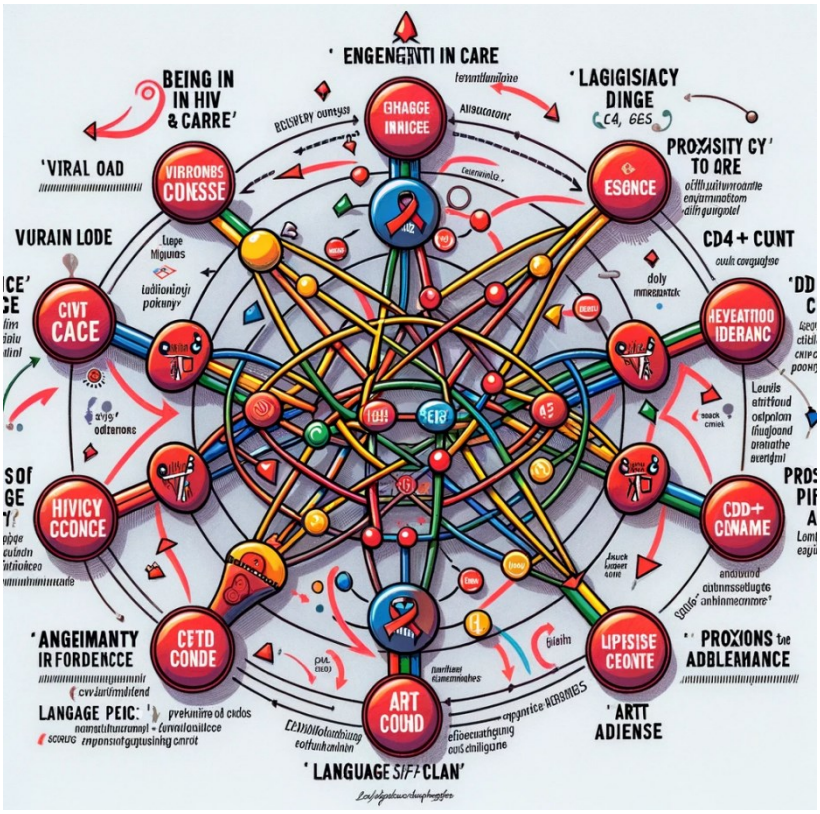


Figure 7-2: ChatGPT-4o output from prompt requesting a directed acyclic graph illustrating relationship between HIV care and ART adherence created on August 10, 2024.

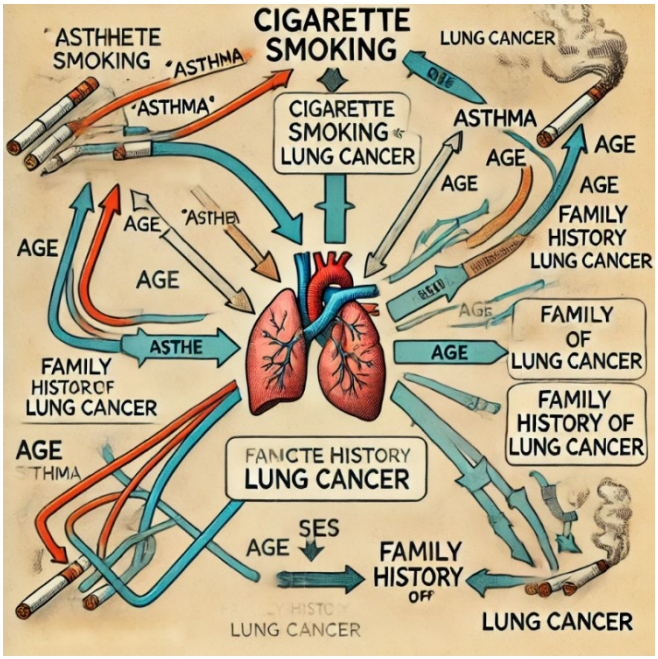


Figure 7-1: ChatGPT-4o output from prompt requesting a directed acyclic graph illustrating relationship between cigarette smoking and lung cancer created on August 10, 2024.

8.3.3 Objective 3: to establish a causal mapping approach of the HIV literature to identify frequently reported HIV-related patient outcomes with the goal of constructing a comprehensive DAG of HIV outcomes reported in the literature.

The third objective was to map the HIV literature to identify HIV-related individual-level outcomes used in HIV studies, with the goal of constructing a DAG illustrating this context. Chapter 6 (Manuscript 3) addressed this with a scoping review [164] of HIV quantitative, mixed methods studies, or RCTs conducted in high-income countries [165]. This scoping review identified a predominance of studies reporting physical and clinical outcomes, some social health-related outcomes, and few mental health outcomes. Consequently, the developed DAG primarily included physical and clinical outcomes, lacking social, mental, and structural factors known (or assumed) to affect health outcomes in this context.

In the case of this review, the objective required a broad research question and search strategy which yielded 9443 unique title and abstracts for screening. After applying focused eligibility criteria, 681 full-text articles were included - an exceptionally large number of included studies. This volume of studies at each stage presented significant time commitments for two reviewers to screen and extract data from. Scoping reviews are generally laborious, resource-intensive, and time-consuming [166], potentially taking up to a year to complete [167]. Such extensive literature data to handle necessitates an automated approach for more efficient processing.

Machine learning algorithms, including LLMs, offer opportunities to automate various review stages [168] such as selection of relevant studies [169], abstract screening [170] and data extraction [171]. Progress is this area is evident, with numerous online tutorials describing AI or ChatGPT use in automating systematic reviews. Tools are also being developed; for instance, Orel et al., [172] published LiteRev, an automated literature review tool using machine learning and natural language processing to streamline literature reviews and provide quick and in-depth overviews of any topics of interest.

8.3.4 Objective 4: to develop and evaluate the feasibility of a novel approach for DAG development with domain experts.

The fourth objective was to design and assess the feasibility of an innovative method for constructing DAGs in collaboration with domain experts. This objective was addressed in Chapter 7 (Manuscript 4) with a feasibility study which iteratively refined the DAG development

approach with domain expert feedback and updated the DAG created in Manuscript 3. The original DAG development approach aimed to have domain experts directly update the DAG created in Manuscript 3, by adding, removing, or adjusting the placement of nodes and edges, with new nodes being added to maintain temporal order. Initially, it was believed that having domain experts update or adapt a baseline DAG, focusing on specific relevant outcomes, would require less time and effort than building a DAG from scratch.

During pilot testing, it quickly became apparent that domain experts could add many nodes and edges, overwhelming the limited screen space available during Zoom sessions and creating a very cluttered and difficult-to-read DAG. Despite the cluttered screen and DAG, domain experts were still able to add more nodes and edges, leading to very length sessions. Initial domain expert feedback indicated that there was a lot of information to keep track of, making the task very mentally taxing. This overwhelming and laborious nature of DAG creation is rarely, if ever, mentioned in the literature [173].

Furthermore, for every new node (v_{n+1}) added to a DAG with n nodes, there are n possible new edges directed from the existing n nodes to the new node v_{n+1}, while not introducing acyclicity [83]. Thus, as the DAG expands, the cognitive load increases. Adding nodes and edges necessitates continuous monitoring for the introduction of potential cycles and re-evaluating the graph's structure, which becomes even more complex with larger graphs. This ongoing need for vigilance and detailed mental effort required to ensure the graph's consistency makes the process cognitively demanding.

Despite being a longstanding tool in epidemiology, use of DAGs in primary care and other applied health research remains scant [173]. A review of health research from 1999 to 2017 found that when DAGs were said to be used in health research studies, there was substantial variation in use and reporting of important details such as target estimands of interest and the implied adjustment sets, with some studies not even reporting their actual DAG [173]. Limited use of DAGs in this research context may be explained by limited availability of practical guidance and supporting materials for implementation that are accessible to primary care health researchers that may have limited knowledge of causal inference.

This suggests that more robust, practical guidelines with straightforward instructions may help increase the use of DAGs in the primary care research context [1]. Though some efforts have

been made to facilitate its use, such as the development of user-friendly software tools for DAG construction such as 'DAGitty' [174] and the publication of guides for researchers new to DAGs [175]. However, there is still need for further research into practical strategies for integrating DAG construction into the workflow of primary care researchers, particularly those working in multidisciplinary teams [176].

8.4 Implications for research, practice and policy

This research advances the science and practice of primary care research by placing causal inference methods at the heart of addressing complex research questions. By making causal inference methods, particularly DAGs, more accessible, this work facilitates a more robust, yet practical approach to health research and policy development. This study contributes to the field in several important ways:

1. It provides guidance for clinicians and researchers on integrating different types of data in the creation of DAGs, including literature, domain expertise, and emerging data sources.
2. It demonstrates how DAGs can be used in practice to address issues with outcome selection, leveraging available knowledge and minimizing biases.
3. This research highlights the importance of causal modelling, especially when informing policy decisions, and addresses the challenges of data quality and completeness in evolving situations such as the COVID-19 pandemic.
4. Most importantly, it introduces a series of integrated approaches that incorporate available data, literature, domain expertise with LLMs to facilitate a more efficient and practical approach DAG development.

Thus, this work may facilitate causal modeling through an integrated approach, ultimately leading to better-informed health research practices and more effective policy recommendations.

8.5 Limitations

One limitation of this dissertation is that the recruitment for domain experts in the DAG development study (Chapter 7) was carried out using convenience sampling and the subsequent small sample size. This potential weakness was potentially offset by a commitment to recruit diverse domain experts, in terms of their expertise in HIV and causal inference, gender, and ethnicity. Ultimately, the included domain experts had varying years of experience, and expertise

in terms of HIV knowledge as well as research designs, including 5 PhD researchers (expertise including: qualitative research, statistics, fuzzy cognitive mapping, migrant health, reproductive rights), 1 PhD candidate (HIV sexual health in men who have sex with men), and 1 MSc student (with extensive experience working in communities with vulnerable populations in HIV). A challenge faced in participant recruiting was the balance of expertise in the clinical area of interest, HIV, and knowledge about causal inference methods and DAGs. Though a minority of participants had familiarity with DAGs through coursework exposure, most were not knowledgeable about the topic. Though, the aim of this study was to assess the feasibility of an alternative approach to DAG development. Thus, this diversity in expertise contributes to the understanding of the utility of this approach in researchers new to DAGs.

8.6 Conclusion

The four studies of this dissertation research provide a comprehensive assessment of the methodological challenges of integrating public data and expert knowledge through a causal inference lens in primary care research. By examining theoretical, technological, and practical challenges of implementation, this research provides a practical and integrated approach to DAG creation, combining public data, LLMs, and domain expertise. An important contribution of this work was the inclusion of experts in the knowledge generation process. By iteratively improving the approach with experts, this research not only identified and contextualized gaps in integrating diverse forms of data in practice. Thus, this approach offers pragmatic and comprehensive approach informing causal modelling.

THESIS REFERENCE LIST

1. Iwata H, Wakabayashi T, Kato R. The dawn of directed acyclic graphs in primary care research and education. Journal of General and Family Medicine. 2023;24(4):274-5.
2. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(Volume 34, 2013):61-75.
3. Janiaud P, Axfors C, Saccilotto R, Hemkens L, Schmitt A, Hirt J. COVID-evidence: a living database of trials on interventions for COVID-19. Published online April. 2020;1.
4. Janiaud P, Hemkens LG, Ioannidis JPA. Challenges and Lessons Learned From COVID-19 Trials: Should We Be Doing Clinical Trials Differently? Can J Cardiol. 2021;37(9):1353-64.
5. Pearl J. The Eight Pillars of Causal Wisdom. Los Angeles, California: UCLA; 2017 April 24, 2017.
6. Pearl J. Interpretability and explainability from a causal lens: Institute for Pure & Applied Mathematics 2019.
7. Hernan MA, Robins JM. Causal Inference: What if. Boca Raton: Chapman & Hall/CRC; 2020.
8. Pearl J. Causal inference in statistics: An overview. Statistics Surveys. 2009;3(none).
9. Pearl J. Statistics and causal inference: A review. Sociedad de Estadistica e Investigacion Operativa. 2003;12(2):281-345.
10. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health. 2016;37:17-32.
11. Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016;352:i189.
12. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry DA, editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. NY: Springer-Verlag; 1999.
13. Greenland S, Pearl J. Causal Diagrams. Encyclopedia of Epidemiology2006.
14. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
15. Hernan MA, Robins JM. Causal Inference: What if Boca Raton: Chapman & Hall; 2020.
16. Raghupathi W, Raghupathi V. Big data analytics in healthcare- promise and potential. Health Information Science and Systems. 2014;2(3):1-10.
17. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. Int J Epidemiol. 2021;49(6):2058-64.
18. Hartnett K. To Build Truly Intelligent Machines, Teach Them Cause and Effect. Quanta. 2018.
19. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58.
20. Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed. Jordan M, Lauritzen SL, Lawless JF, Nair V, editors. New York, NY: Springer; 2000.
21. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1:206-15.
22. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in commercial gender classification. Proceedings of Machine Learning Research 2018;8:1-15.

23. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46.
24. Al-Zaiti SS, Alghwiri AA, Hu X, Clermont G, Peace A, Macfarlane P, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). European Heart Journal-Digital Health. 2022;3(2):125-40.
25. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Transl Vis Sci Technol. 2020;9(2):7.
26. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circ Cardiovasc Qual Outcomes. 2020;13(10):e006556.
27. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Trans Intell Syst Technol. 2024;15(3):Article 39.
28. Gutierrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about gpt-3 in-context learning for biomedical ie? think again. arXiv preprint arXiv:220308410. 2022.
29. Zhu K, Wang J, Zhou J, Wang Z, Chen H, Wang Y, et al. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts2023.
30. Tonmoy S, Zaman S, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:240101313. 2024.
31. Wilcox A. Birth weight and perinatal mortality: the effect of maternal smoking. American Journal of Epidemiology. 1993;137(10):1098-104.
32. Martin DJ, Prabhakaran V, Kuhlberg J, Smart A, Isaac WS. Participatory problem formulation for fairer machine learning through community based system dynamics. ICLR 2020. 2020.
33. Cheng H, Stapleton L, Wang R, Bullock P, Chouldechova A, Wu ZS, et al. Soliciting Stakeholders' Fairness Notions in Child Maltreatment predictive systems. arXiv. 2021.
34. Pearl J. The seven tools of causal inference, with reflections on machine learning. Communications of the ACM. 2019;62(3):54-60.
35. Wallace-Wells D. We Had the Vaccine the Whole Time. NYMag. 2020 December 7.
36. Moderna Announces Primary Efficacy Analysis in Phase 3 COVE Study for Its COVID-19 Vaccine

Candidate and Filing Today with U.S. FDA for Emergency Use Authorization [press release]. Cambridge, Massachusetts2020.
37. Van Norman GA. Drugs, Devices, and the FDA: Part 1. JACC: Basic to Translational Science. 2016;1(3):170-9.
38. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. Trends Pharmacol Sci. 2019;40(8):577-91.
39. Lo AW, Siah KW, Wong CH. Machine Learning with Statistical Imputation for Predicting Drug Approval. Harvard Data Science Review. 2019.
40. Day S, Jonker AH, Lau LPL, Hilgers RD, Irony I, Larsson K, et al. Recommendations for the design of small population clinical trials. Orphanet J Rare Dis. 2018;13(1):195.
41. Adaptive Platform Trials C. Adaptive platform trials: definition, design, conduct and reporting considerations. Nat Rev Drug Discov. 2019;18(10):797-807.

42. Park JJH, Harari O, Dron L, Lester RT, Thorlund K, Mills EJ. An overview of platform trials with a checklist for clinical readers. J Clin Epidemiol. 2020;125:1-8.
43. Thielman NM, Cunningham CK, Woods C, Petzold E, Sprenz M, Russell J. Ebola clinical trials: Five lessons learned and a way forward. Clin Trials. 2016;13(1):83-6.
44. Chow SC, Huang Z. Innovative Thinking on Endpoint Selection in Clinical Trials. J Biopharm Stat. 2019;29(5):941-51.
45. Park JJ, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. Clin Epidemiol. 2018;10:343-51.
46. Canadian Task Force on the Periodic Health Examination. The Periodic Health Examination. CMAJ. 1979;121:1193-254.
47. Sackett DL. Rules of Evidence and Clinical Recommendations on the Use of Antithrombotic Agents. Chest. 1989;95(2):2S-4S.
48. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305-10.
49. Shrier I. Estimating Causal Effect with Randomized Controlled Trial. Epidemiology. 2013;24(5):779-81.
50. Oxford Centre for Evidence-Based Medicine. Levels of Evidence University of Oxford; 2009.
51. Berlin JA, Golub RM. Meta-analysis as evidence: Building a better pyramid.pdf>. JAMA. 2014;312(6):603-5.
52. Wittes J. Sample Size Calculations for Randomized Controlled Trials. Epidemiologic Reviews. 2002;24(1):39-56.
53. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer. 2014;110(3):551-5.
54. Cohen AT, Goto S, Schreiber K, Torp-Pedersen C. Why do we need observational studies of everyday patients in the real-life setting? European Heart Journal Supplements. 2015;17(suppl D):D2-D8.
55. Hoel AW, Kayssi A, Brahmanandam S, Belkin M, Conte MS, Nguyen LL. Under-representation of women and ethnic minorities in vascular surgery randomized controlled trials. J Vasc Surg. 2009;50(2):349-54.
56. Rothberg MB, Class J, Bishop TF, Friderici J, Kleppel R, Lindenauer PK. Underrepresentation of Women, Elderly Patients, and Racial Minorities in the Randomized Trials Used for Cardiovascular Guidelines. JAMA Intern Med. 2014;174(11):1867-8.
57. Konrat C, Boutron I, Trinquart L, Auleley GR, Ricordeau P, Ravaud P. Underrepresentation of elderly people in randomised controlled trials. The example of trials of 4 widely prescribed drugs. PLoS One. 2012;7(3):e33559.
58. Courtright K. POINT: Do Randomized Controlled Trials Ignore Needed Patient Populations? Yes. Chest. 2016;149(5):1128-30.
59. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):29.
60. Meurer WJ, Roger JL, Berry DA. Adaptive Clinical Trials: A partial remedy for the therapeutic misconception? JAMA. 2012;307(22):2377-5.
61. Thorlund K, Haggstrom J, Park JJ, Mills EJ. Key design considerations for adaptive clinical trials: a primer for clinicians. BMJ. 2018;360:k698.

62. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. N Engl J Med. 2017;377(1):62-70.
63. Bogin V. Master protocols: New directions in drug discovery. Contemp Clin Trials Commun. 2020;18:100568.
64. Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. Ann Oncol. 2017;28(1):34-43.
65. Saville BR, Berry SM. Efficiencies of platform clinical trials: A vision of the future. Clin Trials. 2016;13(3):358-66.
66. Pak K, Jacobus S, Uno H. Decision on performing interim analysis for comparative clinical trials. Contemp Clin Trials Commun. 2017;7:224-30.
67. Tharmanathan P, Calvert M, Hampton J, Freemantle N. The use of interim data and Data Monitoring Committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. BMC Med Res Methodol. 2008;8:12.
68. Korn EL, Freidlin B. Adaptive Clinical Trials: Advantages and Disadvantages of Various Adaptive Design Elements. J Natl Cancer Inst. 2017;109(6).
69. Ning J, Huang X. Response-adaptive randomization for clinical trials with adjustment for covariate imbalance. Stat Med. 2010;29(17):1761-8.
70. Vanderweele TJ. Surrogate measures and consistent surrogates. Biometrics. 2013;69(3):561-9.
71. Prentice R. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine. 1989;8:431-40.
72. Elston J, Taylor RS. Use of surrogate outcomes in cost-effectiveness models: a review of United Kingdom health technology assessment reports. Int J Technol Assess Health Care. 2009;25(1):6-13.
73. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015;175(8):1389-98.
74. Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. Eur J Cancer. 2019;106:196-211.
75. Hughes MD, Daniels MJ, Fischl MA, Kim S, Schooley RT. CD4 cell count as a surrogate endpoint in HIV clinical trials: A meta-analysis of studies of the AIDS Clinical Trials Group. AIDS. 1998;12:1823-32.
76. Beauchemin C, Johnston JB, Lapierre ME, Aissa F, Lachaine J. Relationship between progression-free survival and overall survival in chronic lymphocytic leukemia: a literature-based analysis. Curr Oncol. 2015;22(3):e148-56.
77. Cortazar P, Zhang JJ, Sridhara R, Justice RL, Pazdur R. Relationship between OS and PFS in metastatic breast cancer (MBC): Review of FDA submission data. Journal of Clinical Oncology. 2011;29(15_suppl):1035-.
78. Gandhi GY, Murad MH, Fujiyoshi A, Mullan RJ, Flynn DN, Elamin MB, et al. Patient-Important Outcomes in Registered Diabetes Trials. JAMA. 2008;299(21):2543-9.
79. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.
80. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701.

81. Neyman J. On the application of probability theory to agricultural experiments: Essay on principles. Section 9. Statistical Science. 1923;5:465-80.
82. Holland PW. Causal Inference, Path Analysis and Recursive Structural Equation Models. Princeton, New Jersey: Educational Testing Service; 1988.
83. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-710.
84. Greenland S, Brumback B. An overview of relations among causal modelling methods. International Journal of Epidemiology. 2002;31:1030-7.
85. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12(3):313-20.
86. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequesite for confounding evaluation: An application to birth defects epidemiology. American Journal of Epidemiology. 2002;155(2):176-87.
87. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.
88. Bishop CM. Pattern Recognition and Machine Learning. Jordan M, Kleinberg J, Scholkopf B, editors. New York, NY: Springer; 2009.
89. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, Massachusetts

London, England: The MIT Press; 2016.
90. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
91. OpenAI. Language models can explain neurons in language models 2023 [updated May 9, 2023. Available from: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.
92. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
93. Anthropic. Introducing Claude 2023 [updated March 14, 2023. Available from: https://www.anthropic.com/news/introducing-claude.
94. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint arXiv:230318223. 2023.
95. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv:200108361. 2020.
96. Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Is temperature the creativity parameter of large language models? arXiv preprint arXiv:240500492. 2024.
97. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
98. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys. 2023;56(2):1-40.
99. Guo Q, Cao S, Yi Z. A medical question answering system using large language models and knowledge graphs. International Journal of Intelligent Systems. 2022;37(11):8548-64.
100. Zou J, Schiebinger L. AI can be sexist and racist - it's time to make it fair. Nature. 2018;559(7714):324-6.
101. Bolukbasi T, Chang K, Zou J, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 30th Conference on Neural Information Processing Systems (NIPS 2016). 2016.

102. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356:183-6.
103. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences. 2018;115(16):E3635-E44.
104. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Virtual Event, Canada: Association for Computing Machinery; 2021. p. 610–23.
105. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117(23):12592-4.
106. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154(11):1247-8.
107. Sandvig C, Hamilton K, Karahalios K, Langbort C, editors. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. 64th Annual Meeting of the International Communication

Association; 2014 May 22, 2014; Seattle, WA, USA.
108. Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Scholkopf B. Avoiding Discrimination through Causal Reasoning. arXiv. 2018.
109. Richardson B, Gilbert JE. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. arXiv. 2021.
110. Rawls J. Justice as Fairness. The Philosophical Review. 1958;67(2):164-94.
111. Bell NK. Nozick and the Principle of Fairness. Social Theory and Practice. 1978;5(1):65-73.
112. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness Through Awareness. arXiv. 2011.
113. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. arXiv. 2015.
114. Kamiran F, Calders T. Classifying without discriminating. Proceedings of IEEE Xplore. 2009.
115. Calders T, Verwer S. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery. 2010;21(2):277-92.
116. Kamishima T, Akaho S, Asoh H, Sakuma J, editors. Fairness-Aware Classifier with Prejudice Remover Regularizer2012; Berlin, Heidelberg: Springer Berlin Heidelberg.
117. Nabi R, Shpister I. Fair Inference on Outcomes. arXiv. 2018.
118. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys. 2021;54(6):1-35.
119. Kusner M, Loftus J, Russell C, Silva R. Counterfactual fairness. arXiv. 2018.
120. World Health Organization. Global Health Observatory (GHO) data 2020 [Available from: https://www.who.int/gho/hiv/en/.
121. O'Brien N, Chi YL, Krause KR. Measuring Health Outcomes in HIV: Time to Bring in the Patient Experience. Ann Glob Health. 2021;87(1):2.
122. World Health Organization. Constitution of the World Health Organization. Geneva, Switzerland: World Health Organization; 2006.

123. Kay ES, Batey DS, Mugavero MJ. The HIV treatment cascade and care continuum: updates, goals, and recommendations for the future. AIDS Res Ther. 2016;13:35.
124. Fischl MA, Richman DD, Greico MH, Gottlieb MS, Volberding PA, Laskin OL, et al. The efficacy of azidothymidine AZT in the treatment of patients with AIDS and AIDS-related complex. N Engl J Med. 1987;317(4):185-92.
125. The Strategies for Management of Antiretroviral Therapy (SMART) Study Group. CD4+ count-guided interruption of antiretroviral treatment. N Engl J Med. 2006;355(20):2283-97.
126. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC, et al. Effect of Early vs. Deferred Antiretroviral therapy for HIV on Survival. N Engl J Med. 2009;360(18):1815-26.
127. Group ISS, Lundgren JD, Babiker AG, Gordin F, Emery S, Grund B, et al. Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection. N Engl J Med. 2015;373(9):795-807.
128. Deeks SG, Lewin SR, Havlir DV. The end of AIDS: HIV infection as a chronic disease. The Lancet. 2013;382(9903):1525-33.
129. Kobin A, Sheth N. Levels of adherence required for virologic suppression among newer antiretroviral medications. The Annals of Pharmacotherapy. 2011;45:371-80.
130. Iacob SA, Iacob DG, Jugulete G. Improving the Adherence to Antiretroviral Therapy, a Difficult but Essential Task for a Successful HIV Treatment-Clinical Points of View and Practical Considerations. Front Pharmacol. 2017;8:831.
131. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, van Lunzen J, et al. Sexual Activity Without Condoms and Risk of HIV Transmission in Serodifferent Couples When the HIV-Positive Partner Is Using Suppressive Antiretroviral Therapy. JAMA. 2016;316(2):171-81.
132. World Health Organization. Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach. World Health Organization,; 2016.
133. Battistini Garcia SA, Guzman N. Acquired Immune Deficiency Syndrome CD4+ Count.: StatPearls Publishing; 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513289/.
134. Clinical Info HIV. Initiation of Antiretroviral Therapy 2019 [updated December 18, 2019. Available from: https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv/initiation-antiretroviral-therapy?view=full.
135. World Health Organization. Adherence to long-term therapies: Evidence for action. Switzerland: World Health Organization,; 2003.
136. Mugavero MJ, Amico KR, Horn T, Thompson MA. The state of engagement in HIV care in the United States: from cascade to continuum to control. Clin Infect Dis. 2013;57(8):1164-71.
137. Haider MR, Brown MJ, Harrison S, Yang X, Ingram L, Bhochhibhoya A, et al. Sociodemographic factors affecting viral load suppression among people living with HIV in South Carolina. AIDS Care. 2021;33(3):290-8.
138. Blank AE, Fletcher J, Verdecias N, Garcia I, Blackstock O, Cunningham C. Factors associated with retention and viral suppression among a cohort of HIV+ women of color. AIDS Patient Care STDS. 2015;29 Suppl 1:S27-35.
139. Wawrzyniak AJ, Rodriguez AE, Falcon AE, Chakrabarti A, Parra A, Park J, et al. Association of individual and systemic barriers to optimal medical care in people living with HIV/AIDS in Miami-Dade County. J Acquir Immune Defic Syndr. 2015;69 Suppl 1:S63-72.

140. Castilho JL, Melekhin VV, Sterling TR. Sex differences in HIV outcomes in the highly active antiretroviral therapy era: a systematic review. AIDS Res Hum Retroviruses. 2014;30(5):446-56.
141. Aidala AA, Wilson MG, Shubert V, Gogolishvili D, Globerman J, Rueda S, et al. Housing Status, Medical Care, and Health Outcomes Among People Living With HIV/AIDS: A Systematic Review. Am J Public Health. 2016;106(1):e1-e23.
142. Engler K, Lenart A, Lessard D, Toupin I, Lebouche B. Barriers to antiretroviral therapy adherence in developed countries: a qualitative synthesis to develop a conceptual framework for a new patient-reported outcome measure. AIDS Care. 2018;30(sup1):17-28.
143. Long S, Loutfi D, Kaufman JS, Schuster T. Limitations of Canadian COVID-19 data reporting to the general public. J Public Health Policy. 2022;43(2):203-21.
144. Fremont A, Lurie N. Appendix D, The Role of Racial and Ethnic Data Collection in Eliminating Disparities in Health Care. In: National Research Council (US) Panel on DHHS Collection of Race and Ethnic Data, editor. Eliminating Health Disparities: Measurement and Data Needs. Washington (DC): National Academies Press (US); 2004.
145. Ramraj C, Shahidi FV, Darity W, Kawachi I, Zuberi D, Siddiqi A. Equally inequitable? A cross-national comparative study of racial health inequalities in the United States and Canada. Social Science & Medicine. 2016;161:19-26.
146. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. Understanding Eliminating, Racial Ethnic Disparities in Health, Care. In: Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington (DC): National Academies Press (US)

Copyright 2002 by the National Academy of Sciences. All rights reserved.; 2003.
147. Fiscella K, Sanders MR. Racial and Ethnic Disparities in the Quality of Health Care. Annu Rev Public Health. 2016;37:375-94.
148. Stanbrook MB, Salami B. CMAJ's new guidance on the reporting of race and ethnicity in research articles. CMAJ. 2023;195(6):E236-E8.
149. Berry I, Brown KA, Buchan SA, Hohenadel K, Kwong JC, Patel S, et al. A better normal in Canada will need a better detection system for emerging and re-emerging respiratory pathogens. Canadian Medical Association journal. 2022;194(36):E1250-E4.
150. Canadian Institutes of Health Information. Health equity and population health 2024 [Available from: https://www.cihi.ca/en/topics/health-equity-and-population-health.
151. Canadian Institutes of Health Information. Race-based data collection and health reporting In: CHIH, editor. Ottawa2020.
152. Pinto AD, Eissa A, Kiran T, Mashford-Pringle A, Needham A, Dhalla I. Considerations for collecting data on race and Indigenous identity during health card renewal across Canadian jurisdictions. CMAJ. 2023;195(25):E880-E2.
153. Golestaneh L, Neugarten J, Fisher M, Billett HH, Gil MR, Johns T, et al. The association of race and COVID-19 mortality. EClinicalMedicine. 2020;25.
154. Gershengorn HB. Inequitable Resource Allocation Amidst a Pandemic—A Crisis Within a Crisis. JAMA Network Open. 2022;5(3):e221751-e.
155. Long S, Schuster T, Piché A, Research S. Can large language models build causal graphs? arXiv preprint arXiv:230305279. 2023.
156. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

157. Google DeepMind. Introducing Gemini: DeepMind's next-generation AI 2023 [Available from: https://www.deepmind.com
158. Mistral. Introducing Mistral 7B: A high-performance dense language model. 2023 [Available from: https://mistral.ai.
159. Cohere. Introducing Command R: Cohere's New Retrieval-Enhanced LLM. 2023 [Available from: https://cohere.ai.
160. Decoder. OpenAI just opened ChatGPT-API - and its ten times cheaper than GPT-3.5. 2023.
161. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al., editors. Zero-shot text-to-image generation. International conference on machine learning; 2021: Pmlr.
162. OpenAI. Introducing GPT-4o and more tools to ChatGPT free users 2024 [updated May 13, 2024. Available from: https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.
163. Silberling A. Why is AI so bad at spelling? Because image generators aren't actually reading text2024. Available from: https://techcrunch.com/2024/03/21/why-is-ai-so-bad-at-spelling/?guccounter=1

164. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Scoping studies: towards a methodological framework. 2005;8(1):19-32.
165. World Bank. New country classifications by income level: 2019-2020 2019 [cited 2020 March 21]. Available from: https://blogs.worldbank.org/opendata/new-country-classifications-income-level-2019-2020.
166. Mak S, Thomas A. An Introduction to Scoping Reviews. J Grad Med Educ. 2022;14(5):561-4.
167. Western University. Knowledge Synthesis: Systematic & Scoping Reviews 2024 [Available from: https://guides.lib.uwo.ca/knowledgesynthesis.
168. de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. Computing. 2023;105(10):2171-94.
169. van Dinter R, Tekinerdogan B, Catal C. Automation of systematic literature reviews: A systematic literature review. Information and Software Technology. 2021;136:106589.
170. Olofsson H, Brolund A, Hellberg C, Silverstein R, Stenström K, Österberg M, et al. Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. Res Synth Methods. 2017;8(3):275-80.
171. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4:78.
172. Orel E, Ciglenecki I, Thiabaud A, Temerev A, Calmy A, Keiser O, et al. An Automated Literature Review Tool (LiteRev) for Streamlining and Accelerating Research Using Natural Language Processing and Machine Learning: Descriptive Performance Evaluation Study. J Med Internet Res. 2023;25:e39736.
173. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021;50(2):620-32.
174. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. International Journal of Epidemiology. 2017;45(6):1887-94.
175. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol. 2022;142:264-7.

176. Krieger N, Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. Int J Epidemiol. 2016;45(6):1787-808.

APPENDIX A: ETHICS APPROVAL

Centre universitaire de santé McGill | McGill University Health Centre

2023-09-07

Dr. Bertrand Lebouche
Infectious Diseases

c/o: Stephanie Long
email: stephanie.long@mail.mcgill.ca

RE: Final REB Approval of a New Research Project
Leveraging expert knowledge and large language models for the co-creation of causal diagrams to inform clinical trial planning: A feasibility and acceptability study **(DAGs / 2024-9781)**

MUHC REB Co-Chair for the CTGQ panel: Me Marie Hirtle

Dear Dr. Lebouche,

Thank you for submitting your responses and corrections for the research project indicated above, as requested by the McGill University Health Centre (MUHC) Research Ethics Board (REB).

The MUHC REB, more precisely its Cells, Tissues, Genetics & Qualitative research (CTGQ) panel provided conditional approval for the research project after a delegated review provided by its member(s).

On 2023-09-07, a delegated review of your responses and corrections was provided by member(s) of the MUHC REB. The research project was found to meet scientific and ethical standards for conduct at the MUHC.

The following documents were approved or acknowledged by the MUHC REB:

- **Initial Submission Form**
  - (F11H-NIR-113824)
- **REB Conditions & PI Responses Form(s)**
  - (F20-118051)
- **Scientific evaluation report**
  - (Form 3 Oral Comprehensive Exam Final Report - Stephanie Long (1).pdf)
- **Signed commitement**
  - (PI Commitment and Signature_June272023_BL[67659].pdf)
- **Informed Consent form**
  - (Consent form - RIMUHC - Clinician - July 4, 2023.doc) [Date: 2023-07-04, Version: 1.4]
  - (Consent form - RIMUHC - Patient - July 4, 2023.doc) [Date: 2023-07-04, Version: 1.4]
  - (Consent form - RIMUHC - Patient - FR v3 - Sept 7, 2023.doc) [Date: 2023-09-07, Version: v6, September 7, 2023]

    - (Consent form - RIMUHC - Clinician - FR v3 - Sept 7, 2023.doc) [Date: 2023-09-07, Version: v6, September 7, 2023]
- **Information for participants**
    - (Domain Expert Interview Guide v4 - July 4, 2023.docx) [Date: 2023-07-04, Version: 1.4]
    - (Domain Expert Survey - Causal mapping and LLM - v3 June 27, 2023.docx) [Date: 2023-06-27, Version: 1.3]
    - (Domain Expert Survey - Causal mapping without LLM - v3 June 27, 2023.docx) [Date: 2023-06-27, Version: 1.3]
    - (Domain Expert Survey - Causal mapping and LLM - v3 - FR - September 7, 2023.docx) [Date: 2023-08-10, Version: 1.4]
    - (Domain Expert Survey - Causal mapping without LLM - v3 - FR - Sptember 7, 2023.docx) [Date: 2023-08-10, Version: 1.4]
    - (Domain Expert Interview Guide v6 FR Septembre 7, 2023.docx) [Date: 2023-08-21, Version: 1.5]
- **Research protocol**
    - (Causal mapping and LLM experts protocol - final version .docx) [Date: 2023-06-14, Version: 1.5]
- **Approval of the Department / Division Head**
    - (Causal mapping study - department approval CVIS (MR signed)_July 4, 2023.pdf) [Date: 2023-07-04]
- **Supplementary documents**
    - (Coverletter_CausalMapping_July42023.pdf)

This will be reported to the MUHC REB and will be entered accordingly into the minutes of the next CTGQ meeting. Please be advised that you may only initiate the study after all required reviews and decisions are received and documented and you have received the MUHC authorization letter.

The approval of the research project is valid until 2024-09-07.

All research involving human subjects requires review at recurring intervals. To comply with the regulation for continuing review of at least once per year, it is the responsibility of the investigator to submit an *Annual Renewal Submission Form* (F9) to the REB prior to expiry. Please be advised that should be protocol reach its expiry before a Continuing review has been submitted, the data collected after the expiry date may not be considered valid. However, should the research conclude for any reason prior to approval expiry, you are required to submit a *Completion (End of Study) Report* (F10) to the board once the data analysis is complete to give an account of the study findings and publication status.

Furthermore, should any revision to the project or other development occur prior to the next continuing review, you must advise the REB without delay. Regulation does not permit initiation of a proposed study modification prior to its approval by the REB.

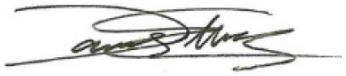
The MUHC REB is registered and works under the published guidelines of the *Tri-Council Policy Statement 2*, in compliance with the *Plan d'action ministériel en éthique de la recherche et en intégrité scientifique* (MSSS, 1998) and the *Food and Drugs Act* (2001.06.07), acting in conformity with standards set forth in the (US) *Code of Federal Regulations* governing human subjects research and functioning in a manner consistent with internationally accepted principles of good clinical practice.

Appendix A

We wish to advise you that the MUHC REB "working procedures" completely satisfies the requirements for Research Ethics Board Attestation (REBA) as stipulated by Health Canada. Neuro and MUHC research that is subject to US Federal Wide Assurance is conducted under FWA00000840.

We trust this will prove satisfactory to you. Thank you for your consideration in this matter.

Best Regards,



James Ellasus
Coordonnateur du CÉR du CUSM | MUHC REB Coordinator
pour: le co-président du CÉR du CUSM | for: MUHC REB Co-Chair

Signed on 2023-09-07 at 16:05