Extracting, Analyzing and Using Patterns of Service Utilization from a Cross Continuum Health Service System to Optimize Care for Patients with Complex Issues – A Methodological Approach.

By

Jonas Bambi Yona

MSc University of Northern British Columbia 2006, MBA University of Northern British Columbia 2008

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In the School of Health Information Science

University of Victoria

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱ SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Extracting, Analyzing and Using Patterns of Service Utilization from a Cross Continuum Health Service System to Optimize Care for Patients with Complex Issues – A Methodological Approach.

By

Jonas Bambi Yona

MSc University of Northern British Columbia 2006, MBA University of Northern British Columbia 2008

Supervisory Committee

Alex Kuo (PhD), Supervisor
School of Health Information Science,
University of Victoria

Abraham Rudnick (MD, PhD), Committee Member
School of Health Information Science,
University of Victoria

Ken Moselle (PhD), Outside Member
Applied Clinical Research Unit,
Vancouver Island Health Authority

Abstract

Introduction:

The clinical service system performs a very large and diverse array of functions associated with a broadly differentiated array of problems. As a result, the service system is broken up functionally and administratively into service units. To provide the best possible care for patients with chronic or complex problems the interoperation of service system components needs to be optimized. Such an optimization requires the enactment of appropriate problem-specific clinical protocols as well as cohort-specific Clinical Practice Guidelines (CPGs). However, alignment of the service system operations with CPGs is very challenging. This is mostly due to the challenge of identifying longitudinal patterns of service utilization (PSUs) in a cross-continuum data to assess adherence to the CPGs. Hence, the research activities aim to address the following questions: (1) what methodology can be employed to extract and identify clinically understandable cohorts-specific patients' PSUs within sparse high-dimensional cross-continuum healthcare datasets? (2) to what extent can one cut across the complexity of a cross-continuum service structure to capture the dynamics of the journey of patients with complex issues that clearly portray their engagement with the service system, to help locate potential operational problems? (3) once identified, to what extent can PSUs be used to inform quality assurance and quality improvement (QA/QI) initiatives?

Methods:

Starting with a semantic layer, referred to as the Clinical Context Coding Scheme, to address data granularity and nomenclature issues, various machine learning approaches were used to extract PSUs. These include the use of nested-iterative graph community detection, directed graph, and Natural Language Processing (NLP) clustering. These steps were followed by the use of various graph metrics and input from clinical and operational subject matter experts to refine the level of resolution for the extracted patterns.

Results:

The results have shown that by using a nested-iterative community detection or NLP clustering, it is possible to extract cohorts-specific high-prevalence functionally integrated PSUs. The results have also shown that directed graphs are well suited to the task of depicting the way that the diverse components of the system are functionally coupled—or remain disconnected—by patients journeys.

Discussion:

These findings have several implications related to the optimization of care for patients with chronic/complex problems, including: (1) the possibility of influencing the reorganization of some services within a health organization service-structure to provide the most optimal connections between services to address patients' needs, and (2) the first step in addressing the challenge of locating potential operational problems for patients with complex issues engaging with a complex healthcare service system. Additionally, the combination of the proposed methodologies with various statistical analysis, have demonstrated that PSUs can play an important role in informing diverse QA/QI initiatives, including: (1) providing a nuanced approach to assess and measure access disparity, based on local reality, across the full care continuum, and a first step in addressing inequities for a healthcare organization, and (2) equipping a healthcare organization with required information, based on local reality, to provide better care for the opioid overdose patients, as well as being pro-active in preventing subsequent overdoses. However, in informing QA/QI initiatives, distal factors such as social determinants of health not captured within the dataset are not considered in the models, limiting the usefulness of the recommendations. This is a limitation that future research will need to address.

Conclusion:

The research activities undertaken provide a first step – the extraction of PSUs, in introducing a novel analytics framework relying on patients' service pathways as a foundation to evaluate conformance of interventions to cohort-specific CPGs. In collaboration with various clinical and operational SMEs, future research will expand on PSUs, to include other elements required for this novel analytical framework to be used as one of the paradigms to the secondary use of data collected by health organizations, to support the implementation of Learning Healthcare Systems.

Table of Contents

List of Figures

Dedication

First, I would like to thank my dear wife Maggie and my kids Emma, Anaïs, and Marc-André for their patience, support and sacrifice during my doctoral program. A lot of family time was foregone for me to be able to complete this program. This is truly a family achievement.
Though thousands of kilometers away, I also would like to thank my family back in Africa: my mother Diodata, my father John-Thomas, bother Dr. Emmanuel Bambi, sisters Claudine and Johanna, and late sister Sophie for always checking on me, to see how my program was progressing.

Secondly, I would like to thank Dr. Ken Moselle for his time and dedication to making this endeavor a success story. The countless ongoing conversations and discussions we have had over the years have opened my eyes to various possibilities and opportunities that are accessible to us, to really transform our healthcare system and make it more efficient and more equitable, for the betterment of our communities. He is truly a forward thinker that deeply cares about those that cannot advocate for them selves with regards to their health and well being.

Finally, I would like to thank Dr. Abraham Rudnick for his concise and prompt responses to my requests and questions, and Dr. Alex Kuo for agreeing to take me in as his supervisee. My gratitude also goes to all my collaborators, especially Dr. Yudi Santoso and Stanley Robertson for their technical expertise and prowess, and Dr. Ernie Chang for this various clinical feedback and insights.

Chapter 0: Foreword and Literature Review

Foreword

Access to Care for Patients with Complex Problems

Timely and appropriately sustained access to health services may be impacted substantially by the capacity of persons to initiate and maintain effective engagement with systems of health services, as well as the structure of the services and the processes encapsulated within the service system to address patients' needs. From a person's perspective, appropriate and effective service access may be associated directly or indirectly to several factors, including homelessness, neuro-cognitive impairment, psychiatric issues, mobility issues and various forms of stigmatization. These stigmatizations may be associated with person demographics such as racial/ethnic status, as well as health conditions such as substance use/addiction (Baker et al., 2018; Christiani et al., 2008; Cradock-O'Leary et al., 2002; Iezzoni et al., 2000; Laursen et al., 2014). From a service system perspective, timely access to care may be impacted by various factors, including system capacity and lack of optimal services integration.

Patients with complex problems may be contending with one or more of the factors previously indicated. For example, a person may be physically homeless and may be contending with a chronic/severe psychiatric illness. Or a person may be neurocognitively impaired and physically disabled. Consequently, there may be services provided that target patients who are contending only with a single factor, but different types of services or different operationally integrated service delivery models may be required to respond effectively to the composite impact of several factors.

Part of this issue around service integration stems from the fact that administratively integrated services, or even integration at the level of robust technical integration (e.g., a full cross-continuum implementation of a clinical information system) does not necessarily translate into operationally integrated service systems that can respond effectively to the needs of persons when those needs span multiple organizational subdivisions within the health region.

For example, case management functions within the continuum of Mental Health & Substance Use (MHSU) services may work well to enable and ensure appropriate levels of engagement with psychiatrist consultation services. However, even the combination of case managers plus psychiatrists may have limited capacity to reach into other dimensions, such as medical diagnostic/investigative services, when that access is initiated by medical specialists outside of MHSU, or the clinical treatment/management services for physical health problems associated with known or identified causes of poor health

outcomes (Laursen et al., 2012). Though well engaged with portions of the service system as a whole, key aspects of care may be orphaned, including those medical services that are associated with known major causes of excess morbidity, and/or early mortality.

Hence, from a QA/QI perspective, building out foundations for improving quality of care for a cross continuum health service-system functioning as a complex system can be a challenge. This challenge is amplified when the service system with limited capacity must provide care to high risk/high-needs populations with complex/chronic problems who are themselves subject to an array of environment factors including a variably poisoned drug supply, pandemic spread of various infectious agents such as SARS-CoV-2, just to name a few.

Use Clinical Protocols and Clinical Process Guidelines to Optimize Care for Patients

The clinical service system performs a very large and diverse array of functions associated with a broadly differentiated array of problems. For example, there are over 69,000 International Classification of Diseases - ICD-10 version (*International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)*, 2023) that map coarsely onto over 70,000 ICD-10 Diagnoses. These functions can only be in multiple service units spread out across a geographically distributed array of service users. Indeed, the full array of functions constituting a health service system could not all be located in a single very large building because the functions themselves are not all facility-based. For example, home and community care takes place in the home of a person. Addictions outreach services will be provided on the streets.

Consequently, the service system is broken up functionally and administratively into service units. However, service requires for any health problems that are complex or chronic and progressively more clinically elaborated will require these various components to work together in functionally integrated manner. As such, to provide the best possible care to patients, interoperation of service system components needs to be optimized.

To achieve this optimization, models of idealized service delivery processes are employed. More specifically, this optimization entails integration of clinically appropriate problem-specific clinical protocols, and optimization of service-system-encompassing clinical pathways. When an intervention is required for a discrete problem and require access to services that are administered in a standard way, then a standard clinical protocol may be applied to optimize the care of a patient. As an example, one

can consider the problem of sepsis protocols for emergency departments (McVeigh, 2020). Sepsis protocols are protocols that are clearly articulated and often coded as clinical decision support tools with clinical information systems. They specify the signs and symptoms that should alert clinicians to the possibility of a patient becoming septic (McVeigh, 2020). Using locally available evidence, they specify the diagnostic and the investigation that need to be carried out to perform differential diagnosis and recommend interventions that provide a protocol-based care (McVeigh, 2020). Clinical information systems usually capture and supply the data drivers necessary to populate sepsis clinical decision support protocols (McVeigh, 2020). Additionally, conformance of actual clinical decisions with the sepsis protocols can be seen within the local data (McVeigh, 2020). Hence, clinical operations can be optimized around circumscribed protocols, such as sepsis protocols. The extraction of aggregated information from transactional clinical information systems can support the effort by providing visibility into the relationship between usual practices and cohorts who share a similar clinical need/risk profile – at a particular point in the history of their interactions with the service system – their "journey".

On the other hand, when caring for patients with complex evolving chronic conditions, such as heart failure or schizophrenia, a clinically effective and appropriate response requires coordinated timely access to a diverse and overtime changing array of services. In such a case a clinical practice guideline (CPG) is required. CPGs consist of arrays of clinical protocols whose branching execution is contingent on etiology, pathophysiology, and treatment response (Panteli et al., 2019). These guidelines are imposed on service system operations by parties directing and delivering care, and supported by technology, e.g., orders or order sets in Clinical Information Systems (CIS). They are expressed in generic terms that need to be translated into clinical pathways that uses local service system terms (Rotter et al., 2019).

Because the activities that constitute adherence to a CPG for complex conditions may be distributed longitudinally across a large array of service units, and because CPG-adherent decisions at any point in time may be dictated by both the CPG and by the current clinical/functional/behavioral status of the patient, achieving even moderately complete visibility into processes that represent CPG-adherence may be quite challenging. As such, the use of local data to optimize clinical care for persons who may benefit from process compliance with CPGs may will be quite challenging. These measurement difficulties at the level of the basic unit of analysis will accumulate in generation of any cohort-level measures of association between local service system operations and CPGs and outcomes associated with CPG adherence.

Using Patterns of Service Utilization (PSUs) for Health-Service System Optimization for Patients with Complex Problems

Patients Journeys - Dynamics of Patients Interacting with Healthcare Service System

A diverse array of factors influences an individual or a population health status, healthcare access and health outcome. Some of these include socio-economic, behavioral, and biological factors. These factors are classified as proximal or distal determinant of health (Kawachi & Berkman, 2000). These distal determinants are also often referred to as "social determinants of health (Commission on Social Determinants of, 2008), or as "non-medical" factors that influence health outcomes (Hacker et al., 2022). Factors such as education, employment status, housing availability, financial resources are classified as distal determinant of health. Importantly, these distal determinants generally fall outside the sphere of influence of a health service system – though they may have a powerful impact on health risk and outcomes, as highlighted in the work of Centers for Disease Control and Prevention or the World Health Organization.

Behavioral risk and biological factors are classified as proximal determinant of health. Behavioral risk factors include alcohol use, illicit drug use, healthcare service use, behavior, and compliance with medical treatment. Biological factors refer to aspects such as an individual biological predisposition that can potentially affect their clinical or functional status. These factors, whether distal or proximal influence one another, and have a strong impact on how individuals engage with healthcare services and

their ultimate health outcome (see figure 1).

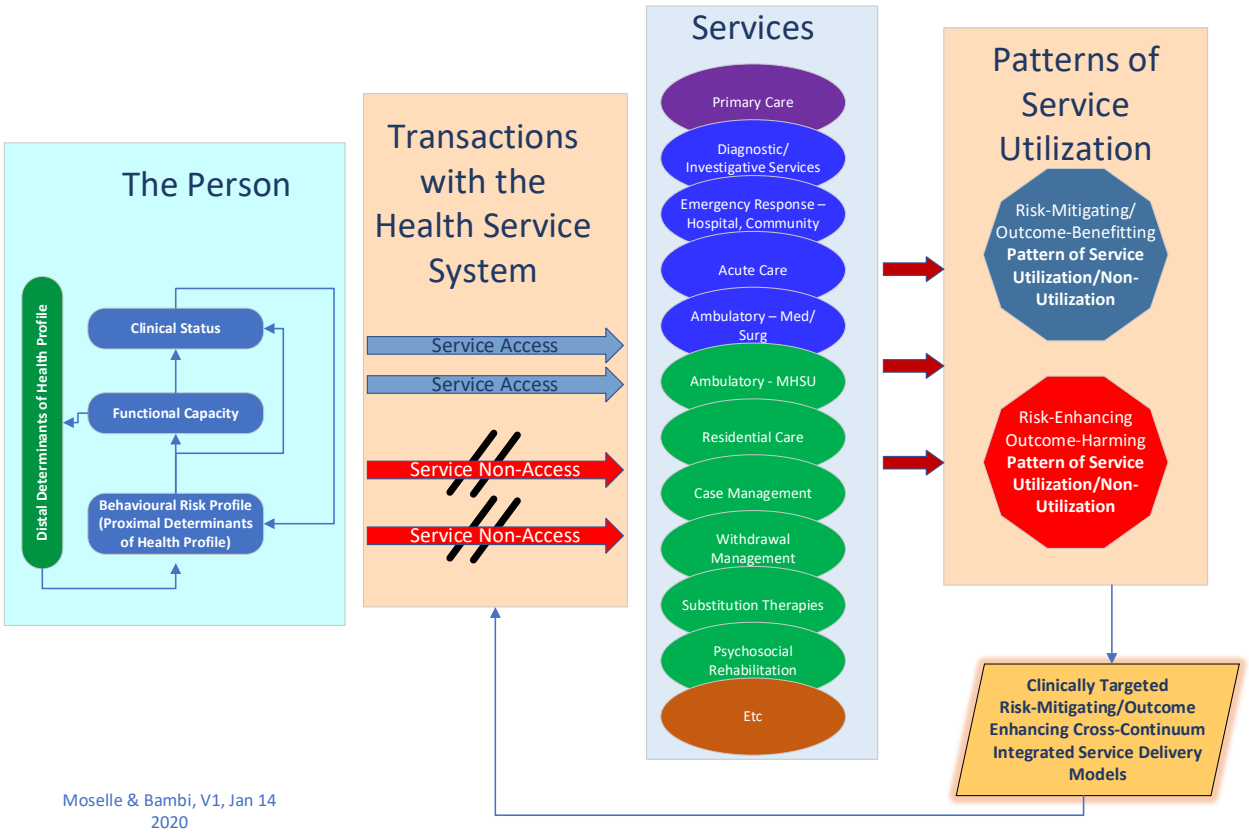


Figure 1 -Individual Engagement with Healthcare Services

If we picture individuals' transactions with the healthcare system over time as a set of trajectories, we can partition these into segments that reflect the dynamics of persons over time, as well as service pathways enacted by the service system over time (see figure 1). The first is a person's clinical trajectory, expressed in terms of longitudinal bodies of data that characterize attributes of the person, such as clinical status, functional capacity, and behavioral risk profile. The second is a service access trajectory, reflecting activities that the service system performs in the context of interactions with the person. This includes access to service such as emergency care, acute care, withdrawal management, primary care, and residential care. Both trajectories are so intertwined and interdependent, creating a patient journey within the healthcare service system. The "Patient Journey" is assembled from one or multiple Service Pathways.

A "patient journey" may be broken into multiple segments. These represent phases that a patient may go through. They include pre-clinical phase (no symptoms), where an individual may only present bio-

medical markers, or behavioral risks for a certain disease (Ahlgrim et al., 2019). This may be followed by a prodromal phase, where the patient is showing sign of illness before full emergence of symptoms (George et al., 2017). Early intervention may be initiated at this stage to reduce the likelihood that a clinically significant form of the illness will emerge or to reduce the severity of the illness or to bias odds in favor of less severe impact on chronic or recurring conditions. A good example of such an initiative is the EPI (Early Psychosis Intervention), introduced by the province of BC and implemented by all the health authorities to help young people with early psychosis (*EPI in British Columbia*, 2020).

The next phase is the active clinical phase where symptoms require a response from the person experiencing those symptoms, and where the service system may enhance outcomes in short-term and longer-term by providing care or performing interventions that the person cannot perform on his/her own behalf. These interventions are intended to reduce suffering, protect capacity to function adaptively, and alter longer-term illness trajectories. The "right interventions" instituted in a timely fashion may become critical at this stage in the "patient journey."

In the active clinical phase, the dynamics that govern the achievement of health outcomes for point-in-time treatment of discrete disease entities, as well as over-time management of chronic disease are a function of the pathophysiology of disease progression and response to treatment, as well as the dynamics of a system of health services that is positioned in between the person and treatments or services that are keyed to the health conditions with which the person is contending – over time. These dynamics do not emerge in a vacuum, imposed by the governing mechanism within the service system on the populations accessing services. They emerge as the reciprocal interplay of people initiating access and the operational dynamics of the service system.

Use of PSUs in Optimizing Care for Persons and Population Contending with Complex Problems

Persons and populations contending with chronic and typically progressively more complex chronic health conditions will interact with an increasingly diverse array of health services. Clinical trajectories for such persons, and the experience of sentinel clinical events such as organ failure or loss of functional capacity, is an emergent characteristic of the dynamics of persons interacting with a complexly structured health service system. In this program of research activities, we demonstrate a methodology for discovering and depicting the dynamics that characterize the interoperation of multiple components of a complex health service system, in relationship to cohorts contending with increasingly impactful chronic conditions. These dynamics appear as longitudinal patterns of service utilization (PSUs), sparsely

distributed in a high-dimensional array of services spanning a full array of secondary and tertiary services, including hospital and community-based services, outreach, residential care, case management, and numerous others, for medical/surgical problems such as cardiovascular disease, as well as mental health and addictions services.

As stated previously, when caring for patients with complex evolving chronic conditions that requires coordinated timely access to a diverse and overtime changing array of services, cohorts-specific clinical practice guidelines (CPG) are required. However, optimization encompassing cross-continuum service system clinical pathways is more challenging than optimization that target a problem-specific protocol.

To illustrate such a challenge one can consider the acute care hospitalization and ambulatory care follow-up for persons with schizophrenia. Also, consider an optimally interoperating cross-continuum service models with a set of CPGs covering complexly unfolding chronic conditions. An optimal care would include (1) an array of services that covers the prodromal phases of a chronic often relapsing condition such as schizophrenia, (2) the acute care hospitalization, (3) an array of post-discharge stabilization and rehabilitation options, including services such as mobile crisis response, and psychiatric consultation, (4) an array of progressively more staffing-intensive case management models, (5) various secondary or tertiary residential care options, (6) psychosocial rehabilitation services, and (7) addictions harm reduction or rehab/recovery services for persons with a co-morbid substance use disorder. Moreover, various services will need engagement to address the various medical comorbidities usually associated with schizophrenia condition, including risk for kidney disease associated with side-effects of psychiatric medications via their attendant risk for metabolic syndromes (Laursen et al., 2014), or the heightened risk for cardiovascular disease (Laursen et al., 2012). With more than 50 CPGs outlined in the BC guidelines (BC_Ministry_of_Health, 2024), addressing high prevalence problems with various degrees of complexities, it should be noted that this level of complexity is not unique to the schizophrenia cohorts only.

Challenges at the level of modeling and measurement stem from three sources: the first is related to the breadth of information required to know whether the CPG is being enacted in a clinically appropriate manner. This is especially true if the CPG recruits services that span a full continuum of services, e.g., medical/surgical services for various physical health concerns, mental health services (acute care, ambulatory, residential care, etc.), addictions services, possibly out-reach for homeless persons, given the downward socio-economic mobility of persons with problem such as schizophrenia. Hence, access to cross-continuum encounters data from one or more systems is required.

Secondly, even if such data are accessible, there is the foundational challenge when trying to align service system operations around CPGs at a population level. The challenge of identifying longitudinal patterns of service utilization in the data, both to know what was done, and to know whether it should have been done, and to know whether outcomes intended by CPGs are being achieved, and if not – why not. Given the number of service entities involved in providing coverage for a complex CPG, relative to the number of people who require those services, the relevant data are likely to be distributed quite sparsely in a high-dimensional space.

The third challenge is about selecting the most optimal approach and methodology to extract meaningful PSUs that are readily and correctly interpretable by persons who do not have a background in statistics, research, or data science. Additionally, once these PSUs are generated, what other additional products can be generated to address QA/QI needs to support the optimization of a complex healthcare service-system.

Each of the published manuscripts outlined in this document, address both the second and third challenges. The first challenge was addressed through the access to a data source consisting of retrospective longitudinal transactional data contents extracted from a single instance of a Clinical Information System (CIS) deployed across the continuum of services provided by one of the Health Authorities within Canada. The span of the health service organization includes almost all secondary and tertiary services for all ages, for persons contending with medical/surgical issues and/or mental health/substance use issues. This includes acute care/intensive care services, hospital and community-based emergency response, ambulatory services, residential care ser-vices for older adults or persons contending with mental health issues, case management services, and a range of addictions harm reduction or rehab and recovery-oriented services. The encounter data accessed by this program of research activities consists of approximately 10 million encounters of 7 years for approximately 1 million patients. With the exception of a small number of restricted services where data are strictly embargoed (e.g., services for persons who are victims of sexual assault) this represents data for all service recipients. A certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following British Columbia, Canada Ethics harmonization guideline. The REB number is H21-02817.

Literature Review

The encounters data capturing the interaction of patients with complex issues interacting with a complex healthcare system is complex by nature. Given this complexity, a careful consideration needed to be given to the choice of the approach and method to use to do the analysis. Additionally, a thorough literature review needed to be conducted to acquire an understanding of what has been done in the space of healthcare data analysis and role of Machine Learning in conducting the analysis. The literature review conducted aimed at addressing both issues.

Graph Network Modelling and NLP

Analytical methods must be fit for data structures and contents. The selection of analytical methods is determined by core datasets for the study. In particular, the source data for a large fraction of the modeling/visualization activities set out in this study are sparse high-dimensional extracts from a clinical information system. Bzdok et al. (2018) refer to such clinical datasets as "rich and unwieldy". As well, core data sets (e.g., encounter data) are irregularly structured, both in scope of content and timing.

On the patient side - patterns of service utilization and the services of different providers reflect the actions taken by patients to initiate or sustain relations with the service system. As well, people have distinguishing distal determinants of health profiles which have pervasive impacts on disease, access to treatment and response. On the service system/provider side - providers must exercise judgment and adapt standard treatment protocols (where they exist) or interpret care guidelines to focal problems experienced by persons person who have their own preferences and are repositories, over time, of multiple possibly interacting focal problems.

The resulting record of service activity and changing characteristics of people are variable in content and timing – even for people who have similar clinical/functional/behavior profiles and are at similar stages in their health "journeys". Stated in slightly different terms: healthcare encounter data may be "noisy" if for no other reason that the actions taken by individuals to access care, and by direct consequence, the records of those interactions in clinical information systems, may be quite variable. Georgousis et al. (2021) further discuss this issue of irregularly structured data, and Lasko and associates (Lasko et al., 2013; Lasko & Mesa, 2019) provides details on their use of deep learning procedures with what they refer to as "noisy, sparse and irregular clinical data" which they also refer to as "dirty clinical data". Yuan & Deng (2021) further discuss the issue of data sparseness in health datasets and the use of graph

methods to enable generation of recommendations to medical specialists using such data. Finally, Lin et al. (2020) took on the challenge of using graph methods to identify relationships between diagnoses in what they term "noisy" data. On a related note, Pivovarov et al. (2015) used graph machine learning methods to conduct analysis of patient record data contents that are heterogeneous with respect to structure and content, including notes, laboratory tests, medications, and diagnostic codes.

Practically, this means that a large fraction of what are often referred to as "classic" statistical procedures or “statistical learning” methods, e.g., linear regression, are not suited to the task of building out the basic units of analysis (Patterns of Service Utilization or "PSUs") from the core set of data employed in this program of research activities. As noted in (Eckart et al., 2021), classic statistical learning methods are not well suited to the task of finding pattern in large high dimensional datasets. Whereas various graph ML methods, NLP and related approaches are better suited to the task of working with the "rich and unwieldy" data contents employed in the study. As a result, this literature review focused on work that employs methods that fall under the broad umbrella of ML procedures, with an emphasis on methods that employ algorithms that build and tune models iteratively.

Abundance and Scarcity of Published Work in ML-Derived Supports for Effective Service System Operations

The objectives of the work presented in this paper are ultimately practical. However, the research also seeks to advance methodological knowledge more broadly. The goal is to supply a methodology that addresses a pronounced gap in an otherwise very large body of work that employs various Machine Learning (ML) methods with health datasets, to promote better care.

This gap in the literature is covered in (K. Moselle et al., 2024), that proposes a simplified model within the health domain that loosely groups a diverse array of Machine Learning-derived information products (ML “Knowables”) into nine layered elements that extend from the intracellular “omic” layer up to the population epidemiological level – see Figure 2. Noting the positioning of CPG-relevant analytics in this scheme (layers 6), the research work reported in this document is located within layer 6, 7, and 8, where the most prominent gap can be noticed. The scheme depicted in figure 1 is abstracted from a review of roughly 270 studies employing Machine Learning with health data.

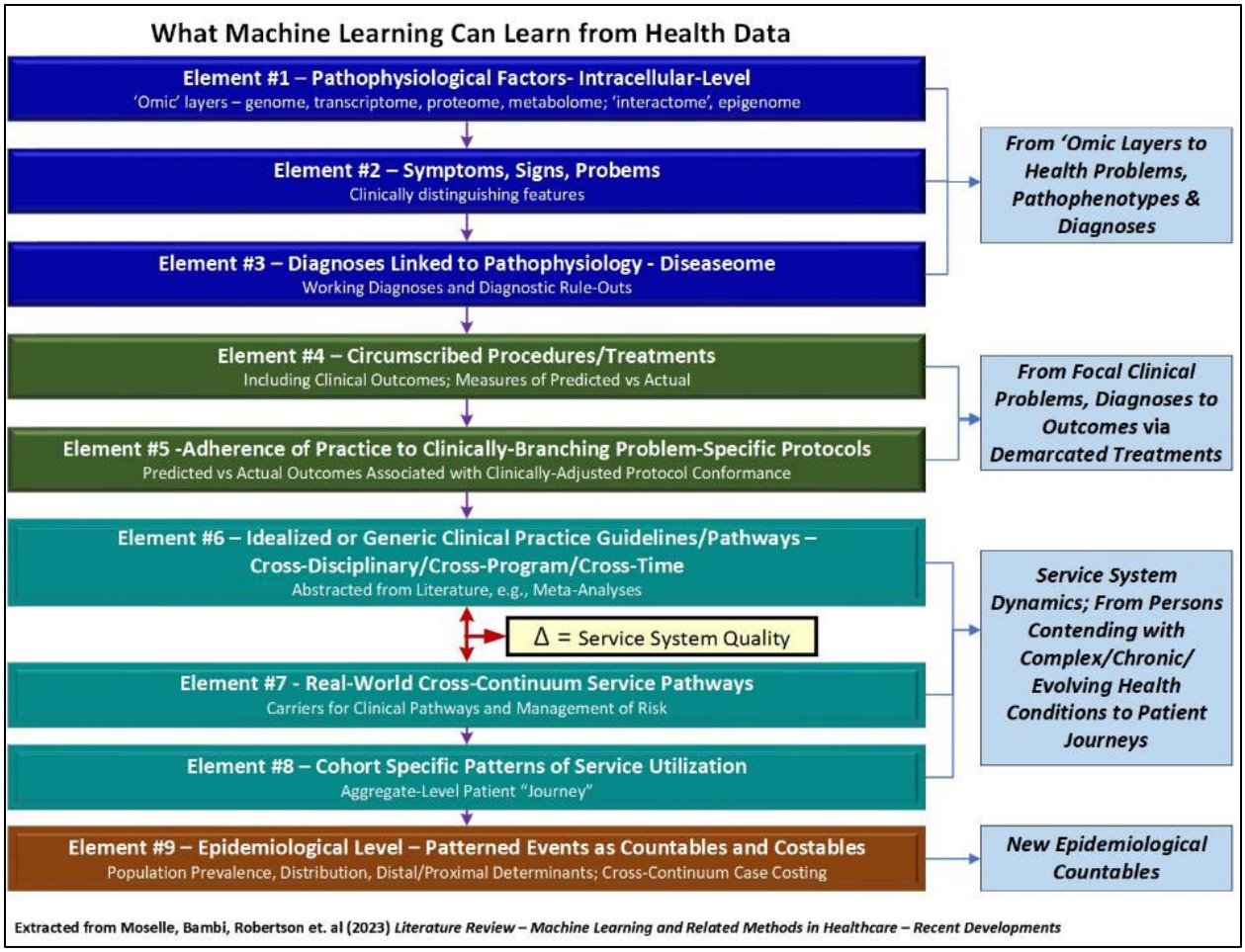


Figure 2 - What can we learn from health data using Machine Learning

To summarize (with a small number of illustrative references):

Element # 1 '-Omic' layers:

These refer to the full range of molecular interactions that can occur at a cellular level, either between or within families (e.g., protein-protein interactions; protein-DNA interactions). Trans-omic models are constructed from contents located at multiple '-omic' layers (e.g., genome, proteome, transcriptome) and describe the connections between genotype and the expression of genotypes in far more complexly structured phenotypic entities, ranging from body structures to disease entities. Graph/network modeling methods are distinctively well-suited to pattern recognition and clinical taxonomic efforts that span the omic levels (Barabási et al., 2017). Details on the use of graph/network methods employed to construct these "trans-omic" entities is provided in (Barabási et al., 2017). The methods outlined in this layer do not relate directly to macro-level QA/QI processes in a health service system. However, the methodological challenges taken up by this body of work are quite substantial, and the graph/network methods developed to address those problems are well tested and understood. Further, this body of

work delivers a mature, and exceptionally generalizable methodology that can be applied to the full range of "knowables" in Figure 1. Some of those methods figure centrally in the analytical workflows associated with some of the studies, e.g., those that employ various graph ML procedures (community detection) to locate cohort-specific patterns of service utilization in sparse high-dimensional service encounter datasets.

Element # 2– Symptoms, signs, problems:

These contents include subjective experiences of the patient (symptoms) and impacts of those symptoms, together with externally observable features that are directly accessible to the diagnostician. A major body of work employing graph-based deep learning is concerned with extracting clinically relevant signals from a large array of sources relating to a diverse array of diagnostic entities. A thorough analysis of this body of work and assessment of potential and future directions is provided in (Ahmedt-Aristizabal et al., 2021). While much of the work compares performance against interpretations made by clinical experts to train algorithms, some of the work is concerned with relative performance of humans vs machine, e.g., (Jaremko et al., 2021).

Element # 3 – Working Diagnoses and Rule-Outs

There are two relevant bodies of published material: (1) work that seeks to extract diagnoses from free text-based documents, and (2) work that seeks to establish a diagnosis or identify cases based on material contained in a patient record. Regarding the first category, there is a large literature employing Natural Language Processing (NLP) methods. Many of these works are concerned with extracting discrete diagnoses or creating labelled datasets for supervised machine learning procedures from free-text radiology reports (Banerjee et al., 2017; Elkin et al., 2008). Similar work has been carried out with other types of source free-text documents to extract categories of information that are quite distinct from what would be featured in radiology reports, e.g., health-risk behaviors from mental health records (Stewart & Velupillai, 2021). From a purely pragmatic standpoint – even if humans can outperform various NLP or related machine learning procedures in extracting diagnoses or diagnostically-relevant information from text, real-world "accuracy" of a diagnostic impression formed by a clinician in the course of delivering care is conditional upon the proportion of text-based documents that are read. In evaluating the potential for NLP-based methods applied to free-text to promote most informed-possible care decisions, the capacity for such methods to read and report out on the entire record is perhaps a relevant factor. This body of work provides good proof-of-concept support for the NLP methods used in one of our studies.

Element # 4– A Procedures, Treatments, Expected Outcomes

Moving up from Element # 3 to Element #4, NLP methods may be used to identify procedures or treatments that were performed, using free text or other source documents, NLP and other ML methods may also be used to determine effectiveness of procedures, or to identify treatments (e.g., molecular-level interventions) that are more/less likely to produce clinical benefit. Additionally, there is a substantial body of work undertaken and reported recently that employs network medicine methods to support the personalized medicine agenda. This agenda seeks to create clinical phenotypes anchored in processes taking place at a molecular level or organ or body level, and target interventions to those process. Work in the field spans a range, from precision medicine at a pathophysiological/molecular level e.g. (Rost et al., 2016), to work focused on specific conditions, including a large literature on machine-learning-based approaches to cancer care (Alabi et al., 2022; Carlisle et al., 2015; Ge et al., 2019; Hase et al., 2017; Nakagawa & Fujita, 2018), celiac disease (Piccialli et al., 2021), diabetes (Gökçay Canpolat & Şahin, 2021), and allergic disease (Shamji et al., 2023).

With regard to treatment successes vs failures - successful treatment for a problem may be explained in terms of a relatively small number of factors associated with the health condition and the actions that the treatment has been shown to have on that condition. The space of factors to be considered in understanding treatment failures is far more open and unbounded. Further to this issue – when it comes to failures associated with protocol-based care (see Element #5 directly below) the problem may be compounded by the fact that protocols typically cover an array of procedures that are enacted over time, often in response to response to earlier stages in the unfolding of a condition. Machine learning procedures, most notably unsupervised relatively more "assumption-free" procedures may be particularly well-suited to the task.

Element # 5– Problem-Specific Protocols – and Expected Outcomes:

The focus here is on problems which may require an array of interventions, particularly when there are multiple etiologic factors involved in the production of arrays of related diagnostic entities. Outcomes associated with care that conforms/does not conform to protocols have been extensively studied using various classic statistical methods, e.g., (Pike et al., 2013) for work concerned with protocol-based care for sepsis. However, the literature becomes quite thin with regard to use of ML approaches to determine whether care conforms to protocols, or to evaluate outcomes associated with care that conforms to protocols. With regard to outcomes, ML methods are being used to estimate risk for outcomes or predict outcomes, including risk for re-hospitalization (Norman et al., 2017; Orangi-Fard et al.), and psychiatric readmission (Rumshisky et al., 2016).

Limited application of ML methods to look at protocol-based care may reflect factors unrelated to the methods. In particular, given the possibly multiple components associated with protocols (or guidelines, see directly below), it may be unclear what constitutes conformance to a protocol and what constitutes a departure. As well, given that the protocols may be applied to conditions that are chronic, where trajectories may change or different co-morbidities may emerge, it may not be clear what constitutes success. So long as the research agenda using ML approaches is driven largely by researchers external to health service systems that encompass the range of services and outcomes associated with such populations, the clinical expert knowledge may be slow to connect with very high levels of data scientist expertise – outside of healthcare.

Element # 6 – Clinical Guidelines/Clinical Pathways

Clinical practice guidelines consist of structured sequences of clinical interventions (Panteli et al., 2019). (Rotter et al., 2019) further stipulate that a clinical pathway consists of a translation of generic clinical practice guidelines into processes taking place with local health service system structures. In other word, clinical pathways are clinical practice guidelines translated into local service system terms (Rotter et al., 2019). ML and related procedures have been used to provide visibility into factors located on care pathways that predict key interventions located on the pathway, e.g., use and speed of thrombolysis in acute response to stroke (Allen et al., 2019).

In theory, ML procedures could be employed with health data to evaluate impacts of guideline-based care, or to work backwards from outcomes to guidelines. However, ML studies of clinical pathways are subject to the same set of challenges referenced above with regard to clinical protocols. Indeed, the challenges are probably greater because protocols are coupled to problems, treatments, and outcomes. Guideline-based care is methodologically and conceptually a more elusive constructs to try to operationalize. To determine which particular pathway through a protocol would be appropriate for a given patient, large volumes of patient information might be required. Further – the relevant data may be contained in multiple sources that are not necessarily linked – or the clinically meaningful data required to operationalize the application of protocol-based care or evaluate impacts may not be readily available.

As well, people who are receiving some variant of guideline-based care, particularly when it relates to chronic disease, may be contending with a progressively more elaborate set of problems which may reflect quite distinctive pathophysiological mechanisms, calling for the introduction of other clinical guidelines.

Element # 7 – Service Pathways

Service pathways are "real-world" depictions of activities that actually take place following a clinical pathway within a local array of health services. These pathways are keyed to problems that do not lend themselves to complete resolution at any particular service unit and are therefore embodied as networks of interactions of patients with networks of providers who are associated with service units. These Service Pathways may then be assembled into collections at a patient-level to reflect their point-in-time and longitudinal health profiles, the local contexts of their lives (including environ-mental factors and distal/non-medical/social determinants of health), local service system capacity and operational characteristics, and possibly changing population-level "competition" for access to scarce services. Hence, service pathways consist of cohort-specific predictable recurring patterns of service utilization that actually take place within a local service system (Aggarwal et al., 2020; Carroll & Richardson, 2019; Huo et al., 2017).

Use of graph modeling approaches to characterize networks of interacting providers is well-studied. What is generally missing from the work is an effort to relate these networks of interacting providers to clinical outcomes. As well, the emphasis is largely on characteristic patterns of interaction among provider entities – patterns of service utilization. What is more sparsely represented in the literature is an effort to characterize groups of persons – or communities of persons if derived via graph community detection methods – or an effort to then discern what are the underlying clinical features of those people who track to distinctive service system "epigenetic pathways" across the service system landscape.

As such, the clinical relevance or significance of the work for the systems-level quality agenda is limited. Where the researchers take on that challenge, the results are methodologically significant. See for example (Huo et al., 2017): "*Guidelines of CRC, such as those published in the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology, do exist. However, demonstrated by our study, real world practices are vastly different. Understanding these derivations is critical for reducing time needed for patients to receive appropriate care. SNA is a promising approach in exploring patterns of care pathways.*" What these authors are pointing to is the potential for graph ML (e.g., Social Network Analysis) methods to highlight the separation between real-world Service Pathways and more idealized Clinical Pathways. Service Pathways, reflecting real-world dynamics of service system operation, are positioned as carriers for clinical pathways. Reducing the gap between Service Pathways and Clinical Pathways is one way of operationally defining "quality improvement" at a health service system level.

Element # 8 – Patient Journeys

Assembled from one or more Service Pathways. They reflect the interaction of the person with a service system as they contend with possibly multiple problems, associated with bounded episodes of care or changing personal need (Lin et al., 2020).

Element # 9 - – Epidemiological aspects

Treating processes (e.g., PSUs) as "countables" to estimate demand and measure impacts of efforts to alter service system dynamics (Rose, 2020).

There are very large numbers of studies covering Elements # 1-5, where the focus is on discrete diagnostic entities and associated procedures or protocols. The picture changes when the focus shifts to Element # 6, where the core unit of analysis is CPG adherence. There will generally be large numbers of clinical trials supporting each of the component recommended practices associated with each stage in the treatment of a chronic condition or with different branches in an array of trajectories common to a disease. These clinical trials form the evidentiary foundations for evidence-based CPGs. However, what is largely missing in the ML literature is work that operationalizes the construct "CPG-adherence" and evaluates the impacts of such adherence.

This thinning of the ML literature is equally apparent within the domains set out by Elements # 7 and 8, where the focus is on locating patterns of service events that span the health service system. This is also the case for Element # 9, which requires products of Elements # 7 and 8 to supply new trans-diagnostic "countables".

One factor can at least be identified that could contribute to this clearly discernible trailing off of work in an otherwise very comprehensive literature: if benefits of CPG-based care for complex or chronic problems are at least partially emergent characteristic of adherence at all stages of disease progression within clinically complex entities, then studies would need to access very diverse longitudinal bodies of clinical features of persons, treatments and procedures, related longitudinally to a broad array of service entities, linked at a person level. Within this inevitably sparse and very high-dimensional space, every case is likely to be distinguishable. Based on well-established principles of statistical disclosure control (El Emam & Arbuckle, 2013) virtually every case would be regarded in principle as a carrier of risk for re-identification. Use of perturbative methods such as differential privacy (Bambauer et al., 2013; Xu et al., 2017), that alter truth of the source data must be ruled out because they require the results of analyses of unperturbed data to demonstrate that analytical integrity has not been compromised (Bambauer et al., 2013). Given the above, and associated limitations in real-world public access to the required data (Malin & Goodman, 2018) the literature covering Elements # 6-9 is very thin.

Chapter 1: Collection Narrative

Research Objectives

For persons with chronic and typically increasingly complex and impactful conditions, an increasingly diverse array of services must engage, while the person's capacity to surmount access challenges is becoming progressive more limited. For such persons, "quality" emerges dynamically as an attribute of service pathways that span the continuum. In these cases, evidence-based quality improvement requires tools that can look out across the continuum and discern patterns of service utilization in very large bodies of cross-continuum service encounter data.

One cannot improve processes if one cannot model them. By that same token, decision-makers in service systems cannot make evidence-informed decisions if they are not supplied with models that embody and render transparent the dynamics that govern outcomes for populations who depend on the system for needed services. One cannot model processes and communicate those models in understandable form to various clinical oversight and governance bodies if one cannot see them in the data. Hence, the pragmatic intent of this work is to supply ready-for-use evidence from local service system operations to support more effective service for persons with complex/chronic health problems, under conditions of expanding population demand but limited budget. Focusing on PSUs, the work to be presented relies on extracting useful information, from contents accumulating clinical information systems to optimize service system structure and function on behalf of patients contending with chronic disease. The intent is to build out a set of methods and generate a set of PSUs-related products that can be put in the hands of parties within a healthcare organization who perform planning and clinical quality oversight functions.

The work presented in the various manuscripts attached to this program of research activities (see figure 3), are organized around operationalization of the construct “clinical quality” for a given cohort as the degree of separation of real-world PSUs from idealized CPG-adherent clinical pathways. Hence, locating PSUs within inevitably sparse, high-dimensional arrays of patient-service encounter data becomes a cornerstone for the research activities conducted herein. Following this, are questions raised in each of the manuscripts that are closely applicable to quality assurance/quality improvement activities within a complex service system working under conditions of fiscal constraint to meet the needs of populations with complex problems. Below is a summary of each of the manuscripts with focus

on the objective of the manuscript, the research questions, methods, and key findings. More details can be found in each corresponding manuscript.

Prior to providing the summary of each manuscript, it is important to emphasize the role that data pre-processing played in making the data analysis ready. Given the complexity and spread of services that constitute a longitudinal cross continuum dataset used for this program of research activities, the issue of nomenclature and data granularity needed to be addressed. Given 2000+ Service Units represented in the CIS of the organization whose data was used for the research, a scheme organized around six set of codes constituting a semantic layer was used. This scheme, referred to as the Clinical Context Coding Scheme (CCCS) (Koval & Moselle 2018) was applied to the 2000+ Service Units to generate approximately 200 equivalent Service Classes to address both the nomenclature and data granularity issues. Each Service Class was assigned a name that bears some discernible relationship to the functions performed by the component Service Units. Hence, all the modelling activities perform in all the studies reported in any of the manuscript used Service Classes. It should be noted that this scheme was not empirically generated but was rather based on input from Clinical Subject Matter Expert (SMEs), as well as Service System Operation Experts (SSOEs).

BioMedInformatics and Knowledge journals were chosen the publication of the manuscripts below. This choice was discussed and supported by the supervisory committee. They offer open access publication model, which will improve the dissemination of our work to the academic community.

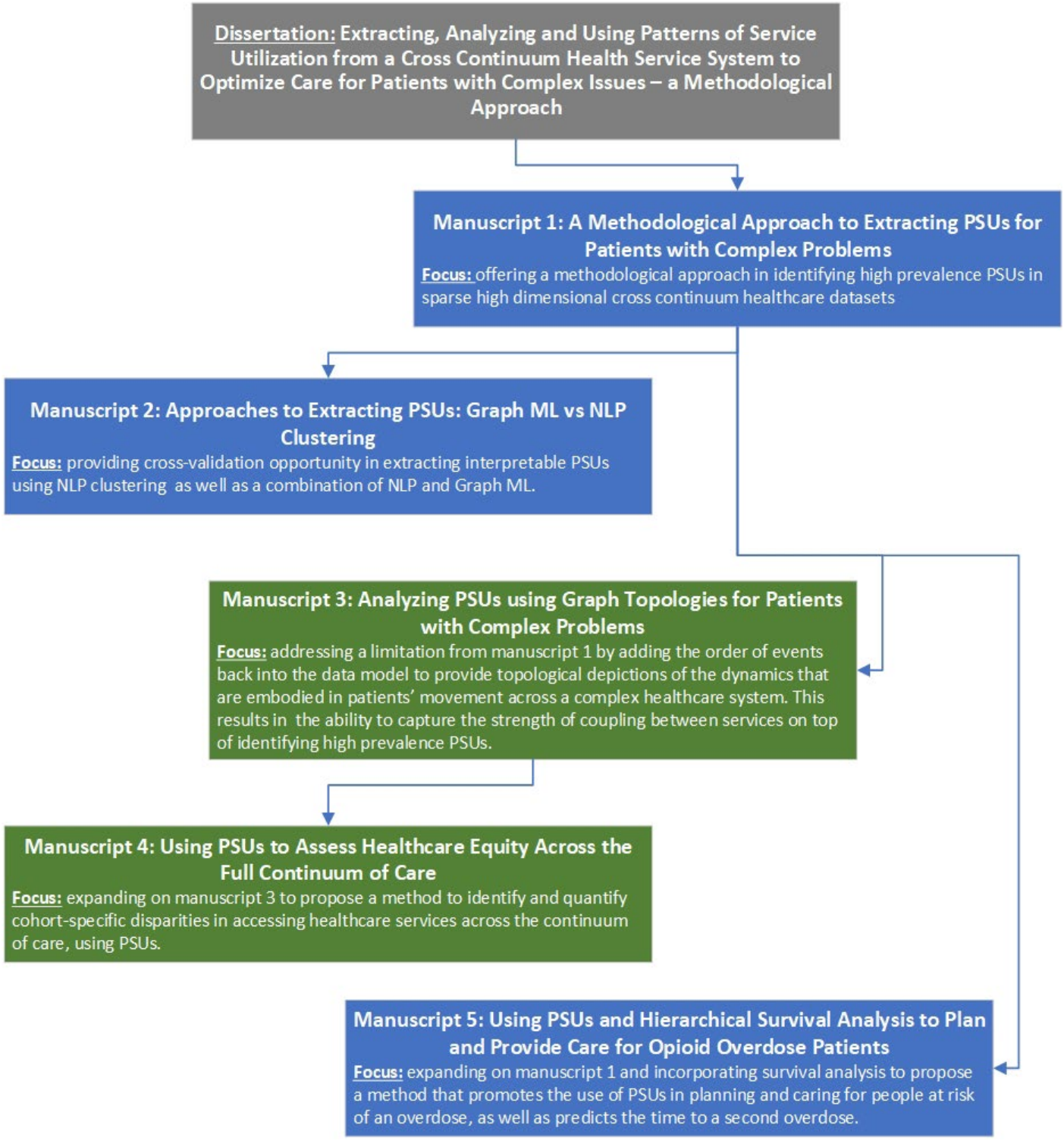


Figure 3 - Relationship Between Manuscripts in Addressing Care to Patients with Complex Problems Using PSUs

Manuscripts Summaries and Key Findings

Manuscript 1: Methodological Approach to Extracting Patterns of Service Utilization from Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems

In this manuscript, the focus was on offering a methodological approach in identifying high prevalence PSUs in sparse high dimensional cross continuum healthcare datasets. The goal was to extract meaningful functionally integrated PSUs to influence the reorganization of some services to provide better care for patients with complex problems. To achieve this, the following questions were asked:

1. What mechanism can be used to address the cross-continuum data granularity and nomenclature issues to generate intelligible dataset that can be analyzed?
2. For cohorts with large volumes of interactions with diverse arrays of services spanning the continuum, can graph machine learning methods (community detection) be employed to extract clinically understandable clusters of services (PSUs), which reflect distinctive needs?
3. Methodologically, what mechanism can be used to determine the optimal number of communities?
4. Within a given community of services, can one separate out those services that reflect common features of cohorts, such as need or risk, versus those services that are keyed to variable features of persons within cohorts? Stated in slightly different terms, can one separate out services that "belong" in communities versus ser-vices that are forced into one community or another by the community detection algorithms?
5. Can one generate results that are readily and correctly interpretable by persons who do not have a background in statistics, research, or data science?

Using a sample cohort of patients with complex problem, the proposed methodology demonstrated that the use of a semantic layer -CCCS to reduce data dimension, resulted in generating dimensionally -reduced and functionally-named Service Classes as equivalent representations of services locations. This was followed by the encounter data being re-engineered as a bi-partite graph consisting of nodes with edges connecting patient to Service Classes. Next a bi-partite projection into services was applied to the bi-partite graph, followed by the application of a "nested iterative" graph community detection, to

generate communities of services. Finally, using graph metrics, combined with input from clinical and operational subject matter experts, it was possible to extract meaningful functionally integrated PSUs.

However, one limitation of the proposed model was the fact that it is atemporal, the events are collapsed across time and the order of the events are not taken into consideration. Although it was possible to highlights the prevalence of connection between services, using the edge weights to assess the strength of prevalence, the proposed model failed to capture the strength of coupling between services as well as the order of events. This is what the manuscript 3 aimed at addressing.

Additionally, this research introduced a novel analytical framework relying on patients' service pathways -PSUs as a foundation to generate the basic entities required to evaluate conformance of interventions to cohort-specific clinical practice guidelines. This was further explored in subsequent manuscripts. Finally, given the novelty of the construct relying on PSUs to support and drive QA/QI initiatives to optimize services system structures and functions to address the needs for patients with complex problems, this introduced the need for the extraction of PSUs to be validated using other approaches. Manuscript 2 addressed this issue.

Manuscript 2: Approaches to Extracting Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs NLP Clustering

As patients interact with a healthcare service system, patterns of service utilization (PSU) emerge. These PSUs are embedded in sparse high-dimensional arrays of longitudinal cross-continuum health service encounter data. As previously mentioned, once extracted, PSUs have the potential to provide quality assurance/quality improvement (QA/QI) efforts with information required to optimize service system structures and functions, to improve outcomes for complex patients with chronic diseases. Using the method proposed in the first manuscript, it was possible to extract meaningful functionally integrated PSUs. This paper built on the findings of the first manuscript to provide an opportunity for cross-validation in extracting these interpretable PSUs. In this manuscript, approaches using (1) Natural Language Processing (NLP) clustering method, and (2) a hybrid NLP-Graph method, are compared to the "nested iterative" graph community detection methods used in manuscript one to check if the extracted PSUs are comparable.

To address the above, the following questions were asked:

1. To what extend can NLP methods be used to extract PSUs from longitudinal heterogeneous cross continuum healthcare data? How do the data need to be modeled and what data pre-processing needs to be done to generate the base data upon which the NLP methods can be applied?
2. Are the results from NLP clustering for Service Classes similar to those obtained using graph community detection? Are they judged to be similar by clinical subject matter experts (SMEs) or clinical/administrative service system operations experts (SSOEs)?
3. Does a hybrid NLP-Graph community detection approach generate meaningful results, and how do the results compare to (a) community detection results, with simple frequency-based edge-weighted projections of service-service interactions, and (b) results obtained using NLP-based clustering approaches, employing measures of cosine similarity between vectors reflecting patient journeys?

The result demonstrated that the various approaches produced comparable results, from a clinical perspective, as judged by clinical subject matter experts and service system operations experts. The similarity in results provided an opportunity for cross-validation in generating interpretable PSUs.

Moreover, similar to the finding in manuscript one, the finding in this manuscript stressed (1) the need to use the semantic layer -CCCS to reduce the dimension of the data to conduct any meaningful analysis and (2) the need to engage with clinical or service system operations experts, both in interpreting the results, but also in determining what level of granularity is the most effective in producing useful results. Finally, given the capabilities and rapid expansion of NLP-based methods, this provided future opportunities to take advantage of NLP capabilities, such as prediction using PSUs. These will be explored in future studies. However, the limitation encountered in manuscript one with regards to the models failing to capture the strength of coupling between services as well as the order of events still applied. This limitation was addressed in manuscript three.

Manuscript 3: Analyzing Patterns of Service Utilization Using Graph Topology to Understand the Dynamic of Engagement of Patients with Complex Problems with Health Services

In manuscript one and two, graph machine learning and NLP were used to capture the dynamics governing interactions between health services and persons with complex problems. However, the data model used collapsed patients' encounters events across time, representing the service-system

dynamics as atemporal entities. This resulted in a model that was able to highlight the prevalence of connection between services but failed to capture the strength of coupling between services as well as the order of events. In this manuscript, the order of evens was put back into the data model to provide topological depictions of the dynamics that are embodied in patients' movement across a complex healthcare system. The goal was to reveal patients' dynamics from a cross continuum complex healthcare dataset in such a way that their depictions is clear in informing the effort to change patients' outcome by altering those dynamics. To address this, the questions below were asked:

1. Can patient journeys be recorded as sequences of services encounters, using a directed graph that can then be used to identify high-prevalence sequences within and across persons?
2. With the proposed methodology, to what extend can one cut across the complexity of a cross continuum service structure to capture the dynamics of the journey of patients with complex issues that clearly portray their engagement with the ser-vice system, to help locate potential operational problems?

With this method, once the semantic layer was applied to the encounter data to address the data granularity and nomenclature issues, the generated equivalent Service Classes could be arranged using a "next service" relationship. This refers to a scenario whereby if patient A is accessing Service Class 1 (SC1), followed by Service Class 2 (SC2), then 3, and again 3; such a sequence will be depicted as directed graph connecting with directed edges connecting SC1 to SC2 to SC3, and a loop to SC3. When this is applied to all the patients within a cohort, a dense connected graph can be generated. Following this, various resolutions can be applied to highlight the most prevalent connections.

The results showed that directed graphs are well suited to the task of depicting the way that diverse components of the system are functionally coupled – or remain disconnected - by patient journeys. Additionally, by setting the resolution on the graph topology visualization, important characteristics can be highlighted, including high prevalence repeating sequences of service events. These can be interpreted readily by clinical subject matter experts.

With the proposed method, areas of improvement can be identified, by providing visibility into the dynamics of patients' engagement, to optimize care for patients with complex problems, or chronic conditions that may rely on multiple services that may not necessarily be optimally connected. Finally, the methodology provided a first step in addressing the challenge of locating potential operational problems for patients with complex issues engaging with a complex healthcare service-system.

Note in particular, the three visualizations on page 103 of this dissertation:

- The first depicts the most prevalent next-service-event sequences for a cohort of persons contending with severe psychiatric illnesses such as schizophrenia. The services are interconnected in this visualization, in keeping with service models that effectively "wrap" multiple services around the patient to promote effective adherence to medication regimens, management of potential challenges around diet/nutrition, maintaining a place of living, and effective management of other challenges to health and safe independent living outside of facility settings. See information available from Island Health regarding the full range of services provided by Island Health (Mental Health & Substance Use Resources, 2024). Many of these services appear as nodes in the first visualization on page 103.
- The third visualization on that page depicts the dense interconnection of services that are all accessed by a cohort of persons in Island Health who are contending with heart failure or other forms of very serious cardiovascular illness. The interconnections are fully in keeping with CPGs related to heart failure, e.g., CCS/CHFS heart failure guidelines (McDonald et al., 2021). Note that two of the authors of those heart failure CPGs are physician leaders in Island Health (Dr. Elizabeth Swiggum and Dr. Michael Chan). They are well positioned to provide clinical leadership to ensure that multiple components of the service system conform to the CPGs as set out in those heart failure guidelines.
- The second visualization reflects the interconnection of services within Island Health for a cohort of persons who have survived an opioid overdose, as per documentation of Island Health Emergency Department encounters. There may very well be clinical practice guidelines that relate to operation of services located on the periphery of that hub and spoke (with no rim) model. The absence of a rim in this graph topology is a direct indicator of a relatively low degree of coupling of services along the periphery. This may reflect absence of clinical practice guidelines to dictate how the service should operate. It may reflect inadequate resourcing to meet demand, reflected in delays in accessing services on that periphery, and resulting in access to other "fallback" services – such as low-barrier Emergency Departments. It may reflect both.

When quality of care is impacted by the dynamics of patients' movement through the healthcare service system, the methods outlined in this manuscript may supply information and products in supporting the effort in addressing quality issues. They can be very useful tools in the hand of QA/QI to help inform services delivery changes and optimization initiatives, based on local realities. In manuscript 4 and 5, we expanded on the concept of PSUs to: (1) provide a nuanced approach in assessing healthcare equity using patterns of service utilization across the full continuum of care, and (2) plan and provide care for overdose patients and predicting the time to a second overdose using patterns of service utilization and hierarchical survival analysis.

Manuscript 4: Disparities in Access to Services, as Evident in Patients Journeys: Illustrating a Nuanced Approach in Assessing Healthcare Equity Using Patterns of Service Utilization Across the Full Continuum of Care

In this manuscript the concept of PSUs was expanded upon to assess healthcare equity. Healthcare organizations have a contractual obligation to the public to address population-level inequities to health services access and shed light on them. Various studies have focused on achieving equitable access to healthcare services for vulnerable patients. However, these studies do not provide a nuanced perspective based on local reality across the full continuum of care. In our previous work, graph topology was used to provide visual depictions of the dynamics of patients' movement across a complex healthcare system. Using patients' encounters data represented as a graph, this study expanded on the work in manuscript 3 to proposes a methodology to identify and quantify cohort-specific disparities in accessing healthcare services across the continuum of care. The goal was to provide a more nuanced understanding of healthcare access disparity.

Using a cross-continuum healthcare dataset from a regional health authority, and focusing on a chosen set of cohorts of patients with varied levels of vulnerabilities. For this use case scenario, the degree of vulnerability refers specifically to differential capacity to initiate access to any services other than low-barrier access services, e.g., emergency departments. The term "vulnerability" also refers to limited capacity to remain connected to services over time. If a cohort is "vulnerable" in these two related senses, then they would also be vulnerable to the health risks associated with inadequate access to those services.

The following related questions were considered:

1. Determinants of access: what pattern of access to health services are associated with the chosen cohorts of interest?
2. To what extent can we use PSUs to identify healthcare access inequities and provide a more refined approach in assessing the cause of the inequity?

For this study, once the semantics layer CCCS was applied on the encounter data, and the generated equivalent Service Classes modelled using a graph, logistic regression was applied to estimate transition probabilities in state space models. Using a cohort comparison approach, followed by consultation with SMEs, the review of the results demonstrated that a more nuanced approach to assessing access-to-care-disparity was doable using patients' patterns of service utilization from a longitudinal cross continuum healthcare dataset.

Hence, any organization that wishes to structure their services to provide better care to its vulnerable population, based on local realities, can use the proposed approach as part of their toolkit in assessing the services that need to be restructured or integrated. Moreover, this provided a first step in addressing inequities for vulnerable patients in accessing healthcare services.

Although the information generated from the proposed methodology can be valuable knowledge in the hands of Quality Assurance/Quality Improvement in guiding organizations to implement a more equitable healthcare system, additional steps need to be considered to fully address these inequities. These include (1) the understanding additional factors that may affect timely access to appropriate healthcare services, such as patient's distal determinant of health, and social determinant of health, and (2) the undertaking of specific, localized, and measurable actions developed and sustained through ongoing engagement with the communities, supportive leadership, dedicated resources, accountability, and transparency. To achieve this, further analyses need to be conducted involving various clinical and social SMEs, and additional dataset beyond patients encounters data. This is a potential future study.

Manuscript 5: Use of Patterns of Service Utilization and Hierarchical Survival Analysis in Planning and Providing Care for Overdose Patients and Predicting the Time-to-Second Overdose

In this manuscript, a method that promotes the use of PSUs in caring for people at risk if an overdose was proposed. This was a continuation of previous works where Graph Machine Learning was used to model the products for planning and evaluating the treatment of patients with complex issues. The goal was to rely on PSUs to provide a more targeted response that goes beyond the traditional taxonomical diagnosis approach, in caring for people at risk of an opioid overdose and prevent subsequent overdoses.

Using graph community detection and survival analysis, and relying on encounters data collected from a host organization Clinical Information System, the following questions were asked:

1. Using PSU, to what extent can we determine whether the opioid overdose cohort is homogeneous or not with respect to determinants of risk?
2. How many communities constitute the opioid overdose cohort, based on how patients within this cohort interact with the host organization service system?
3. To what extent can we determine the risk of a subsequent opioid overdose, based on the community an opioid overdose patient belongs to, and quantify it, using survival analysis?

Similar to previous studies, using a semantic layer -CCCS, a dimensionally-reduced and functionally-named Service Classes were generated. These are equivalent representations of services locations. This was followed by the encounter data being re-engineered as a bi-partite graph consisting of nodes with edges connecting patient to Service Classes. Next a bi-partite projection onto persons was applied to the bi-partite graph. Applying a graph community detection using Louvain algorithm on the projected graph clustered opioid overdose cohort into various communities, based on their patterns of service utilization across the continuum of care. These communities demonstrated that the overdose cohort is not homogeneous with respect to risk. Moreover, focusing on the generated communities of patients and applying survival analysis showed that the risk for a second overdose among these communities is not only different, but could also be quantified.

The proposed method can help inform a treatment heterogeneity approach that is likely to be more efficient for a cohort made of diverse individuals such as the opioid overdose cohort. This can provide an opportunity for the healthcare service system to apply a more targeted approach to care that is likely to be more efficient for each of the communities constituting the overdose cohort. It can also guide targeted support for patients at risk of subsequent overdoses. Hence, providing such information to a healthcare organization will not only equip the organization with required information to provide better

care for the opioid overdose patients, but also be pro-active in preventing subsequent overdoses, by providing targeted support for the patients at a risk of subsequent overdoses.

Other distal and social determinant of health play a role in determining the risk of a second overdose. However, data representing these determinants are distributed across multiple systems, and a given health service system may have access to only a portion of those data. Additionally, information supplied by medical anthropologists may be important - to provide more information on the behavioural features of persons that engender risk, e.g., how/where they acquire drugs, how they use drugs. Combining such dataset with PSUs is likely to provide a more complete picture of the risk factors to provide an even better prevention mechanism for subsequent overdoses. Such potential study is worth a consideration.

Afterword

Concluding Remarks

If one operationalizes the construct “clinical quality” for a given cohort as the degree of separation of real-world PSUs from idealized CPG-adherent clinical path-ways, then one would need to accomplish the following in order to build out a foundation for evidence-supported QA/QI: (1) locate PSUs within inevitably sparse, high-dimensional arrays of patient-service encounter data; (2) demonstrate predictable relationships between PSUs and outcomes; and (3) generate counts of both CPG-compliant and non-compliant processes for demand estimators and performance monitors. The work presented in this study was able to lay out methodological foundations for such a program of health-service-system analytics by identifying service pathways that are clinically interpretable. In future studies, the methodology will need to be expanded to devise (1) an approach that can be used to associate the interpretable extracted PSUs with outcomes for persons and for service system operations, and (2) an approach to accurately identify and count CPG-compliant and CPG-non-compliant processes.

Moreover, In the process of developing a methodology for extracting and identifying service pathways that are clinically interpretable, several key findings emerged. These are detailed in the respective manuscript, and below is a short summary:

1. *The importance of a semantic layer*: to produce an analysis-ready version of encounter data, a semantic layer needs to be applied to the dataset. The roles that such a semantic layer play

include: (1) perform the initial phase of the dimension reduction to address data granularity issues which is likely to be the case for most healthcare organizations, given the spread and high number of location involved in providing services that cover the full continuum of care, and (2) attach consistent functional name to the derived dimensionally-reduced location names, to address nomenclature issues. It should be noted that the above steps are not empirically derived but are generated in collaboration with service systems experts. For the host organization whose data was used for the study, the semantic layer – CCCS was created with the help SSOEs and was used to reduce for number of Service Units from 2000+ units to about 200 equivalent Service Classes that were used throughout the analysis conducted for this study. A semantic layer, similar to the CCCS is likely to be require for any organization that want to adopt the methodologies proposed herein.

2. *Macroscope analogy:* There is no underlying truth regarding a given PSUs associated with cohorts that the methods are correctly or incorrectly detecting. It is helpful to think of the methods presented in this study as a macroscope that provides visibility into patterns that are located in sparse high-dimensional datasets – patterns located in a space that is too complex for them to be detected by SMEs without the assistance of pattern detection/construction tools.
3. *The importance of partnership with clinical representatives:* referred herein as SMEs and SSOEs, they represent a crucial component of the proposed methodology. Their involvement is not sporadic, but more of a partnership. From the creation of a semantic layer for data pre-processing, to the selection of the cohorts of interest, to the analysis and interpretation of the results, they play a central role, along with the data scientists, statisticians, and computer science members of the team.

Additionally, the extraction and identification of high-prevalence and repeating PSUs, along with other associated finding such as (1) the use of PSUs to illustrate a more nuanced approach in assessing healthcare equity, and (2) the use of PSUs in planning and providing a more heterogeneous and targeted care for patients affected by the Opioid calamity have significant implication on healthcare organization operations support and planning. These implications were highlighted in the respective manuscripts and include: (1) the possibility of influencing the reorganization of some services within healthcare organization service structure, in order to provide better care for vulnerable patients with complex problem, (2) supporting the initial step in addressing the challenge of locating potential operational problems for patients with complex issues engaging with a complex healthcare service system, (3) supporting the identification of areas of improvement for healthcare services delivery, by providing

visibility into the dynamics of patients' engagement, to optimize care for patients with complex problems, or chronic conditions that may rely on multiple services that may not necessarily be optimally connected, (4) supporting a first step in addressing inequities for a healthcare organization, by providing visibility into services accesses disparities that impact vulnerable patients or patients with complex problems, (5) providing information to support a more heterogeneous and targeted care for Opioid overdose patients, and (6) providing information that can equip healthcare organizations to be pro-active in preventing subsequent overdose for opioid overdose patients.

Several limitations were also identified that will need to be addressed in future studies. These were also captured in the respective manuscript and only a summary is included here. The focus of the study is methodological. These methodologies are universal. However, the application of the methodologies proposed in this study need to be adjusted to local realities. Hence, the engagement of local SMEs, QA/QI, and buy-in from senior leadership is a vital requirement for the uptake and implementation of the proposed methodologies. Other limitations are directly related to the fact that more work need to be done to expand on various finding identified during the study. These include: (1) Expanding on the findings from the topological depiction of patients' dynamic engagement with services to identify potential operational problems for patients with complex problems engaging with a complex healthcare system. Such a study will need to engage various clinical SMEs and SSOEs, as well as incorporate information collected from various clinical guidelines to identify the problems and potential solutions. (2) Expanding on the findings from the use of PSUs to identify and measure cohort-specific disparities in service access, and including additional steps, to address inequities in healthcare, based on local realities. Such a study will need to consider additional factors such as distal and social determinant of health, as well as ongoing engagement with the communities, supportive leadership, dedicated resources, accountability, and transparency. (3) Building on the findings of using PSUs combined with survival analysis, a study that incorporate distal and social determinant of health into the analysis may provide a more complete picture in providing the most optimal care and support for Opioid overdose patients.

A final potential future study that was not captured is any of the manuscripts is the re-expression of patient journeys in terms of interactions with Communities of services, one level up from the Service Classes, to identify phases in journeys, e.g., persons who are actively using and making use of harm reduction services, persons who are engaged in rehab/recovery services. For chronic/relapsing

conditions such as addictions to many drugs, this level of dimensional reduction of service data may help to identify factors that position persons at or close to critical junctures in their journeys, e.g., the transition from harm-reduction "journey" to a rehab/recovery journey.

Next Steps – Implementation Recommendations and Considerations for PSUs Methodologies

The methodologies outlined in this work are intended to be scalable and extensible to a range of different health issues that fall within the scope for any health service system that is responsible for the full range of major health issues that affect populations, e.g., any regional health authority.

However, their application needs to be tailored to the needs, risks and services provided to different cohorts. Their application also needs to be adjusted to local realities within the service system that is responding to those cohort-specific needs. These local realities include the structures and associated functions of a local health service system, the standard nomenclatures or labels used to refer to service entities and services provided, as well as the technologies that are used to capture information about patients and the services that are provided. Full implementation of the methodologies entails implementation of a host of activities that extend well beyond the scope of work covered in this dissertation.

Note in particular that the dissertation focuses on methods that rely heavily on formative engagement of subject matter experts to determine when pattern recognition tools have generated products that accurately reflect the characteristics of cohort members and the services provided – by the local system. Though the subject matter expert substantive engagement is critical and necessary, it does not provide a sufficient set of assurances that the products derived via machine learning methods have been correctly interpreted and are suitable for application. A more complete set of procedures would need to ensure the following: (1) appropriate cohort definition and (2) appropriate validation of results.

Cohorts Definition.

Cohorts may be defined in terms of features that reflect characteristics of persons or features that reflect the interaction of persons with a service system. For example, a cohort could be defined in terms of a diagnostic profile, e.g., heart failure with comorbid kidney disease. This approach may be effective when there is a fairly consistent mapping between the diagnostic profile and the services provided by a demarcated set of services.

In cases where populations accessing a service are heterogeneous with respect to diagnostic profile, or where service access does not map predictably onto access to a demarcated set of services, it may then be more effective to define the cohort in terms of access to services. For example, a person could be admitted to a detox facility for an addiction to diverse array of substances, e.g., alcohol, or opioids. As well, many people who are heavy users of alcohol or opioids will never enter a detox facility. In this case, depending on the intended use of the products, e.g., demand estimation for detox facilities, the modeling would be done in terms of a cohort that all shared one feature in common – access to detox. A relatively complete methodology would need to be able to work with cohorts defined in either of these two ways.

Validation of results.

As noted above, the input from SMEs is a necessary but not sufficient condition for deeming a set of products as an accurate reflection of a process that exists in the real world, or for interpreting the processes highlighted by the data, e.g., deeming a process as a problem, or deeming a process as one to be promoted. A methodology that provides more substantive validation would consist of a set of components that would address the following issues: (1) are the derived product methodologically cogent, (2) are the products accurately reflecting some underlying reality, and (3) are the products useful. Usefulness can be judged from various perspective, including patients care, service system structure and alignment with organizational strategic priorities. Specifically, to determine if the derived products are methodologically cogent and reflect some underlying reality, a chain of activities outlined below need to occur:

- Engagement with clinical informatics: these are the resources within the organization that have a good understanding of how patients and care providers interact at the point of care, what technologies they use and how events and their knowledge of patients and those events is captured in source systems, including clinical information systems. This is a crucial step, as the source data, and the data structure used by the methodologies outlined in this work, need to be well understood to adjust the semantic layer to reflect the local reality.
- Engagement with the technical team in charge of data infrastructure: these are the resources in charge of extracting the data out of the clinical information systems used to capture patients' interaction with the service system, transform them and load them into an environment that is accessible and usable for secondary use. The Extract Transform Load (ETL) processes need to be

documented and reviewed, and any alteration of the data at source needs to be well understood. This is important, as the source data used by the methodology (as part of secondary use of the clinical data) need to be tracked back to the source (primary use of clinical data) for validity testing. Stated in slightly different terms – the method must be able to track the lineage of the data from source to the data in the form that it exists in warehoused environments, and the form in which the data emerge in the final analytical products.

- Engagement of clinical SMEs and SSOEs: an iterative engagement with this group in analyzing the data and generating products is one of the corner stones of the proposed methodology. This engagement needs to be both formative and summative, as the products generated need to be clinically meaningful and understandable, both in detail and as a composite set of products that reflect critical features of cohorts and cross-continuum response.
- Engagement with people with lived experience: any products concerned with health processes involving patients requires engagement with people with lived experience. This is critical to ensure that SMEs evaluation of the processes corresponds with their experience, as patients or stakeholders.
- Use of validation frameworks/techniques: when applicable, a construct validation framework, such as multi-trait, multi-method approaches, concurrent or predictive validation (Cronbach & Meehl, 1955) can be applied. Moreover, if applicable, multiple technical approaches/algorithms can be used to generate the targeted products and assess their similarities.
- Engagement with leadership: the products need to be presented to leadership to evaluate understandability and clarity, in term of the implication and impacts of the products/recommendations on the service system structure. These consumers of products include parties responsible for services demand estimation and services planning. These leaders need to be supplied both with information that describes processes, and with information that provides insights into the factors that govern the dynamics that underly persistent processes. This information about causes or determinants (or at least correlates) is critical if leadership is going to take evidence-informed steps to alter processes and outcomes.
- Engagement with the privacy and data protection representatives: given that the proposed methodology and associated products is introducing a new paradigm shift, sharing of local experience between various healthcare organizations and the academic community need to be promoted and sustained. This is the only way the new paradigm can slowly transition into the new normal, for the benefit of patients with complex/chronic problems, and the overall health

of our community. Persons with expertise around privacy legislation need to be engaged, together with persons who have expertise in the area of statistical disclosure control, using the process outlined in (Hundepool et al., 2010) as an example.

Next Steps - Foundation for a Learning Health System Framework

There are many definitions attached to the learning healthcare system (LHS) paradigm. One of the definitions states: “A learning healthcare system is one that is designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care” (Olsen et al., 2007). Announced about two decades ago, much research (Etheredge, 2007; Friedman et al., 2015; Friedman et al., 2010; Greene et al., 2012; Menear et al., 2019) have discussed the promises of this novel paradigm. However, to date, there is no comprehensive report on its implementation to transform healthcare services into an agile and adaptable learning healthcare systems (Budrionis & Bellika, 2016).

In our attempt to embrace LHF paradigm, we are proposing the concept of a Target Information Architecture (TIA), as a missing link for LHS framework in facilitating its implementation. The health service system being too big and too complex, modelling it may too challenging. However, by properly modelling the layered structures and information-dependent functions of the health service system, we can learn how integrated patterns of service system operations can be modified to free the system from its self-imposed restrictions. Hence, the TIA becomes the mind for a LHS, a way of constructing, describing, and modelling the paradigm. It becomes the blueprint that can be translated into building plan to construct and implement an LHS.

In this model that we are promoting, we are putting forward something that scales out to the challenges of bringing dynamics of cross-continuum service system operations into focus. Hence, the work conducted in this study provides one of the crucial building blocks in building the foundation required for a learning healthcare system. It provides one of the analytical workflows that is used to populate the target information architecture with local evidence-based results.

When it comes to the secondary use of data collected in health organizations CIS, the analytical workflows can travel along two pathways. They reflect two broad paradigms associated with large classes of products. They also reflect analytical approaches, and importantly, they imply a certain way of thinking about what is being modeled by the data. As shown in figure 4 (k. Moselle et al., 2024) , one of

the paradigms relies on pre-existing, externally derived coding standards that are grafted onto the local data source. This approach has been adopted as the de-facto approach in analyzing healthcare data and has been widely adopted by many healthcare organizations. The second paradigm is an assumption-free approach that assumes no prior knowledge transaction and relies on learning via pattern recognition methods applied to local data. This approach is a novel approach in consuming secondary healthcare data to generate knowledge for healthcare organization and has been the focus of our study.

The two approached are not opposite to one another, but rather complimentary. We are at the beginning of the journey in enabling the population of the TIA with local evidence-based result to facilitate the implementation of an LHS. The work conducted in this study provides an important starting point.

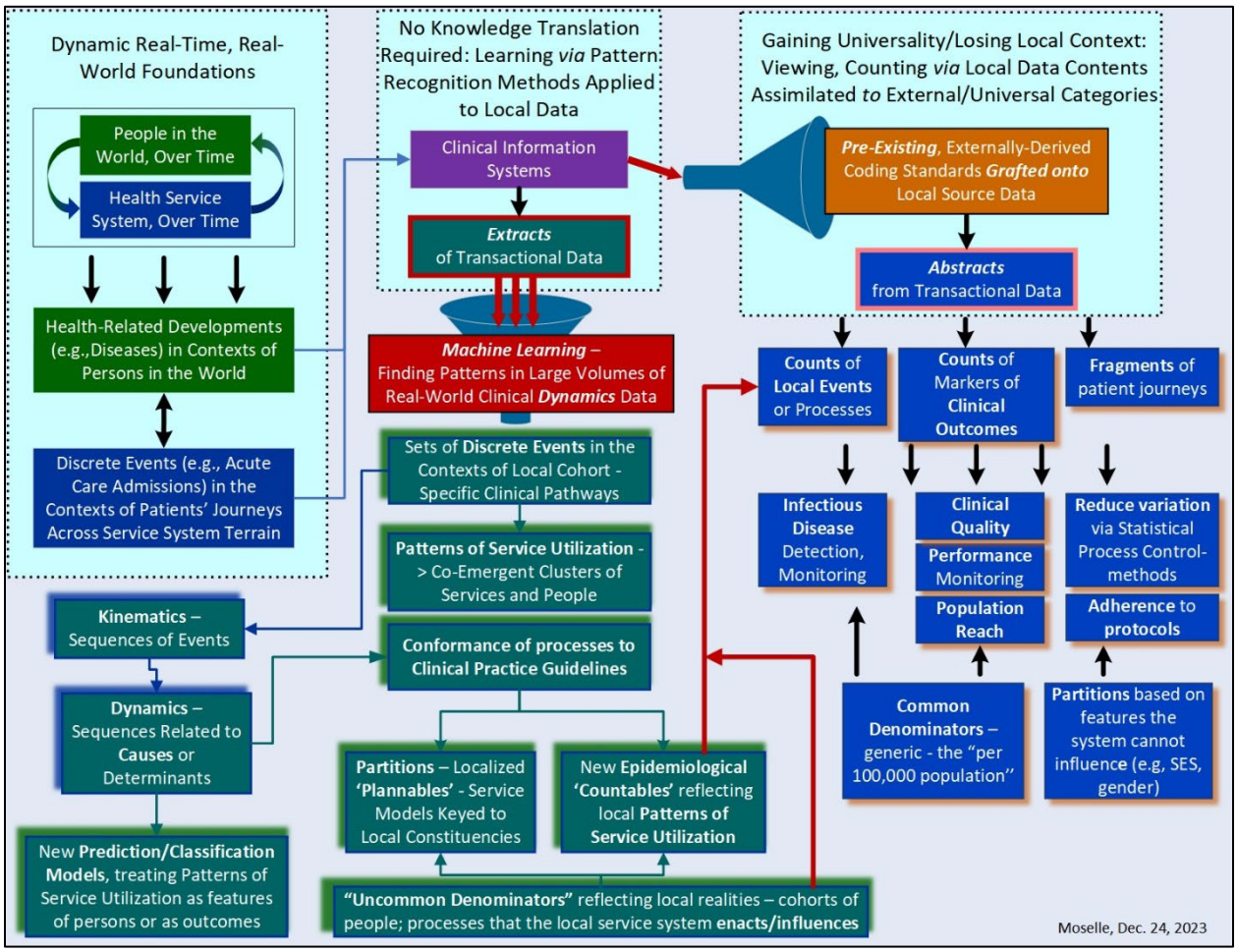


Figure 4 - Two Analytical Paradigms for Health Service Data

Reference

Aggarwal, N., Ahmed, M., Basu, S., Curtin, J. J., Evans, B. J., Matheny, M. E., . . . Shah, R. U. (2020). Advancing artificial intelligence in health settings outside the hospital and clinic. *NAM perspectives*, *2020*.

Ahlgrim, N. S., Garza, K., Hoffman, C., & Rommelfanger, K. S. (2019). Prodromes and Preclinical Detection of Brain Diseases: Surveying the Ethical Landscape of Predicting Brain Health. *eNeuro*, *6*(4), ENEURO.0439-0418.2019. https://doi.org/10.1523/ENEURO.0439-18.2019

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., & Petersson, L. (2021). Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors*, *21*(14), 4758.

Alabi, R. O., Almangush, A., Elmusrati, M., & Mäkitie, A. A. (2022). Deep machine learning for oral cancer: from precise diagnosis to precision medicine. *Frontiers in Oral Health*, *2*, 794248.

Allen, M., Pearn, K., Monks, T., Bray, B. D., Everson, R., Salmon, A., . . . Stein, K. (2019). Can clinical audits be enhanced by pathway simulation and machine learning? An example from the acute stroke pathway. *BMJ open*, *9*(9), e028296.

Baker, J., Travers, J. L., Buschman, P., & Merrill, J. A. (2018). An Efficient Nurse Practitioner-Led Community-Based Service Model for Delivering Coordinated Care to Persons With Serious Mental Illness at Risk for Homelessness [Formula: see text]. *Journal of the American Psychiatric Nurses Association - Journal Article*, *24*(2), 101.

Bambauer, J., Muralidhar, K., & Sarathy, R. (2013). Fool's gold: an illustrated critique of differential privacy. *Vand. J. Ent. & Tech. L.*, *16*, 701.

Banerjee, I., Madhavan, S., Goldman, R. E., & Rubin, D. L. (2017). Intelligent word embeddings of free-text radiology reports. AMIA annual symposium proceedings,

Barabási, A.-L., Loscalzo, J., & Silverman, E. K. (2017). *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Harvard University Press.

BC_Ministry_of_Health. (2024). *BC Guidelines*. Retrieved from https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines

Budrionis, A., & Bellika, J. G. (2016). The learning healthcare system: where are we now? A systematic review. *Journal of biomedical informatics*, *64*, 87-92.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature methods*, *15*(4), 233-234. https://doi.org/10.1038/nmeth.4642

Carlisle, A., Caceres, I., Mehta, S., Schindler, J., & Sharma, J. (2015). A combined machine learning and bioinformatic analysis approach identifies biological pathways that predict clinical stage and survival outcome in neuroblastoma patients. *Cancer Research*, *75*(15_Supplement), 3758-3758.

Carroll, N., & Richardson, I. (2019). Mapping a careflow network to assess the connectedness of connected health. *Health informatics journal*, *25*(1), 106-125.

Christiani, A., Hudson, A. L., Nyamathi, A., Mutere, M., & Sweat, J. (2008). Attitudes of Homeless and Drug-Using Youth Regarding Barriers and Facilitators in Delivery of Quality and Culturally Sensitive Health Care. *Journal of child and adolescent psychiatric nursing*, *21*(3), 154-163. https://doi.org/10.1111/j.1744-6171.2008.00139.x

Commission on Social Determinants of, H. (2008). *Closing the gap in a generation: health equity through action on the social determinants of health: final report of the commission on social determinants of health*. World Health Organization.

Cradock-O'Leary, J., Young, A. S., Yano, E. M., Wang, M., & Lee, M. L. (2002). Use of General Medical Services by VA Patients With Psychiatric Disorders. *Psychiatric Services*, *53*(7), 874-878. https://doi.org/10.1176/appi.ps.53.7.874

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Eckart, L., Eckart, S., & Enke, M. (2021). A brief comparative study of the potentialities and limitations of machine-learning algorithms and statistical techniques.

Eckart, L., Eckart, S., & Enke, M. (2021, 2021). A brief comparative study of the potentialities and limitations of machine-learning algorithms and statistical techniques.

El Emam, K., & Arbuckle, L. (2013). *Anonymizing health data: case studies and methods to get you started*. " O'Reilly Media, Inc.".

Elkin, P. L., Froehling, D., Wahner-Roedler, D., Trusko, B., Welsh, G., Ma, H., . . . Brown, S. H. (2008, 2008). NLP-based identification of pneumonia cases from free-text radiological reports.

EPI in British Columbia. (2020). BC Early Psychosis Intervention Program. Retrieved October 2020 from https://www.earlypsychosis.ca/epi-in-british-columbia/

Etheredge, L. M. (2007). A rapid-learning health system: what would a rapid-learning health system look like, and how might we get there? *Health affairs*, *26*(Suppl1), w107-w118.

Friedman, C., Rubin, J., Brown, J., Buntin, M., Corn, M., Etheredge, L., . . . Stead, W. (2015). Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *Journal of the American Medical Informatics Association*, *22*(1), 43-50.

Friedman, C. P., Wong, A. K., & Blumenthal, D. (2010). Achieving a nationwide learning health system. *Science translational medicine*, *2*(57), 57cm29-57cm29.

Ge, L., Chen, Y., Yan, C., Zhao, P., Zhang, P., & Liu, J. (2019). Study progress of radiomics with machine learning for precision medicine in bladder cancer management. *frontiers in Oncology*, *9*, 1296.

George, M., Maheshwari, S., Chandran, S., Manohar, J., & Sathyanarayana Rao, T. (2017). Understanding the schizophrenia prodrome. *Indian Journal of Psychiatry*, *59*(4), 505-509. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_464_17

Georgousis, S., Kenning, M. P., & Xie, X. (2021). Graph deep learning: State of the art and challenges. *IEEE Access*, *9*, 22106-22140.

Greene, S. M., Reid, R. J., & Larson, E. B. (2012). Implementing the learning health system: from concept to action. *Annals of internal medicine*, *157*(3), 207-210.

Gökçay Canpolat, A., & Şahin, M. (2021). Glucose lowering treatment modalities of type 2 diabetes mellitus. *Diabetes: from Research to Clinical Practice: Volume 4*, 7-27.

Hacker, K., Auerbach, J., Ikeda, R., Philip, C., & Houry, D. (2022). Social determinants of health—an approach taken at CDC. *Journal of public health management and practice*, *28*(6), 589-594.

Hase, T., Ghosh, S., Palaniappan, S. K., & Kitano, H. (2017). Cancer network medicine. *Netw Med*, 294-323.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J.,…Wolf, P. (2010). Handbook on statistical disclosure control. *ESSnet on Statistical Disclosure Control*.

Huo, T., George Jr, T. J., Guo, Y., He, Z., Prosperi, M., Modave, F., & Bian, J. (2017). Explore Care Pathways of Colorectal Cancer Patients with Social Network Analysis. *Studies in Health Technology and Informatics*, *245*, 1270-1270.

Iezzoni, L. I., McCarthy, E. P., Davis, R. B., & Siebens, H. (2000). Mobility impairments and use of screening and preventive services. *American Journal of Public Health*, *90*(6), 955-961. https://doi.org/10.2105/AJPH.90.6.955

International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). (2023). Centers for Disease Control and Prevention. Retrieved April 1, 2024 from https://www.cdc.gov/nchs/icd/icd-10-cm.htm

Jaremko, J. L., Felfeliyan, B., Hareendranathan, A., Thejeel, B., Vanessa, Q.-L., Østergaard, M., . . . Maksymowych, W. P. (2021). Volumetric quantitative measurement of hip effusions by manual versus automated artificial intelligence techniques: An OMERACT preliminary validation study.

Jaremko, J. L., Felfeliyan, B., Hareendranathan, A., Thejeel, B., Vanessa, Q.-L., Østergaard, M., . . . Maksymowych, W. P. (2021, 2021). Volumetric quantitative measurement of hip effusions by manual versus automated artificial intelligence techniques: An OMERACT preliminary validation study.

Kawachi, I., & Berkman, L. (2000). Social cohesion, social capital, and health. *Social epidemiology*, *174*(7), 290-319.

Ken Moselle, J. B., Stan Robertson, Yudi Santoso, Abraham Rudnick. (2024). *Target Information Architecture (TIA) - the Missing Link in Learning Health Systems Frameworks*.

Koval , A., & Moselle , K. (2018). *Clinical Context Coding Scheme - Describing Utilisation of Services of Island Health between 2007-2017* Conference of the International Population Data Linkage Association, Banf, Alberta.

Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, *8*(6), e66341.

Lasko, T. A., & Mesa, D. A. (2019). Computational phenotype discovery via probabilistic independence. *arXiv preprint arXiv:1907.11051*.

Laursen, T. M., Munk-Olsen, T., & Vestergaard, M. (2012). Life expectancy and cardiovascular mortality in persons with schizophrenia. *Current opinion in psychiatry*, *25*(2), 83-88. https://doi.org/10.1097/yco.0b013e32835035ca

Laursen, T. M., Nordentoft, M., & Mortensen, P. B. (2014). Excess Early Mortality in Schizophrenia. *Annual Review of Clinical Psychology*, *10*(1), 425-448. https://doi.org/10.1146/annurev-clinpsy-032813-153657

Lin, Z., Yang, D., & Yin, X. (2020). Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access*, *8*, 156663-156676.

Malin, B., & Goodman, K. (2018). Between access and privacy: challenges in sharing health data. *Yearbook of medical informatics*, *27*(01), 055-059.

McDonald, M., Virani, S., Chan, M., Ducharme, A., Ezekowitz, J. A., Giannetti, N.,…Lepage, S. (2021). CCS/CHFS heart failure guidelines update: defining a new pharmacologic standard of care for heart failure with reduced ejection fraction. *Canadian Journal of Cardiology*, *37*(4), 531-546.

Mental Health & Substance Use Resources. (2024). Island Health. Retrieved 2024 from https://www.islandhealth.ca/learn-about-health/mental-health/mental-health-substance-use-resources

McVeigh, S. E. (2020). Sepsis management in the emergency department. *Nursing Clinics*, *55*(1), 71-79.

Menear, M., Blanchette, M.-A., Demers-Payette, O., & Roy, D. (2019). A framework for value-creating learning health systems. *Health research policy and systems*, *17*, 1-13.

Moselle, k., Bambi, J., Robertson, s., Santoso, Y., Rudnick, A., & Chang, E. (2024). *Target Information Architecture (TIA) - the Missing Link in Learning Health Systems Frameworks*.

Moselle, K., Bambi, J., Santoso, Y., Sadri, H. S., Robertson, S., Howie, J., . . . Chang, E. (2024). Abundance and Scarcity of Published Work in Machine Learning DerivedSupports for Effective Service System Operations. In. published: unpublished.

Nakagawa, H., & Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer science*, *109*(3), 513-522.

Norman, C., Van Nguyen, T., & Névéol, A. (2017). Contribution of natural language processing in predicting rehospitalization risk. *Medical care*, *55*(8), 781.

Olsen, L., Aisner, D., & McGinnis, J. M. (2007). The learning healthcare system: workshop summary.

Orangi-Fard, N., Akhbardeh, A., & Sagreiya, H. (2022). Predictive model for icu readmission based on discharge summaries using machine learning and natural language processing.

Panteli, D., Legido-Quigley, H., Reichebner, C., Ollenschläger, G., Schäfer, C., & Busse, R. (2019). Clinical practice guidelines as a quality strategy. *Improving healthcare quality in Europe*, 233.

Piccialli, F., Calabrò, F., Crisci, D., Cuomo, S., Prezioso, E., Mandile, R., . . . Auricchio, R. (2021). Precision medicine and machine learning towards the prediction of the outcome of potential celiac disease. *Scientific Reports*, *11*(1), 5683.

Pike, F., Yealy, D. M., Kellum, J. A., Huang, D. T., Barnato, A. E., Eaton, T. L., . . . Weissfeld, L. A. (2013). Protocolized care for early septic shock (ProCESS) statistical analysis plan. *Critical Care and Resuscitation*, *15*(4), 301-310.

Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., & Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of biomedical informatics*, *58*, 156-165.

Rose, S. (2020). Intersections of machine learning and epidemiological methods for health services research. *International Journal of Epidemiology*, *49*(6), 1763-1770.

Rost, B., Radivojac, P., & Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. *FEBS letters*, *590*(15), 2327-2341.

Rotter, T., de Jong, R. B., Lacko, S. E., Ronellenfitsch, U., & Kinsman, L. (2019). Clinical pathways as a quality strategy. *Improving healthcare quality in Europe*, 309.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, *6*(10), e921-e921.

Shamji, M. H., Ollert, M., Adcock, I. M., Bennett, O., Favaro, A., Sarama, R., . . . Fontanella, S. (2023). EAACI guidelines on environmental science in allergic diseases and asthma–leveraging artificial intelligence and machine learning to develop a causality model in exposomics. *Allergy*, *78*(7), 1742-1757.

Stewart, R., & Velupillai, S. (2021). Applied natural language processing in mental health big data. *Neuropsychopharmacology*, *46*(1), 252.

Xu, C., Ren, J., Zhang, Y., Qin, Z., & Ren, K. (2017). DPPro: Differentially Private High-Dimensional Data Release via Random Projection. *IEEE Transactions on Information Forensics and Security*, *12*(12), 3081-3093. https://doi.org/10.1109/TIFS.2017.2737966

Yuan, H., & Deng, W. (2021). Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning. *Internet Research*, *32*(2), 454-476.

Chapter 2: 1st Manuscript

BioMedInformatics

MDPI

Article

A Methodological Approach to Extracting Patterns of Service Utilization from a Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems

Jonas Bambi [1], Yudi Santoso [2], Hanieh Sadri [2], Ken Moselle [3], Abraham Rudnick [4,*], Stan Robertson [5], Ernie Chang [6], Alex Kuo [1], Joseph Howie [2] , Gracia Yunruo Dong [7,8], Kehinde Olobatuyi [8], Mahdi Hajiabadi [2] and Ashlin Richardson [9]

Citation: Bambi, J.; Santoso, Y.; Sadri, H.; Moselle, K.; Rudnick, A.; Robertson, S.; Chang, E.; Kuo, A.; Howie, J.; Dong, G.; et al. Methodological Approach to Extracting Patterns of Service Utilization from a Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems. *Biomedinformatics* **2024**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Received: 22 February 2024
Revised: 10 March 2024
Accepted: 25 March 2024
Published: date

[1] Department of Health Information Science, Faculties of Human and Social Development, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; jonasbambi@uvic.ca (J.B.); akuo@uvic.ca (A.K.)
[2] Department of Computer Science, Faculties of Engineering and Computer Science, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; y.santoso8@gmail.com (Y.S.); haniehsadri@uvic.ca (H.S.); joehowie@uvic.ca (J.H.); m.hajiabadi67@gmail.com (M.H.)
[3] Department of Clinical Psychology, Faculty of Social Science, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; kmoselle@uvic.ca
[4] Departments of Psychiatry and Bioethics, School of Occupational Therapy, Faculties of Medicine and Health, Dalhousie University, Halifax, B3H 4R2, Canada; Abraham.Rudnick@nshealth.ca
[5] Independent Researcher, Victoria, V8Y 2W3, Canada; stanrobertson@shaw.ca
[6] Retired Physician and Independent Computer Scientist, Victoria, V9C 4B1, Canada; ecsendmail@gmail.com
[7] Department of Statistical Sciences, Faculties of Arts and Science, St. George Campus, University of Toronto, Toronto, M5S 1A1, Canada; gracia.dong@utoronto.ca
[8] Departments of Mathematics and Statistics, Faculty of Science, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; olobatuyikenny@uvic.ca
[9] Predictive Services Unit, Wildfire Service, Province of British Columbia, Victoria, V8W 9V1, Canada; ashlin.richardson@gov.bc.ca
* Correspondence: abraham.rudnick@nshealth.ca

Abstract: Background: Optimizing care for patients with complex problems entails the integration of clinically appropriate problem-specific clinical protocols, and the optimization of service-system-encompassing clinical pathways. However, alignment of service system operations with Clinical Practice Guidelines (CPGs) is far more challenging than the time-bounded alignment of procedures with protocols. This is due to the challenge of identifying longitudinal patterns of service utilization in the cross-continuum data to assess adherence to the CPGs. Method: This paper proposes

a new methodology for identifying patients' patterns of service utilization (PSUs) within sparse high-dimensional cross-continuum health datasets using graph community detection. Result: The result has shown that by using iterative graph community detections, and graph metrics combined with input from clinical and operational subject matter experts, it is possible to extract meaningful functionally integrated PSUs. Conclusions: This introduces the possibility of influencing the reorganization of some services to provide better care for patients with complex problems. Additionally, this introduces a novel analytical framework relying on patients' service pathways as a foundation to generate the basic entities required to evaluate conformance of interventions to cohort-specific clinical practice guidelines, which will be further explored in our future research.

Keywords: clinical pathways; clinical practice guidelines; decision support; graph community
detection; Louvain algorithm; health information management; health service system; machine learning algorithms

---

1. Introduction

1.1. Patterns of Service Utilization (PSUs) for Health-Service-System Optimization

To provide the best possible care to patients with complex needs over time, the service system needs to be optimized. This optimization entails the integration of clinically appropriate problem-specific clinical protocols, and the optimization of service-system-encompassing clinical pathways. With regard to problem-specific clinical protocols, consider the problem of sepsis protocols for emergency departments [1]: these are protocols that are clearly articulated and often coded as clinical decision support tools with clinical information systems. They specify the signs and symptoms that should alert clinicians to the possibility of a patient becoming septic [1]. Using locally available evidence, they specify the diagnostic and the investigation that need to be carried out to perform differential diagnosis and recommend interventions that provide a protocol-based care [1]. Clinical information systems usually contain the data necessary to populate sepsis clinical decision support protocols [1]. Additionally, the sepsis protocols being enacted/or not can be seen within the local data [1]. Hence, clinical operations can be optimized around circumscribed protocols, such as sepsis protocols, and the extraction of aggregated information from transactional clinical information systems can support the effort.

To illustrate optimization encompassing service system clinical pathways, consider the acute care hospitalization and ambulatory care follow-up for persons with schizophrenia. Also, consider optimally interoperating cross-continuum service models that scale up to complexly unfolding chronic conditions that are covered by clinical practice guidelines (CPGs). Following the above, examples of the cross-continuum services that could be required would include (1) an array of services that covers the prodromal phases of a chronic often relapsing condition such as schizophrenia, (2) the acute care hospitalization, (3) an array of post-

discharge stabilization and rehabilitation options, including various arrangements of services including mobile crisis response and psychiatric consultation, (4) an array of progressively more staffing-intensive case management models, (5) various secondary or tertiary residential care options, (6) psychosocial rehabilitation services, and (7) addictions harm reduction or rehab/recovery services for persons with a co-morbid substance use disorder. Also, various services will need engagement to address the various medical comorbidities or emergent conditions usually associated with the schizophrenia condition, such as the engagement of various services to address the risk for kidney disease associated with side-effects of psychiatric medications via their attendant risk for metabolic syndromes [2], or the heightened risk for cardiovascular disease [3]. This level of complexity is not unique to schizophrenia cohorts. There are more than 50 CPGs in the BC guidelines to address high prevalence problems with various degrees of complexities [4].

Optimizing clinical operations around circumscribed protocols may be possible via access to service encounters and related information to determine whether a protocol is indicated, e.g., problems and diagnosis, lab results, together with information about what procedures were performed. With this information input into clinical governance bodies in the service system, operations can be optimized around circumscribed protocols. Standard methods such as statistical process control [5] can also be applied. Optimizing service system operations around CPGs, on the other hand, is far more involved. These CPGs may involve a diverse array of services, assembled into a branching array of protocols whose enactment is conditional upon the clinical, functional and behavioral risk profiles of persons, at any point in time, over time.

Alignment of service system operations with CPGs is far more challenging than time-bounded alignment of procedures with protocols. The challenges arise from at least two sources. The first is concerned with the breadth of information required to know whether the CPG is being enacted in a clinically appropriate manner. If the CPG recruits services that span a full continuum of services, e.g., medical/surgical services for various physical health concerns, mental health services (acute care, ambulatory, residential care, etc.), addictions services, and possibly outreach for homeless persons, given the downward socio-economic mobility of persons with problems such as schizophrenia—access to cross-continuum encounters data from one or more systems is required.

Secondly, even if such data are accessible, there is the foundational challenge when trying to align service system operations around CPGs at a population level: the challenges of identifying longitudinal patterns of service utilization in the data. This include: (1) knowing what was carried out, (2) knowing whether it should have been carried out, and (3) knowing whether outcomes intended by CPGs are being achieved, and if not, why not. Given the number of service entities involved in providing coverage for a complex CPG relative to the number of people who require those services, the relevant data are likely to be distributed quite sparsely in a high-dimensional space.

If we cannot optimize processes we cannot see in the data, then pattern recognition methods must be employed with these sparse, high-dimensional

arrays of continuum and time-spanning health service data in order to identify the patterns. This paper illustrates a method for identifying high-prevalence patterns of service utilization (PSUs) in high-dimensional health service datasets associated with clinically specified sub-populations, e.g., persons with a confirmed diagnosis of schizophrenia. The method is built on a foundation of well-understood graph community detection machine learning methods – Louvain [6]. However, importantly, the methodology employs these community detection methods in a nested, iterative way to yield PSUs that are relatively homogeneous with respect to function and are tied in clearly clinically discernable ways to clinical cohorts.

1.2. Abundance and Scarcity of Published Work in ML-Derived Supports for Effective Service System Operations

The objectives of the work presented in this paper are ultimately practical. However, the research also seeks to advance methodological knowledge more broadly. The goal is to supply a methodology that addresses a pronounced gap in an otherwise very large body of work that employs various machine learning (ML) methods with health datasets, to promote better care.

This gap in the literature is covered in [7], who proposes a simplified model within the health domain that loosely groups a diverse array of machine learning-derived information products (ML "Knowables") into nine layered elements that extend from the intracellular "omic" layer up to the population epidemiological level—see Figure 1. Noting the positioning of CPG-relevant analytics in this scheme (layers 6), the research work reported in this document is located within layer 6, 7, and 8, where the most prominent gap can be noticed. The scheme depicted in Figure 1 is abstracted from a review of roughly 270 studies employing machine learning with health data.

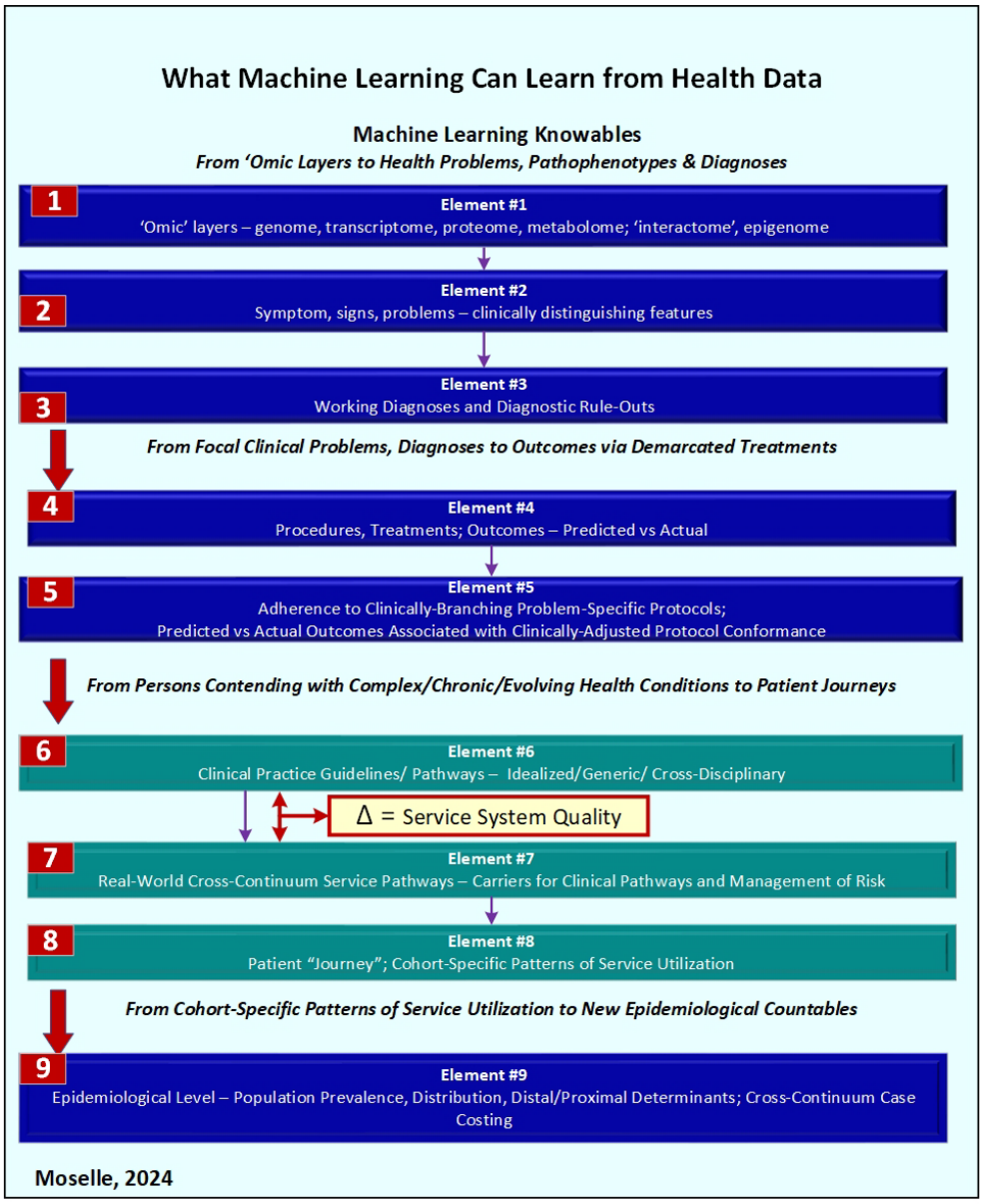


Figure 1. Machine learning "knowables" within the health arena—from 'omics' to epidemiology.

To summarize (with a small number of illustrative references):

Element # 1—'-Omic' layers:

These refer to the full range of molecular interactions that can occur at a cellular level, either between or within families (e.g., protein–protein interactions; protein-DNA interactions). Trans-omic models are constructed from contents located at multiple '-omic' layers (e.g., genome, proteome, transcriptome) and describe the connections between genotype and the expression of genotypes in far more complexly structured phenotypic entities, ranging from body structures to disease entities. Graph/network modeling methods are distinctively well-suited to pattern recognition and clinical taxonomic efforts that span the omic levels [8]. Details on the use of

graph/network methods employed to construct these "trans-omic" entities is provided in [8].

Element # 2—Symptoms, signs, problems:

These contents include subjective experiences of the patient (symptoms) and impacts of those symptoms, together with externally observable features that are directly accessible to the diagnostician. A major body of work employing graph-based deep learning is concerned with extracting clinically relevant signals from a large array of sources relating to a diverse array of diagnostic entities. A thorough analysis of this body of work and assessment of potential and future directions is provided in [9]. While much of the work compares performance against interpretations made by clinical experts to train algorithms, some of the work is concerned with the relative performance of humans vs. machine, e.g., [10].

Element # 3—Working Diagnoses and Rule-Outs:

There are two relevant bodies of published material: (1) work that seeks to extract diagnoses from free text-based documents, and (2) work that seeks to establish a diagnosis or identify cases based on material contained in a patient record. Regarding the first category, there is a large literature employing Natural Language Processing (NLP) methods. Many of these works are concerned with extracting discrete diagnoses or creating labelled datasets for supervised machine learning procedures from free-text radiology reports [11–14]. Similar work has been carried out with other types of source free-text documents to extract categories of information that are quite distinct from what would be featured in radiology reports, e.g., health-risk behaviors from mental health records [15].

Element # 4—Procedures, Treatments, Expected Outcomes:

Moving up from Element # 3 to Element # 4, NLP methods may be used to identify procedures or treatments that were performed, using free text or other source documents; NLP and other ML methods may also be used to determine effectiveness of procedures, or to identify treatments (e.g., molecular-level interventions) that are more/less likely to produce clinical benefit. Additionally, there is a substantial body of work undertaken and reported recently that employs network medicine methods to support the personalized medicine agenda. This agenda seeks to create clinical phenotypes anchored in processes taking place at a molecular level or organ or body level, and target interventions to those processes. Work in the field spans a range, from precision medicine at a pathophysiological/molecular level, e.g., [16], to work focused on specific conditions, including a large body of literature on machine-learning-based approaches to cancer care [17–21], celiac disease [22], diabetes [23], and allergic disease [24].

Element # 5—Problem-Specific Protocols—and Expected Outcomes:

The focus here is on problems which may require an array of interventions, particularly when there are multiple etiologic factors involved in the production of arrays of related diagnostic entities. Outcomes associated with care that conforms/does not conform to protocols have been extensively studied using various classic statistical methods, e.g., [25] for work concerned with protocol-based care for sepsis. However, the literature becomes quite thin with regard to the use of ML approaches to determine whether care

conforms to protocols, or to evaluate outcomes associated with care that conforms to protocols. With regard to outcomes, ML methods are being used to estimate risk for outcomes or predict outcomes, including risk for rehospitalization [26,27], and psychiatric readmission [28].

Element # 6—Clinical Guidelines/Clinical Pathways

Clinical practice guidelines consist of structured sequences of clinical interventions [29]. Rotter et al. [30] further stipulate that a clinical pathway consists of a translation of generic clinical practice guidelines into processes taking place with local health service system structures. In other word, clinical pathways are clinical practice guidelines translated into local service system terms [30]. ML and related procedures have been used to provide visibility into factors located on care pathways that predict key interventions located on the pathway, e.g., the use and speed of thrombolysis in acute response to stroke [31].

Element # 7—Service Pathways

Service pathways are "real-world" depictions of activities that actually take place following a clinical pathway within a local array of health services. These pathways are keyed to problems that do not lend themselves to complete resolution at any particular service unit, and are therefore embodied as networks of interactions of patients with networks of providers who are associated with service units. These Service Pathways may then be assembled into collections at a patient-level to reflect their point-in-time and longitudinal health profiles, the local contexts of their lives (including environmental factors and distal/non-medical/social determinants of health), local service system capacity and operational characteristics, and possibly changing population-level "competition" for access to scarce services. Hence, service pathways consist of cohort-specific predictable recurring patterns of service utilization that actually take place within a local service system [32–34].

Element # 8—Patient Journeys

Assembled from one or more Service Pathways. They reflect the interaction of the person with a service system as they contend with possibly multiple problems, associated with bounded episodes of care or changing personal need [35].

Element # 9—Epidemiological aspects

Treating processes (e.g., PSUs) as "countables" in order to estimate demand and measure impacts of efforts to alter service system dynamics [36].

There are very large numbers of studies covering Elements # 1–5, where the focus is on discrete diagnostic entities and associated procedures or protocols. The picture changes when the focus shifts to Element # 6, where the core unit of analysis is CPG adherence. There will generally be large numbers of clinical trials supporting each of the component recommended practices associated with each stage in the treatment of a chronic condition or with different branches in an array of trajectories common to a disease. These clinical trials form the evidentiary foundations for evidence-based CPGs. However, what is largely missing in the ML literature is work that operationalizes the construct "CPG-adherence" and evaluates the impacts of such adherence.

This thinning of the ML literature is equally apparent within the domains set out by Elements # 7 and 8, where the focus is on locating patterns of service events that span the health service system. This is also the case for Element # 9, which requires products of Elements # 7 and 8 to supply new trans-diagnostic "countables".

One factor can at least be identified that could contribute to this clearly discernible trailing off of work in an otherwise very comprehensive literature: if benefits of CPG-based care for complex or chronic problems are at least partially emergent characteristics of adherence at all stages of disease progression within clinically complex entities, then studies would need to access very diverse longitudinal bodies of clinical features of persons, treatments and procedures, related longitudinally to a broad array of service entities, linked at a person level. Within this inevitably sparse and very high-dimensional space, every case is likely to be distinguishable. Based on well-established principles of statistical disclosure control [37], virtually every case would be regarded in principle as a carrier of risk for re-identification. The use of perturbative methods such as differential privacy [38,39], that alter the truth of the source data, must be ruled out because they require the results of analyses of unperturbed data to demonstrate that analytical integrity has not been compromised [38]. Given the above, and associated limitations in real-world public access to the required data [40], the literature covering Elements # 6–9 is very thin.

1.3. Objectives

The work presented in this paper is organized around the following questions that are directly relevant to quality assurance/quality improvement activities in a complex service system working under conditions of fiscal constraint to meet the needs of populations with complex problems:

What mechanism can be used to address the cross-continuum data granularity and nomenclature issues to generate intelligible dataset that can be analyzed?

For cohorts with large volumes of interactions with diverse arrays of services spanning the continuum, can graph machine learning methods (community detection) be employed to extract clinically understandable clusters of services (PSUs), which reflect distinctive needs?

Methodologically, what mechanism can be used to determine the optimal number of communities?

Within a given community of services, can one separate out those services that reflect common features of cohorts, such as need or risk, versus those services that are keyed to variable features of persons within cohorts? Stated in slightly different terms, can one separate out services that "belong" in communities versus services that are forced into one community or another by the community detection algorithms?

Can one generate results that are readily and correctly interpretable by persons who do not have a background in statistics, research, or data science?

2. Methodological Approach

2.1. Source Data

Source data for the work consists of retrospective longitudinal transactional data contents extracted from a single instance of a Clinical Information System (CIS) deployed across the continuum of services provided by one of the Health Authorities within Canada (hereinafter referred to as "the health service organization" or "host organization"). The span of the health service organization includes almost all secondary and tertiary services for all ages, for persons contending with medical/surgical issues and/or mental health/substance use issues. This includes acute care/intensive care services, hospital and community-based emergency response, ambulatory services, residential care services for older adults or persons contending with mental health issues, case management services, and a range of addictions harm reduction or rehab and recovery-oriented services. The encounter data accessed by this study consists of approximately 10 million encounters over 7 years for approximately 1 million patients. With the exception of a small number of restricted services where data are strictly embargoed (e.g., services for persons who are victims of sexual assault) this represents data for all service recipients. To access the source data, a certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada Ethics harmonization guideline.

2.2. Features Selection

The data used for this study consist of patients encounters data collected over several years by the host organization. The data collected for this study included the following: (1) demographic data: gender, and (2) encounters data: patient identifier (Patient ID), encounter number (encrypted), encounter type, age at encounter, service code, entry code (e.g., via emergency), admit date, discharge date, transfer date, transfer-to, transfer-from, discharge disposition, admit facility, admit building, admit unit name, admit location, and location classifiers.

There are three main activities required to conduct the analysis for the work reported in this document: (1) addressing the nomenclature and data granularity issue, (2) cohort selection, and (3) graph analytics. With regard to addressing nomenclature and data granularity issues, all the location and service-related data including service code, admit facility, admit building, admit unit name, admit location, location classifiers were used to generate Service Class Names and Service Class IDs. At the end of this step all location and service-related fields were replaced by the equivalent Service Class IDs and Service Class Names. More details on this step are provided in the section below. To select the cohort of interest for this study, the Service Class ID, the demographic data and remaining encounters data, including patient identifier, encounter number, encounter type, age at encounter, service code, entry code, admit date, discharge date, and discharge disposition were used. For the cohort of interest chosen for this research, the transfer details were not needed. As a result, the transfer date, transfer-to, and transfer-from fields were not used. During cohort selection process, any of the chosen demographics or encounters data fields can be used as a filter to fine-tune the

cohort selection criteria. More details on the cohort selection process are provided in a subsequent sub-section under the analysis and results section. Finally, to create the bi-partite graph to conduct graph analytics, once the cohort selection was completed, only the patient identifier and Service Class ID were used.

2.3. Data Pre-Processing and Data Re-Engineering—Addressing Nomenclature and Data Granularity Issues

The health service organization consists of an array of programs and services that is architected as roughly 2000+ Service Units within the implementation of their CIS. In modeling the structure and dynamics of patient interaction with services, meaningful distinctions between functions performed by services must be preserved. However, there are issues of spurious or unnecessary granularity that need to be addressed in the raw source encounter data. The term "spurious granularity" refers to Service Units that show up in the data as different entities when they perform identical functions on behalf of cohorts of similar persons. The term "unnecessary granularity" refers to Service Units that have three features: (1) they are identified as distinctive Service Units in the CIS location build; (2) although they are not functionally identical, the distinguishing features are not germane to a particular modeling task at hand; and (3) given the sparseness of the data, it is unlikely that various machine-learning-based clustering procedures will group these services together.

An example of "unnecessary granularity" would be an Emergency Department, which will show up in the CIS location build as an ambulatory treatment area, a trauma bay, a treatment bed, an area named "general", and a checkout area. There might be modeling purposes that require this level of granularity. However, for a more cross-continuum macro-level view of patient encounter histories, this level of granularity may break otherwise-homogeneous patterns of service utilization into fragments, based on where the patient was located for a single Emergency Department encounter.

An example of "spurious granularity" is the presence of 90+ homecare service units in the host organization's CIS location build. For operational and contracts management purposes, these locations need to be preserved as unique entities. However, for the analysis required for this study, these represent only one functional entity that is responsible for dealing with homecare-related services.

Additionally, Unit Names associated with Service Units in the CIS location build are often opaque or uninterpretable. For example, an addiction post-withdrawal stabilization unit appears in this location build as "Holly", or there is a Service Unit named "Clinics" which provides ambulatory services for children and youths with physical disabilities. There are large numbers of Service Units where the Unit Names are uninterpretable, or interpretation is a matter of guessing.

The Clinical Context Coding Scheme (CCCS) [41] was designed as a flexible solution to issues of data granularity and nomenclature. This scheme is organized around six sets of codes, constituting a semantic layer applied to all 2000+ Service Units. The roughly 200 Service Classes employed for the modeling in this paper consist of equivalence classes formed by the application of these code sets to those Service Units. Also, each Service Class

has a name that bears some discernible relationship to the functions performed by the component Service Units. This enables visualizations of patterned entities to be understood, and it also supports the use of any supervised machine learning procedures that require meaningfully labeled data. The modeling activities reported in this paper are performed on patient–service encounters with Service Classes.

2.4. Creating Cohorts to Locate Service System Structures and Functions

The community detection algorithms employed in this work will generate clusters of Service Classes, without regard for the underlying reason for groups of services to co-occur in multiple patient journeys within a cohort. There are two classes of reasons for this co-occurrence; there are cases where services appear as interoperating units because the services collectively perform a distinctive function in a consistent fashion for diverse groups of patients. For example, laboratory services, medical imaging, and emergency departments will co-occur in the records of large numbers of patients who are contending with a very diverse array of problems.

However, the clustering of some services reflects the dynamics of service access by groups of patients, even if the functions performed by component Service Units are not dependent on one another, and/or the component services are located under distinct administrative or clinical management structures within the health service organization. For example, a cluster might emerge that consists of three services: hospital-based emergency departments, addictions medicine specialist consultation services in emergency departments or medical/surgical acute care units, and a community-based-maximum-23-hours-stay shelter for persons who are under the influence of drugs or alcohol. The services are not linked by diagnosis and are not located under a single administrative unit within the health service organization.

The connections between these services represent recurring patterns of cross-continuum service access on the part of select groups of patients, such as homeless persons who are heavy users of various substances and experience a host of physical health problems. When graph methods are employed for other cohorts, e.g., older adults contending with heart failure, the emergency department may show up in a different cluster that includes electro-diagnostic, cardio-vascular treatment and rehabilitation services.

To detect cohort-specific clusters of services, the starting point is the identification of a cohort of concern using an array of clinically characterizable features. These cohorts of concern are identified and defined by Subject Matter Experts (SMEs). Graph community detection algorithms are then executed on the cohort. This enables the identification of services that reflects characteristic functions of the services, compared to cohort-specific clusters that reflect the needs of cluster members and the efforts made by those members or providers to connect those persons with services.

2.5. Generating Communities of Services

The raw data for the work presented in this paper consists of encounter histories for every patient with a history of access to services within the health service organization since 2016—one million people and ten million

encounters. Each encounter contains an anonymized patient identifier, a unique encounter identifier key, date and time stamps, a unique CCCS Service Class ID and Service Class name.

The encounter data are engineered as a bipartite graph consisting of patients and encounters, using nodes with edges connecting patients to Service Classes. A patient is connected to a Service Class when he/she uses the service. Given a bipartite graph, one can perform a bipartite projection onto services. A given pair of Service Classes is connected by a patient when they are both accessed by the same patient. The number of patients who use both Services Classes becomes the weight of the edge connecting those two Service Classes (see Figure 2).

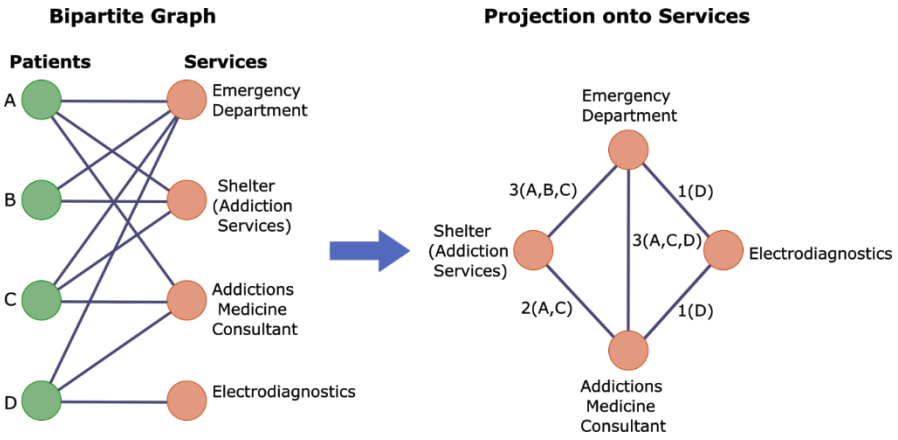


Figure 2. Bipartite graph projection.

After completing the bipartite projection onto Service Classes, we use the Louvain graph community detection algorithm [6] to uncover the grouping of Service Classes that reflect relatively high-prevalence PSUs by patients. There are other well-known clustering algorithms such as Fast-Greedy, Edge-Betweenness, and Leading-Eigen [42]. However, through the analysis conducted, it was found that the Louvain algorithm often produces the most intelligible results.

The Louvain algorithm works by maximizing the modularity value which is defined as

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where A_{ij} is the weight of the edge between node i and node j, k_i is the sum of the edge weights over all the edges that are connected to node i, and m is the total edge weights in the graph. Here, c_i is the label of the community in which node i belongs to. At the beginning each node has its own community. The algorithm starts by randomly choosing a node, then checks other nodes attached to that node to see if merging the communities would result in higher Q.

This algorithm works with the weights associated with all pairs of Service Classes in the bipartite projected graph. They create clusters that maximize the weighted degree of interconnection of Service Classes within a community (in-degrees), while minimizing the degree of interconnection

with other Service Classes (out-degrees). Modularity is a measure that reflects the success of this conjoint optimization of in-degrees and out-degrees.

Community detection methods may be applied in a nested fashion, iteratively within communities generated at a previous iteration (see Figure 3). While conducting the analysis, it was noticed from the clinical perspective that the results were often still too coarse, with many heterogeneous Service Classes clustered together, when nested iterative community detection was not applied. It should be emphasized that the concept of iteration referred to here is not the same as the number of passes in the Louvain algorithm. In the proposed approach, once the Louvain algorithm has generated the first set of communities, each community is isolated and treated as a new graph and the Louvain algorithm is applied again on each of the isolated graphs. This means that each community, once generated, can be considered as a graph by its own and therefore we can apply the Louvain algorithm to it, resulting in smaller sub-communities. Because each iteration results in a finer-grained delineation of service system structures, the total number of communities will increase until communities can no longer be divided any further, i.e., further iterations do not yield a finer-grained delineation of the community structure in the data.

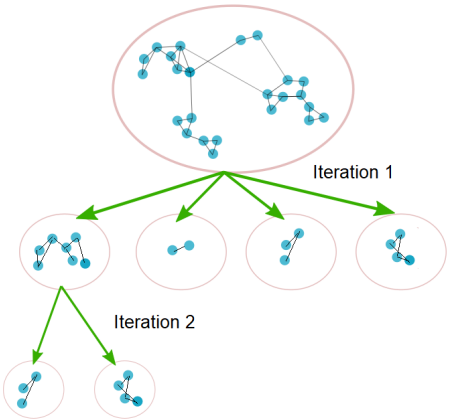


Figure 3. Iterative community detection process.

To demonstrate the iterative community detection process, Figure 4 provides a snapshot of results for an illustrative community detection iteration process. As an example, let us consider a sub-community with Service Classes ID: 1, 2, 3, 14, 15, 30, 42, 81, 150, 168, 238, 248, 249, 251. Suppose this sub-community is one of the communities that were generated as the result of iteration 2. At iteration 3, this sub community splits into two, including sub-community 1, 2, 3, 14, 30, 42, 168, 238, 251, and sub-community 15, 81, 150, 248, 249. At iteration 4, only sub-community 1, 2, 3, 14, 30, 42, 168, 238, 251 is broken further into sub-community 1, 2, 3, 14, 30, 230, and sub-community 168, 151. However, it can be noticed that sub-community 15, 81, 150, 248, 249 remains unchanged at iteration 4. Finally, at iteration 5, we can observe that the community detection algorithm is no longer able to break the last sub communities any further. As a result, the iteration stops at this level and the result is reviewed with a clinical subject matter expert (SME) and a service system operation expert (SSOE) to determine the iteration that provides the result that is most clinically meaningful.

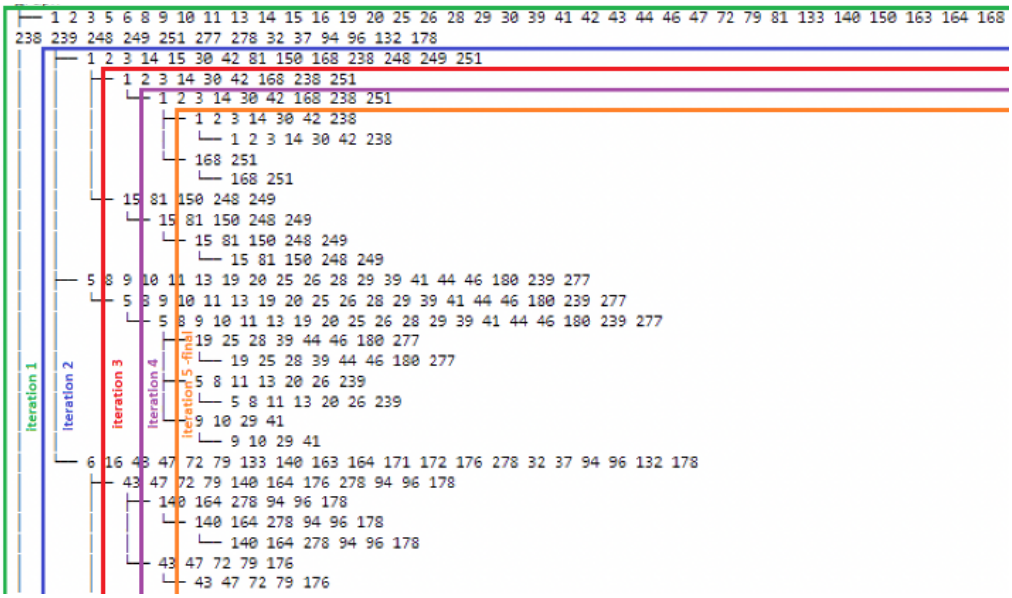


Figure 4. Sample Community Detection Iteration Process.

2.6. Extracting PSUs from Communities of Services

Communities of service that are generated from cross-continuum health service data by unsupervised graph clustering procedures will typically include services that are used by almost all persons within a cohort who interact with any of the services in the cluster. They may also include services that are associated with variable features of fractions of the total group of people who use the services within a community. To identify a core set of services within a community that embody a clinically meaningful function that relates to the needs of a clinically characterizable cohort, the following heuristics was employed:

1. Quantitative criteria using metadata: graph metrics including the graph internal weighted degree, the external weighted degree, and the weighted degree (sum of internal and external weighted degrees) were used, to determine the cut-off point.
2. Qualitative criteria: these include judgments from clinical cohort-specific subject matter experts regarding the characteristics of the cohorts within which the community detection has been run.

3. Analysis and Results

3.1. Analysis Setup: Cohort Creation

As described previously, the CCCS codes were layered onto the raw location data to yield a re-engineered, analysis-ready version of the encounter data. To illustrate the methods and distinctive products associated with the proposed method, a cohort of people contending with schizophrenia was created. This was based on their access to Schizophrenia Services. The schizophrenia cohort was composed of 2008 patients (772 females, 1233 males, and 3 unknown gender), aged between 12 and 87 years, with a range between 1 and 200 interactions. For graph analytics, only two columns are required. Hence, as shown in Table 1, the data representing this cohort used

for graph analytics is only composed of the Patient ID and Service Class ID representing the patients encounters.

Table 1. Fields with sample data representing the schizophrenia cohort format required for graph analytics (2 Nodes).

Patient_ID	Service Class ID
P1	22
P2	34
P3	161
P4	22
P1	13
…	…
P5	243

The tools and languages that were used include R 4.1.3 and Python 3.10.4. The description of the custom R Shinny tool was used to generate/analyze cohorts, as well as the Python code, which can be provided upon request.

3.2. Generating Communities of Services

Going through the iterative community detection process is analogous to the process of separating wheat from chaff. Using the schizophrenia cohort as an example and applying the iterative community detection, several communities of services related to several areas of patients' needs were generated. Some of the communities are made of services that are functionally connected and some are knit together by the features of cohort members.

A total of three iterations were performed: at the first iteration, 4 communities were generated, followed by 10 communities at the second iteration, then 22 communities at the third iteration. After three iterations, the number of communities did not increase above 22, hence meeting the stop criteria.

Tables 2–4 highlight the refinement process. The information presented in the table include Service Class ID (SC_ID), service class name, community ID (CID), internal weighted degree (IWD), external weighted degree (EWD) and weighted degree (WD). One of the four communities (community '1-2') generated at the first iteration was used as an example. As illustrated in Table 2, at the first iteration resolution, community '1-2' is made of a mix of various heterogeneous services. At the second iteration, as illustrated in Table 3, community '1-2' is broken into three communities ('2-2', '2-3', and '2-4') that are gradually becoming homogeneous with regard to rehab recovery and harm reduction treatment services.

Table 2. At the first iteration, community '1-2' (one of the communities chosen for illustration) is made of a mix of various heterogeneous services.

SC_ID	Service Class Name	CID	IWD	EWD	WD
22	MHSU-Addictions-Clinic-Adult-Ambulatory	1-2	369	2158	2527
34	MHSU-Addictions-Clinical Intake-Adult	1-2	367	1615	1982
161	Addictions Medicine Specialist Consultation to Acute Care	1-2	293	2318	2611

23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	1-2	201	682	883
13	MHSU-Assertive Community Treatment (ACT)-Adult	1-2	196	1209	1405
203	Overdose-Related Services	1-2	185	812	997
243	MHSU-Addictions-Rapid/High-Intensity Assessment and Follow-Up	1-2	185	895	1080
21	MHSU-Addictions-Sobering and Assessment Centre	1-2	162	592	754
14	MHSU-Addictions-Outreach and Intensive Case Management-Adult	1-2	144	552	696
29	MHSU-Residential Care-Licensed	1-2	113	1040	1153
24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	1-2	108	351	459
26	MHSU-Residential Care-Lower-Level Support	1-2	108	707	815
10	Tertiary Specialized Residential Care-Adult	1-2	75	338	413
20	MHSU-Rehab Services-Adult-Moderate Intensity	1-2	45	256	301
270	COVID-19 Outreach Assessment	1-2	29	136	165
272	COVID-19 Outreach Assessment Team-Provider	1-2	28	43	71
81	MHSU-Crisis Response-Walk-In	1-2	26	192	218
171	MHSU-Developmental Disabilities-Adults-Assessment and Support-Ambulatory	1-2	23	211	234
175	MHSU-Addictions-Supervised Consumption-Ambulatory	1-2	21	70	91
30	MHSU-Crisis-Residential	1-2	20	87	107
3	MHSU-Adult Community Outreach-Moderate to High Risk	1-2	17	136	153
275	COVID-19 MHSU Health Monitoring	1-2	15	40	55
74	Adjunctive Therapies in Acute Care-Respiratory	1-2	4	15	19
158	Telehealth-Miscellaneous	1-2	2	12	14

Table 3. At the second iteration, community '1-2' is broken into three communities ('2-2', '2-3', and '2-4') that are gradually becoming homogeneous, with '2-4' especially becoming homogeneous with regard to harm-reduction and rehab recovery services.

SC_ID	Service Name	CID	IWD	EWD	WD
13	MHSU-Assertive Community Treatment (ACT)-Adult	2-2	63	1342	1405
26	MHSU-Residential Care-Lower-Level Support	2-2	53	762	815
29	MHSU-Residential Care-Licensed	2-2	52	1101	1153
10	Tertiary Specialized Residential Care-Adult	2-2	34	379	413
20	MHSU-Rehab Services-Adult-Moderate Intensity	2-2	28	273	301
81	MHSU-Crisis Response-Walk-In	2-2	11	207	218
3	MHSU-Adult Community Outreach-Moderate to High Risk	2-2	8	145	153
74	Adjunctive Therapies in Acute Care-Respiratory	2-2	3	16	19
14	MHSU-Addictions-Outreach and Intensive Case Management-Adult	2-3	32	664	696
243	MHSU-Addictions-Rapid/High-Intensity Assessment and Follow-Up	2-3	31	1049	1080
270	COVID-19 Outreach Assessment	2-3	13	152	165
272	COVID-19 Outreach Assessment Team-Provider	2-3	11	60	71
275	COVID-19 MHSU Health Monitoring	2-3	7	48	55
30	MHSU-Crisis-Residential	2-3	6	101	107
171	MHSU-Developmental Disabilities-Adults-Assessment and Support-Ambulatory	2-3	6	228	234
175	MHSU-Addictions-Supervised Consumption-Ambulatory	2-3	6	85	91
34	MHSU-Addictions-Clinical Intake-Adult	2-4	256	1726	1982
22	MHSU-Addictions-Clinic-Adult-Ambulatory	2-4	247	2280	2527
161	Addictions Medicine Specialist Consultation to Acute Care	2-4	189	2422	2611

23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	2-4	148	735	883
203	Overdose-Related Services	2-4	113	884	997
21	MHSU-Addictions-Sobering and Assessment Centre	2-4	103	651	754
24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	2-4	86	373	459
158	Telehealth-Miscellaneous	2-4	2	12	14

At the third iteration, only the services that were in community '2-2', '2-3' at the second iteration are broken into two communities each ('3-2', '3-3', '3-4', and '3-5'). However, community '2-4' from the second iteration remained unchanged. Subsequent iterations are not able to yield any additional breaking of the communities, hence the algorithm stops. Working with team members with a clinical and health services system operations background, it was determined that the third iteration provided an appropriate resolution with an interpretable community of services. With their help, as illustrated in Table 4, the various generated communities were reviewed and labeled as follows:

(1) High intensity community-based treatment (13, 10): this is the community of services that provide high intensity community-based treatment and support for people with severe psychiatric illnesses. (2) Lower intensity community-based treatment (26, 29, 20, 81): this is the community of services that provide lower intensity community-based treatment and support for people with severe psychiatric illnesses. (3) Addiction-outreach focused support (14, 243, 270): these are the services that provide support for people with high risk/high needs addiction problems. This is a linkage-focused set of services, not a treatment-focused set of services and a link to rehab recovery/harm reduction services. People using these services are mostly disfranchised, likely homeless and weakly connected to other services, and potentially high users of low barriers services such as the emergency. (4) Additions ongoing support (34, 22, 161, 23, 203, 21, 24): providing rehab recovery and harm reduction services. Under these services, people receive structured ongoing support to help with addictions problems. These services can be wrapped around patients to help manage various risk related to addictions problems.

Table 4. At the third iteration, community '2-2' breaks into '3-2' and '3-3', whereas community '2-3' breaks into '3-4' and '3-5'. However, community '2-4' from the second iteration remains unchanged at the third iteration.

Category	SC_ID	Service Name	CID	IWD	EWD	WD
High intensity community-based treatment for people with severe psychiatric illness	13	MHSU-Assertive Community Treatment (ACT)-Adult	3-2	24	1381	1405
	10	Tertiary Specialized Residential Care-Adult	3-2	20	393	413
	~~3~~	~~MHSU-Adult Community Outreach-Moderate to High Risk~~	~~3-2~~	~~4~~	~~149~~	~~153~~
Lower intensity community-based treatment for people with severe psychiatric illness	26	MHSU-Residential Care-Lower-Level Support	3-3	33	782	815
	29	MHSU-Residential Care-Licensed	3-3	29	1124	1153
	20	MHSU-Rehab Services-Adult-Moderate Intensity	3-3	18	283	301
	81	MHSU-Crisis Response-Walk-In	3-3	8	210	218
	~~74~~	~~Adjunctive Therapies in Acute Care-Respiratory~~	~~3-3~~	~~2~~	~~17~~	~~19~~

Addiction-outreach focused support for high risk/high needs addictions problems	14	MHSU-Addictions-Outreach and Intensive Case Management-Adult	3-4	24	672	696
	243	MHSU-Addictions-Rapid/High-Intensity Assessment and Follow-Up	3-4	23	1057	1080
	270	COVID-19 Outreach Assessment	3-4	11	154	165
	~~175~~	~~MHSU-Addictions-Supervised Consumption-Ambulatory~~	~~3-4~~	~~6~~	~~85~~	~~91~~
X	30	MHSU-Crisis-Residential	3-5	3	104	107
	171	MHSU-Developmental Disabilities-Adults-Assessment and Support-Ambulatory	3-5	3	231	234
	272	COVID-19 Outreach Assessment Team-Provider	3-5	3	68	71
	275	COVID-19 MHSU Health Monitoring	3-5	3	52	55
Additions ongoing support: harm reduction and/or rehab recovery.	34	MHSU-Addictions-Clinical Intake-Adult	3-6	256	1726	1982
	22	MHSU-Addictions-Clinic-Adult-Ambulatory	3-6	247	2280	2527
	161	Addictions Medicine Specialist Consultation to Acute Care	3-6	189	2422	2611
	23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	3-6	148	735	883
	203	Overdose-Related Services	3-6	113	884	997
	21	MHSU-Addictions-Sobering and Assessment Centre	3-6	103	651	754
	24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	3-6	86	373	459
	~~158~~	~~Telehealth-Miscellaneous~~	~~3-6~~	~~2~~	~~12~~	~~14~~

Also, notice the removal of some services – represented in the table using strike-through texts (3, 74, 175 and 158) from within some of the communities due to a relatively lower internal weighted degree. Finally, notice the exclusion of the community made of service classes (30, 171, 272 and 275), labelled as "X". These services were forced into one community by the community detection algorithm that must fit every service class into a community. In consultation with team members with clinical and health services system operations background, it was determined that these services do not display any interpretable characteristic or perform any function as a group. Also, notice their overall relatively low internal weighted degree across the entire community. Referring to the analogy previously described and equating this sorting and labelling process to "separating wheat from chaff", the community of services (30, 171, 272 and 275) can be referred to as "chaff" and can be safely discarded.

4. Discussion

Figure 5 outlines the end-to-end process for extracting PSUs from a longitudinal, sparse/high-dimensional encounters data. Given the methodological nature of the paper, only the results of one of the branches of the iterative communities (community 2) were reported. However, the process outlined for community 2 applies to the other communities (1, 3, and 4) as well, and their corresponding sub-communities. Although not reported, at the end, a total of 22 communities of services were extracted for the schizophrenia cohort.

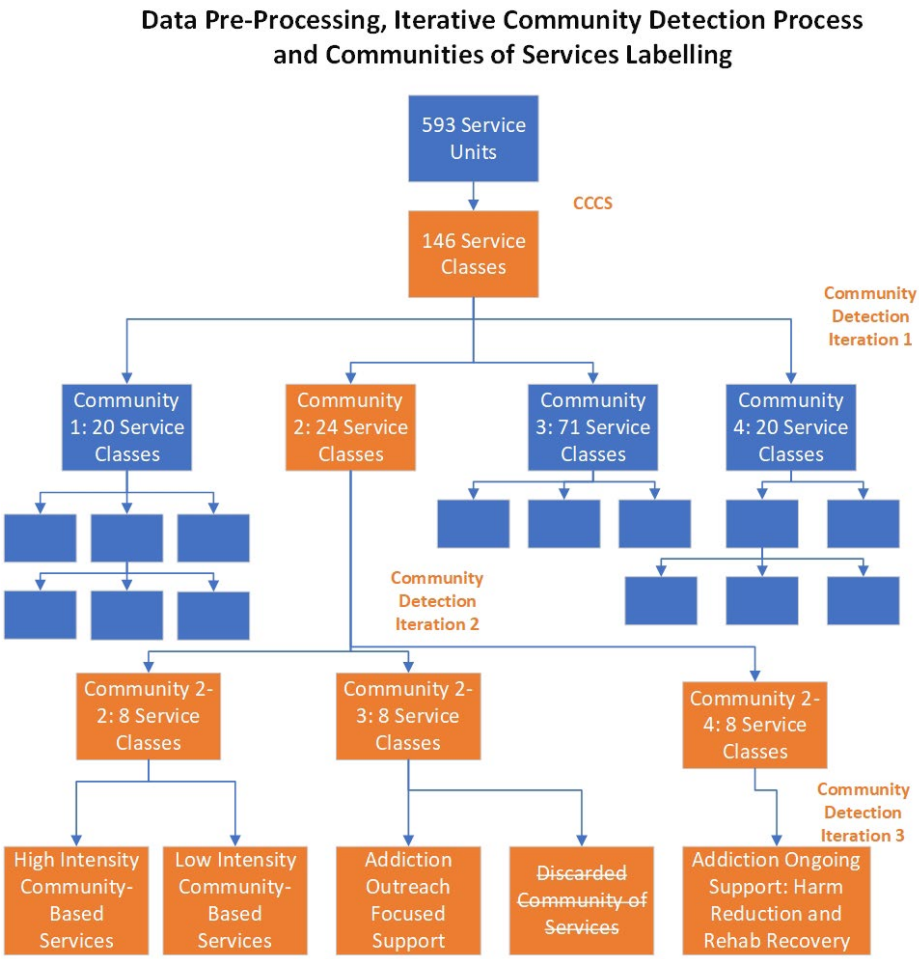


Figure 5. PSUs extraction end-to-end process.

The cohort that was chosen to illustrate the methodology proposed in this paper interacted with a total of 593 Service Units. These represent services that spawn across the continuum of care and were documented as encounters in the host organization CIS. To extract any meaningful PSUs at this level of granularity (both "spurious granularity" and "unnecessary granularity") is not feasible, regardless of the ML algorithm used. The application of CCCS provided an opportunity to address the granularity and nomenclature issues and reduce the dimension of the data and make it analysis ready. This step converted the 593 Service Units into analyzable 146 Service Classes, as illustrated in Figure 5. With the nomenclature issue addressed, and the data granularity reduced, the data is ready for the application of an iterative community detection algorithm. At the end of the iterative community detection, a total of 22 communities of services were generated. A sample of those communities of services, as shown in Figure 5, have demonstrated that meaningful patterns of service utilization can emerge from this process with the help of SMEs and SSOEs combined with the use of various graph metrics. Graph/network models are well suited for pattern recognition, and have been used in many domains [8,43–45]. However, there is no work, to our knowledge, that has used a pattern recognition approach

to a cross continuum multi-dimensional dataset to extract meaningful patterns of services utilization.

Hence, from a methodological perspective, the strength and importance of this paper is the ability to demonstrate that working from a large body of longitudinal, sparse/high-dimensional encounter data spanning a full continuum of secondary and tertiary health services, it is possible to generate intelligible patterns of service utilization. The work in this paper has demonstrated that graph community detection methods and metrics, when combined with the application of an appropriate semantic layer and engagement of relevant SMEs, have the potential to generate face-valid intelligible results from initially sparse, high-dimensional patient–service system encounter data.

The methodology featured in this paper starts with the use of a semantic layer, CCCS, to perform the initial phase of the dimensional reduction. This coding is both generated and applied to the more granular Service Units by service system experts. It is not derived empirically. The next stage in the analysis involved team members with both an analytical and clinical background in the selection of the cohort of interest, i.e., the schizophrenia cohort. As shown in Figure 5, the schizophrenia cohort was engaged with 593 Service Units. The application of CCCS permitted the 593 Services Units to be reduced to 146 Service Classes, hence setting the stage for carrying out graph community detection. Graph community detection was carried iteratively on the cohort of interest. The models were refined by using graph metrics such as modularity to set cut-offs and eliminate Service Classes that are only weakly associated with other elements within PSUs. Subject matter experts provided feedback on the level of resolution and applied labels to the resulting communities. Those labels relate directly to the functions performed by services constituting the communities. The community of services that failed to demonstrate interpretable characteristics was discarded.

Community detection and related methods are being used as a means for providing visibility (literally) into patterns in the data. The objective is not to produce a definitive answer to questions such as “how is this cohort partitioned?”. There is no underlying truth regarding a given patient journey or PSUs associated with cohorts that the methods are correctly or incorrectly detecting. It is helpful to think of the methods presented in this paper as a macroscope that provides visibility into patterns that are located in sparse high-dimensional datasets—patterns located in a space that is too complex for them to be detected by SMEs without the assistance of pattern detection/construction tools. The objective is to produce information that can be used by parties with expert knowledge of cohorts and service system operations to develop tactics to solve problems based on patterns that are identified and depicted by the tools.

There are several important contributions from this study. First, the methods set out in this paper generate a foundation set of observables that can be used with various other methods to generate actionable results. One important set of results that will be featured in other papers makes use of these basic observables to predict sentinel events, such as overdoses or falls in older adults. Methodologically, some of that work entails bipartite projections onto people, rather than services, yielding clusters of people who

are relatively homogeneous with respect to PSUs, and are then shown to be relatively homogeneous with respect to sentinel events via prediction models using community membership as features.

Second, the methods set out in this paper provide a foundation set of observables that are directly applicable to the task of evaluating the impacts of services on patient journeys and on outcomes. The PSUs can be used to attach features to people that can then be employed to generate risk-adjusted measures of outcomes. Moreover, PSUs can themselves be regarded as outcome measures in a straightforward pre- vs. post-design, e.g., PSUs for older adults before a fall that resulted in an acute care admission vs. PSUs for those same persons after the fall and resulting hospitalization.

Third, the work presented in this paper constitutes an initial depiction of an innovative set of methods demonstrating the ability to produce clinically understandable results that span the service continuum and go well beyond more common metrics of service system operations such as frequency of visits to emergency departments or acute care readmissions. Further work is required to determine how/when/whether other methods such as Natural Language Processing produce similar results. Such work is underway.

There are also a couple of limitations with the proposed methodology. First, the model proposed in this research is atemporal, the events are collapsed across time and the order of the events are not taken into consideration. The intent is to highlight the prevalence of connection between services, using the edge weights to assess the strength of prevalence. However, this model fails to capture the strength of coupling between services as well as the order of events. This is a limitation that will be addressed in a subsequent study. Second, given the data that was used for this analysis, the findings are limited to the host organization, and hence not immediately generalizable/transferable to other jurisdictions. This is especially important, as the host organization is a relatively self-contained jurisdiction, compared to other healthcare jurisdictions where patients typically move between different jurisdictions for their care needs. However, the methods outlined in this paper are generalizable to other healthcare jurisdictions.

5. Conclusions

The proposed methodology in this paper for analyzing complex healthcare data has enabled the identification of patterns in patient–service encounter data that are difficult to detect via classic statistical methods and deeply resistant to interpretation given the names attached to Service Units in the CIS location build. The CCCS, together with graph community detection methods, set the foundation to generate the basic entities required to evaluate conformance of complex sequences of interventions to cohort-specific clinical practice guidelines (CPGs). In the literature to date, we have not come across work that "drills up" to the level of full cross-continuum patterns of service utilization in a data space that incorporates a very broad array of hospital and community-based acute care, ambulatory, case management and residential services. The use of a patients' services pathway as a foundation in evaluating the conformance of intervention to cohort specific CPGs will be the focus of future research.

Ultimately, we expect there are considerable implications related to the generated communities of services. This includes the possibility of influencing the reorganization of some services within the host organization service structure, in order to provide better care for vulnerable patients with mental and other complex healthcare challenges. These organizational/systems/process impacts would require the engagement of quality assurance/quality improvement processes in the organization, as well as support from the host organization's senior leadership for uptake and use of the results.

Author Contributions: Conceptualization, J.B., K.M., A.R. (Abraham Rudnick) and A.K.; data curation, J.B. and S.R.; formal analysis, J.B., Y.S., H.S., K.M. and A.R. (Abraham Rudnick); investigation, J.B. and K.M.; methodology, J.B., Y.S., H.S., K.M., A.R. (Abraham Rudnick) and A.K.; project administration, J.B.; resources, J.B., K.M. and S.R.; software, J.B., Y.S., H.S., S.R., E.C., J.H., M.H. and A.R. (Ashlin Richardson); supervision, J.B., K.M., A.R. (Abraham Rudnick) and A.K.; validation, J.B., Y.S., K.M., A.R. (Abraham Rudnick) and E.C.; visualization, J.B. and K.M.; writing—original draft, J.B.; writing—review and editing, J.B., Y.S., H.S., K.M., A.R. (Abraham Rudnick), S.R., A.K., J.H., G.D., K.O. and A.R. (Ashlin Richardson). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: A certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada Ethics harmonization guideline. The REB number is H21-02817.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are unavailable because of privacy or ethical restrictions. Requests to access the datasets require a certificate of approval by the University of Victoria Research Ethics Board, following the British Columbia, Canada Ethics harmonization guideline.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McVeigh, S.E. Sepsis management in the emergency department. *Nurs. Clin.* **2020**, *55*, 71–79.
2. Laursen, T.M.; Nordentoft, M.; Mortensen, P.B. Excess Early Mortality in Schizophrenia. *Annu. Rev. Clin. Psychol.* **2014**, *10*, 425–448. https://doi.org/10.1146/annurev-clinpsy-032813-153657.
3. Laursen, T.M.; Munk-Olsen, T.; Vestergaard, M. Life expectancy and cardiovascular mortality in persons with schizophrenia. *Curr. Opin. Psychiatry* **2012**, *25*, 83–88. https://doi.org/10.1097/yco.0b013e32835035ca.
4. BC Guidelines. 2024. Available online: https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines (accesse on 7 March 2024).
5. Thor, J.; Lundberg, J.; Ask, J.; Olsson, J.; Carli, C.; Härenstam, K.P.; Brommels, M. Application of statistical process control in healthcare improvement: Systematic review. *BMJ Qual. Saf.* **2007**, *16*, 387–399.
6. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008.
7. Moselle, K.; Bambi, J.; Santoso, Y.; Sadri, H.S.; Robertson, S.; Howie, J.; Rudnick, A.; Chang, E. Abundance and Scarcity of Published Work in Machine Learning Derived Supports for Effective Service System Operations.*2024, unpublished*.
8. Barabási, A.-L.; Loscalzo, J.; Silverman, E.K. *Network Medicine: Complex Systems in Human Disease and Therapeutics*; Harvard University Press: Cambridge, MA, USA, 2017.
9. Ahmedt-Aristizabal, D.; Armin, M.A.; Denman, S.; Fookes, C.; Petersson, L. Graph-based deep learning for medical diagnosis and analysis: Past, present and future. *Sensors* **2021**, *21*, 4758.

10. Jaremko, J.L.; Felfeliyan, B.; Hareendranathan, A.; Thejeel, B.; Vanessa, Q.-L.; Østergaard, M.; Conaghan, P.G.; Lambert, R.G.W.; Ronsky, J.L.; Maksymowych, W.P. *Volumetric Quantitative Measurement of Hip Effusions by Manual Versus Automated Artificial Intelligence Techniques: An Omeract Preliminary Validation Study*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2021; Volume 51, pp. 623–626.
11. Banerjee, I.; Madhavan, S.; Goldman, R.E.; Rubin, D.L. Intelligent word embeddings of free-text radiology reports. In Proceedings of the AMIA Annual Symposium Proceedings, American Medical Informatics Association, Washington, DC, USA, 4–8 November 2017; p. 411.
12. Elkin, P.L.; Froehling, D.; Wahner-Roedler, D.; Trusko, B.; Welsh, G.; Ma, H.; Asatryan, A.X.; Tokars, J.I.; Rosenbloom, S.T.; Brown, S.H. *NLP-Based Identification of Pneumonia Cases from Free-Text Radiological Reports*; American Medical Informatics Association: Bethesda, MD, USA, 2008; p. 172.
13. Garla, V.; Taylor, C.; Brandt, C. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *J. Biomed. Inform.* **2013**, *46*, 869–875.
14. Martinez, D.; Ananda-Rajah, M.R.; Suominen, H.; Slavin, M.A.; Thursky, K.A.; Cavedon, L. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *J. Biomed. Inform.* **2015**, *53*, 251–260.
15. Stewart, R.; Velupillai, S. Applied natural language processing in mental health big data. *Neuropsychopharmacology* **2021**, *46*, 252.
16. Rost, B.; Radivojac, P.; Bromberg, Y. Protein function in precision medicine: Deep understanding with machine learning. *FEBS Lett.* **2016**, *590*, 2327–2341.
17. Alabi, R.O.; Almangush, A.; Elmusrati, M.; Mäkitie, A.A. Deep machine learning for oral cancer: From precise diagnosis to precision medicine. *Front. Oral Health* **2022**, *2*, 794248.
18. Carlisle, A.; Caceres, I.; Mehta, S.; Schindler, J.; Sharma, J. A combined machine learning and bioinformatic analysis approach identifies biological pathways that predict clinical stage and survival outcome in neuroblastoma patients. *Cancer Res.* **2015**, *75*, 3758.
19. Ge, L.; Chen, Y.; Yan, C.; Zhao, P.; Zhang, P.; Liu, J. Study progress of radiomics with machine learning for precision medicine in bladder cancer management. *Front. Oncol.* **2019**, *9*, 1296.
20. Hase, T.; Ghosh, S.; Palaniappan, S.K.; Kitano, H. Cancer network medicine. *Netw. Med.* **2017**, 294–323, doi:10.4159/9780674545533-014
21. Nakagawa, H.; Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci.* **2018**, *109*, 513–522.
22. Piccialli, F.; Calabrò, F.; Crisci, D.; Cuomo, S.; Prezioso, E.; Mandile, R.; Troncone, R.; Greco, L.; Auricchio, R. Precision medicine and machine learning towards the prediction of the outcome of potential celiac disease. *Sci. Rep.* **2021**, *11*, 5683.
23. Gökçay Canpolat, A.; Şahin, M. Glucose lowering treatment modalities of type 2 diabetes mellitus. *Diabetes Res. Clin. Pract.* **2021**, *4*, 7–27.
24. Shamji, M.H.; Ollert, M.; Adcock, I.M.; Bennett, O.; Favaro, A.; Sarama, R.; Riggioni, C.; Annesi-Maesano, I.; Custovic, A.; Fontanella, S. EAACI guidelines on environmental science in allergic diseases and asthma–leveraging artificial intelligence and machine learning to develop a causality model in exposomics. *Allergy* **2023**, *78*, 1742–1757.
25. Pike, F.; Yealy, D.M.; Kellum, J.A.; Huang, D.T.; Barnato, A.E.; Eaton, T.L.; Angus, D.C.; Weissfeld, L.A. Protocolized care for early septic shock (ProCESS) statistical analysis plan. *Crit. Care Resusc.* **2013**, *15*, 301–310.
26. Norman, C.; Van Nguyen, T.; Névéol, A. Contribution of natural language processing in predicting rehospitalization risk. *Med. Care* **2017**, *55*, 781.
27. Orangi-Fard, N.; Akhbardeh, A.; Sagreiya, H. *Predictive Model for Icu Readmission Based on Discharge Summaries Using Machine Learning and Natural Language Processing*, 1st ed.; MDPI: Basel, Switzerland, 2022; p. 10.
28. Rumshisky, A.; Ghassemi, M.; Naumann, T.; Szolovits, P.; Castro, V.M.; McCoy, T.H.; Perlis, R.H. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl. Psychiatry* **2016**, *6*, e921.
29. Panteli, D.; Legido-Quigley, H.; Reichebner, C.; Ollenschläger, G.; Schäfer, C.; Busse, R. Clinical practice guidelines as a quality strategy. *Improv. Healthc. Qual. Eur.* **2019**, 233. https://doi.org/10.1787/b11a6e8f-en
30. Rotter, T.; de Jong, R.B.; Lacko, S.E.; Ronellenfitsch, U.; Kinsman, L. Clinical pathways as a quality strategy. *Improv. Healthc. Qual. Eur.* **2019**, 309. https://doi.org/10.1787/b11a6e8f-en

31. Allen, M.; Pearn, K.; Monks, T.; Bray, B.D.; Everson, R.; Salmon, A.; James, M.; Stein, K. Can clinical audits be enhanced by pathway simulation and machine learning? An example from the acute stroke pathway. *BMJ Open* **2019**, *9*, e028296.
32. Huo, T.; George Jr, T.J.; Guo, Y.; He, Z.; Prosperi, M.; Modave, F.; Bian, J. Explore Care Pathways of Colorectal Cancer Patients with Social Network Analysis. *Stud. Health Technol. Inform.* **2017**, *245*, 1270.
33. Carroll, N.; Richardson, I. Mapping a careflow network to assess the connectedness of connected health. *Health Inform. J.* **2019**, *25*, 106–125.
34. Aggarwal, N.; Ahmed, M.; Basu, S.; Curtin, J.J.; Evans, B.J.; Matheny, M.E.; Nundy, S.; Sendak, M.P.; Shachar, C.; Shah, R.U. Advancing artificial intelligence in health settings outside the hospital and clinic. *NAM Perspect.* **2020**, *2020*, doi: 10.31478/202011f
35. Lin, Z.; Yang, D.; Yin, X. Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access* **2020**, *8*, 156663–156676.
36. Rose, S. Intersections of machine learning and epidemiological methods for health services research. *Int. J. Epidemiol.* **2020**, *49*, 1763–1770.
37. El Emam, K.; Arbuckle, L. *Anonymizing Health Data: Case Studies and Methods to Get You Started*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2013.
38. Bambauer, J.; Muralidhar, K.; Sarathy, R. Fool's gold: An illustrated critique of differential privacy. *Vand. J. Ent. Tech. L.* **2013**, *16*, 701.
39. Xu, C.; Ren, J.; Zhang, Y.; Qin, Z.; Ren, K. DPPro: Differentially Private High-Dimensional Data Release via Random Projection. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 3081–3093. https://doi.org/10.1109/TIFS.2017.2737966.
40. Malin, B.; Goodman, K. Between access and privacy: Challenges in sharing health data. *Yearb. Med. Inform.* **2018**, *27*, 55–59.
41. Koval, A.; Moselle, K. Clinical Context Coding Scheme-Describing Utilisation of Services of Island Health between 2007–2017. In Proceedings of the Conference of the International Population Data Linkage Association, Banf, AB, Canada, 12–14 September 2018.
42. Chejara, P.; Godfrey, W.W. *Comparative Analysis of Community Detection Algorithms*; IEEE: Minneapolis, MN, USA, 2017; pp. 1–5.
43. Niyirora, J.; Aragones, O. Network analysis of medical care services. *Health Inform. J* **2020**, *26*, 1631–1658. https://doi.org/10.1177/1460458219887047.
44. Palmer, R.; Utley, M.; Fulop, N.J.; O'Connor, S. Using visualisation methods to analyse referral networks within community health care among patients aged 65 years and over. *Health Inform. J.* **2020**, *26*, 354–375.
45. Khazaee, A.; Ebrahimzadeh, A.; Babajani-Feremi, A. Application of pattern recognition and graph theoretical approaches to analysis of brain network in Alzheimer's disease. *J. Med. Imaging Health Inform.* **2015**, *5*, 1145–1155.

Chapter 3: 2nd Manuscript

BioMedInformatics

MDPI

Article

Approaches to Extracting Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs. Natural Language Processing Clustering

Jonas Bambi [1], Hanieh Sadri [1], Ken Moselle [1], Ernie Chang [2], Yudi Santoso [1], Joseph Howie [1], Abraham Rudnick [3,*], Lloyd T. Elliott [4] and Alex Kuo [1]

[1] University of Victoria, Victoria, BC V8P 5C2, Canada; jonasbambi@uvic.ca (J.B.); haniehsadri@uvic.ca (H.S.); kmoselle@uvic.ca (K.M.); y.santoso8@gmail.com (Y.S.); joehowie@uvic.ca (J.H.); akuo@uvic.ca (A.K.)
[2] Independent Reseracher, ecsendmail@gmail.com
[3] Dalhousie University, Halifax, NS B3H 4R2, Canada
[4] Simon Fraser University, Burnaby, BC V5A 1S6, Canada; lloyd_elliott@sfu.ca
* Correspondence: abraham.rudnick@nshealth.ca

Citation: Bambi, J.; Sadri, H.; Moselle, K.; Chang, E.; Santoso, Y.; Howie, J.; Rudnick, A.; Elliott, L.; Kuo, A. Approaches to Extracting Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs. Natural Language Processing Clustering. *Biomedinformatics* **2024**, *4*, x. https://doi.org/10.3390/xxxx

Academic Editor(s): Name

Received: 4 June 2024
Revised: 13 July 2024
Accepted: 5 August 2024
Published: date

Abstract: *Background:* As patients interact with a healthcare service system, patterns of service utilization (PSUs) emerge. These PSUs are embedded in the sparse high-dimensional space of longitudinal cross-continuum health service encounter data. Once extracted, PSUs can provide quality assurance/quality improvement (QA/QI) efforts with the information required to optimize service system structures and functions. This may improve outcomes for complex patients with chronic diseases. *Method:* Working with longitudinal cross-continuum encounter data from a regional health service system, various pattern detection analyses were conducted, employing (1) graph community detection algorithms, (2) natural language processing (NLP) clustering, and (3) a hybrid NLP–graph method. *Result:* These approaches produced similar PSUs, as determined from a clinical perspective by clinical subject matter experts and service system operations experts. *Conclusions:* The similarity in the results provides validation for the methodologies. Moreover, the results stress the need to engage with clinical or service system operations experts, both in providing the taxonomies and ontologies of the service system, the cohort definitions, and determining the level of granularity that produces the most clinically meaningful results. Finally, the uniqueness of each approach provides an opportunity to take advantage of the various analytical capabilities that each approach brings, which will be further explored in our future research.

Keywords: clinical pathways; clinical practice guideline; clustering; decision support; electronic healthcare; graph community detection; health information management; health service system; machine learning algorithms; natural language processing

1. Introduction

1.1. Clinical Practice Guidelines, Clinical Pathways, and Services Pathways

The intent of this work is to extract useful information from data that have been accumulating in clinical information systems in order to optimize service system structure and function on behalf of patients contending with chronic/complex diseases. To achieve this, three constructs will be considered, including (1) generic clinical practice guidelines (CPGs), (2) idealized clinical pathways for those CPGs within a local service system, and (3) real-world cohort-specific service pathways located within local service system encounter data. For this study, we employ various machine learning (ML) methods to identify real-world service pathways in cross-continuum (i.e., across all services) longitudinal encounter data.

Clinical practice guidelines (CPGs) are evidence-informed recommendations intended to optimize patient care [1]. They consist of structured sequences of clinical interventions [1]. These generic guidelines are disease-class-specific but service-system-agnostic. As an example, CPGs for chronic diseases such as heart failure [2] provide evidence-based support for branching arrays of decisions that are keyed to the patient's clinical, functional, or behavioral status. Clinical pathways, on the other hand, translate generic service system-agnostic CPGs into idealized local service-system-specific terms. As patients interact with a healthcare service system, patterns of service utilization (PSUs) emerge [3]. Hence, real-world service pathways consist of cohort-specific predictable recurring PSUs that take place within a local service system.

Perfect conformance of PSUs and of idealized service pathways is conditional upon an array of counterfactual conditions, for example an adequate supply of affordable services, and accessibility for all members of at-risk or clinically impacted populations. In the real world, pathways tracked by persons across the continuum of services are subject to the influence of at least four sets of potent factors: (1) factors attributable to the service system (e.g., limited capacity) [4]; (2) stigma associated with disease (e.g., addictions) [5–7]; (3) impacts arising from disease pathophysiology (e.g., difficult-to-predict emergence of comorbidities) [8]; and (4) patient factors that may impact treatment effectiveness and limit capacity to initiate and sustain requisite levels of service system engagement as a disease progresses [9–11]. The combined effect of these factors is real-world cohorts tracking to service pathways that may be positioned in the health service space at some distance from the idealized clinical pathways keyed to CPGs.

From this vantage point, the construct "quality" may be defined operationally in terms of the distance between idealized clinical pathways and real-world PSUs that are etched into the service terrain by cohorts of patients. If (1) PSUs are to be used to measure conformance of practice to complexly structured CPGs, and (2) those PSUs are based on machine learning methods applied to large volumes of variable-quality transactional

data extracted from real-world transactional clinical information systems, then an organization using those PSUs to assure or improve quality must have confidence that they provide an accurate view of the local service system operations. This paper describes a method for supplying that assurance by applying three machine learning methods to those data and generating results that are directly comparable between methods.

1.2. Use of Graph Analytics for Healthcare Data

Numerous graph or network algorithms and methods have been applied to health-relevant data in recent years to examine diverse systems, including intracellular processes that relate clinical signs and symptoms to pathophysiological mechanisms [12], online social networks [13], biological networks [14], disease networks [15], and others. Moreover, a large body of work in the areas of disease, treatment, and health service system operations has been built upon graph analytics [16]. Examples include the following: (1) supporting medical predictive tasks such as discovering unknown disease associations for drug repositioning or comprehending disease progression [17,18]; (2) promoting drug discovery and molecular mechanism exploration in bioinformatics [18]; (3) improving critical care prediction [19]; (4) supporting diagnosis prediction, patient clinical outcome prediction, and readmission prediction [20]; (5) detecting patterns of care for patients [21]; and (6) exploring, analyzing, and understanding patterns in community referrals for elderly patients, and their use of multiple services through data visualization [22].

Node clustering is a topic that has garnered a great deal of attention in the field of graph computation [23,24]. In the context of graph analytics, clusters are often called communities. Many algorithms and methods with which to discover communities have been proposed. Some focus on the performance and some on the quality of the result. Here, quality means whether the partition of the nodes among the communities makes sense from the experts' point of view. Well-known clustering algorithms include Fast-Greedy [25], Edge-Betweenness [25], Leading-Eigen [26], and Louvain [27].

The work conducted in [3] found that the Louvain algorithm often produces the most interpretable results. Nevertheless, while conducting some analysis in [3], we discovered issues with the Louvain method that prompted a modification to the procedure. In particular, Louvain is constrained by its resolution; given a graph, the smallest cluster or community that can be detected is bounded from below by the size of the graph; the larger the graph, the larger this minimum size.

1.3. Use of NLP in Analyzing Healthcare Data

Natural language processing (NLP) is a major branch of machine learning (ML). In recent years, NLP tools have been used extensively in healthcare as a method for extracting clinically meaningfully coded data from free text. Examples include (1) effective knowledge extraction from patient records using NLP [28,29], (2) extraction of symptoms from unstructured clinical information system data to be used in COVID-19 prognostics [30], and (3) use of NLP techniques to support clinical decisions on patients' health outcomes [31].

NLP methods were originally designed to process texts in natural human languages. It has been known that many of the NLP methods are also applicable to many kinds of data that can be represented as strings. What is largely absent in the literature is the notion that a patient healthcare journey, consisting of a series of encounters with a large array of service entities, can also be treated as a string (after encoding the sequence of service utilization as a string of tokens). Therefore, healthcare encounters data are subject to many of the same types of analytical procedures employed with text documents or samples of human speech. Through methods such as TF-IDF (term frequency-inverse document frequency), documents can be represented as vectors in a word-space and hence can then be clustered. We - to take advantage of these NLP capabilities to extract PSUs.

1.4. Objectives

As previously stated, at a cohort level, the construct "quality" can be defined operationally as the distance between idealized clinical pathways and the real-world PSUs. With this definition, evidence-based quality assurance/quality improvement (QA/QI) requires a method for locating PSUs within sparse high-dimensional arrays of cross-continuum health service encounter data sourced from records in the clinical information systems. In the work conducted in [3], graph community detection methods were employed to detect communities of services that reflect PSUs. With graph community detection, encounter data are viewed as a bipartite graph of persons interacting with services, which is then projected to form a network of services. In this paper, a method for providing concurrent cross-validation of solutions derived from graph representations of the source data and NLP-based approaches is described.

Hence, the work in this paper is organized around the following questions:

To what extent can NLP methods be used to extract PSUs from longitudinal heterogeneous cross-continuum healthcare data? How do the data need to be modeled and what data pre-processing needs to be conducted to generate the base data upon which the NLP methods can be applied?

Are the results from NLP clustering for Service Classes similar to those obtained using graph community detection? Are they judged to be similar by clinical subject matter experts (SMEs) or clinical/administrative service system operations experts (SSOEs)?

Does a hybrid NLP–graph community detection approach generate meaningful results, and how do the results compare to (a) community detection results, with simple frequency-based edge-weighted projections of service-service interactions, and (b) results obtained using NLP-based clustering approaches, employing measures of cosine similarity between vectors reflecting patient journeys?

2. Methodological Approach

2.1. Concurrent Validation via Application of Multiple Methods to the Same Body of Data, Modeled in Different Ways

The methodology reported in this document is roughly analogous to the multitrait–multimethod approach to constructing validation within a classic

test and measurement paradigm, tracing back to the work of Campbell and Fiske [32]. For the work in this document, the constructs to be validated are PSUs. Our intent is to identify underlying functions expressed in terms of functions that span service system structures and emerge over the course of potentially large numbers of interactions with different services that address different needs and risks.

2.2. Source Data

The source data used in this paper consist of retrospective longitudinal transactional service encounter data extracted from a single instance of a clinical information system (CIS) deployed across the continuum of services provided by one of the health authorities within Canada (hereinafter referred to as "host organization"). The host organization provides a comprehensive array of secondary and tertiary health services, for all ages, for persons contending with medical/surgical issues and/or mental health/substance use issues. This includes acute care/intensive care services, hospital- and community-based emergency response, ambulatory services, residential care services for older adults or persons contending with mental health issues, case management services, and a range of addiction harm reduction or rehab and recovery-oriented services. The encounter data accessed by this study consist of approximately 10 million encounters over 7 years for approximately 1 million patients. This represents data for all service recipients, except a few restricted services where the data are strictly prohibited (e.g., services for persons who are victims of sexual assault). To access the source data, a certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada, ethics harmonization guideline.

2.3. Data Pre-Processing

Two main branches of analysis are presented in this paper: a graph method and an NLP method. However, before any analysis is undertaken, as part of the data preparation, we consider the granularity of the data. The services provided by the host organization are encapsulated into an array of roughly 2000 Service Units within a location built for the CIS used to support care delivery. Service unit names within the CIS are often opaque, rendering them unsuitable for supervised machine learning methods that require meaningfully labeled data. Moreover, the service units may vary widely with regard to granularity; for example, multiple beds will appear as a single unit within an acute care facility, but multiple beds in a large array of family care homes for frail elderly will show up as multiple service units. To address these issues, a clinical context coding scheme (CCCS) was developed [33].

The CCCS is organized around six sets of codes, constituting a semantic layer applied to all of the Service Units to generate clinically functional Services Classes with meaningful names. By applying the CCCS to source data as previously proposed in [3,34,35], the 2000 service units in the host organization's CIS are converted into approximately 200 clinically functional Service Classes. This activity was conducted in collaboration with SSOES.

These Service Classes are the codes that are used to conduct the various analyses in this study.

2.4. Use of Community Detection in Extracting PSUs

The work conducted in [3] proposes a methodology for extracting PSUs from cross-continuum longitudinal healthcare data using graph community detection. The data consist of encounter data, where each row is a record of a service accessed by a patient. We can view these data as a bipartite graph between patients and Service Classes.

To determine which service classes cluster together based on their pattern of utilization we can perform bipartite projection onto the Service Classes, as illustrated in Figure 1. Two Service Classes are connected by an edge when there are patients who use both. The number of such patients then becomes the weight of the edge. One can then observe that if a pair of Service Classes tend to co-occur in the longitudinal encounter histories of numerous patients, they will have a strong connection between them as measured by the edge weight.

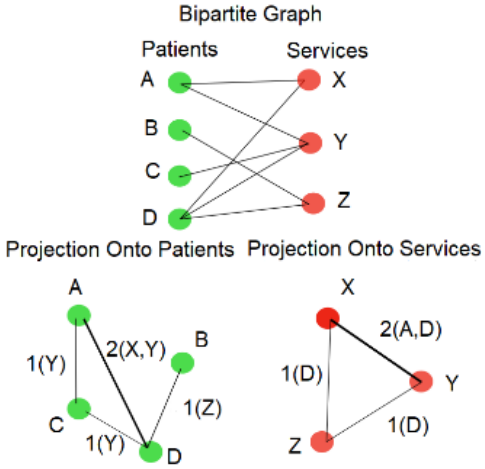


Figure 1. Bipartite graph and bipartite projection.

Next, the Louvain community detection [27] can be applied to the projected graph. However, while conducting the analysis in [3], it was discovered that from a clinical perspective, the results are often still too coarse, with many heterogeneous Service Classes clustered together. To solve this problem, [3] proposed to iteratively apply the Louvain algorithm in a nested fashion. Note that in this case, iteration is not the same as the number of passes in the Louvain algorithm (instead, iteration refers to the level of nesting). In the approach proposed in [3], once the Louvain algorithm has generated the first set of communities, each community is isolated and treated as a new graph and the Louvain algorithm is applied again on each of the isolated graphs. The process is repeated with subsequent set of communities (see Figure 2).

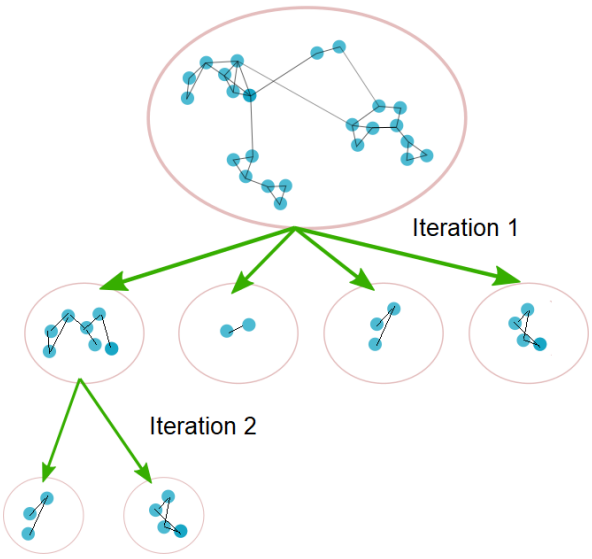


Figure 2. Iterative community detection.

With each iteration, the number of communities increases, and the size of each resulting community is reduced. At a certain point, Louvain no longer breaks the communities any further (i.e., the number of communities remains unchanged, with further iterations) and the algorithm stops. The output communities for each iteration are collected, and the results are compared using modularity values and by engaging with clinical SMEs. Additionally, for each node, the internal weighted degree (the weighted degree inside the community), and the external weighted degree (the weighted degree of a node to nodes outside of its community), are computed. The internal and external weighted degrees can be used to measure the strength of the bond of each node to its community.

Figure 3 provides a snapshot of the results for an illustrative community detection iteration process. For this illustration, the sub-community with Service Classes 1, 2, 3, 14, 30, 42, 81, 150, 168, 238, 248, 249, and 251 is considered. This is one of the communities that were generated as the result of iteration 2. At iteration 3, this sub community will split into two, including sub-communities 1, 2, 3, 14, 30, 42, 168, 238, and 251 and sub-communities 15, 81, 150, 248, and 249. Additionally, at iteration 5, we observe that the community detection algorithm is no longer able to break the last sub-communities any further. As a result, the iteration stops at this level, and the result is reviewed with SMEs and SSOEs to determine the iteration that provides the result that is most clinically meaningful.

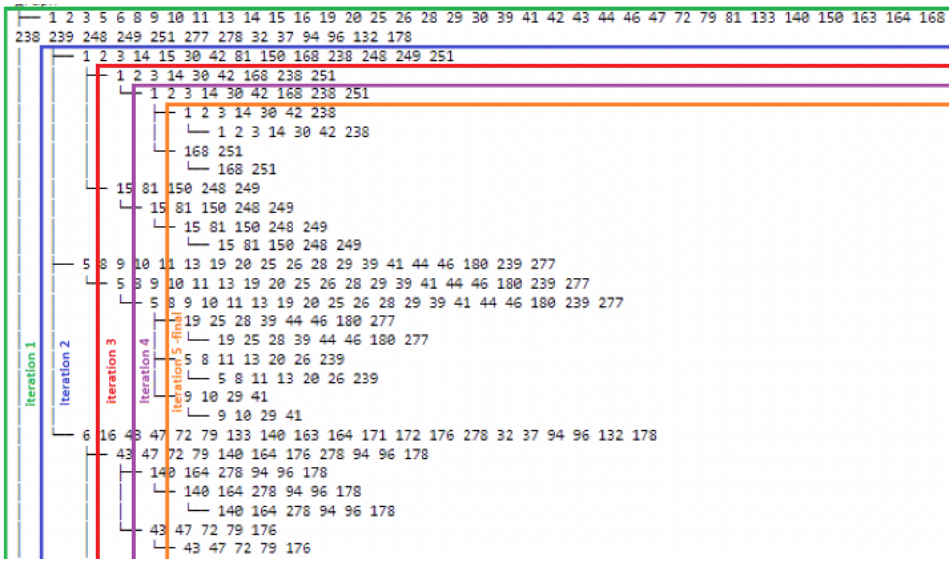


Figure 3. Sample community detection iteration process.

2.5. Use of NLP in Extracting PSUs

Natural language processing (NLP) is a rapidly expanding field in computer science. NLP methods begin by systematically tokening elements within a document (e.g., words) and mapping the tokens into a vector space. Thus, every document in a collection of documents (a corpus) has a vector representation. The process of tokenization can be conducted in several ways, i.e., using frequency counts or considering the proportion of occurrences throughout the corpus (TF-IDF) [36]. First, the tokens must be defined: this can be simply the words in the document or patterns of words. Once the tokens have been defined, vectors can be created for each document in the corpus. TF-IDF is used to generate a normalized count of the tokens in a document weighted by the logarithmic ratio of the entire corpus. Hence, if a token only appears in one of the documents in the corpus, it will have a very small weighting. Conversely, if a token appears in most of the documents in the corpus, its weight will be much larger.

After each document has a TF-IDF vector representation, document vectors can be compared pairwise using a dot product. If two documents are highly similar in token weights, then their corresponding dot product (a.k.a. cosine similarity) will be close to one. Likewise, if two documents are dissimilar in their token weights, then the dot product will be close to zero. Therefore, with the vector representation, one can measure similarity between documents in a given corpus. Moreover, the similarity scores allow algorithms such as K-means to measure the closeness of documents in a corpus and generate clusters of related documents from a given corpus.

This method can be translated quite directly to patient encounter histories (patient journeys) embedded in vectors reflecting the history of encounters with roughly 200 Service Classes. Now, for every patient one can create "sentences" composed of the Service Class IDs.

We then tokenize the sentences with a uni-gram, bi-gram, or skip-gram vectorization. Here, a k-gram is a sequence of k words in a document, and a skip-gram is a sequence of words in a document that are separated by other

intervening words. Once tokenized, we apply TF-IDF. After this tokenization, the TF-IDF vectors can be used to measure the patient's journey similarity using various similarity metrics with regards to the other patients in the "corpus".

Cosine similarity is a metric used to measure how similar the documents are irrespective of their sizes. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space by taking a dot product of the two vectors [37]. In the case of patients' journey similarities, a dot product between two patient's vectors provides a measure of similarity on a patient-to-patient level. Finally, one can apply the K-means algorithm to the patients' dot products and generate a clustering of similar patients in the "corpus".

To create a cluster of services classes, a similar approach to the above is used. First, each patient's history of service utilization is generated as a sentence. The words in such sentences are Service Class ids that a patient had interacted with. For example, if patient "A" engaged with Service Classes "X", "Y", "Z", and "X" again, one would illustrate the sequence of service engagement as "X Y Z X". Second, the sentences are tokenized using frequency counts to transform a sentence showing the history of service utilization into a vector. From this, a matrix that illustrates the frequency of each Service Class utilization for a patient is generated. In other words, each row describes a patient and each column describes a Service Class. Third, unlike the process used for patient clustering, to cluster Service Classes, the matrix is transposed such that each column corresponds to patients that had an interaction with a Service Class, and rows correspond to Service Classes that patients engaged with. Fourth, TF-IDF is applied on the resulting transposed matrix to have normalized counts. Fifth, cosine similarity is applied to the service vectors to measure the degree of similarity between the services classes. Finally, a clustering algorithm is applied to create a cluster of services classes based on degree of similarity with regards to patients' engagement with the services.

Figure 4 provides an illustrative summary of part of the process used in using NLP to generate PSUs. Note that for purposes of privacy protection, the data in the above tables consist of simulated patient journeys.

1	FA6LC3	80	80	80	34	34	23	22	154	80	149	1
2	H5GH13	22	80	154	80	146	154	113	138	173		
3	1DA8BF3	80	80	154	134	33	80	80				
4	C3HK64	80	80	110	80	14	80	173	88	80	111	8
5	FLLA824	148	167	173	154							
...	...	...	...	...	...	...	...	...	...	...	...	
1591	FKL78K9	167	154	155	138	173						

Patient' history of services utilization as a sentence

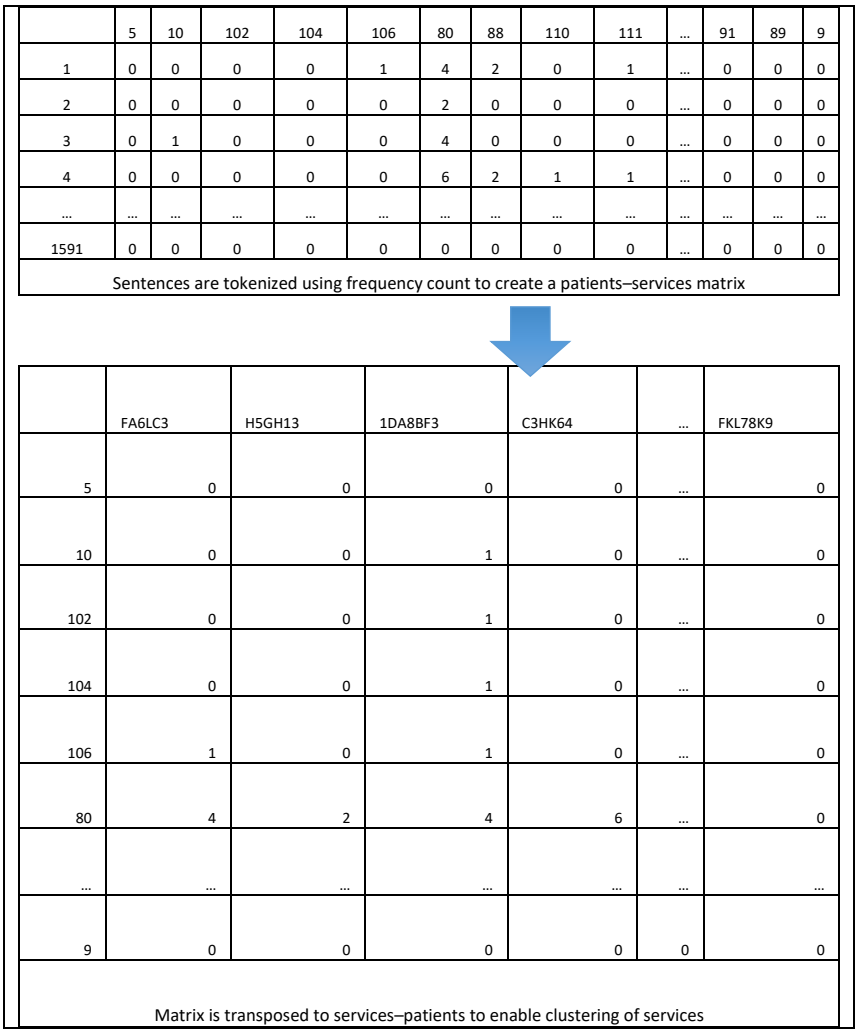


Figure 4. PSUs generation using NLP.

2.6. Combining Community Detection with the NLP to Extract PSUs

In this approach, both the capabilities of NLP and graph community detection are combined. Instead of creating a bipartite graph with patients and Service Classes, we used a TF-IDF matrix to create a projected graph for Service Classes. To accomplish this, first, each patient's history of service utilization was created as a sentence. Then, using frequency counts, these sentences were tokenized. As a result, we formed a matrix in which the columns indicate the frequency count of Service Classes for each patient and the rows represent the patients. Then, to produce normalized counts, TF-IDF was applied to the resulting transposed matrix.

To create a projected graph of Service Classes, the weights between two Service Classes were calculated by computing the dot product of each service vector with other Service Classes. This results in a measure of similarity on a service-to-service level. Hence, services that are utilized by many of the same patients will have a high dot product and a correspondingly high weight. Similarly, the services that are less accessed by the same patients will have a low dot product, resulting in low weight. Once the service-to-service graph

is created, the Louvain algorithm can be applied iteratively, as previously described in generating PSU using graph community detection.

3. Analysis

In collaboration with clinical SMEs, a cohort of patients who have taken an opioid overdose (OD-cohort) was considered. The data used represent anonymized cross-continuum patients' data, extracted from the host organization's CIS. The OD-cohort was analyzed, applying the methods described in the previous sections. The data contained 5279 patients (1606 females, 3672 males, and 1 unknown sex), aged between 14 and 92 years, with a range between 1 and 200 interactions.

For the analysis, three approaches were used. First, we performed community detection using weights from the bipartite projection. Second, we applied NLP using TF-IDF, cosine similarity, and clustering algorithms. Note that for NLP, two clustering approaches were used, including K-means and hierarchical clustering. Third, NLP, using TF-IDF and cosine similarity, was combined with the community detection algorithm.

The analyses were conducted in Python using the libraries igraph [38], scikit-learn [39], and scipy [40]. We also used the pandas and numpy libraries for data pre-processing.

3.1. Analysis Using Community Detection

The Louvain community detection was run iteratively (as described in Section 2.4). Focusing on the OD-cohort, we found that the number of communities did not increase after five iterations. The number of communities ranged from four communities at one iteration to thirty-one at five iterations. With one iteration, the resulting communities were too functionally heterogeneous to be meaningful. On the other hand, with five iterations, the communities that were generated were "too small". By "too small", we mean that in an analysis of another sample of patients contending with the same cohort-defining characteristics, we are likely not to reproduce the same communities at that more granular level of resolution.

A plot of modularity value versus the number of iterations is displayed in Figure 5. Based on this figure and the elbow heuristic, the optimal number of iterations is between two and three (as the modularity plateaus after three). In consultation with clinical SMEs and SSOEs, we found that iteration three provided the most optimal clinically meaningful communities of services,in relationship to the key characteristics shared by all members of the cohort.

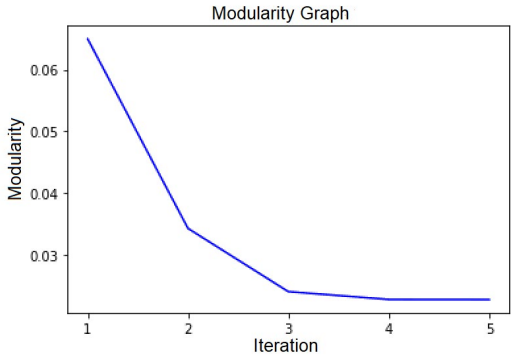


Figure 5. Plot of modularity values versus the number of iterations.

Even when the number of iterations in the community detection solution has been determined, any given community may still include services that are not related in a clearly discernable way to other services in the community, or to the cohort-defining characteristics. To trim these Service Classes from the solutions, we computed the internal and external weighted degrees of the nodes. We noticed that nodes with large internal weighted degrees tended to form communities that are relatively stable and more understandable from the SMEs and SSOEs point of view. Therefore, we focused on nodes with large internal weighted degrees. In consultation with clinical SMEs and SSOEs, a cut point was drawn on the result table listing the services classes in a community to separate and discard the Service Classes with low internal weighted degrees from the others.

3.2. Analysis Using NLP Methods with K-Means and Hierarchical Clustering

With regard to the NLP solutions, first, similarity measures among the Service Classes were built, as explained in Section 2.5. Then, various clustering algorithms were applied to the resulting matrix of similarity values. For K-means, one must choose a number k indicating the number of clusters. The elbow method provides a systematic method of determining the best k. Using this method, we plotted an average score for all clusters versus k. The score that is commonly used is the sum of square distances from each point to the centroid to which the cluster belongs. The elbow point is the inflection point on the curve [41]. The value of k for this point is regarded as the best value for k. This elbow point is not always obvious, and sometimes it is not easy to pinpoint visually. The KneeLocator function [42] was used to find the elbow point. Figure 6 shows the plot of this score. It can be noticed that for the OD-cohort, the elbow point is at around k = 9. This differs from the optimal number of communities that were generated at iteration three using the community detection approach.

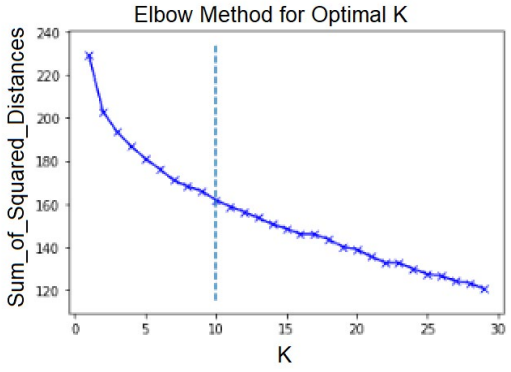


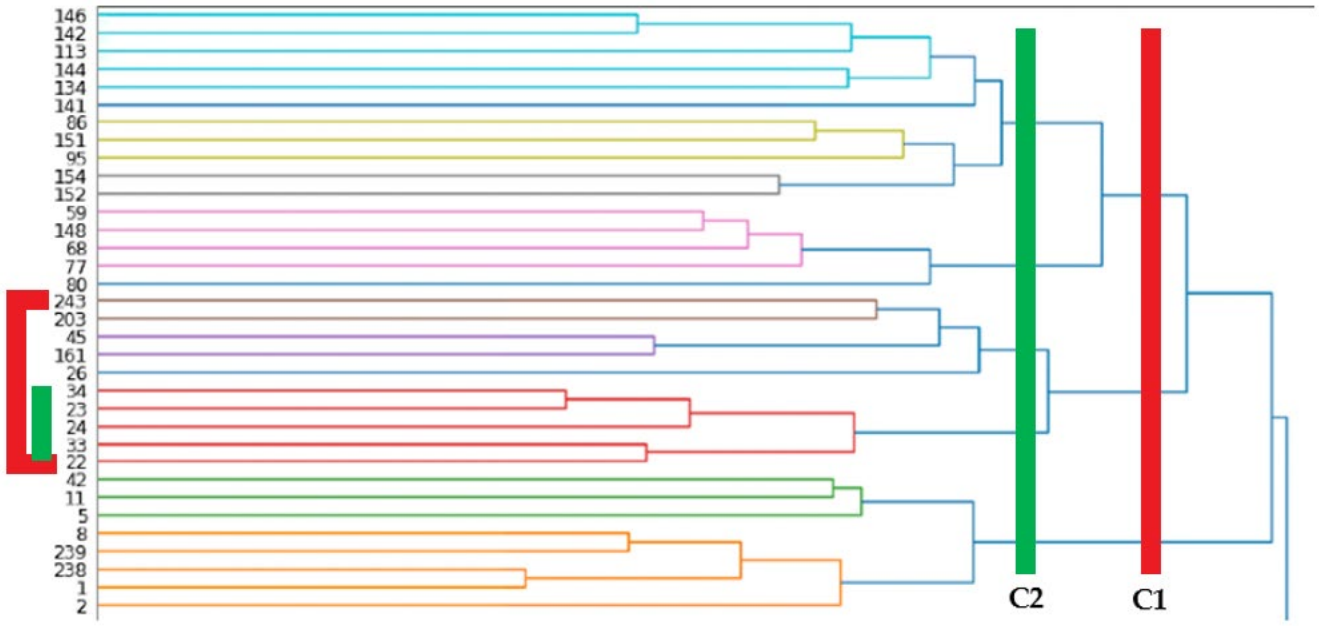
Figure 6. K-means elbow method and KneeLocator function.

Using the elbow method and setting k to 9, we generated clusters that were too big and contained Service Classes that were not related from a clinical perspective. Hence, for the purposes of the exploratory analyses

reported in this paper, we decided to set the value of k to 19, which is the number of communities that was generated in the graph community detection using iteration three. (Recall that iteration three was judged overall to be the most optimal from clinical SME/SSOE perspectives; i.e., it was judged to be the array of services most clearly related to features of clinically understandable and characterizable cohorts.)

In the result comparison section, several major clusters of K-means on the OD-cohort with k = 19 can be seen. The similarity percentage column (in Table 3) shows the probability of similarity between each Service Class and other Service Classes in the cosine similarity matrix. So, the NLP clusters were ordered based on this similarity percentage to determine the most important nodes in each cluster. This approach is similar to ordering the communities based on weighted degrees inside the community to identify the most important nodes in each community.

Next, a hierarchical clustering algorithm was applied. A sample of the results for the OD-cohort are shown in Section 4. With hierarchical clustering, the number of clusters was not set beforehand. This number was decided once the results were generated by choosing a cut-off line on the horizontal axis (Figure 7). This determined which services classes needed to be included in which clusters. Using this approach, several similarities with the communities from the graph community detection were observed, as well as similarities with the NLP clustering using K-means.



C1: 243,203,45,161,26,34,23,24,33,22

C2: 34,23,24,33,22

Figure 7. Addiction-related services for OD-cohort—NLP and hierarchical clustering.

4. Results

Several clusters of related services were generated during the analysis of the OD-cohort. To name a few, they included emergency and acute care-related services, addictions-related services, and psychiatry-related services. For this paper, the addictions-related services were chosen to highlight the similarity of results among the different approaches. In reviewing the results of the different approaches reported in this paper, we note that none of the solutions have the status of "truth"; all contribute heuristically, together with input from SMEs and SSOEs, to a working judgment of what can be treated as "true enough".

The Tables 1-5 below are organized by groupings of related clusters using different approaches. A cut-off line was included to separate the strongly connected services from the weakly connected services within clusters. For the hierarchical clustering, the markings within the diagram were used to visually separate the different cut-off points and indicate the Service Classes that were used together for comparison. Two cut-off lines (C1 and C2) were added to illustrate the flexibility of interpretation that this approach provides. Finally, at the end of the cluster results, a similarity matrix was added to capture all the Service Classes that are similar across of the different approaches.

In order to compare the different solutions, note that each Service Class has an associated Service Class ID. For the graph community detection results, the NLP plus K-means clustering results, and the NLP plus community detection solutions, each Service Class ID is paired with a Service Class label. For the NLP plus hierarchical clustering, as well as the similarity matrix, to make most effective use of space, only the Service Class IDs are displayed.

Table 1. Addiction-related services for OD-cohort—graph community detection. The red band represents the cut-off line included to separate the strongly connected services from the weakly connected services within clusters.

Service Class ID	Service Class Label	Internal Weighted Degree	External Weighted Degree
33	MHSU-Clinical Intake-Adult	2724	24,011
22	MHSU-Addictions-Clinic-Adult-Ambulatory	2655	18,690
23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	1934	10,699
24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	1439	6762
275	COVID-19 MHSU Health Monitoring	398	2607
165	MHSU-Shared Care or Collaborative Care	325	1814
40	MHSU-Personality Disorders Therapy (DBT)	10	40
284	Surgery-Day Care Antimicrobial Therapy	6	18
78	Med/Surg Intensive Acute Care-Neo-Natal	3	9

Table 2. Addiction-related services for OD-cohort—NLP and K-means clustering. .

Service Class ID	Service Class Label	Similarity Percentage
33	MHSU-Clinical Intake-Adult	9.36
22	MHSU-Addictions-Clinic-Adult-Ambulatory	7.92
23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	6.72
203	Overdose-Related Services	6.19
34	MHSU-Addictions-Clinical Intake-Adult	5.88
24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	5.44

Table 3. Addiction-related services for OD-cohort—NLP and community detection. The red band represents the cut-off line included to separate the strongly connected services from the weakly connected services within clusters.

Service Class ID	Service Class Label	Internal Weighted Degree	External Weighted Degree
23	MHSU-Addictions-Withdrawal Management (Detox)-Adults	2.3489	8.0714
34	MHSU-Addictions-Clinical Intake-Adult	2.3551	6.6388
33	MHSU-Clinical Intake-Adult	2.141	12.9558
22	MHSU-Addictions-Clinic-Adult-Ambulatory	1.9415	10.6506
24	MHSU-Addictions-Post-Withdrawal Stabilization-Residential-Adults	1.8488	5.9565
165	MHSU-Shared Care or Collaborative Care	0.8368	4.3058
21	MHSU-Addictions-Sobering & Assessment Centre	0.5759	1.642
67	MHSU-Perinatal Mental Health	0.2168	0.6528

Table 4. Solution similarities matrix. The gray band cover cells in which a Service Class is missing from respective cluster.

	Graph Community Detection	NLP + K-Mean Clustering	NLP + Hierarchical Clustering	NLP + Community Detection
Common Service Classes	22	22	22	22
	23	23	23	23
	24	24	24	24
	33	33	33	33
		34	34	34
	165			165
		203	203	

Note that the chosen cluster for illustration does not features all addiction services but addiction services with a rehabilitation/recovery orientation such as withdrawal management (Service Class 23) and post-withdrawal stabilization (Service Class 24).

5. Discussion

The purpose of this paper is to show the similarity of results across the different approaches for cross validation. Analogous to Campbell and Fiske's multitrait–multimethod approach in examining construct validity [32], this paper supplies a method for validating PSUs that were generated previously using iterative graph community detection [3]. The different approaches used

in this paper produced results that were similar across methods, where that similarity was manifest as overlaps in the Service Classes that, in effect, load most heavily on a given cluster. The slight differences in grouping of Service Classes that can be noticed among the approaches mostly affect the Service Classes that are not strongly connected to a given cluster regardless of method. This similarity in results provides cross validation for the PSUs, demonstrating that they are not artifacts of the method employed to produce the solutions.

In addition to graph community detection methods, the methodology in this paper explores how we can take advantage of NLP capabilities to extract PSUs. To do this, spurious or variable granularity of the services needs to be reduced. This was accomplished by using a clinical context coding scheme (CCCS) [33]. This is a semantic layer that groups Service Units into a reduced set of equivalence classes (Service Classes) that are relatively homogeneous with regard to their clinical functions. Then, a patient journey can be viewed as a sentence, or a string of words, in which the words are made of series of encounters with the CCCS-based Service Classes, arranged in the order in which they occurred. One can then apply the TF-IDF method [36] and cosine similarity to identify similarities among patients in a chosen cohort. Based on these similarity measures, one can cluster the Service Classes that are commonly used by similar patients. These clusters can then qualify as PSUs upon review by clinical SMEs and SSOEs. In the analysis conducted in this paper, several clustering algorithms were employed, and the results were compared. These include K-means [43] and hierarchical clustering [44]. To our knowledge, focusing on longitudinal heterogeneous cross-continuum healthcare data to extract PSUs, this is the first time NLP has been used in this manner.

Furthermore, the uniqueness of each approach provides an opportunity to take advantage of the various other capabilities that each approach brings. In future work, as the concepts of predictions using patient journeys similarities are expanded, both NLP capabilities and graph various metrics, or a combination of both, will be used.

The methods outlined in this paper are applied to around 200 equivalence classes (i.e., Service Classes) generated by applying the six sets of CCCS codes to the source data, represented as patients' encounters with roughly 2000 Service Units. Because some significant portion of the granularity of the data at the Service Unit level is not related to clinical purpose or function, given the sparseness and high dimensionality of the data, it is unlikely that the methods used in this paper would generate meaningful or usable results without prior aggregation via the use of the CCCS scheme.

The methodology outlined in our previous work [3] stresses the need to engage with clinical SMEs and SSOEs in (1) providing the taxonomies and ontologies of the Service Classes, as well as the cohort definitions, and (2) determining the level of granularity that produces the most clinically meaningful result. This still holds true for the proposed NLP method. Both the CCCS scheme and the cohort definition preceded the application of any NLP methods. Additionally, it was demonstrated that the optimal value for k in K-means, as computed via the elbow method, was not sufficient for the purposes of generating the most clinically interpretable clusters of service. In

other words, the application of purely objective/quantitative criteria will not enable these methods to converge on the "best" solutions. They provide information of heuristic value that can be used by SMEs and SSOEs to arrive at solutions that are "true enough" to employ the results for other purposes, e.g., prediction models.

In support of the above, the capabilities of visually assessing the results and picking the appropriate iteration for community detection or cut-off point for NLP plus hierarchical clustering can demonstrate the power of visual analytics. This provides a platform that makes the collaboration between data scientists, ML specialists, and SME/SSOE easier, more efficient, and transparent.

Though K-means and hierarchical clustering represent the most frequently used algorithms [45], there are various other existing clustering algorithms. Moreover, there are other types of approaches to accessing vectors similarities using NLP. We have described a popular protocol in NLP that provides a document-level view of a corpus; however, NLP offers finer views of text as well. The algorithm Word2Vec [46] takes a corpus of documents and produces a vectorization of each word in the vocabulary of the corpus. Besides the ones described in this paper, other clustering algorithms or NLP approaches were not used. This is a limitation of this study.

In future studies we plan to use Word2Vec combined with other dimensionality reduction techniques such as tSNE, and we will apply other types of clustering algorithm, including Gaussian mixture models and structural clustering, to extract PSUs.

6. Conclusions

Using a cohort of patients contending with addictions, a set of analyses using anonymized cross-continuum patients' data, extracted from the host organization's clinical information system (CIS), was performed. The analysis consisted of three different approaches: (1) graph community detection; (2) NLP using TF-IDF (term frequency inverse document frequency), cosine similarity, and clustering algorithms; and (3) a combination of both approaches. The analyses produced comparable results from a clinical perspective, especially for services that are strongly connected. Hence, this paper begins to take on the challenge of providing what is, in effect, construct validation [32,47] for ML-derived entities.

Moreover, with the rapid advancement of transformer-based models, such as ChatGPT, used for processing NLP tasks, the work outlined in this paper provides a first step in, and a basis for, generating cohort-specific simulated data. This has the potential to provide high-quality synthetic data that still maintain the statistical characteristics of the real data. Additionally, the work outlined in this paper opens the door for combining the capabilities of NLP with prediction using patients' encounters data. Using PSUs, algorithms such as RNN (recurrent neural networks) and random forest can be combined with NLP to predict patients who are likely to experience a certain outcome, such as an overdose, based on their pattern of engagement with the healthcare services. Such an endeavor can help address the challenge of proactivity in preventing a certain outcome, e.g., overdose, as well as

potential demand estimate, in looking after patients at risk of a certain outcome. These works are currently underway.

Author Contributions: Conceptualization, J.B., K.M., A.R., and A.K.; Data curation, J.B. and H.S.; Formal analysis, J.B., K.M., and E.C.; Investigation, J.B.; Methodology, J.B. and E.C.; Project administration, J.B.; Resources, J.B. and K.M.; Software, J.B., H.S., J.H., and Y.S.; Supervision, J.B., K.M., A.R., and A.K.; Validation, J.B.; Visualization, J.B.; Writing—original draft, J.B., L.T.E.; Writing—review and editing, J.B., Y.S., H.S., K.M., A.R., A.K., J.H., and L.T.E.; Funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: A certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada, Ethics harmonization guideline. The REB number is H21-02817.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The datasets presented in this article are unavailable because of privacy or ethical restrictions. Requests to access the datasets require a certificate of approval by the University of Victoria Research Ethics Board, following the British Columbia, Canada, Ethics harmonization guideline.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Panteli, D.; Legido-Quigley, H.; Reichebner, C.; Ollenschläger, G.; Schäfer, C.; Busse, R. Clinical practice guidelines as a quality strategy. In *Improving Healthcare Quality in Europe*; OECD Publishing: Paris, France, 2019; p. 233.
2. Howlett, J.G.; McKelvie, R.S.; Costigan, J.; Ducharme, A.; Estrella-Holder, E.; Ezekowitz, J.A.; Giannetti, N.; Haddad, H.; Heckman, G.A.; Herd, A.M. The 2010 Canadian Cardiovascular Society guidelines for the diagnosis and management of heart failure update: Heart failure in ethnic minority populations, heart failure and pregnancy, disease management, and quality improvement/assurance programs. *Can. J. Cardiol.* **2010**, *26*, 185–202.
3. Bambi, J.; Santoso, Y.; Sadri, H.; Moselle, K.; Rudnick, A.; Robertson, S.; Chang, E.; Kuo, A.; Howie, J.; Dong, G.Y. A methodological approach to extracting patterns of service utilization from a cross-continuum high dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems. *BioMedInformatics* **2024**, *4*, 946–965.
4. Dawkins, B.; Renwick, C.; Ensor, T.; Shinkins, B.; Jayne, D.; Meads, D. What factors affect patients' ability to access healthcare? An overview of systematic reviews. *Trop. Med. Int. Health* **2021**, *26*, 1177–1188.
5. Stangl, A.L.; Earnshaw, V.A.; Logie, C.H.; Van Brakel, W.C.; Simbayi, L.; Barré, I.; Dovidio, J.F. The Health Stigma and Discrimination Framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Med.* **2019**, *17*, 31.
6. Cradock-O'Leary, J.; Young, A.S.; Yano, E.M.; Wang, M.; Lee, M.L. Use of general medical Services by VA patients with psychiatric disorders. *Psychiatr. Serv.* **2002**, *53*, 874–878. https://doi.org/10.1176/appi.ps.53.7.874.
7. Christiani, A.; Hudson, A.L.; Nyamathi, A.; Mutere, M.; Sweat, J. Attitudes of homeless and drug-using youth regarding barriers and facilitators in delivery of quality and culturally sensitive health care. *J. Child Adolesc. Psychiatr. Nurs.* **2008**, *21*, 154–163. https://doi.org/10.1111/j.1744-6171.2008.00139.x.
8. De Groot, V.; Beckerman, H.; Lankhorst, G.J.; Bouter, L.M. How to measure comorbidity: A critical review of available methods. *J. Clin. Epidemiol.* **2003**, *56*, 221–229.
9. UNAIDS: Joint United Nations Programme on HIV/AIDS. Protocol for the identification of discrimination against people living with HIV. In *Protocol for the Identification of Discrimination against People Living with HIV*; UNAIDS: Genava, Switzerland, 2000; p. 40.

10. Nyblade, L.; Stockton, M.A.; Giger, K.; Bond, V.; Ekstrand, M.L.; Lean, R.M.; Mitchell, E.M.H.; Nelson, L.R.E.; Sapag, J.C.; Siraprapasiri, T. Stigma in health facilities: Why it matters and how we can change it. *BMC Med.* **2019**, *17*, 25.
11. Iezzoni, L.I.; McCarthy, E.P.; Davis, R.B.; Siebens, H. Mobility impairments and use of screening and preventive services. *Am. J. Public Health* **2000**, *90*, 955–961. https://doi.org/10.2105/AJPH.90.6.955.
12. Barabási, A.-L.; Loscalzo, J.; Silverman, E.K. *Network Medicine: Complex Systems in Human Disease and Therapeutics*; Harvard University Press: Cambridge, MA, USA, 2017.
13. Mislove, A.; Marcon, M.; Gummadi, K.P.; Druschel, P.; Bhattacharjee, B. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, CA, USA, 24–26 October 2007; pp. 29–42.
14. Pavlopoulos, G.A.; Secrier, M.; Moschopoulos, C.N.; Soldatos, T.G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P.G. Using graph theory to analyze biological networks. *BioData Min.* **2011**, *4*, 1–27.
15. Wysocki, K.; Ritter, L. Diseasome. *Annu. Rev. Nurs. Res.* **2011**, *29*, 55–72.
16. Rostami, M.; Oussalah, M.; Berahmand, K.; Farrahi, V. Community detection algorithms in healthcare applications: A systematic review. *IEEE Access* **2023**, *11*, 30247–30272.
17. Toor, R.; Chana, I. Network Analysis as a Computational technique and its benefaction for predictive analysis of healthcare data: A systematic review. *Arch. Comput. Methods Eng.* **2021**, *28*, 1689–1711.
18. Yi, H.-C.; You, Z.-H.; Huang, D.-S.; Kwoh, C.K. Graph representation learning in bioinformatics: Trends, methods and applications. *Brief. Bioinform.* **2021**, *23*, bbab340.
19. Wanyan, T.; Kang, M.; Badgeley, M.A.; Johnson, K.W.; De Freitas, J.K.; Chaudhry, F.F.; Vaid, A.; Zhao, S.; Miotto, R.; Nadkarni, G.N. Heterogeneous graph embeddings of electronic health records improve critical care disease predictions. In Proceedings of the Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, 25–28 August 2020; pp. 14–25.
20. Wu, T.; Wang, Y.; Wang, Y.; Zhao, E.; Yuan, Y. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Sci. Rep.* **2021**, *11*, 5858.
21. Niyirora, J.; Aragones, O. Network analysis of medical care services. *Health Inform. J.* **2020**, *26*, 1631–1658. https://doi.org/10.1177/1460458219887047.
22. Palmer, R.; Utley, M.; Fulop, N.J.; O'Connor, S. Using visualisation methods to analyse referral networks within community health care among patients aged 65 years and over. *Health Inform. J.* **2020**, *26*, 354–375.
23. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174.
24. Yin, H.; Benson, A.R.; Leskovec, J.; Gleich, D.F. Local higher-order graph clustering. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 555–564.
25. Clauset, A.; Newman, M.E.J.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111.
26. Newman, M.E.J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **2006**, *74*, 036104.
27. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008.
28. Stewart, R.; Velupillai, S. Applied natural language processing in mental health big data. *Neuropsychopharmacology* **2021**, *46*, 252.
29. Souili, A.; Cavallucci, D.; Rousselot, F. Natural Language Processing (NLP)—A Solution for Knowledge Extraction from Patent Unstructured Data. *Procedia Eng.* **2015**, *131*, 635–643. https://doi.org/10.1016/j.proeng.2015.12.457.
30. Silverman, G.M.; Sahoo, H.S.; Ingraham, N.E.; Lupei, M.; Puskarich, M.A.; Usher, M.; Dries, J.; Finzel, R.L.; Murray, E.; Sartori, J. NLP methods for extraction of symptoms from unstructured data for use in prognostic covid-19 analytic models. *J. Artif. Intell. Res.* **2021**, *72*, 429–474.
31. Reyes-Ortiz, J.A.; González-Beltrán, B.A.; Gallardo-López, L. Clinical decision support systems: A survey of NLP-based approaches from unstructured data. In Proceedings of the 2015 26th International Workshop on Database and Expert Systems Applications (dexa), Valencia, Spain, 1–4 September 2015; pp. 163–167.
32. Campbell, D.T.; Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **1959**, *56*, 81.

33. Koval, A.; Moselle, K. Clinical Context Coding Scheme—Describing utilisation of services of Island Health between 2007–2017. In Proceedings of Conference of the International Population Data Linkage Association, Banff, AB, Canada, 12–14 September 2018.
34. Bambi, J.; Santoso, Y.; Moselle, K.; Robertson, S.; Rudnick, A.; Chang, E.; Kuo, A. Analyzing patterns of service utilization using graph topology to understand the dynamic of the engagement of patients with complex problems with health services. *BioMedInformatics* **2024**, *4*, 1071–1084.
35. Bambi, J.; Dong, G.Y.; Santoso, Y.; Moselle, K.; Dugas, S.; Olobatuyi, K.; Rudnick, A.; Chang, E.; Kuo, A. Patterns of service utilization across the full continuum of care: Using patient journeys to assess disparities in access to health services. *Knowledge* **2024**, *4*, 252–264.
36. Ramos, J. Using TF-IDF to determine word relevance in document queries. In Proceedings of the first Instructional Conference on Machine Learning, Los Angeles, CA, USA, 23–24 June 2003; Volume: 242, pp. 29–48.
37. Lahitani, A.R.; Permanasari, A.E.; Setiawan, N.A. Cosine similarity to determine similarity measure: Study case in online essay assessment. In Proceedings of the 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 26–27 April 2016; pp. 1–6.
38. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **2006**, *1695*, 1–9.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Nunez-Iglesias, J.; Van Der Walt, S.; Dashnow, H. *Elegant SciPy: The Art of Scientific Python*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
41. Cui, M. Introduction to the K-means clustering algorithm based on the elbow method. *Account. Audit. Financ.* **2020**, *1*, 5–8.
42. Satopaa, V.; Albrecht, J.; Irwin, D.; Raghavan, B. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops, Minneapolis, MN, USA, 20–24 June 2011; pp. 166–171.
43. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Society. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108.
44. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.
45. Karthikeyan, B.; George, D.J.; Manikandan, G.; Thomas, T. A comparative study on K-means clustering and agglomerative hierarchical clustering. *Int. J. Emerg. Trends Eng. Res.* **2020**, *8*, 1600–1604.
46. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
47. Cronbach, L.J.; Meehl, P.E. Construct validity in psychological tests. *Psychol. Bull.* **1955**, *52*, 281.

Chapter 4: 3rd Manuscript

BioMedInformatics

MDPI

Article

Analyzing Patterns of Service Utilization Using Graph Topology to Understand the Dynamic of the Engagement of Patients with Complex Problems with Health Services

Jonas Bambi [1], Yudi Santoso [2], Ken Moselle [1], Stan Robertson [3], Abraham Rudnick [4,*], Ernie Chang [5] and Alex Kuo [1]

Citation: Bambi, J.; Santoso, Y.; Moselle, K.; Robertson, S.; Rudnick, A.; Chang, E.; Kuo, A. Analyzing Patterns of Service Utilization Using Graph Topology to Understand the Dynamic of the Engagement of Patients with Complex Problems with Health Services. *Biomedinformatics* **2024**, *4*, x. https://doi.org/10.3390/xxxxx
Academic Editor(s): Name

Received: 5 February 2024
Revised: 3 March 2024
Accepted: 29 March 2024
Published: date

[1] University of Victoria, Victoria, BC V8P 5C2, Canada; jonasbambi@uvic.ca (J.B.); kmoselle@uvic.ca (K.M.); akuo@uvic.ca (A.K.)
[2] No Affiliation; y.santoso8@gmail.com
[3] No Affiliation; stanrobertson@shaw.ca
[4] Dalhousie University, Halifax, NS B3H 4R2, Canada
[5] Retired Physician; ecsendmail@gmail.com

* Correspondence: abraham.rudnick@nshealth.ca

Abstract: Background: Providing care to persons with complex problems is inherently difficult due to several factors, including the impacts of proximal determinants of health, treatment response, the natural emergence of comorbidities, and service system capacity to provide timely required services. Providing visibility into the dynamics of patients' engagement can help to optimize care for patients with complex problems. Method: In a previous work, graph machine learning and NLP methods were used to model the products of service system dynamics as atemporal entities, using a data model that collapsed patient encounter events across time. In this paper, the order of events is put back into the data model to provide topological depictions of the dynamics that are embodied in patients' movement across a complex healthcare system. Result: The results show that directed graphs are well suited to the task of depicting the way that the diverse components of the system are functionally coupled—or remain disconnected—by patient journeys. Conclusion: By setting the resolution on the graph topology visualization, important characteristics can be highlighted, including highly prevalent repeating sequences of service events readily interpretable by clinical subject matter experts. Moreover, this methodology provides a first step in addressing the challenge of locating potential operational problems for patients with complex issues engaging with a complex healthcare service system.

Keywords: clinical pathways; clinical practice guideline; decision support; electronic healthcare; graph topology; health information management; health service system; machine learning algorithms; hub and spoke; patients' engagement dynamics

1. Introduction

1.1. Engagement of Patients with the Healthcare Service System

Persons and populations contending with chronic and progressively more complex chronic health conditions will typically interact with an increasingly diverse array of health services over time [1,2]. The clinical trajectories for such persons, and the experience of sentinel clinical events, such as organ failure or the loss of functional capacity, are an emergent characteristic of the dynamics of persons interacting with a complexly structured health service system.

Providing care to patients with complex problems, while relying *solely* on the traditional taxonomical diagnostic approach to care management, can be a challenge for the healthcare system and may not be of benefit to the patients [3]. This approach is also inherently limited. Chronic disease treatment and management consume over 42% of the total direct medical care expenditure in Canada [4], over 50% in the US [5,6], and is a key factor driving the overall growth in spending in the traditional Medicare program [7]. Most notably, outcomes will be impacted by the dynamics of patients with complexly emerging comorbidities interacting with a changing array of services that respond to these emerging problems and diagnoses. In this case, outcomes are an emergent characteristic of the dynamic interplay of the pathophysiology of chronic diseases, the proximal determinants of the health profile of the patient (e.g., health risk behavior; adherence to treatment protocols), and the accessibility of services, which itself is complexly determined by funding, geography, and public health emergencies that compete for resources, just to name a few.

In [8,9], graph machine learning (ML) and natural language processing (NLP) methods were used to model the products of service system dynamics as atemporal entities, using a data model that collapsed patient encounter events across time. The result is clusters of services where the clustering reflects functional proximity and the co-emergence of access to multiple services within individual patient records. The resulting static models show how closely positioned services are to one another. However, this model failed to capture the strength of the coupling between services as well as the order of service access events. One can optimize service system operations from such models when the outcomes for patients are a simple linear combination of the services or impacts or effects of each component. In this case, optimization is a direct extrapolation from the population

incidence/prevalence of focal conditions. The structures and functional organization of the components of the service system map cleanly onto discrete diagnostic entities.

To relate outcomes to the dynamics of patients interacting with different components of the service system over time, the order of events needs to be incorporated back into the models created in [8,9]. The objective is to provide visibility into the interoperation of the multiple components of the service system based on the prevalent longitudinal features of patient journeys through the health service system. To the extent that outcomes for a complex chronic patient level are coupled with those dynamics, these models of service system dynamics are essential for service system optimization.

This paper presents a methodology for discovering and depicting the dynamics that characterize the interoperation of multiple components of a complex health service system, in relation to cohorts contending with increasingly impactful chronic conditions. These dynamics appear as longitudinal patterns of service utilization, sparsely distributed in a high-dimensional array of services spanning a full array of secondary and tertiary services. Secondary and tertiary services include hospital and community-based services, outreach, residential care, case management, and numerous others, for medical/surgical problems such as cardiovascular disease as well as mental health and addiction services.

1.2. Graph/Network Modeling and Analysis

Since the last century, network/graph analysis has become an indispensable tool for analyzing systems whose structures and dynamics are embodied in patterns of discrete interactions among large arrays of elements [10,11]. The network models of service system dynamics provide a basis for designing and staging interventions that will alter those dynamics and associated outcomes. This is compared to interventions that increase the supply of select components in a system, without altering the dynamics. More of the same dynamics can be expected to engender more of the same outcomes.

What links together a very large body of work using graph analysis for a very diverse array of problems is a base representation of the source data as a set of entities (nodes) and connections between the nodes (edges) that reflect the dynamics of the system. Those edges may reflect asymmetrical processes where something associated with one node either precedes some other nodal event in time, or there is some form of causal relationship between one node and another. An example would be pathways through the service system that originate in one location, e.g., an emergency department, and connect at some

future point in time to another service, e.g., an electro-diagnostic procedure, or admission to an acute care bed, or a residential care facility, or a morgue encounter. These nodes and edges form directed graphs.

As well, the edges may reflect symmetrical relationships among events that may not be directly related but are instead linked to some common factor. An example would be two health services that are not functionally connected but are linked by patients who access both. In healthcare, the mediating or connecting function may arise from the patients, not the service protocols that are embodied by programs in the system. In such a case, the nodes and edges will form an undirected graph.

Both undirected and directed graph models and analytics have a role in understanding the clinical dynamics of patients and the service system dynamics that emerge as patients interact over time with the service system. If one is concerned with the ways in which the components of the service system are related to one another, one may employ a static representation of the service system represented by an undirected graph. That is what was illustrated in [8,9]. In that case, the goal is to know how functionally proximal (and therefore accessible) are different services from one another. For example, how "far" is a withdrawal management service from a post-withdrawal stabilization bed vs a residential care facility for persons contending with severe psychiatric illness, such as schizophrenia.

If one is concerned with how the clinical/functional/behavioral risk profile of patients (over time) impact the service system operations, performing graph analytics within some clinically characterizable cohorts may be needed. An example would be a cohort of persons contending with a severe psychiatric illness with a comorbid substance use disorder, or persons contending with a chronic-relapsing condition that is not associated with an underlying severe psychiatric condition. The components of the service system may interoperate differently for these two cohorts. If the concern is with dynamics, the underlying representation of the data may consist of a directed graph.

In healthcare, network analysis has been used in many areas: (1) in precision medicine, by linking intracellular structures and processes to diagnostically relevant features [12]; (2) in the extraction and interpretation of clinically relevant signs and symptoms from a large array of sources relating to a diverse array of diagnostic entities [13]; (3) in the understanding of disease–gene associations, with a focus on understanding the construction of human disease network, diseasome [14]; (4) in the area of precision diagnostics and the facilitation in accelerating the diagnosis of rare or previously unrecognized diseases [12]; (5) in supporting various predictions,

including diagnosis, patient clinical outcomes, and readmission [15]; (6) in improving healthcare quality through clinical practice guideline, consisting of protocols relying on diverse bodies of clinical information and treatments provided [16,17], and depicted as cohort-specific recurring patterns of service utilization that actually take place within a network of local services—service pathways [18]; and (7) in the depiction of patient journeys, assembled from one or more service pathways, in response to patients contending with possibly multiple co-occurring or emerging problems [19].

There is a very large body of works that cover the first six areas referenced above. However, the literature becomes far sparser when it comes to the seventh area, which is concerned with the *de facto* clustering of services based on the interaction of the cross-continuum system with diverse clinical populations. To construct these cross-continuum longitudinal models, the source data must have a good coverage of the full space of services, so that key services contributing to the dynamics are contained within the set of events out of which the models are constructed. Because patient journeys in these high-dimensional spaces are almost invariably distinguishable, the source data carries privacy risks due to risk of re-identification [20]. This militates against the public release of the datasets required to construct these models, and that is a likely contributor to the sparseness of the published results.

The work set out in this paper was undertaken by a team located inside the firewalled boundaries of a regional health authority. This team had access to a complete patient journey dataset, de-identified to the point where the risk of identity disclosure is managed, but without the perturbative changes to the core clinical contents that would be required to render the source data in a form that would, at least, arguably be suitable for public release—see, for example, the differential privacy for one such approach [21,22].

Using these data, in this paper, a method is proposed for viewing longitudinal healthcare encounter data, spanning a cross-continuum service system, as a directed network, to uncover some network topologies, which highlight the way services are in effect coupled by patients as they access the services over time. The objective is to understand the dynamic of the engagement of patients with complex problems with the healthcare services. Therefore, the analogy of physical network topology [23] to highlight these dynamic features is used.

2. Objectives

The focus of this study was on revealing the dynamics of service system interoperations, reflecting the movement of patients with complex issues through the service system. Revealing patient

dynamics from a cross-continuum complex healthcare dataset in such a way that their depictions are clear in informing the effort to change patients' outcome by altering those dynamics is challenging. Hence, to address this, the following questions are answered in this work:

- Can patient journeys be recorded as sequences of service encounters, using a directed graph that can then be used to identify high-prevalence sequences within and across persons?
- With the proposed methodology, to what extent can one cut across the complexity of a cross-continuum service structure to capture the dynamics of the journey of patients with complex issues that clearly portray their engagement with the service system, to help to locate potential operational problems?

3. Methods

3.1. Addressing Data Granularity Issues

The health organization whose data were being used for this study provides a comprehensive array of secondary and tertiary health services. This includes acute care/intensive care services, hospital and community-based emergency response, ambulatory services, residential care services for older adults or persons contending with mental health issues, case management services, and a range of addiction harm reduction or rehab and recovery-oriented services. A certificate of approval was provided by the Research Ethics Board to conduct this research project, under the study number H21-02817.

One or more of these services are encapsulated into an array of roughly 2000 Service Units within a location built for a clinical information system used to support the delivery of care. Service unit names within the local clinical information system are often opaque, rendering them unsuitable for supervised machine learning methods that require meaningfully labeled data. As well, the service units may vary widely with regard to granularity—for example, multiple beds will appear as a single unit within an acute care facility, but multiple beds in a large array of family care homes for frail elderly will show up as multiple service units.

To address these issues, a clinical context coding scheme (CCCS) was developed [24]. This scheme is organized around six sets of codes, constituting a semantic layer applied to all of the 2000 Service Units. The roughly 200 Service Classes employed for the modeling in this paper consist of equivalence classes formed by the application of these code sets to those service units.

3.2. Representing Patient Access to Healthcare Service Systems as a Graph

In order to study the dynamics of the healthcare system, service utilization is visualized as a directed graph through the following method. In the encounter dataset, there are patients and records of

their interactions with service classes. One patient typically uses several service classes throughout his/her journey. A given patient may also access a given service class on more than one occasion, e.g., repeat admissions to an Emergency Department or routine repeated blood work. If the records of a patient are arranged sequentially in chronological order, one service can be connected to another by a "next service" relationship. Suppose patient A used service class 1, and then 2, and then 2 again, and then 3, and then 2. This sequence can be depicted as a directed graph with the service classes as nodes and directed edges that reflect the transitions. This is illustrated in Figure 1.

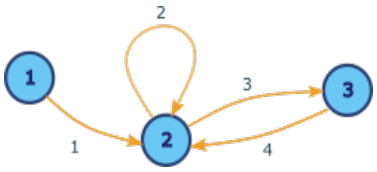


Figure 1. A directed graph of three nodes and four edges, depicting a patient journey from service class 1 to 2, to 2 again, to 3 and back to 2. Here, the edge label is for the sequence order.

For each patient, a directed graph can be drawn based on their healthcare journey. Over the course of successive service events, the graph becomes more complex, with more nodes associated with services, and more edges connecting nodes by virtue of their co-occurrence within the longitudinal record of the patient.

If a set of patients or a cohort, rather than just a single patient, is considered, the directed graphs can be aggregated to generate a single-weighted directed graph. The weights of the edges can be defined through several formulations. First, the total number of transition instances can be used as the weight, which can be referred to as the raw weight. Second, the total number of patients within the cohorts that have ever undergone such transition can be used as the weight. A third method is to generate and use transition probabilities for each node, either 'to' or 'from'.

For a node x as an example, 'probability-from' can be viewed as the empirical transition probability from x for each edge, defined by the raw weight, divided by the total number of transitions from x—Tf(x). Similarly, 'probability-to' can be computed by gathering the raw weight for each edges 'to' x divided by the total number of transitions to x—Tt(x). This is illustrated in Figure 2. In Figure 2a, there is a partial view of the graph centering on x, with raw edge weight. With respect to x, there is Tf(x) = 11 and Tt(x) = 25. In Figure 2b, the edge weight is replaced by 'probability-to', while in Figure 2c, the edge weight is replaced by 'probability-from'. To compute the new weights for all edges in the graph, all nodes need to be considered one by one.

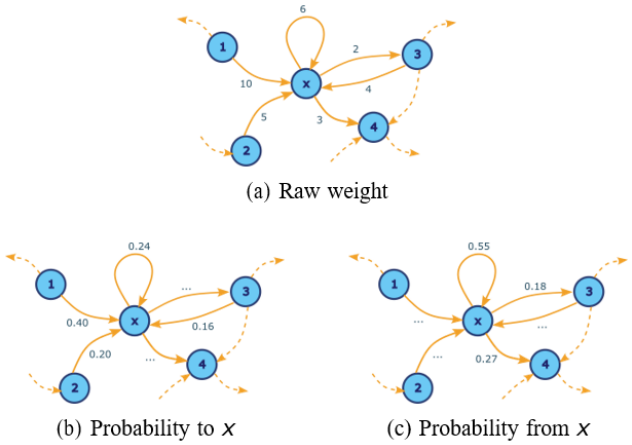


Figure 2. An example of switching from raw weight to probabilistic weight.

If there are many patients in the cohort, the aggregated graph can be very dense, with the number of edges approaching the square of the number of nodes. It would be difficult to digest such a graph visually, especially when the number of nodes is also quite large. Therefore, there is a need to take a further step to illuminate the core topology of the graph.

Transitions that are of low prevalence are less likely to reflect features that are shared by the members of the cohort. As a result, (a) one may not know when that edge appears, and (b) one may not know whether one can extrapolate from that edge to another sample of persons who share the same cohort-defining properties. Hence, if one wants to extrapolate from this particular sample of persons with feature X, the selection of the most prevalent edges is needed. In such a case, an edge with a large weight is more important than an edge with a small weight. To see the features of service system dynamics that reflect the common characteristics of the cohort members, a filter for the highest prevalent next-service-event edges is applied to the graph, by limiting the number of edges shown. This number is set to a value such that the edges that reflect only a variation in the features of the cohort members and are not directly associated with the characteristics that define the cohort are filtered out. Clinical subject matter experts may be required to make that determination.

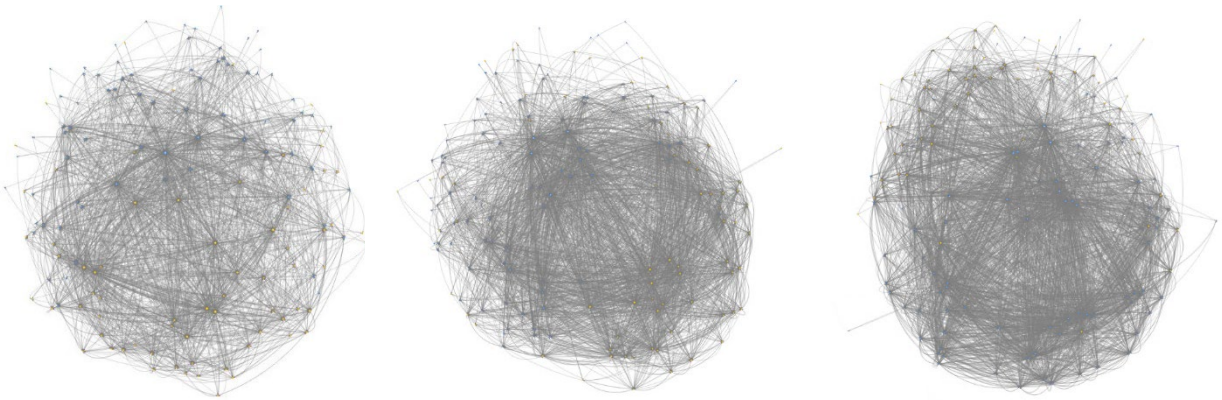
4. Analysis and Results

4.1. Cohort Selection, Analysis Setup, and Visualization Adjustment

For the analysis, the encounter data from one of the regional Health Authorities within Canada, covering the period from 2016 to early 2023, were used. Three cohorts were selected. These cohorts are quite diverse with regards to the underlying pathophysiologies. The proximal determinants of health that are intrinsic to people and their health conditions, including the pathophysiology of diseases and

reasonably predictable clusters of co-emergent conditions, were some of the characteristics that were used for selecting the cohorts of interest. Hence, Cohort A had 2008 patients with a cohort definition based on diagnosis. Cohort B had 5397 patients, reflecting a different diagnostic profile. Cohort C had 15,063 patients, where the cohort-defining feature was a combination of clinical and demographic features. For the purpose of the work presented in this paper, it was not necessary to supply clinical details.

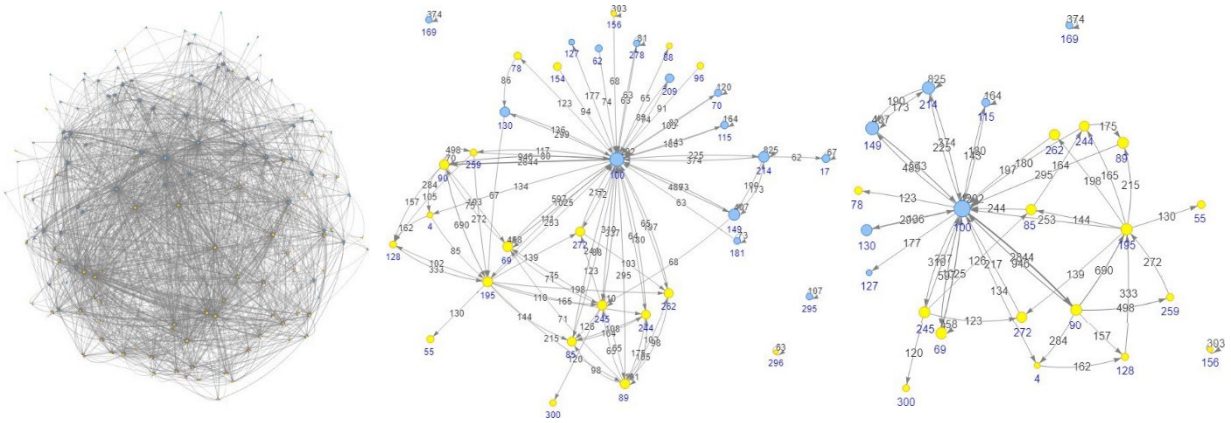
Figure 3a depicts the graph with a full set of edges for Cohort A. In Figure 3b,c, we see comparable graphs for Cohort B and Cohort C, with a full set of edges. At this level of resolution, it is difficult (at least visually) to detect any features of the graph topologies that would distinguish the products based on the three cohorts.



(**a**) Cohort A—all edges. (**b**) Cohort B—all edges. (**c**) Cohort C—all edges.

Figure 3. Comparing the visualizations for all three cohorts with full sets of edges.

Focusing on Cohort A, moving from the picture of the graph with all edges (in Figure 4a) to the picture with 100 largest raw weight edges (in Figure 4b) and 50 edges (in Figure 4c), we see that the picture becomes clearer and clearer as the number of edges decreases. However, it should be noted, as previously mentioned, that the tuning for the most suitable number of edges will need to be determined by a clinical subject matter expert (SME) and/or a healthcare service system operation expert (SSOE)—i.e., those parties that have excellent cross-continuum knowledge of the mechanisms involved in navigating across service units.



(**a**) Cohort A—all edges. (**b**) Cohort A—100 edges. (**c**) Cohort A—50 edges.

Figure 4. Demonstrates the need to adjust the number of the largest weight edges included in the visualization.

A patient may contribute more than once to an edge weight. If the concern is only on the number of patients in each transition, the weight can be changed from raw weight to the number of patients. As the numbers of edges shown are dialed down, the apparent topology of the graph might look slightly different. The graph for Cohort A with fifty edges, using the number of patients, is shown in Figure 5. This is to be compared to Figure 4c.

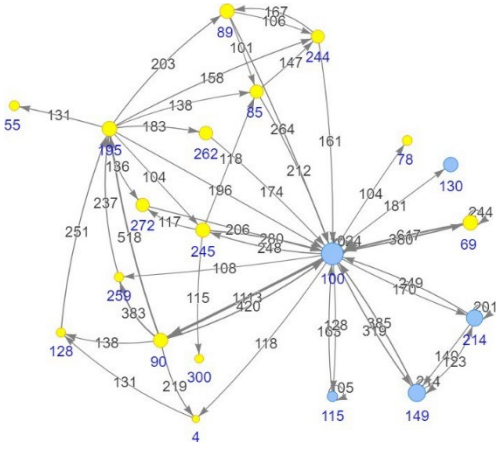
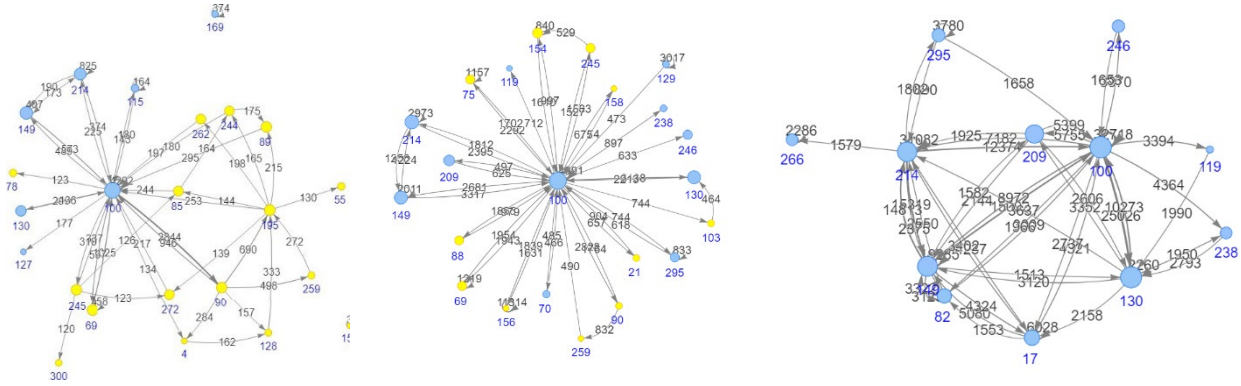


Figure 5. The graph for Cohort A using the number of patients as the edge weight and including only the 50 largest weight edges.

4.2. Cohort Topologies' Comparison

Each node represents a service class. De-identified numeric IDs were used to simplify the picture. The yellow nodes are for MHSU (mental health substance use)-related service classes, while the blue nodes are for non-MHSU or medical/surgical service classes. What is striking for Cohort A, represented in Figure 6a, is that there is a close-knitted network among the MHSU service classes, while the non-MHSU service classes are only mainly connected to node 100. It can

be noted that node 100 plays a central role in this graph. This happens to be the case for the other cohorts as well. By studying this plot, one can obtain some insights on how the service classes inter-operate for this cohort. For example, it can be noticed that many patients move from 90 to 195, either directly or through 128 or 259. Then, they move from 195 to some other MHSU service classes.



(**a**) Cohort A—50 edges. (**b**) Cohort B—50 edges. (**c**) Cohort C—50 edges.

Figure 6. Comparing the core topologies among the three cohorts.

Next, Cohort A was compared to the other two cohorts, Cohort B and Cohort C. For this comparison, the number of edges shown was fixed to 50. Note that the optimal number of edges to be shown in each cohort may differ from each other. However, for such diverse cohorts as the ones represented here, the differences in the topologies can still be seen without tuning to the most optimal value. Figure 6 is used to show the plots for these cohorts.

Here, the differences between the topologies of Cohort A and the Cohort B can be clearly seen. While in Cohort A there is a closely connected network among the many MHSU service classes, in Cohort B, the MHSU service classes are largely un-connected to each other, but rather, only connected to node 100. This forms a topology that can be referred to as a hub and spoke topology [25].

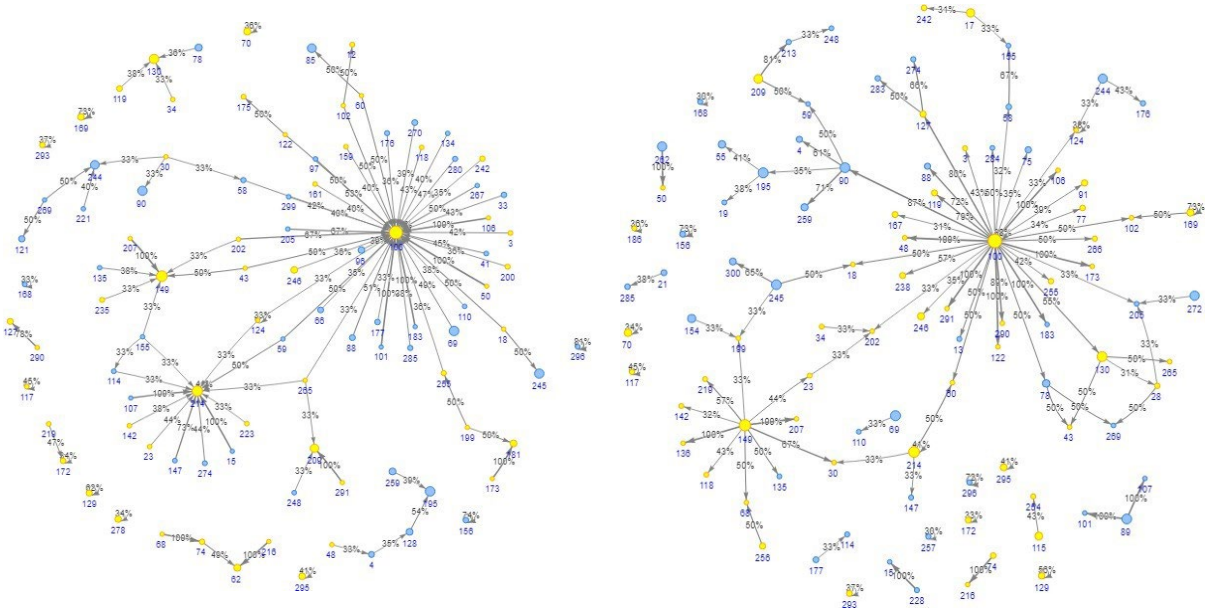
In Cohort C, the MHSU service classes are minor, used by only a small subset of patients. For this reason, they are not seen in Figure 6c, which includes only the 50 largest weight edges. Notably, the non-MHSU service classes (i.e., the medical/surgical service classes) form a closely connected network in this cohort. The node 214 and node 149 play strong roles in this network, followed by 130 and 17.

4.3. Representing a Cohort Topology Using Transition Probability

So far in this work, the number of instances and the number of patients have been used as the edge weight. Using other weights provides a way to look at the network from different perspectives. Focusing on Cohort A, the transition probabilities are now used as the

weight. Figure 7 shows two plots, one with probability-from-edge-weight and another with probability-to-edge-weight, both with one hundred edges shown. They look different from those in Figure 4b, which uses the number of instances as the edge weight. Note that these probabilistic weights are local measures, i.e., they are node-centered. Thus, they are seen from a node point of view.

In the case of 'probability-from', having chosen a node, all arrows from that node are considered. In the case of 'probability-to', having chosen a node, all arrows to that node are considered instead. For example, in Figure 7a, for many nodes, the most probable 'next' node is 100, while in Figure 7b, for many nodes, the most probable 'previous' node is 100. This reflects the fact that 100 is the most used service class for this cohort. However, it can also be noticed that some nodes are more attached to 214 or 149 instead of 100. Note that these plots with a probabilistic edge weight hide the actual number of transitions.



(**a**) 'Probability-from'. (**b**) 'Probability-to'.

Figure 7. Cohort A with 'probability-from' and 'probability-to' edge weights.

5. Discussion

This paper has demonstrated that the dynamics of a patient journey at the local level can be characterized via a next-service-event model. Following the work that was performed in [8,9], this paper injected the order of events into the model to provide a topological depiction of the service system as a whole, reflecting the dynamics of patients' movement across a complex health service system. As well, the strength of coupling between services can be assessed and

measured using metrics such as ‘probability-to’ and ‘probability-from’.

As was stated in [8,9], this exercise still relies on the engagement of SMEs and/or SSOEs to select the most appropriate cohorts of interest and determine the appropriate resolution (i.e., the number of edges) of the visualization that emerges from the graph topology. This is due to their understanding of factors that affect service access, such as waiting lists, easiness of accessibility to a service, and service functional integration or lack of it.

Moreover, similar to the manipulation of a macroscope, different cohorts may require different resolutions, depending on: (1) the size of the cohort, (2) the total numbers of interactions between the cohort members and the service system, and (3) the chosen unit of measurement for the weight—number of interactions vs number of patients vs probability of transition. Additionally, the level of homogeneity of patients in a cohort can influence the choice of the resolution. If they share some core features but are highly variable on others, there will be many edges that have low weights, and these will need to be filtered out. In some limit cases where everyone was exactly the same and used the same services in the same order, the filtering will have very little impact on the shape of the topology.

The results show that both Cohort A and Cohort C demonstrate a strong and effective service interconnectivity. For Cohort A, a strong interconnection among some MHSU services can be seen, whereas Cohort C shows an interconnection among certain medical/surgical services. Both of these cohorts demonstrate the evidence of an overarching service delivery model that wraps the services around the patients. This is not the case for Cohort B, where many pathways that move ‘from’ and ‘to’ service class 100 can be seen, with very little interconnection among themselves. Cohort B interactions represent a hub and spoke model. Considering that service class 100 represents the Emergency Department, the hub and spoke model shows how dependent Cohort B is on the emergency service to access other required services.

The hub and spoke topologies may be optimal from the standpoint of the efficient use of resources—it requires fewer edges to enable access to all points connected to the hub. However, the distance traveled may be greater than is the case in a point-to-point topology. If distance and time are correlated, and if timely access to services is important, as is the case for healthcare, the hub and spoke model may not be effective. As well, the hub may not be equipped to facilitate transport to all important nodes on the rim. Moreover, a question may arise on why is a service, such as 100, located at the center. Is it a reflection of the easy accessibility to that central service? Is it a reflection of the service range provided at the service location? These

two factors, namely easy accessibility and the service range provided, can explain the emergence of 100 as a hub and the subsequent loops. In the case of an emergency, represented in this paper as service 100, it is the hub for reasons of "least action / easy access", not "best access to a full range of services". The emergence of such a topology in a system may be indicative of problems. This may result in the overcrowding of the emergency departments, a reality that many healthcare organizations around the world are facing [26,27].

These cohort comparison visualizations show how services interoperate for clinical problems that are covered by explicitly articulated clinical practice guidelines. The notion of the "least action", as previously stated, influences the dynamics of patients' engagement with a complex service system. These guidelines introduce the notion of injecting some "force" into influencing the service system and ordering providers to behave in a certain manner. These guidelines are imposed on service system operations by parties directing and delivering care, and are supported by technology. The differences in the visualizations are immediate and striking, as shown in the topological representation of service engagement for Cohort A and Cohort C.

When the quality of care is impacted by the dynamics of patients' movement through the healthcare service system, the methods outlined in this paper may supply information and products to support the efforts in addressing quality issues. They can be very useful tools in the hand of QA/QI to help to inform service delivery changes and optimization initiatives, including the cost–benefit analysis of the various service delivery models.

Integrating service delivery is an important feature of any health service organization. However, it is hard to assess the dynamics of this integration or quantify its effectiveness. Using visualization, the method outlined in this paper can be used to assess the dynamics of an integrated service delivery structure. Additionally, using the concept of graphs and sub-graphs, such a structure's effectiveness can also be quantified.

In a subsequent paper, the findings of this paper will be expanded to analyze the quality of care implication, including treatment response/non-response, for each of the topology. In addition, the dynamic interplay of services and the access (or lack thereof) associated with those dynamics will be explored from a cost and system capacity perspective. Moreover, regarding service delivery integration, the concept of graph and sub-graph will be explored from the perspective of quantifying the effectiveness of the various topologies. Finally, the hub and spoke phenomenon will be explored in more detail to determine to what extent the emergence of such a topology is indicative of problems.

The proximal determinants of health can be distinguished into two broad categories: (1) those that are intrinsic to people and their health conditions, and (2) those that are embodied in the person's interaction with the service system. In the cohort creation and selection phase, these intrinsic person characteristics were considered. However, beyond the consideration of intrinsic person characteristics in the cohort selection, one of the contributions of the methodology proposed in this paper is to provide a way to factor in the dynamic of patients interacting with health services and treat it as a potential proximal determinants of health. However, this can be challenging to identify and measure. This will be explored in more detail in a subsequent study.

The study conducted in this paper tremendously benefited from inputs from team members with a clinical background during the analysis of the cases that were chosen for illustration. For the local implementation of the proposed methodology, it would be ideal to involve patients or persons with lived experience as well at every stage of the analysis. The characteristics of the chosen cohorts of concern drove some of the characteristics of the topologies. The people in these cohorts experience problems directly associated with behaviors that take place within larger social contexts. In order to characterize people in terms of these behaviors, methods such as community engagement or direct observational methods may need to be applied. Also, given the broader social implication, as previously stated, the inclusion of distal determinants of health into the model may be a good extension of the methodology. However, this will require access to a dataset that may not be readily available within a healthcare organization.

This methodology is geared toward identifying the core/most prevalent features, in terms of service engagement, of patients within a cohort. While this represents the usual questions that may need to be answered from an operational perspective, there may be cases where the non-core/least prevalent features may need to be known about patients within a cohort. When the number of edges is very high, the topological depiction of the cohort does not provide any visibility into the dynamics of engagement with the service for the patients in the cohort. The number of edges needs to be reduced for the resolution to provide a clearer/interpretable image of the topology. For a cohort that shares a core set of features, but that are highly variable in others, by limiting the number of edges, many edges with a low weight will be excluded. If the purpose of the analysis is to know more about the non-core characteristics of the cohort, the methodology outlined in this paper cannot provide an answer. This is a limitation of the proposed methodology.

6. Conclusions

This paper has demonstrated that directed graphs can be used to represent a sequence of patient encounter events with the healthcare service system, and the dynamics of patients' journeys with a complex cross-continuum healthcare system can be depicted using various visualizations. Additionally, by carefully selecting various resolutions for the graph topology visualizations, the various characteristics of the patient cohorts can be uncovered, including highly prevalent sequences of patient–service engagement.

The work presented in this paper is the first step in addressing the challenge of locating potential operational problems for patients with complex issues engaging with a complex healthcare service system. Following the analogy of a macroscope, the visualization of the topology, combined with the ability to adjust the resolution of a topological image, by modifying the number of edges in the graph, can help to locate the potential problematic features of the service system. However, more work will need to be conducted in collaboration with various clinical SMEs and SSOEs, incorporating information collected from various clinical guidelines, to expand on the finding of this work to uncover potential operational issues and potential solutions. This will be a focus of future works.

The dynamics of patient engagement will always be driven by multiple factors. These include the healthcare organization's capacity to provide care [28], the capacity of the patients to sustain or not sustain the engagement with the services [29,30], and the structure of the service system [31,32]. The division of the service system into sub-systems that are functionally distinctive is a useful way to provide better care for patients. It is a necessary way of organizing a large and complex dynamic system. Additionally, some sub-systems are more routinely or invariably connected than others. Moreover, patients will always be closer to some services than others, depending on their needs. The goal behind the proposed method is to identify areas of improvement, by providing visibility into the dynamics of patients' engagement and to optimize care for patients with complex problems or chronic conditions that may rely on multiple services that may not necessarily be optimally connected. The methodology is universal, but any implementation of the methodology has to be localized to the reality of the health organization.

Author Contributions: Conceptualization, J.B., Y.S., K.M., A.R., and A.K.; data curation, S.R.; formal analysis, J.B., Y.S., and K.M.; investigation, J.B., Y.S., and K.M.; methodology, J.B., Y.S., and K.M.; project administration, J.B.; resources, J.B., K.M., and S.R.; software, Y.S., S.R., and E.C.; supervision, J.B., K.M., A.R., and A.K.; validation, J.B., Y.S., K.M., and E.C.; visualization, J.B., Y.S., and K.M.; writing—original draft, J.B. and Y.S.; writing—review and editing, J.B., Y.S., K.M., and A.R.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: A certificate of approval was provided by the University of Victoria's Research Ethics Board (REB), following the British Columbia, Canada Ethics harmonization guideline. The REB number is H21-02817.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are unavailable because of privacy or ethical restrictions. Requests to access the datasets require a certificate of approval by the University of Victoria's Research Ethics Board, following the British Columbia, Canada Ethics harmonization guideline.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This section provides additional details for the tools and algorithms used to generate the visualizations outlined in this paper. This provide an in-depth technical understanding of the methodology used in this paper. It also makes it easier to implement approaches similar to the ones outlined in this paper to analyze complex healthcare dataset to generate topologies representing the dynamics of patients' engagements.

A zip file containing the R source code used for graph manipulation and visualization, along with supportive documentation (Readme File), and a sample dataset have been submitted along with this manuscript. Anyone is permitted to use the code. As previously mentioned, the data used for the study will not be made available due to privacy concerns. The sample dataset provided, as part of the appendix, is for illustrative purposes only. However, the methodology described in the manuscript, combined with the source code, can be applied to any dataset similar to the one used for the study.

References

1. Yu, W.; Ravelo, A.; Wagner, T.H.; Phibbs, C.S.; Bhandari, A.; Chen, S.; Barnett, P.G. Prevalence and costs of chronic conditions in the VA health care system. *Med. Care Res. Rev.* **2003**, *60*, 146S–167S.
2. Hoffman, C.; Rice, D.; Sung, H.-Y. Persons with chronic conditions: Their prevalence and costs. *JAMA* **1996**, *276*, 1473–1479.
3. Dalgleish, T.; Black, M.; Johnston, D.; Bevan, A. Transdiagnostic approaches to mental health problems: Current status and future directions. *J. Consult. Clin. Psychol.* **2020**, *88*, 179–195. https://doi.org/10.1037/ccp0000482.
4. Mirolla, M. *The Cost of Chronic Disease in Canada*; GPI Atlantic Glen Haven: Glen Haven, NS, Canada, 2004.
5. Anderson, G.F. *Chronic Care: Making the Case for Ongoing Care*; Robert Wood Johnson Foundation: 2010.
6. Charlson, M.; Charlson, R.E.; Briggs, W.; Hollenberg, J. Can disease management target patients most likely to generate high costs? The impact of comorbidity. *J. Gen. Intern. Med.* **2007**, *22*, 464–469.
7. Thorpe, K.E.; Ogden, L.L.; Galactionova, K. Chronic conditions account for rise in Medicare spending from 1987 to 2006. *Health Aff.* **2010**, *29*, 718–724.
8. Bambi, J.; Moselle, K.; Santoso, Y.; Sadri, H.; Robertson, S.; Hajiabadi, M.; Howie, J.; Hawkins-Seagram, A.; Richardson, A.; Rudnick, A.; et al. Methodological Considerations in Extracting and Analyzing Patterns of Service Utilization for Patients with Complex Problems to Optimize Care Delivery. **2024**.
9. Bambi, J.; Santoso, Y.; Sadri, H.; Moselle, K.; Howie, J.; Robertson, S.; Rudnick, A.; Chang, E.; Eliott, L.; Kuo, A. Approaches to Generating Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs Natural Language Processing Clustering. **2024**.

10. Prasse, B.; Van Mieghem, P. Predicting network dynamics without requiring the knowledge of the interaction graph. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2205517119.
11. Zhong, C.; Arisona, S.M.; Huang, X.; Batty, M.; Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 2178–2199.
12. Barabási, A.-L.; Loscalzo, J.; Silverman, E.K. *Network Medicine: Complex Systems in Human Disease and Therapeutics*; Harvard University Press: Cambridge, MA, USA, 2017.
13. Ahmedt-Aristizabal, D.; Armin, M.A.; Denman, S.; Fookes, C.; Petersson, L. Graph-based deep learning for medical diagnosis and analysis: Past, present and future. *Sensors* **2021**, *21*, 4758.
14. Wysocki, K.; Ritter, L. Diseasome. *Annu. Rev. Nurs. Res.* **2011**, *29*, 55–72.
15. Wu, T.; Wang, Y.; Wang, Y.; Zhao, E.; Yuan, Y. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Sci. Rep.* **2021**, *11*, 5858.
16. Panteli, D.; Legido-Quigley, H.; Reichebner, C.; Ollenschläger, G.; Schäfer, C.; Busse, R. Clinical practice guidelines as a quality strategy. *Improv. Healthc. Qual. Eur.* **2019**, 233.
17. Ellrodt, G.; Cook, D.J.; Lee, J.; Cho, M.; Hunt, D.; Weingarten, S. Evidence-based disease management. *JAMA* **1997**, *278*, 1687–1692.
18. Aggarwal, N.; Ahmed, M.; Basu, S.; Curtin, J.J.; Evans, B.J.; Matheny, M.E.; Nundy, S.; Sendak, M.P.; Shachar, C.; Shah, R.U. Advancing artificial intelligence in health settings outside the hospital and clinic. *NAM Perspect.* **2020**, *2020*.
19. Lin, Z.; Yang, D.; Yin, X. Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access* **2020**, *8*, 156663–156676.
20. El Emam, K.; Arbuckle, L. *Anonymizing Health Data: Case Studies and Methods to Get You Started*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2013.
21. Bambauer, J.; Muralidhar, K.; Sarathy, R. Fool's gold: An illustrated critique of differential privacy. *Vand. J. Ent. Tech. L.* **2013**, *16*, 701.
22. Xu, C.; Ren, J.; Zhang, Y.; Qin, Z.; Ren, K. DPPro: Differentially Private High-Dimensional Data Release via Random Projection. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 3081–3093. https://doi.org/10.1109/TIFS.2017.2737966.
23. Alhanani, R.A.; Abouchabaka, J. An overview of different techniques and algorithms for network topology discovery. In *2014 Second World Conference on Complex Systems (WCCS)*; IEEE: Piscataway, NJ, USA, 2014; pp. 530–535.
24. Koval, A.; Moselle, K. Clinical Context Coding Scheme—Describing Utilisation of Services of Island Health between 2007–2017. In Proceedings of Conference of the International Population Data Linkage Association, Banf, AB, Canada, 12–14 Sepetember 2018.
25. Pósfai, M.; Barabási, A.-L. *Network Science*; Cambridge University Press: Cambridge, UK, 2016.
26. Durand, A.-C.; Palazzolo, S.; Tanti-Hardouin, N.; Gerbeaux, P.; Sambuc, R.; Gentile, S. Nonurgent patients in emergency departments: Rational or irresponsible consumers? Perceptions of professionals and patients. *BMC Res. Notes* **2012**, *5*, 525.
27. Uscher-Pines, L.; Pines, J.; Kellermann, A.; Gillen, E.; Mehrotra, A. Emergency department visits for nonurgent conditions: Systematic literature review. *Am. J. Manag. Care* **2013**, *19*, 47–59.
28. Johannes, B.; Graaf, D.; Blatt, B.; George, D.; Gonzalo, J.D. A multi-site exploration of barriers faced by vulnerable patient populations: A qualitative analysis exploring the needs of patients for targeted interventions in new models of care delivery. *Prim. Health Care Res. Dev.* **2019**, *20*, e61.
29. Christiani, A.; Hudson, A.L.; Nyamathi, A.; Mutere, M.; Sweat, J. Attitudes of Homeless and Drug-Using Youth Regarding Barriers and Facilitators in Delivery of Quality and Culturally Sensitive Health Care. *J. Child Adolesc. Psychiatr. Nurs.* **2008**, *21*, 154–163. https://doi.org/10.1111/j.1744-6171.2008.00139.x.
30. Cradock-O'Leary, J.; Young, A.S.; Yano, E.M.; Wang, M.; Lee, M.L. Use of General Medical Services by VA Patients With Psychiatric Disorders. *Psychiatr. Serv.* **2002**, *53*, 874–878. https://doi.org/10.1176/appi.ps.53.7.874.
31. Calder, R.; Dunkin, R.; Rochford, C.; Nichols, T. Australian health services: Too complex to navigate. **2019**.
32. Clarfield, A.M.; Bergman, H.; Kane, R. Fragmentation of care for frail older people—An international problem. Experience from three countries: Israel, Canada, and the United States. *J. Am. Geriatr. Soc.* **2001**, *49*, 1714–1721.

Chapter 5: 4th Manuscript

knowledge

MDPI

Article

Patterns of Service Utilization Across the Full Continuum of Care: Using Patient Journeys to Assess Disparities in Access to Health Services

Jonas Bambi [1,*], Gracia Yunruo Dong [2,5], Yudi Santoso [3], Ken Moselle [4], Sophie Dugas [5], Kehinde Olobatuyi [5], Abraham Rudnick [6], Ernie Chang [7], and Alex Kuo [1]

[1] Department of Health Information Science, Faculties of Human and Social Development, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; jonasbambi@uvic.ca; akuo@uvic.ca
[2] Department of Statistical Sciences, Faculties of Arts and Science, St. George Campus, University of Toronto, Toronto, M5S 1A1, Canada; gracia.dong@utoronto.ca
[3] Independent Researcher, Victoria, BC, V8R 5B4, Canada,; y.santoso8@gmail.com;
[4] Department of Clinical Psychology, Faculty of Social Science, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; kmoselle@uvic.ca
[5] Departments of Mathematics and Statistics, Faculty of Science, Victoria Campus, University of Victoria, Victoria, V8P 5C2, Canada; sophiedugas@uvic.ca; olobatuyikenny@uvic.ca
[6] Departments of Psychiatry and Bioethics, School of Occupational Therapy, Faculties of Medicine and Health, Dalhousie University, Halifax, B3H 4R2, Canada; Abraham.Rudnick@nshealth.ca
[7] Retired Physician and Independent Computer Scientist, Victoria, V9C 4B1, Canada; ecsendmail@gmail.com
* Correspondence: jonasbambi@uvic.ca or jonasbambi@gmail.com; Tel.: +1-250-507-4262

Citation: Bambi, J.; Dong, G.; Santoso, Y.; Moselle, K.; Dugas, S.; Olobatuyi, K.; Rudnick, A.; Chang, E.; Kuo, A. Patterns of Service Utilization Across the Full Continuum of Care: Using Patient Journeys to Assess Disparities in Access to Health Services. *Knowledge* **2024**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Received: 29 February 2024
Revised: 18 April 2024
Accepted: 24 April 2024
Published: date

Abstract: Healthcare organizations have a contractual obligation to the public to address population-level inequities to health services access and shed light on them. Various studies have focused on achieving equitable access to healthcare services for vulnerable patients. However, these studies do not provide a nuanced perspective based on the local reality across the full continuum of care. In previous work, graph topology was used to provide visual depictions of the dynamics of patients' movement across a complex healthcare system. Using patients' encounters data represented as a graph, this study expands on previous work and proposes a methodology to identify and quantify cohort-specific disparities in accessing healthcare services across the continuum of care. The result has demonstrated that a more nuanced approach to assessing access-to-care disparity is doable using patients' patterns of service utilization from a longitudinal cross-continuum healthcare dataset. The proposed method can be used as part of a toolkit to support healthcare organizations that wish to structure their services to provide better care to their vulnerable populations based on the local realities. This provides a first step in

addressing inequities for vulnerable patients in accessing healthcare services. However, additional steps need to be considered to fully address these inequities.

Keywords: public healthcare; mental health; healthcare equity; healthcare access; electronic healthcare; health information management; graph topology; access disparities

1. Introduction

1.1. Disparities in Accessing Healthcare Services

According to Whitehead and Dahlgren, echoing the World Health Organization constitution [1], "equity in health implies that everyone could ideally attain their full health potential, and that no one should be disadvantaged from achieving this potential because of their social position or other socially determined circumstance". Healthcare organizations operate on the assumption that they have a contractual obligation to the public, requiring them to address population-level inequities and shed light on them. Furthermore, these organizations are committed to addressing the methodological challenges associated with establishing the causal mechanisms, determinants, or correlations of these inequities. The aim is to identify factors within the control of the healthcare system that can be influenced. Should there be disparities in accessing healthcare services, the hosting organization bears the responsibility to rectify them. If healthcare service organizations wish to become effective for health interventions, organizations need to prioritize locally relevant strategies that are oriented to working with vulnerable groups [2]. Consequently, the challenge lies in tracing these inequities back to their root causes or determinants and relating them to factors that fall within the organization's purview.

Certain groups of patients considered vulnerable have been extensively examined, with considerable research focusing on achieving equitable access to healthcare services. Health inequality affects all Canadians, but for vulnerable populations, it has a much stronger impact on their health [3]. Several studies, including [4,5], have demonstrated that vulnerable patients tend to have an overall lower quality of life and health. Additionally, access to care is one of the most significant challenges affecting the quality of life of vulnerable populations [6]. The lack of proper access to healthcare can potentially lead to a reduced life span compared to the general population. Such population groups include individuals affected by, e.g., psychiatric issues such as schizophrenia [7,8]. Research has shown that individuals with schizophrenia have an exceptionally short life expectancy, averaging approximately 20 years below that of the general population [8]. Moreover, high mortality is found in all age groups, with two-thirds of this excess mortality caused by natural deaths [7,8]. One of the factors that influence this reduction in life expectancy include the fact that common physical illness are diagnosed late and do not receive sufficient treatment, in spite of the fact that people with schizophrenia tend to live an unhealthy lifestyle with regards to diet, smoking, alcohol consumption, and lack of physical exercises [8].

People with mental health and substance use (MHSU) [9] and those suffering from homelessness [9,10] as well as physical mobility [11] and

neurocognitive disorders [12] tend to have less access to healthcare services. Lack of access to proper services for serious/chronic mental health issues has also been shown to be one of the major challenges in treating those mental health and addictions issues that lead to homelessness and an excessive use of some services such as emergency departments [10,13]. Moreover, mobility limitations is especially true for older adults [14,15] and may be due to chronic conditions, including arthritis and chronic lung problems [16,17]. Access to physicians and sub-specialty care has also been shown to be more difficult for patients with mobility impairment [18], hence compromising their quality of care delivery. Furthermore, despite the fact that neurocognitive disorders represent a growing major concern in the world, with a growing aging population, treatment and management of neurocognitively impaired patients remains suboptimal due to inadequate treatment and poor quality of care, increasing the risks of adverse outcome [12].

Various stigmatization are also known to negatively impact access to healthcare services [16,19,20]. There are various types of stigmas, including health condition-related stigma and non-health-related stigma. Health condition-related stigma affects people living with a specific disease or health condition, including leprosy, epilepsy, mental health disorders, cancer, HIV, and obesity/overweight [21]. Non-health-related stigma include socio-economic status, age, gender, race, and sexual orientation [22]. A stigma can be a barrier to care for patients seeking support to maintain a healthy quality of life, seeking disease prevention service, or treatment for acute or chronic healthcare conditions [23]. In health facilities, the manifestation of stigma may result in subtle forms of deprivation such as longer waits for services or being referred to junior providers [23,24]. It may also manifest in an outright denial of services, physical or verbal abuse, or provision of sub-standard care [23]. As a result, stigma within healthcare may undermine access to diagnosis and treatment and result in poor or unsuccessful health outcomes [23,25–27].

The above-cited studies capture the typical reality of cohorts of patients accessing specific services. The challenge is the fact that these studies do not provide a nuanced perspective based on the local reality across the full care continuum. It is common to look at differences in access to certain services, such as how often a cohort of patients accesses the emergency department. It is also common to look at differential access to a specific service within a medical specialty such as cardiovascular. However, a view of patient access from a cross-continuum, including secondary and tertiary services, is what is lacking from these studies. If the goal is to address disparities in access to services for vulnerable patients that may be affected by a diverse array of needs, there is no need to restrict one's view to a limited number of services, but one should rather provide visibility to all services across the continuum of care. Using local data to capture the local reality, the advocated approach provides the opportunity to (1) determine whether the disparity shows up consistently across areas or is specific to some services and not others and (2) facilitate the determination of whether the disparity in access lies with the patients' capacity to advocate for services and/or initiate/maintain access to services or with the service system structure or a combination of the two factors. This provides a potential for a tangible solution to address the disparity.

Moreover, if we are concerned with disparities in access, there is an immediate problem that need to be addressed. Addressing disproportionately high rates of access to a given service, e.g., large numbers of emergency department visits for persons in a specific demographic or contending with a specific underlying condition, is methodologically straightforward. However, when addressing disparities in access, our efforts frequently involve compiling aggregate counts for events that have yet to occur. To do this, we need some sort of reference standard or expected value, enabling us to interpret measured rates effectively and transform disparities into inequities. The problem is that for many cohorts and for many services, such standards may not exist, or they may be imprecise and therefore poorly positioned as a basis for determining when rates of service access fall "out of range". For example, how many cardiovascular-related investigative procedures should be performed for persons with schizophrenia who are not yet displaying obvious signs/symptoms of cardiovascular illness? Such information may or may not be found in the literature. If we further partition that group into persons with schizophrenia and a comorbid substance use disorder vs. persons with schizophrenia without a comorbid substance use disorder, will the reference standard be readily available in the literature as well? Operating on the assumption that externally supplied reference standards are generally not going to be available for a method that must be general in scope, we may resort to cohort comparison design—using rates for other groups as a proxy for external reference standards.

Prior works [28–30] have developed methodologies to extract patterns of service utilization (PSUs) from longitudinal electronic healthcare records to optimize healthcare services. Expanding on this by (1) using source data consisting of longitudinal transactional service encounter data provided by one of the health authorities in Canada and (2) relying on a cohort comparison design, this paper proposes a methodology that uses patients' encounter data represented as a graph to analyze access to healthcare services for a set of patient cohorts with varied levels of vulnerabilities, as chosen by clinical subject-matter experts (SMEs). The goal is to corroborate previous findings and offer a more nuanced perspective on access inequities. Providing this type of analysis will help facilitate the identification of the factors that are within the control of the host organization to address the inequities. Moreover, the focus of the paper is methodological, and a host organization can apply the proposed methodology to any cohort of interest.

1.2. Objectives

This paper illustrates a method for locating cohort-specific disparities in health service access within large and sparse high-dimensional, full cross-continuum health service datasets that span hospital and community sectors for medical/surgical as well as mental health and/or substance use/addictions services. These disparities are important when they are not commensurate with need or risk, as they then become markers for inequities in health service access. Ethically, healthcare organizations are obligated to acknowledge and address these inequities by shedding light on them and identifying factors within their control. This study illustrates a specific instance of a broader methodology that utilizes a set of longitudinal health service encounter data

to identify and quantify cohort-specific disparities in access to healthcare services.

Using a cross-continuum healthcare dataset from a regional health authority, the goal is to provide a more nuanced understanding of healthcare access disparity, focusing on a chosen set of patient cohorts with distinguishing levels of vulnerabilities. Also, the work does not assume that the identified disparities arise solely from clinical cohort-defining characteristics of persons, in which case disparities would be expected to show up consistently across the service area. In keeping with the working hypothesis that disparities are a joint function of features of persons and features of service system structures and functions, the study looks for disparities in both medical/surgical and MHSU service areas.

To summarize, the work presented in the paper is organized around three overarching substantive questions:

Are there disparities in service access or patterns of service utilization (PSUs) that are associated with cohort membership? For the work in this paper, cohorts are distinguishable on the basis of expected differential capacity to initiate access to anything other than low-barrier access services, e.g., emergency departments, and remain connected to services over time. The term "vulnerable" is used in this paper to refer to that capacity.

Do identified disparities in access show up consistently across service areas?

To what extent can we use PSUs to identify healthcare access inequities and provide a more refined approach in assessing the cause of the inequity?

To address the above questions, from a methodological perspective, the following approaches are used:

1. Cohort comparison: Patients in the chosen cohorts of interest are users of medical/surgical and/or MHSU services. However, one of the cohorts is made of patients that are considered more vulnerable than the other. Using PSUs, the cohort of more vulnerable patients will be compared to a cohort of patients with less severe vulnerability in terms of access to various healthcare services. When measuring access disparities for the above cohorts, both empirical and statistical approaches will be used, and the results will be compared;

Consultation with SMEs will be conducted to review the results and to answer the questions that were raised earlier regarding determinants of access.

2. Methodology

2.1. Source Data

The source data used in this paper consist of retrospective longitudinal transactional service encounter data contents extracted from a single instance of a Clinical Information System (CIS) deployed across the continuum of services provided by one of the health authorities within Canada (hereinafter referred to as "host organization"). It provides a comprehensive array of secondary and tertiary health services for all ages, for persons contending with medical/surgical issues and/or mental health/substance use issues. This includes acute care/intensive care services, hospital and community-based emergency response, ambulatory services, residential care services for older adults or persons contending with mental health issues, case management

services, and a range of addiction harm reduction or rehab and recovery-oriented services. The encounter data accessed by this study consist of approximately 10 million encounters over 7 years for approximately 1 million patients. Except for a few restricted services where the data are strictly embargoed (e.g., services for persons who are victims of sexual assault), this represents data for all service recipients. To access the source data, a certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada, ethics harmonization guideline.

2.2. Data Pre-Processing: Addressing Data Granularity and Nomenclature Issues

The services provided by the host organization are encapsulated into an array of roughly 2000 service units within a location built of the host organization CIS. The service units may vary widely with regard to granularity; for example, multiple beds will appear as a single unit within an acute care facility, but multiple beds in a large array of family care homes for frail elderly will show up as multiple service units. Additionally, service unit names within the local clinical information system are often opaque. For example, an addiction post-withdrawal stabilization unit appears in this location as "Holly", or there is a service unit named "Clinics" that provides ambulatory services for children and youths with physical disabilities. There are large numbers of service units where the unit names are uninterpretable, or interpretation is a matter of guessing. This renders them unsuitable for any analysis that requires meaningfully labelled data. In modeling the structure and dynamics of patient interaction with services, meaningful distinctions between functions performed by services must be preserved.

The clinical context coding scheme (CCCS) [31] is a semantic layer that was developed as a flexible solution to address issues of data granularity and nomenclature. This scheme is organized around six sets of codes, constituting a semantic layer applied to all 2000+ service units. The roughly 200 service classes employed for the modeling in this paper consist of equivalence classes formed by the application of these code sets to those service units. Also, each service class has a name that bears some discernible relationship to the functions performed by the component service units. The modeling activities reported in this paper are performed on patient–service encounters with service classes.

2.3. Cohort Selection

Cohort selection is based on access to services that target specific classes of diagnoses as well as the capacity to seek out and access services. The classes of diagnoses can be distinguished on the basis of broad-based impact on psychosocial functioning and on chronicity. In this paper, the term "vulnerable" relates to the impact of a condition on the person's functioning. The term "more vulnerable" is used to refer to the cohort of persons contending with the more severe MSHU chronic and pervasively impactful class of conditions. These impacts include limited capacity to seek out and maintain access to services. The term "less vulnerable" is used to refer to the cohort of persons contending with a class of MHSU problems that are episodic and less inherently pervasive in their impact on the person and the ability to seek out and maintain access to services. Members of the "less

vulnerable" cohort are expected to be more capable of seeking out and maintaining accessing to services.

These cohorts are not mutually exclusive. There are some patients with membership in both cohorts, reflecting the co-occurrence of qualifying diagnoses for the two different cohorts. For these patients, if they met the criteria for inclusion in the more vulnerable cohort, they were treated as members of that cohort only.

Two cohorts were considered: One consists of patients with a chronic MHSU disorder. The second consists of patients with a less serious episodic MHSU disorder. Cohort 1 was defined as the group of more vulnerable patients and cohort 2 as the group of less vulnerable patients. The less vulnerable cohort comprises 21,180 patients, and the more vulnerable cohort comprises 1829 patients. Of these, 182 patients belong to both cohorts and were thus assigned to the more vulnerable cohort.

2.4. Representing Patient Access to Healthcare Services System as a Graph and Estimating Transition Probabilities Between Services

In this paper, patients' encounter data are modeled using a directed graph, and logistic regression is used to estimate transition probabilities in state space models. In the encounter dataset, there are patients and records of their interactions with service classes. One patient typically uses several service classes throughout his/her journey. A given patient may also access a given service class on more than one occasion, e.g., repeat admissions to an emergency department or routine repeated blood work. If the records of a patient are arranged sequentially in chronological order, one service can be connected to another by a "next service" relationship. Suppose patient A used service class 1, then 2, then 2 again, then 3, and then 2. This sequence can be depicted as a directed graph with the service classes as nodes and directed edges that reflect the transitions. This is illustrated in Figure 1, where the edge labels are for the sequence order. Patient A's journey can then be coded as 12232.

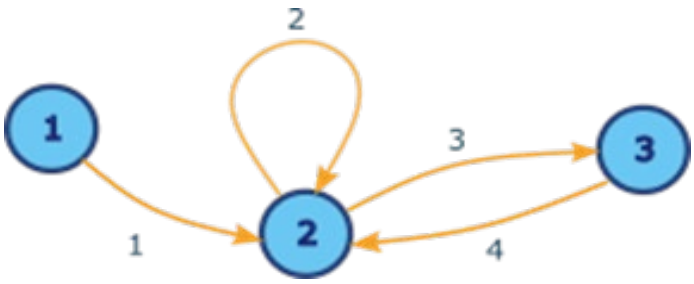


Figure 1. A directed graph of three nodes (representing service class 1,2, and 3) and four edges, depicting a patient journey from service class 1 to 2, to 2 again, to 3, and back to 2. Here, the edge labels are for the sequence order.

Now, among the service classes, we may be interested in only some of them. Thus, we can label them as 1 if it is a service class of interest or 0 otherwise. Suppose that in the example shown in Figure 1, we are interested in service class 2 only; then, we relabel service class 1 as 0, service class 2 as 1, and service class 3 as 0. This is shown in Figure 2. The patient's journey can now be read as 01101.

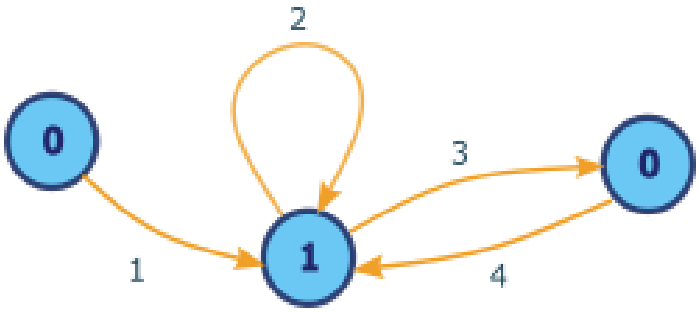


Figure 2. The directed graph from Figure 1 with the nodes relabeled as either 1 (if it is a service class of interest – service class 1) or 0 (if it is not a service class of interest – service class 0).

Graphs can be thought of as state space models, where each node is a state. This has been carried out in several fields, including rainfall data [32], transportation systems [33], and in forestry [34], and, more recently, patterns of service utilization [28,30]. Logistic regression is used to model binary outcomes encoded as 0 or 1 by assuming a linear relationship between the dependent variables and the log-odds of the outcome. That is, if Y is the outcome variable, we use the following model:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k.$$

Here, it is used to model transition probabilities between services.

Using the direct graph where service classes are arranged by the date and time of the interaction so that each patient's journey through the healthcare system is encoded as a sequence of service classes, we pick service classes of interest and then model the probability of transition to these services. Given this set of service classes, we can encode each patient's journey as a sequence of 0 s (if not in the group of interest) or 1 s (if inside the group of interest). Logistic regression is used to model the transition probabilities. Fitting of the logistic regression is performed via the glm() function, which fits generalized linear models, including logistic regression (see [35] for details) in R v4.1.3 [36]. The demographic information we use includes the gender and age of the patients. In this analysis, age is treated as a categorical variable with three categories: 18–40, 41–64, and 65+. In some of the analyses that follow, values of NA are reported when there were no observations within that age group for the specific service class. However, for consistency between analyses, age groupings are kept the same for all service classes. Specifically, the data pre-processing is carried out following the steps below:

1. Obtain the list of patient transitions from the patient analysis tool from each of the cohorts that need to be compared;

Encode all the services in the class of interest with a 1 and all other services with a 0. For example, if a patient has a service history of "0 0 0 1 1 0 0", the patient accessed the service of interest at their fourth and fifth encounters only;

To account for differing underlying morbidity rates between the two cohorts, we consider transitions after and including the first transition to a service of interest. These biases have the potential to affect the trajectory

that patients follow when engaging with the health services. The above would then become "0 1 1 0 0". That is, we remove all except one leading 0;

We count the number of 2-g. The above string would yield 01, 11, 10, 00;

The number of 2-g from all individuals in each cohort is summed; grouping by demographic information such as age and gender is possible if the effect of these factors is of interest.

We then fit a logistic regression model for transition probabilities. That is, we aim to predict the probability of transition from each state Y from the previous state Y_{prev}. The covariates for prediction are the previous state Y_{prev}, cohort membership, and optional demographic information.

Without demographic information, the transition probabilities can be modeled as the probability of a 0-to-1 transition, denoted as the following:

$$P(0 \to 1) = \text{expit}\big(b_0 + b_1 I(c)\big)$$

and the probability of a 1-to-1 transition, denoted as the following:

$$P(1 \to 1) = \text{expit}\big(b_2 + b_3 I(c)\big)$$

where $I(c)$ is 1 if the patient is in the cohort c of interest and 0 otherwise; the expit function is as follows:

$$\text{expit}(x) = \frac{1}{1 + e^{-x}}$$

An alternative formulation requiring only one equation is as follows:

$$P(Y) = \text{expit}\Big(c_0 + \big(c_1 Y_{prev}\big) + \big(c_2 I(c)\big) + \big(c_3 I(c) Y_{prev}\big)\Big)$$

To show that the above formulations are equivalent, we can use the following definitions to go between the two formulations. We set the following:

$$b_0 = c_0\,,\ b_1 = c_2\,,\ b_2 = c_0 + c_1\,,\ and\ b_3 = c_2 + c_3$$

With demographic information, we simply write the following:

$$P(Y) = \text{expit}\Big(c_0 + \big(c_1 Y_{prev}\big) + \big(c_2 I(c)\big) + \big(c_3 I(c) Y_{prev}\big) + \big(c_4(\text{gender} = \text{Male})\big) + c_5(\text{age})\Big)$$

Additional demographic information can be added similarly. Model selection for choosing which demographic information to include is performed with backwards stepwise selection using AIC, using the step() function in R. For the purposes of this paper, since multiple comparisons are made for a variety of service classes, two models are used for the sake of consistency: (1) the model where predictors are Y_{prev}, cohort membership, gender, and age and (2) the model with neither gender or age, using only Y_{prev} and cohort membership as predictors. In practice, model selection should be done and additional demographic information included, if relevant, to ensure the best model fit.

In the case where we only consider the transitions and the cohort membership (and not any demographic information), if we assume an interaction term between cohort and transition, then the probability will be equal to the empirical probability, which is just the ratio of the transition event counts, as there are four transition probabilities ($P(0 \to 1)$ and $P(1 \to$

1) for the two cohorts) to estimate and four parameters for estimation ($c0, c1, c2,$ and $c3$). In this case, if the estimated probabilities are the only values of interest (rather than, say, the coefficients themselves), it can be simpler to calculate the empirical probabilities rather than fitting the model. So, for example, the probability of a 0→1 transition would be as given:

$$P(0 \to 1) = \frac{c\{01\}}{c\{00\} + c\{01\}}$$

where $c(ab)$ is the number of transitions from state a to state b. This is the empirical probability of a transition if currently we are in state 0 to be in state 1 after the transition.

3. Analysis and Results

As previously stated, two cohorts were considered, comprising 21,180 patients from the less vulnerable cohort and 1829 patients from the more vulnerable cohort. Additionally, the 182 patients belonging to both cohorts were assigned to the more vulnerable cohort. It should be noted that not all patients accessed every service, and as such, not all patients are included in the analyses below.

First, a collection of cardiovascular services was taken as the focus and includes age and gender as features in addition to cohort and previous state. It was found that age and gender are not significant, as shown by the result below in Table 1:

Table 1. Cardiovascular.

Title 1	Estimate	Std. Error	z Value	P(|z|)
Intercept	−3.2298	0.2438	−13.25	$<210^{-16}$
Cohort 2	0.8568	0.2404	3.56	0.00037
Y_{prev}	1.5427	0.0835	18.47	$<210^{-16}$
Age 18–40	0.2092	0.1189	1.76	0.07850
Age 41–64	0.2018	0.0801	2.52	0.01181
Age 65+	NA	NA	NA	NA
Gender Male	0.0186	0.0754	0.25	0.80561

For this reason, age and gender were excluded in the analysis of the result shown in Table 2. Taking only previous state and cohort membership, we obtain the following:

Table 2. Cardiovascular transition probabilities, including previous state and cohort only.

	Cohort 1	Cohort 1	Cohort 2	Cohort 2
$Y_{prev} \to$ Y	P	Std.	P	Std.
0→ 0	0.9527	0.0097	0.9047	0.0037
0→ 1	0.0427	0.0097	0.0953	0.0037
1→ 0	0.8257	0.0356	0.6678	0.0159
1→ 1	0.1743	0.0356	0.3322	0.0159

From Table 2, the probabilities of both 0-to-1 and 1-to-1 transitions are higher in cohort 2 compared to cohort 1, which corresponds with a positive fitted effect for cohort 2 membership. Next, the focus of service classes was shifted onto residential care (Table 3), assertive community treatment -ACT (Table 4), psychiatric emergency services -PES (Table 5), and MHSU acute care (Table 6). For these service classes, there is a negative fitted effect corresponding with cohort 2 membership, implying that transitions to the service class of interest are more likely to occur for patients in cohort 1.

Table 3. Residential Care.

	Estimate	Std. Error	z Value	P(|z|)
Intercept	−2.7424	0.1352	−20.28	$<2 10^{-16}$
Cohort 2	−0.8268	0.0753	−10.98	$<2 10^{-16}$
Y_{prev}	−0.0520	0.1557	−0.33	0.73848
Age 18–40	0.4032	0.1216	3.32	0.00091
Age 41–64	0.1874	0.1292	1.45	0.14710
Age 65+	NA	NA	NA	NA
Gender Male	0.1032	0.0760	1.36	0.17432

Table 4. ACT.

Title 1	Estimate	Std. Error	z Value	P(|z|)
Intercept	−3.512	0.252	−13.92	$<2 10^{-16}$
Cohort 2	−0.395	0.145	−2.72	0.0065
Y_{prev}	0.750	0.237	3.17	0.0015
Age 18–40	0.655	0.270	2.43	0.0151
Age 41–64	0.182	0.302	0.60	0.5461
Age 65+	NA	NA	NA	NA
Gender Male	0.242	0.148	1.64	0.1019

Table 5. PES.

Title 1	Estimate	Std. Error	z Value	P(|z|)
Intercept	−1.2181	0.0733	−16.62	$<2 10^{-16}$
Cohort 2	−0.3288	0.0174	−18.93	$<2 10^{-16}$
Y_{prev}	0.2831	0.0179	15.78	$<2 10^{-16}$
Age 18–40	−0.2842	0.0733	−3.88	0.00011
Age 41–64	−0.4852	0.0739	−6.56	$5.310\ 10^{-11}$
Age 65+	−0.8272	0.0766	−10.80	$<2 10^{-16}$
Gender Male	0.0943	0.0140	6.75	$1.5\ 10^{-11}$

Table 6. MHSU Acute Care.

Title 1	Estimate	Std. Error	z Value	P(|z|)
Intercept	−1.7653	0.1325	−13.32	$<2 10^{-16}$
Cohort 2	−0.4491	0.0237	−18.92	$<2 10^{-16}$
Y_{prev}	0.7004	0.0290	24.13	$<2 10^{-16}$
Age 18–40	−0.2152	0.1327	−1.62	0.1048
Age 41–64	−0.3122	0.1334	−2.34	0.0193

Age 65+	−0.4083	0.1362	−3.00	0.0027
Gender Male	0.1682	0.0218	7.71	$1.2\ 10^{-14}$

In our next analyses, we compare the transition probabilities for each cohort considered in our study. To have a fair comparison among the focus sets, we removed age and gender to be able to list the transition probabilities for each cohort. Table 7 shows that for residential care, ACT, PES, and MHSU acute care, the probabilities of both 0-to-1 and 1-to-1 transitions are lower in cohort 2 compared to cohort 1.

Table 7. Transition probabilities.

	Cohort 1	**Cohort 1**	**Cohort 2**	**Cohort 2**
Focus	$P(0 \rightarrow 1)$	$P(1 \rightarrow 1)$	$P\ (0 \rightarrow 1)$	$P\ (1 \rightarrow 1)$
Cardiovascular	0.0427	0.1743	0.0953	0.3322
Residential Care	0.08721	0.08445	0.03765	0.03640
ACT	0.05371	0.11348	0.03343	0.07236
PES	0.1795	0.2299	0.1263	0.1647
MHSU Acute Care	0.12870	0.23125	0.08123	0.15258

The results demonstrate that there is a disparity in access to cardiovascular service between cohort 1 and cohort 2. The probability for a patient in cohort 2 of having a follow-up cardiovascular service is thrice as much as for a patient in cohort 1.

Moreover, as previously mentioned, the goal is to provide a more nuanced understanding of healthcare access disparities affecting vulnerable patients. Using a cross-continuum dataset and expanding the analysis to other services, including residential services, ACT, PES, and MHSU acute care, the following was noted:

1. For residential service, the probability of a patient in cohort 1 to access this service is more than double the probability for patients in cohort 2;

For ACT, we notice the same trend: a patient in cohort 1 is more likely to access ACT than a patient in cohort 2;

Access to PES is also higher for patients in cohort 1 than patients in cohort 2;

For MHSU acute care, once again, patients in cohort 1 have a higher probability of accessing this service than patients in cohort 2.

Based on the above, patients in cohort 1 tend to benefit from MSHU wrap-around services more than patients in cohort 2. It should also be noted that within MHSU, there are services where there is virtually no disparity in access between the two cohorts. An example of such a service is case management.

4. Discussion

The purpose of this paper is to supply a demonstration of a more general method to identify and measure cohort-specific disparities in service access. Focusing on disparities in access to one set of services may produce a compelling set of results, but they may not reproduce across service sectors. We cannot necessarily assume that access within one sector can be treated as

a proxy for full-service system utilization. In other words, looking for disparities in one service sector is only half the method. The full method entails an examination of service utilization across the continuum of care.

Analysis of PSUs within and across cohorts paints a clear picture of a health service system composed of functionally distinguishable clusters of services and service pathways that connect them. The channels that are etched, sometimes quite deeply, across the service system landscape will have an impact on who accesses what services within a given timeframe. Patient factors have some say as to where those pathways are located. Also, service system factors, e.g., span of influence, clinical practice guidelines, and administrative leaders, within program areas will partially determine where those channels are located and how deeply they are etched.

Graph transition probabilities were modelled with logistic regression, where covariates explain variation in transition probabilities of individuals within different classes over time. These covariates often account for differences in the prevalence of transition probabilities over time. This approach is often employed when assessing possible influences on the transition probabilities and is of great importance. Previously, ref. [37] incorporated categorical covariates to model transition probabilities. Also, transition probabilities LR formulated by [37] took into account continuous or time-varying covariates in the analysis of their transitions. They further outlined regression models in which covariates are used to predict individuals' class membership and their next transitions.

This paper proposes a flexible methodology using longitudinal cross-continuum healthcare data where data are treated as a graph, and then, transition probabilities are modeled using logistic regression. The proposed method allows for patients' encounter date to be used and provides a simple method that is easily adaptable to use with other data sources or other cohorts or to find different disparities of interest.

Based on the results of the analysis, we can see that there are differences in access between the group of vulnerable patients and the group of less vulnerable patients. However, these disparities in access do not show up consistently across service areas. For those services where vulnerable patients have less access, it is not clear whether this reflects limited capacity for this group to navigate access the services, is due to processes within the health services system, or is a combination of the two. On one hand, there is a vast amount of literature concerned with one of the categories of vulnerable patients that is consistent with the factor of limited capacity to navigate access to services or sustaining relationships with the service system. On the other hand, the MSHU wrap-around service that this category of vulnerable patients benefits from is a clear demonstration that it is possible for a service system to be structured in a way that addresses the limitations of a vulnerable population. This could be a model that could be replicated outside MHSU as well.

The results have demonstrated that there are group differences regarding access to some of the services. However, the results do not explain why these differences are occurring in the first place. Is this a reflection of limited capacity to navigate access to services on the part of persons within the vulnerable cohort? Alternatively, does the difference emanate in some fashion from persons or processes within the health service system? Or is the

difference a reflection of both factors? Moreover, one must consider the possibility that the cohorts may differ regarding factors other than those that one thinks may be "causing" the inequity. These could be distal factors not captured in the dataset. This is a limitation of this study. To address this, further analyses need to be conducted taking into consideration other potentially contributing factors and involving various clinical and social SMEs and additional datasets beyond patients encounter data, which may not be readily available within a health service clinical information system.

The work laid out in this paper is methodological, demonstrating a method to detect potential inequities in access to services. For privacy issues, the exact diagnosis or the nature of the vulnerability was not disclosed. However, the paper demonstrates that a more nuanced approach to assessing disparity in access to care is doable using PSUs from a longitudinal cross-continuum healthcare dataset. Any organization that wishes to structure their services to provide better care to its vulnerable population can use the proposed approach as part of their toolkit in assessing the services that need to be restructured or integrated. The information generated from the proposed methodology can be valuable knowledge in the hands of quality assurance/quality improvement in guiding organizations to implement a more equitable healthcare system.

5. Conclusions

When patient factors and service system factors come together, inequities may arise, even when both patients and the service system are doing their best. That is, these are emergent characteristics that may run counter to the intent of all parties involved. Our paper outlines a general methodology that provides a nuanced approach to assess and measure access disparity using PSUs based on the local reality across the full care continuum. This provides a first step in addressing inequities for a healthcare organization. The methodology introduced in this paper is also easily extensible to other cohorts or services of interest and with data from a variety of data sources.

Additional steps need to be considered to fully address these inequities. These include (1) the understanding of additional factors, including patients' distal determinants of health and social determinants of health that may affect timely access to appropriate healthcare services and (2) the undertaking of specific, localized, and measurable actions developed and sustained through ongoing engagement with the communities, supportive leadership, dedicated resources, accountability, and transparency.

Author Contributions: Conceptualization, J.B., Y.S., G.D., A.R., and A.K.; methodology, J.B., Y.S., and G.D.; software, Y.S. and G.D.; validation, J.B., K.M., and E.C.; formal analysis, J.B., Y.S., G.D., and K.M.; investigation, J.B., Y.S., and G.D.; resources, J.B. and K.M.; data curation, Y.S. and G.D.; writing—original draft preparation, J.B., Y.S., and G.D.; writing—review and editing, J.B., Y.S., G.D., S.D., K.O., K.M., A.R., E.C., and A.K.; visualization, J.B., Y.S., G.D., and K.M.; supervision, J.B., K.M., A.R., and A.K.; project administration, J.B.;. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: A certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following the British Columbia, Canada, ethics harmonization guideline. The REB number is H21-02817.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are unavailable because of privacy or ethical restrictions. Requests to access the datasets require a certificate of approval by the University of Victoria Research Ethics Board, following the British Columbia, Canada, ethics harmonization guideline.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Whitehead, M.; Dahlgren, G.R.; Organization, W.H. *Levelling Up (Part 1): A Discussion Paper on Concepts and Principles for Tackling Social Inequities in Health*; WHO Regional Office for Europe: Copenhagen, Denmark, 2006.
2. Browne, A.J.; Varcoe, C.M.; Wong, S.T.; Smye, V.L.; Lavoie, J.; Littlejohn, D.; Tu, D.; Godwin, O.; Krause, M.; Khan, K.B.; et al. Closing the health equity gap: Evidence-based strategies for primary health care organizations. *Int. J. Equity Health* **2012**, *11*, 59.
3. Gilliland, J.A.; Shah, T.I.; Clark, A.; Sibbald, S.; Seabrook, J.A. A geospatial approach to understanding inequalities in accessibility to primary care among vulnerable populations. *PLoS ONE* **2019**, *14*, e0210113.
4. Laursen, T.M.; Munk-Olsen, T.; Vestergaard, M. Life expectancy and cardiovascular mortality in persons with schizophrenia. *Curr. Opin. Psychiatry* **2012**, *25*, 83–88. https://doi.org/10.1097/yco.0b013e32835035ca.
5. Waisel, D.B. Vulnerable populations in healthcare. *Curr. Opin. Anaesthesiol.* **2013**, *26*, 186–192.
6. Riley, W.J. Health disparities: Gaps in access, quality and affordability of medical care. *Trans. Am. Clin. Climatol. Assoc.* **2012**, *123*, 167.
7. Auquier, P.; Lançon, C.; Rouillon, F.; Lader, M. Mortality in schizophrenia. *Pharmacoepidemiol. Drug Saf.* **2007**, *16*, 1308–1312. https://doi.org/10.1002/pds.1496.
8. Laursen, T.M.; Nordentoft, M.; Mortensen, P.B. Excess Early Mortality in Schizophrenia. *Annu. Rev. Clin. Psychol.* **2014**, *10*, 425–448,. https://doi.org/10.1146/annurev-clinpsy-032813-153657.
9. Narendorf, S.C. Intersection of homelessness and mental health: A mixed methods study of young adults who accessed psychiatric emergency services. *Child. Youth Serv. Rev.* **2017**, *81*, 54–62. https://doi.org/10.1016/j.childyouth.2017.07.024.
10. Baker, J.; Travers, J.L.; Buschman, P.; Merrill, J.A. An Efficient Nurse Practitioner-Led Community-Based Service Model for Delivering Coordinated Care to Persons With Serious Mental Illness at Risk for Homelessness. *J. Am. Psychiatr. Nurses Assoc.* **2018**, *24*, 101.
11. Robine, J.M.; Ritchie, K. Healthy life expectancy: Evaluation of global indicator of change in population health. *BMJ* **1991**, *302*, 457–460.
12. Amjad, H.; Carmichael, D.; Austin, A.M.; Chang, C.-H.; Bynum, J.P.W. Continuity of Care and Health Care Utilization in Older Adults With Dementia in Fee-for-Service Medicare. *JAMA Intern. Med.* **2016**, *176*, 1371–1378,. https://doi.org/10.1001/jamainternmed.2016.3553.
13. Narendorf, S.C.; Cross, M.B.; Santa Maria, D.; Swank, P.R.; Bordnick, P.S. Relations between mental health diagnoses, mental health treatment, and substance use in homeless youth. *Drug Alcohol Depend.* **2017**, *175*, 1–8. https://doi.org/10.1016/j.drugalcdep.2017.01.028.
14. Courtney-Long, E.A.; Carroll, D.D.; Zhang, Q.C.; Stevens, A.C.; Griffin-Blake, S.; Armour, B.S.; Campbell, V.A. Prevalence of Disability and Disability Type Among Adults—United States, 2013. *MMWR. Morb. Mortal. Wkly. Rep.* **2015**, *64*, 777–782. https://doi.org/10.15585/mmwr.MM6429a2.
15. Jefferis, B.J.; Merom, D.; Sartini, C.; Wannamethee, S.G.; Ash, S.; Lennon, L.T.; Iliffe, S.; Kendrick, D.; Whincup, P.H. Physical Activity and Falls in Older Men. *Med. Sci. Sports Exerc.* **2015**, *47*, 2119–2128. https://doi.org/10.1249/MSS.0000000000000635.
16. Iezzoni, L.I.; McCarthy, E.P.; Davis, R.B.; Siebens, H. Mobility impairments and use of screening and preventive services. *Am. J. Public Health* **2000**, *90*, 955–961. https://doi.org/10.2105/AJPH.90.6.955.
17. Zhao, G.; Ford, E.S.; Li, C.; Crews, J.E.; Mokdad, A.H. Disability and its correlates with chronic morbidities among U.S. adults aged 50–<65 years. *Prev. Med.* **2008**, *48*, 117–121. https://doi.org/10.1016/j.ypmed.2008.11.002.
18. Lagu, T.; Hannon, N.S.; Rothberg, M.B.; Wells, A.S.; Green, K.L.; Windom, M.O.; Dempsey, K.R.; Pekow, P.S.; Avrunin, J.S.; Chen, A.; et al. Access to subspecialty care for patients with mobility impairment. *Ann. Intern. Med.* **2013**, *158*, 441–446. https://doi.org/10.7326/0003-4819-158-6-201303190-00003.

19. Christiani, A.; Hudson, A.L.; Nyamathi, A.; Mutere, M.; Sweat, J. Attitudes of Homeless and Drug-Using Youth Regarding Barriers and Facilitators in Delivery of Quality and Culturally Sensitive Health Care. *J. Child Adolesc. Psychiatr. Nurs.* **2008**, *21*, 154–163. https://doi.org/10.1111/j.1744-6171.2008.00139.x.
20. Cradock-O'Leary, J.; Young, A.S.; Yano, E.M.; Wang, M.; Lee, M.L. Use of General Medical Services by VA Patients With Psychiatric Disorders. *Psychiatr. Serv.* **2002**, *53*, 874–878. https://doi.org/10.1176/appi.ps.53.7.874.
21. Stangl, A.L.; Earnshaw, V.A.; Logie, C.H.; Van Brakel, W.; Simbayi, L.C.; Barré, I.; Dovidio, J.F. The Health Stigma and Discrimination Framework: A global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Med.* **2019**, *17*, 31.
22. Joint United Nations Programme on HIV/Aids. Protocol for the identification of discrimination against people living with HIV. In *Protocol for the Identification of Discrimination against People Living with HIV*; 2000; pp. 40.
23. Nyblade, L.; Stockton, M.A.; Giger, K.; Bond, V.; Ekstrand, M.L.; Mc Lean, R.; Mitchell, E.M.H.; Nelson, L.R.E.; Sapag, J.C.; Siraprapasiri, T.; et al. Stigma in health facilities: Why it matters and how we can change it. *BMC Med.* **2019**, *17*, 25.
24. Ross, C.A.; Goldner, E.M. Stigma, negative attitudes and discrimination towards mental illness within the nursing profession: A review of the literature. *J. Psychiatr. Ment. Health Nurs.* **2009**, *16*, 558–567.
25. Katz, I.T.; Ryu, A.E.; Onuegbu, A.G.; Psaros, C.; Weiser, S.D.; Bangsberg, D.R.; Tsai, A.C. Impact of HIV-related stigma on treatment adherence: Systematic review and meta-synthesis. *J. Int. AIDS Soc.* **2013**, *16*, 18640.
26. Teixeira, E.M.; Budd, G.M. Obesity stigma: A newly recognized barrier to comprehensive and effective type 2 diabetes management. *J. Am. Assoc. Nurse Pract.* **2010**, *22*, 527–533.
27. Rueda, S.; Mitra, S.; Chen, S.; Gogolishvili, D.; Globerman, J.; Chambers, L.; Wilson, M.; Logie, C.H.; Shi, Q.; Morassaei, S.; et al. Examining the associations between HIV-related stigma and health outcomes in people living with HIV/AIDS: A series of meta-analyses. *BMJ Open* **2016**, *6*, e011453.
28. Bambi, J.; Santoso, Y.; Sadri, H.; Moselle, K.; Rudnick, A.; Robertson, S.; Chang, E.; Kuo, A.; Howie, J.; Dong, G.Y.; et al. A Methodological Approach to Extracting Patterns of Service Utilization from a Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems. *BioMedInformatics* **2024**, *4*, 946–965.
29. Bambi, J.; Santoso, Y.; Moselle, K.; Robertson, S.; Rudnick, A.; Chang, E.; Kuo, A. Analyzing Patterns of Service Utilization Using Graph Topology to Understand the Dynamic of the Engagement of Patients with Complex Problems with Health Services. *BioMedInformatics* **2024**, *4*, 1071–1084.
30. Bambi, J.; Sadri, H.; Moselle, K.; Chang E.; Santoso Y.; Howie, J.; Rudnick A.; Elliott L.; Kuo A. Approaches to Generating Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs Natural Language Processing Clustering. 2024.
31. Moselle, K.; Koval, A. Mapping Clinical Contents onto Longitudinal Depictions of Cross-Continuum Service Events in Island Health. *Int. J. Popul. Data Sci.* **2018**, *3*.
32. Sinha, N.C.; Islam, M.A.; Ahamed, K.S. Logistic regression models for higher order transition probabilities of Markov chain for analyzing the occurrences of daily rainfall data. *J. Mod. Appl. Stat. Methods* **2011**, *10*, 337–348.
33. Kozłowski, E.; Borucka, A.; Świderski, A. Application of the logistic regression for determining transition probability matrix of operating states in the transport systems. *Eksploat. i Niezawodn.-Maint. Reliab.* **2020**, *22*, 192–200.
34. Rößiger, G.; Kulla, L.; Bošeľa, M. Changes in growth caused by climate change and other limiting factors in time affect the optimal equilibrium of close-to-nature forest management. *For. J.* **2019**, *65*, 180–190.
35. Glm: Fitting Generalized Linear Models. Available online: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm (accessed on 17 February 2024).
36. R. C. Team. R: A language and environment for statistical computin. In *R Foundation for Statistical Computing*; 2013.
37. Pfeffermann, D.; Skinner, C.; Humphreys, K. The estimation of gross flows in the presence of measurement error using auxiliary variables. *J. R. Stat. Soc. Ser. A (Statistics Soc.)* **1998**, *161*, 13–32.

Chapter 6: 5th Manuscript

knowledge

MDPI

Article

Use of Patterns of Service Utilization and Hierarchical Survival Analysis in Planning and Providing Care for Overdose Patients and Predicting the Time-to-Second Overdose

Jonas Bambi [1,*], Kehinde Olobatuyi [2], Yudi Santoso [3], Hanieh Sadri [4], Ken Moselle [5], Abraham Rudnick [6], Gracia Yunruo Dong [7,8], Ernie Chang [9] and Alex Kuo [1]

Citation: Bambi, J.; Olobatuyi, K.; Santoso, Y.; Sadri, H.; Moselle, K.; Rudnick, A.; Dong, G.; Chang, E.; Kuo, A. Use of Patterns of Service Utilization and Hierarchical Survival Analysis in Planning and Providing Care for Overdose Patients and Predicting the Time-to-Second Overdose. *Knowledge* **2024**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Received: 29 February 2024
Revised: 5 August 2024
Accepted: 13 August 2024
Published: date

[1] Department of Health Information Science, Faculties of Human and Social Development, Victoria Campus, University of Victoria, Victoria, BC V8P 5C2, Canada
[2] Departments of Mathematics and Statistics, Faculty of Science, Victoria Campus, University of Victoria, Victoria, BC V8P 5C2, Canada; olobatuyikenny@uvic.ca
[3] Independent Researcher, Victoria, BC V8R 5B4, Canada; y.santoso8@gmail.com
[4] Department of Computer Science, Faculty of Engineering and Computer Science, Victoria Campus, University of Victoria, Victoria, BC V8P 5C2, Canada; haniehsadri@uvic.ca
[5] Department of Clinical Psychology, Faculty of Social Science, Victoria Campus, University of Victoria, Victoria, BC V8P 5C2, Canada; kmoselle@uvic.ca
[6] Departments of Psychiatry and Bioethics, School of Occupational Therapy, Faculties of Medicine and Health, Dalhousie University, Halifax, NS B3H 4R2, Canada; abraham.rudnick@nshealth.ca
[7] Department of Statistical Sciences, Faculties of Arts and Science, St. George Campus, University of Toronto, Toronto, ON M5S 1A1, Canada
[8] Departments of Mathematics and Statistics, Faculty of Science, Victoria Campus, University of Victoria, Victoria, BC V8P 5C2, Canada； gracia.dong@utoronto.ca
[9] Retired Physician and Independent Computer Scientist, Victoria, BC V9C 4B1, Canada; ecsendmail@gmail.com
* Correspondence: jonasbambi@uvic.ca; Tel.: +1-250-507-4262

Abstract: Individuals from a variety of backgrounds are affected by the opioid crisis. To provide optimal care for individuals at risk of opioid overdose and prevent subsequent overdoses, a more targeted response that goes beyond the traditional taxonomical diagnosis approach to care management needs to be adopted. In previous works, Graph Machine Learning and Natural Language Processing methods were used to model the products for planning and evaluating the treatment of patients with complex issues. This study proposes a methodology of partitioning patients in the opioid overdose cohort into various communities based on their patterns of service utilization (PSUs) across the continuum of care using graph community detection and applying survival analysis to predict time-to-second overdose for each of the communities. The results demonstrated that the overdose cohort is not homogeneous with respect to the determinants of risk. Moreover, the risk for subsequent overdose was quantified: there is a 51% higher chance of

experiencing a second overdose for a high-risk community compared to a low-risk community. The proposed method can inform a more efficient treatment heterogeneity approach for a cohort made of diverse individuals, such as the opioid overdose cohort. It can also guide targeted support for patients at risk of subsequent overdoses.

Keywords: opioid overdose; opioid crisis; clinical pathways; decision support; graph community detection; survival analysis; health information management; health service system; machine learning algorithms; clustering algorithms

1. Introduction

1.1. Use of Patterns of Service Utilization in Planning and Providing Care to Complex Patients

Regarding factors that govern the achievement of outcomes for patients with complex problems, solely relying on traditional taxonomic diagnostic approaches to care management can be limiting [1]. More challenges arise when the cohort of patients sharing the same diagnosis is not homogeneous with respect to other factors affecting their health. Multiple emerging conditions and distinctive distal and proximal determinants of health profiles can affect how patients respond to the treatment [2].

The overdose crisis has had a devastating impact [3–5]. Individuals affected come from diverse socioeconomic statuses, education levels, and cultural backgrounds [6]. Studies have shown that individuals suffering from substance use disorders, mental health disorders, or homelessness are at a higher risk of overdose [7–12]. However, there are cases where individuals who are not in a high-risk group are impacted by opioid overdoses [13–15]. Hence, determining whether the programs designed to combat this crisis are effective is challenging, as they clearly benefit some patients but not others [16], as the overdose cohort is not homogeneous with respect to determinants of risk [17].

The ultimate objectives are to (1) provide the best care for persons at risk for overdoses, (2) reduce the rate of overdoses, and (3) prevent subsequent overdoses. There is a need to look beyond the traditional taxonomic diagnostic approach to care management. One of the venues that need to be explored is the dynamics of engagement of individuals constituting the opioid overdose cohort with the health service system to understand their Patterns of Service Utilization (PSUs) across the continuum of care. PSUs are fundamentally descriptive and consist of channels that are etched progressively across the service system by groups of individuals who share some common set of needs. This may open an opportunity to effectively respond to the opioid crisis by providing a more targeted response to the individuals affected.

1.2. Evidence-Based Care: Machine Learning and Statistical Analysis

Significant effort has been devoted to supporting evidence-based care through both statistical and machine learning (ML) approaches to analyze healthcare data, generate insights, and inform care delivery. ML focuses on

iterative construction and validation of models using algorithms to find patterns in high-dimensional data [18]. In contrast, statistical models focus on inference based on properties of the datasets as a whole (e.g., measures of central tendency for parametric methods; marginal distributions for some non-parametric tests) [18].

Survival analysis has been a standard method in statistics used to assess risk or survival probability over time [19]. For example, it can be used in cancer studies to compare time from complete remission to relapse among several treatments. Other examples include the use of survival analysis to model medical prognosis [20], model factors associated with length of stay for patients [21], and examine the influence of living arrangements and healthcare utilization on patients' mortality [22].

1.3. Objectives

Individuals that have taken an overdose and/or are at risk of a second overdose present a different constellation of risk factors that bias the odds of overdosing. These different constellations of risk factors may reflect the fact that this group of individuals is not homogeneous. Understanding these differences is a first step in providing better care for opioid overdose patients and/or preventing subsequent overdoses.

In our previous works [23–26], the consideration of PSUs in planning and evaluating the treatment of patients with complex issues was proposed. Using Clinical Information System (CIS) encounters data, PSUs represent pathways etched into the service system terrain by the journeys of a patient or cohort of patients as they interact with a cross-continuum health service system. Various ML algorithms, including graph community detection and Natural Language Processing (NLP), are used to (1) group related health services based on PSUs, (2) compare/contrast the effectiveness/existence of a service model in caring for various cohorts of patients, and (3) evaluate access disparity for vulnerable patients.

To achieve this, different approaches were considered, including the following: (1) The use of an iterative graph community detection, combined with input from clinical subject matter experts (SMEs) to identify patterns in patient–service encounter data that are difficult to detect via classic statistical methods, resulting in a grouping of related health services based on PSUs [23]. In this case, services were connected when used by the same patients. The generated communities of services provide a possibility of influencing the reorganization of services within the health service structure to provide better care for vulnerable patients with mental and other complex healthcare challenges. (2) To show the similarity of results across different approaches for cross-validation and to demonstrate that the grouping of related services demonstrated in [23] was not an artifact of the method employed, NLP clustering was used – where each patient's history of service utilization was generated as a sentence [24] . Following this, term frequency-inverse document frequency and cosine similarity were used to measure similarity between services, and a series of clustering algorithms were used to group similar services. The results in [23,24] were determined from a clinical perspective by clinical SMEs and service system operations experts to be similar. (3) The work in [23,24] modeled products of service system dynamics as temporal entities. In [25], the order of events was added to the data model

to provide topological depictions of the dynamics that are embodied in patients' movement across a complex healthcare system. Using a directed graph and applying various topological visualizations of the graph [25], we identified the way diverse components of the healthcare service system are functionally connected or disconnected by patient journeys. This methodology provided a preliminary step in addressing the challenge of locating potential operational problems for patients with complex problems engaging with a complex healthcare service system. (4) Expanding on [25] and using directed graph and logistic regression, a methodology to identify and quantify cohort-specific disparities in accessing healthcare services across the continuum of care was proposed in [26] . The result in [26] demonstrated that a more nuanced approach to assessing access-to-care disparity is feasible using PSUs from a longitudinal cross-continuum healthcare dataset.

As previously stated, survival analysis is a standard method in statistics that has been used to assess risk or survival probability over time, related to several clinical settings [19–22]. Additionally, regarding opioid overdose risk assessment, many studies have been conducted in areas such as: (1) the understanding of risk factors for a population of patients receiving opioids for pain [27], (2) the assessment of opioid overdose risk using patient data level [28], (3) the assessment of risk and protective factors for repeated overdose [29], and (4) the intersectionality of characteristics such as demographics, socioeconomics, and service use among individuals who experienced opioid overdose [17]. What is missing in the literature is a methodological approach for performing a comparison of risks of survival probability for a cohort of opioid overdose patients, based on their pattern of service engagement with the healthcare system across the full continuum of care, well beyond the emergency department and hospital admission.

Using graph community detection and survival analysis and relying on patients' encounters data collected from a host organization, CIS, the work in this paper will expand on the use of PSUs, as outlined in [23–26], to answer the following questions:

1. Using PSUs, to what extent can we determine whether the opioid overdose cohort is homogeneous or not with respect to determinants of risk?
2. How many communities constitute the opioid overdose cohort, based on how patients within this cohort interact with the host organization's cross-continuum service system?
3. To what extent can we determine the risk of a subsequent opioid overdose based on the community an opioid overdose patient belongs to and quantify it using survival analysis?

Answering the above questions will provide an opportunity for the health service system to effectively respond to the opioid crisis by providing a more targeted response to the individuals affected.

2. Methods

2.1. Addressing Data Granularity Issues

We use data supplied by a health organization that provides a comprehensive array of secondary and tertiary health services [30]. These

services include acute care/intensive care services, hospital and community-based emergency response, ambulatory services, residential care services for older adults or persons contending with mental health issues, case management services, and a range of addictions harm reduction or rehab and recovery-oriented services. A certificate of approval was provided by the Research Ethics Board to conduct this research project.

One or more services provided are encapsulated into an array of roughly two thousand Service Units within the location of the host organization's clinical information system used to support the delivery of care. To address the data granularity issues, following our previous works [23–26], a semantic layer, Clinical Context Coding Scheme [31], consisting of a scheme organized around six sets of codes, was applied to all the two thousand Service Units. The approximately two hundred Service Classes employed for the modeling in this paper consist of equivalence classes formed by the application of these code sets to the Service Units.

2.2. Community Detection

Healthcare encounter data can be viewed as a bipartite graph between patients and health services. This bipartite graph can then be projected either onto patients or services. These bipartite projections are illustrated in Figure 1. In this example, we have four patients (A, B, C, D) and three services (x, y, z). Upon projection onto patients, two patients are connected, or, in other words, there will be an edge between them in the projected graph if they use some common services. Furthermore, the weight of that edge is determined by the number of services that the two patients have in common. For example, Patient A and Patient C are connected by an edge because they both use Service y. In this case, there is only one common service, hence the weight of the edge AC is one. Between A and D, there are two common services (x and y), hence the weight of edge AD is two. As an alternative, we can also consider the number of times each patient used each service. For example, if Patient A used Service x five times and Patient D used Service x three times, then Service x contributes three units to the weight of the edge between A and D. On the other hand, for projection onto services, two services are connected if they have some common patients. In this paper, we group the patients based on their patterns of service utilization. Thus, we project the graph onto patients.

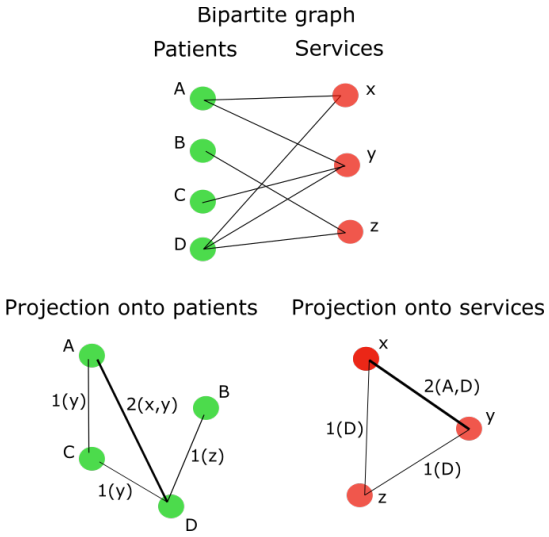


Figure 1. Bipartite projections.

Having a weighted, undirected graph with patients as the nodes, we can then apply the Louvain community detection algorithm [32] to group the patients into communities. Roughly speaking, each community contains patients who have commonalities in their usage of services. Therefore, we can label each community based on the most dominant services used in the community. This becomes the characteristic of the patients in each community. We will show below that, in the case of the opioid overdose cohort, the differences in these characteristics are correlated to different risk levels for repeat overdose. The Louvain algorithm works by maximizing the modularity value, defined as follows:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where A_{ij} is the weight of the edge between node i and node j; k_i is the sum of the edge weights over all the edges that are connected to the node i; and m is the total edge weights in the graph. Here, c_i is the label of the community in which node i belongs to. The delta function has a value equal to one if its two arguments are equal, i.e., if $c_i = c_j$, otherwise the value is 0.

At the initiation phase, each node has its own community. The algorithm starts by randomly choosing a node and then checks other nodes attached to that node to see if merging the communities would result in a higher Q. If yes, then the communities are merged. It would continue iteratively through all the communities until it could not increase Q anymore, and then the algorithm would stop.

2.3. Survival Models

Survival analysis refers to the methods of modeling data where the outcome is the time until an event of interest occurs. One of the main challenges is the presence of instances whose event outcomes become unobservable after a certain time point, or when some instances do not experience any event during the study period. An important feature of survival analysis called censoring is the event of interest that may not have

occurred for all subjects before the completion of the follow-up study. In this study, our main goal is to predict time-to-second overdose. In this case, patients who did not have a second overdose before the entire study period are rightly censored. In this section, we describe the model for longitudinal data with heterogeneous distribution such that the longitudinal data can be clustered into distinct groups.

Let $\boldsymbol{y_i}$ be the longitudinal response for the subject i monitored over a time $t_i; i = 1, \dots, n;$ and n is the number of subsets. Let T_i^* denote the true event time (the time an individual leaves the study or has the second overdose) and C_i be the censoring time. The true event $T_i = \min(C_i, T_i^*)$ represents the estimated survival time for the i^{th} individual. Also, let δ_i^* denote a censoring indicator $I(T^* \leq C_i)$. Therefore, the observed data for the outcome consist of the pairs $(T_i, \delta_i^*), i = 1,2, \dots, n$. The survival function, which represents the probability that the time to the event of interest is not earlier than a specified time t [33,34] is one of the main goals in survival analysis. The survival function is given as follows:

$$S(t) = P(T^* \geq t).$$

The survival function monotonically decreases with time t, and the initial value is 1 when $t = 0$, which signifies that, at the beginning of the observation, 100% of the observed subjects have not experienced a second overdose; in other words, none of the events of interest have occurred. In contrast, the cumulative distribution function, $F(t)$, which represents the probability that the event of interest occurs earlier than t, is $F(t) = 1 - S(t)$. Additionally, hazard function $h(t)$ refers to instantaneous rate [35]. Like $S(t)$, $h(t)$ is a non-negative function. While all the survival functions $S(t)$ decrease over time, the hazard function can have different shapes. The hazard function represents $h(t) = f(t)/S(t)$ where$f(t) = -\delta S(t)/\delta t$.

Survival analysis is generally performed with statistical or ML methods. Both can make predictions of the expected remaining "lifespan" and estimate the survival probability at the estimated survival time. However, the former focuses more on characterizing the distribution of the event times and the statistical properties of the parameter estimation by estimating the survival curves, while the latter focuses primarily on the prediction of the event occurrence at a given time. Depending on assumptions made, traditional statistical methods can be either non-parametric, semi-parametric, or parametric. ML methods are often more efficient in their ability to learn dependencies, including non-linear relationships, between covariates and survival times. In survival analysis, the main challenge facing ML methods is the difficulty of dealing appropriately with censored data. ML methods are effective when there are many instances in a reasonable dimensional feature space, a feat that proves difficult for survival analysis [36]. In non-parametric methods, an empirical estimate of the survival function can be obtained using the Kaplan–Meier (KM) method [37,38]. In the semi-parametric category, the Cox model is the most used regression analysis approach, built on the proportional hazards assumption and employing partial likelihood for the parameter estimation. Parametric methods are more efficient and accurate when the time of event follows a specific, known distribution. It is easier to estimate the time to event with parametric models, while it is impossible with the Cox model [39].

Also, Kaplan and Meier [37] developed the Kaplan–Meier (KM) curve to estimate the survival function using the actual length of the observed time. This method is the most widely used for estimating survival function. Let $T_1 < T_2 < \cdots < T_k$ be a set of distinct ordered event times observed for $n(k \leq n)$ instances. In addition to this, there are censored times for instances whose event times are not observed.

For a given instance i, represented by the triplet (x_i, T_i, δ_i), the hazard function $h(t, x_i)$ in the Cox model follows the proportional hazards assumption, given by

$$h(t, x_i) = h_0(t)\exp(x_i\beta), \text{ for } i = 1, \ldots, n,$$

where the baseline hazard function $h_0(t)$ can be any arbitrary non-negative function of time; $x_i = (x_{i1}, \ldots, x_{ip})$ is the corresponding covariate vector, for instance i; and $\beta^T = (\beta_1, \ldots, \beta_p)$ is the coefficient vector. Based on the assumption of shared baseline hazard function, the survival function is given as follows:

$$S(t) = exp\big(-H_0(t)exp(x\beta)\big) = S_0(t)^{exp(x\beta)}$$

where $H_0(t)$ is the cumulative baseline hazard function, and $S_0(t) = exp\big(-H_0(t)\big)$ is the baseline survival function. Among the parametric models used for survival analysis, the exponential model is characterized by a single parameter, the constant hazard rate λ. In this case, the failure or death is assumed to be a random event that is independent of time. A large value of lambda indicates a higher risk and a shorter survival time. We have $logS(t) = -\lambda t$, in which the relationship between the logarithm of the survival function and time t is linear, with λ as the slope. The Weibull model, a generalized exponential model, is characterized by two parameters $\lambda > 0$ and $\gamma > 0$. The shape of the hazard function is determined using the shape parameter γ, which provides more flexibility compared to the exponential model. If $\gamma = 1$, the hazard function will decrease over time. The scaling of the hazard function is determined by the scaling parameter λ.

2.4. Combining Community Detection and Survival Analysis

As previously mentioned, ML focuses on iterative construction and validation of models using algorithms to find patterns in often rich and unwieldy data, whereas statistics rely on inference to compute various quantitative measures [18]. For this study, graph community detection was used to group patients into communities based on their patterns of service engagement with the health service system. This was followed using survival analysis to quantify the risk of a second overdose for each of the communities of patients. To achieve this, the following steps were followed:

1. The encounter data were engineered as a bipartite graph consisting of nodes with edges connecting patients to Service Classes. A patient (node) is connected to a Service Class (node) when they use a service represented by the Service Class. Recall that roughly two hundred Service Classes employed for modeling in this paper consist of equivalence classes formed by the application of six code sets to the host organization Service Units to reduce granularity.

2. A bipartite projection onto patients was applied (Figure 1) to the bipartite graph to create a weighted graph, where the number of services that were used by two connected patients became the weight of the edge.
3. The Louvain community detection algorithm was applied to the weighted graph to uncover the communities of patients that reflect high-prevalent PSUs by Service Classes.
4. For each of the generated communities, both the service engagement profile and the diagnosis profile were appended.
5. Collaborating with team members with clinical backgrounds, each community of patients was labeled based on their prevalent service engagement and diagnosis profile.
6. Using community belonging as a characteristic of a patient, survival analysis was used to quantify the risk of a second overdose.
7. Using other patient-related characteristics, such as age, gender, and homelessness status, survival analysis was used to further quantify the risk of a second overdose.

3. Analysis

The data that were analyzed contain records of opioid-overdose-related encounters with the emergency department at a regional health authority, from 30 March 2016 to 29 March 2022. The data contain about nine thousand (8975) encounters, of around six thousand (5381) individuals. Out of these individuals, one patient with inconsistent data was excluded. Thus, the number of eligible overdose patients is 5380. From the eligible patients selected, around a quarter (1582) have more than one overdose (OD) and 3798 have only one overdose within the observation period (Figure 2). Furthermore, we also have demographic data, which contain information such as age group, gender, and homelessness. In addition, we have more complete encounter records, which include encounters with all healthcare services within the health authority for those patients. In our analysis, we consider the date of the first overdose event as the zero/start date for each patient. The second overdose is the event of interest. We form a data frame with one row for each patient, and the attributes include the status and the length in days from the first overdose to the second overdose. A patient has status one if observed as having a second overdose; otherwise, the patient has status zero. A patient is censored (i.e., has status zero but no longer contributes to the 'at risk' group) when no longer being observed—i.e., by the latest date of observation (29 March 2022), had not been observed to have a second overdose, or had died before 29 March 2022 and had not had a second overdose before.

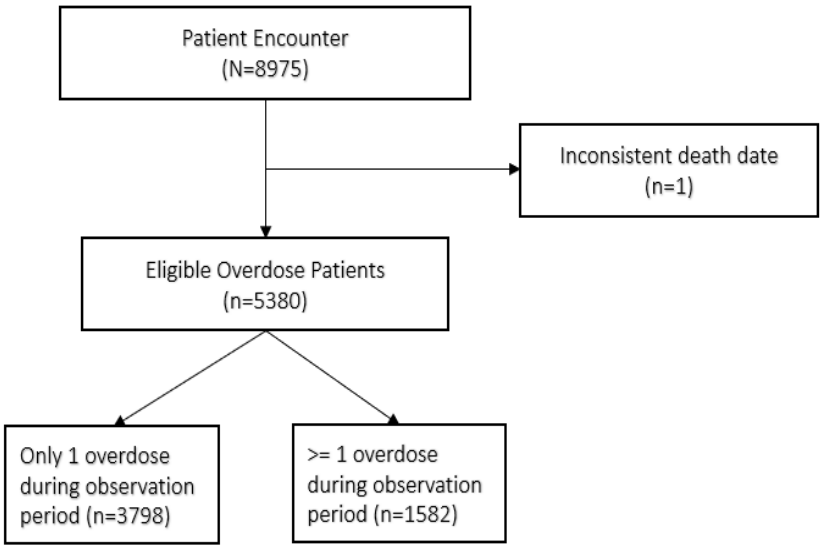


Figure 2. Patient selection process. This is the process used during our data cleaning. We removed any subjects that did not meet the requirement. For example, inconsistent death dates before the start of the observation periods.

3.1. Patient Characteristics

The distribution of the overdose (OD) cohort by grouping attributes is provided in Table 1. From the demographic data, we have age, gender, and homelessness status. We can see that most have an age between twenty and sixty years old. The overall mean age of OD patients was 38 years, the minimum age was 13 years, and the maximum age was 97 years. There are more males (69.65%) than females, and the majority (85.67%) had never been homeless.

Table 1. Distribution of overdose patients by grouping attributes.

Groups	Total OD %	No Second OD	Second OD	Hazard Ratio
	$\%(\boldsymbol{n})$ $(\boldsymbol{N = 5380})$	$\%(\boldsymbol{n})$ $(\boldsymbol{N = 3798})$	$\%(\boldsymbol{n})$ $(\boldsymbol{N = 1582})$	$(\boldsymbol{95\%\ CI})$
Age				
$0-20$	05.84 (314)	05.58 (212)	06.45 (102)	1.00
$20-29$	23.35 (1256)	21.70 (823)	27.37 (433)	0.77 (0.61, 0.97)
$30-39$	27.06 (1456)	25.70 (977)	30.28 (479)	0.77 (0.61, 0.97)
$40-49$	19.33 (1040)	19.50 (742)	18.84 (298)	0.67 (0.52, 0.85)
$50-59$	14.28 (768)	14.90 (565)	12.84 (203)	0.64 (0.50, 0.83)
$60-100$	08.75 (471)	10.60 (404)	04.24 (67)	0.45 (0.32, 0.62)
Gender				
Male	69.65 (3747)	68.14 (2588)	73.30 (1159)	1.34 (1.18, 1.51)
Female	30.29 (1630)	31.78 (1207)	26.70 (423)	1.00
Unknown	00.06 (3)	00.08 (3)	–	–
Community ID				
Community ID 1	20.00 (1076)	15.70 (0595)	30.40 (481)	1.00
Community ID 2	30.72 (1653)	35.00 (1331)	20.40 (322)	0.49 (0.42, 0.58)
Community ID 3	28.23 (1519)	30.70 (1167)	22.20 (352)	0.60 (0.51, 0.70)

Community ID 4	21.04 (1132)	18.60 (750)	27.00 (427)	0.72 (0.62, 0.83)
Ever Homeless				
No (0)	85.67 (4609)	90.18 (3425)	74.80(1184)	1.00
Yes (1)	14.33 (771)	09.82 (373)	25.20 (398)	1.65 (1.45, 1.88)

In addition, we also group the patients by the graph community, which we will discuss more below. There are four communities, and each of the patients belongs to a single community (one to four). Community one has 1076 patients, Community two has 1653 patients, Community three has 1519 patients, and Community four has 1132 patients. The largest is Community two with 30.72%.

3.2. Community Characteristics

In this study, we examined the health service interactions of individuals within a cohort experiencing opioid overdose. Using a bipartite projection on patient data and applying the Louvain algorithm, we generated four distinct communities. For each of the communities, the services across the continuum of care that each community engaged with were reviewed in collaboration with clinical SMEs, and each community was labeled based on the most predominant and distinguishing services. The clinical SMEs that guided the annotation process have extensive experience in health service system operation, healthcare, and computer science. The following are details about the clinical SMEs: (Dr. Ken Moselle (PhD) and Dr. Ernie Chang (MD, PhD) have played a key role in guiding the annotation of the generated communities of patients constituting the overdose cohort, as described in this section. Dr. Ken Moselle is a registered clinical psychologist with extensive experience (25+ years) in health service systems operations. Dr. Ernie Chang is a retired family physician who also holds a PhD in Computer Science. They are both clinical SMEs on the team and co-authors of this manuscript.)

Once each community was labeled, each patient was tied to one of the four communities and assigned a community with a corresponding label. These labels include the following:

- Community one, termed the “reciprocal group”, exhibited a proactive approach to accessing health services, for example, self-referred ambulatory addiction services. They demonstrated higher utilization rates within the service system overall, including Mental Health, and Substance Use (MHSU) and Medical/Surgical (Med/Surg) services. Notably, 80% of patients in this community utilized MHSU clinical intake and addiction clinical intake services. Predominant diagnoses within this group centered around severe addiction issues, with minimal occurrences of schizophrenia-related diagnoses. The average age of patients in this community was 35 years.
- Community two, characterized as the “service-disengaged group”, displayed lower engagement with the service system compared to other communities. They accessed overdose-related and addiction outreach services prior to the overdose events. Diagnostic profiles within this group were not pronounced, with only 8% reporting homelessness and an average age of 36 years.

- Community three, labeled as the "group with complex/serious health problems", exhibited a higher frequency of encounters with Med/Surg services, particularly laboratory and medical imaging procedures. Engagement with MHSU services was comparatively lower, indicating that their engagement with the service system focused on addressing complex medical conditions rather than substance use. Diagnostic data suggested a variety of medical issues, including high rates of palliative care and alterations of awareness. The average age within this community was 46 years, with a considerable number of patients being 60 years or older.
- Community four, characterized as the "group with severe psychiatric issues", demonstrated a high engagement with psychiatric services but low involvement with addiction services. This group exhibited a younger average age of 35 years and a notably high prevalence of schizophrenia diagnoses. Engagement with MHSU services was more prominent than with Med/Surg services.

In Figures 3 and 4, we compared the normalized age distribution (density) of each community. We found that they all have a similar profile, except for Community three, which has a broader and older age distribution. We further showed the density for the age at first overdose of the individuals in the community. We observe that community three has a wider age distribution at first overdose compared to the age distribution of other groups.

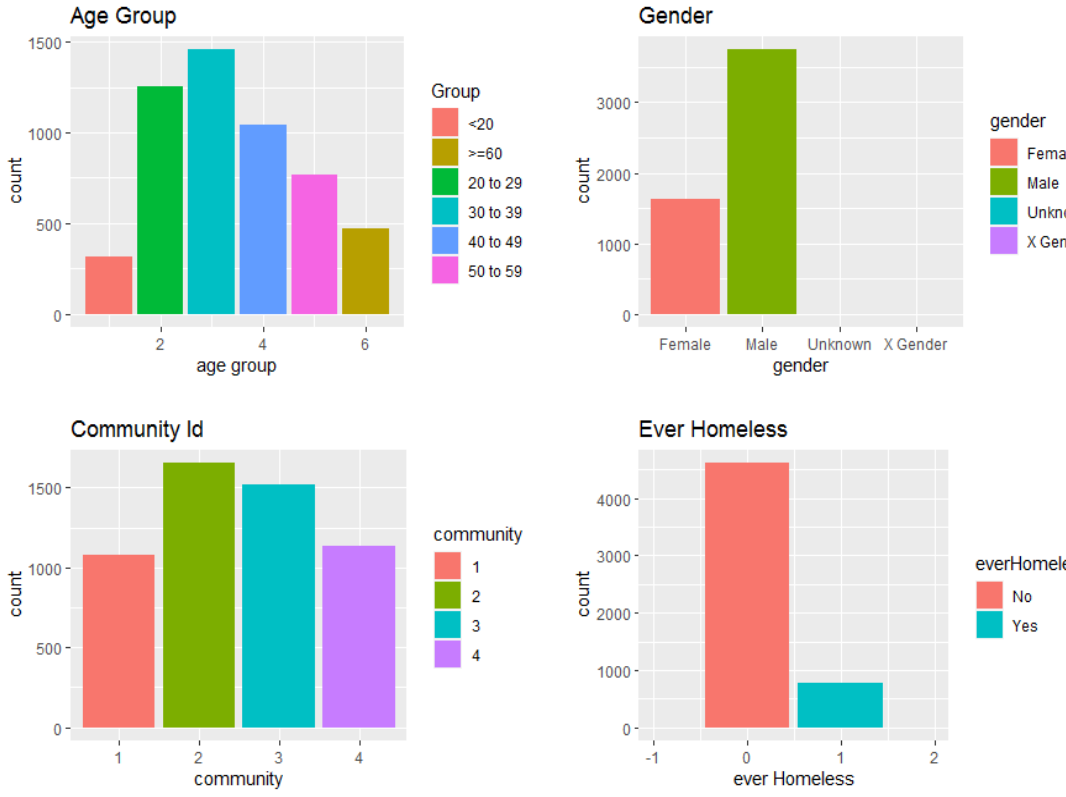


Figure 3. Overall distribution of the grouping attributes. We have different groupings used as covariates, including age group, gender, community grouping, and homelessness status of the overdose patients.

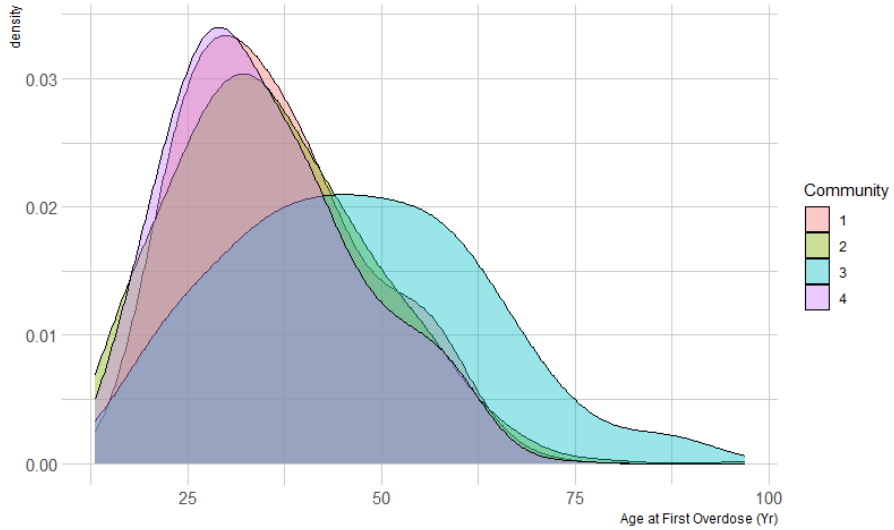


Figure 4. Normalized age distribution plot, grouped by communities. The age distribution of group three is wider compared to other community groups.

3.3. Statistical Analyses

Our analysis began by examining cohort demographics, which included patient grouping based on a community detection algorithm, as well as factors such as gender, age, and homelessness status. Subsequently, Cox proportional hazard models were used to calculate the unadjusted and adjusted hazard ratios (HR and aHR) for patients experiencing a second overdose during the study period. The adjusted hazard model controlled for gender, community ID group, and homelessness status, given previous associations between opioid overdose death rates and older age groups [40] and male gender [41]. Following this, we conducted Kaplan–Meier curve analyses to determine the time to a second overdose event based on gender, age group, homelessness status, and community ID group.

In our analysis, we observed that out of 4515 patients at risk of overdose, 455 experienced a second overdose, resulting in a survival probability of less than 90% within the first hundred study periods. However, when considering the effects of patient grouping, particularly through the community detection algorithm, we identified significant differences in vulnerability within the cohort. Group one, comprising 854 individuals at the start of the study, exhibited the lowest survival probability among the four groups, indicating it as the most vulnerable group despite having a smaller number of at-risk individuals compared to group two, which had the highest number of individuals at risk. All statistical analyses were conducted using the R programming language [42].

4. Results

The cohort under investigation comprised 5380 patients who collectively accounted for 8975 service encounters. Among these patients,

1582 experienced at least one overdose within a span of two thousand days. Notably, in the population experiencing a second overdose, age group three constituted 30.28%, while age group one comprised only 6.45%. Additionally, 73.30% of individuals were male and 25.20% had a history of homelessness.

Regarding community ID grouping, 30.40% belonged to the "reciprocal group". Age groupings were categorized as follows: group one (<20 years), group two (from 20 to 29 years), group three (from 30 to 39 years), group from (40 to 49 years), group five (from 50 to 59 years), and group six (≤60 years) at the time of the first overdose.

Analyzing the grouping by age of patients experiencing a second overdose, we observed that individuals aged 60 years or older had the highest probability of avoiding subsequent overdoses compared to other age groups. Conversely, age groups two and three (from 20 to 29 years and from 30 to 39 years, respectively) exhibited similar, lower survival probabilities.

Figure 5 shows the analysis without grouping. The five-year probability of avoiding a second overdose was approximately 60%. However, after adjusting for various attributes, including gender, homelessness status, age, and community ID, the hazard of experiencing a second overdose increased for male patients (HR = 1.34; 95% CI: 1.18, 1.51) and individuals with a history of homelessness (HR = 1.65; 95% CI: 1.45, 1.88). Conversely, Table 1 shows that the hazard was relatively lower for individuals aged over 20 years, excluding those in community ID group one (e.g., >60 years: HR = 0.45; 95% CI 0.32, 0.62; community ID two: HR = 0.49; 95% CI 0.42, 0.58).

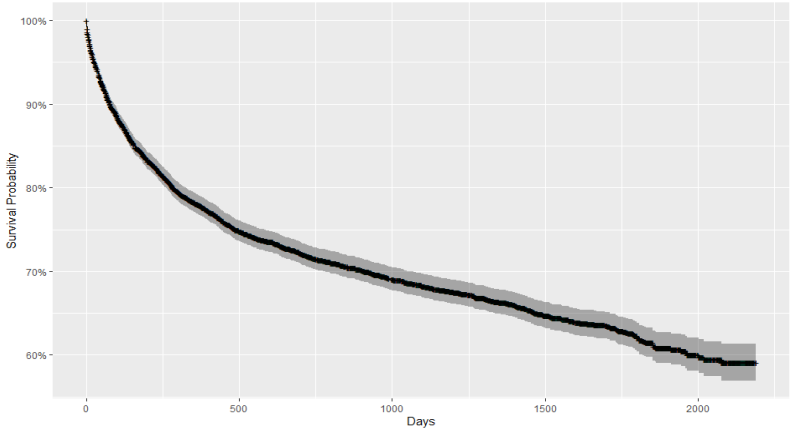


Figure 5. Estimating the survival probabilities over time against second overdose without grouping in the study cohort.

The hazard ratios (HRs) provide crucial insights into the associations between various demographic and contextual factors and the likelihood of experiencing a second overdose within the study period. A hazard ratio represents the relative risk of an event occurring in one group compared to another, with a value greater than one indicating an increased risk and a value less than one indicating a decreased risk. The aHR of 1.34 for male patients suggests that males are 1.34 times more likely to experience a second overdose compared to females, holding all other variables constant. This finding underscores the heightened vulnerability of male individuals within

the cohort to repeat overdose events. The 95% CI (1.18, 1.51) indicates the range within which we can be confident that the true hazard ratio lies, with values above 1 indicating statistical significance. Similarly, an aHR of 1.65 for individuals with a history of homelessness reveals a substantial increase in the likelihood of experiencing a second overdose compared to those who have not experienced homelessness. This result highlights the profound impact of housing instability on the risk of overdose recurrence, suggesting a critical intersection between social determinants of health and substance use outcomes. Again, the narrow 95% confidence interval CI (1.45, 1.88) indicates a statistically significant association.

Conversely, hazard ratios for age groups older than 20 years present interesting findings. Individuals aged 60 years or older exhibit a notably lower hazard of experiencing a second overdose, with aHR of 0.45. This indicates that older individuals are approximately 55% less likely to experience a second overdose compared to younger individuals, after adjusting for other variables. Table 1 shows the confidence interval (95% CI: 0.32, 0.62), which confirms the statistical significance of this effect.

Moreover, individuals in community ID groups other than group one, particularly those in community ID group two, demonstrate a reduced hazard of experiencing a second overdose. The aHR of 0.49 suggests that individuals in community ID group two are approximately 51% less likely to experience a second overdose compared to those in community ID group one, after adjusting for other factors. Again, the narrow confidence interval (95% CI: 0.42, 0.58) underscores the statistical significance of this finding.

Overall, these hazard ratios provide valuable insights into the differential risks associated with demographic and contextual factors, emphasizing the importance of tailored interventions targeting vulnerable subpopulations to mitigate the burden of opioid overdose recurrence.

Figure 6 depicts the survival probability of patients experiencing a second overdose over time, stratified by gender, age group, and community ID, to illustrate these associations. The hazard rate of 1.34 for male gender indicates an 34% increase in the likelihood of a second overdose within two thousand days, while the hazard rate of 1.65 for individuals who have ever experienced homelessness signifies an 65% increase. Notably, there was a significant decrease in the hazard rate among age groups two and three, with over a 50% reduction observed among individuals aged over 60 years. Cumulative proportion plots further illustrate these trends, demonstrating the impact of gender, age, community ID, and homelessness status on the likelihood of experiencing a second overdose. In a comprehensive model encompassing all covariates, male sex (aHR = 1.30; 95% CI 1.15, 1.46), homelessness (aHR = 1.86; 95% CI: 1.63, 2.11), and community ID were identified as significant factors associated with a second overdose, Figure 7.

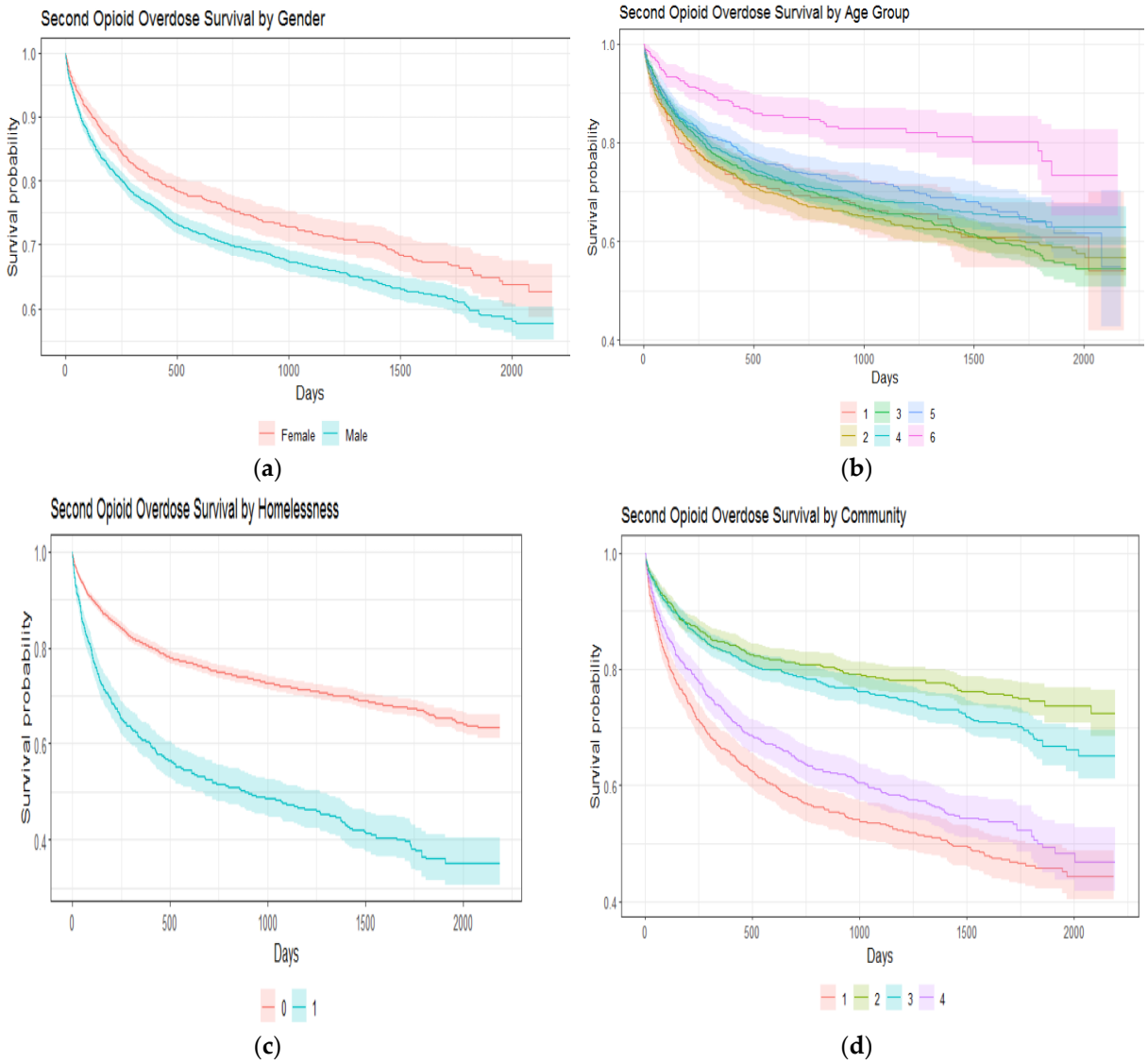


Figure 6. (**a**) Comparing the survival probabilities over time against second overdose for male and female patients. (**b**) Comparing the survival probabilities over time against second overdose among different age groups. (**c**) Comparing the survival probabilities over time against second overdose among different homeless groups. (**d**) Comparing the survival probabilities over time against second overdose among different community groups.

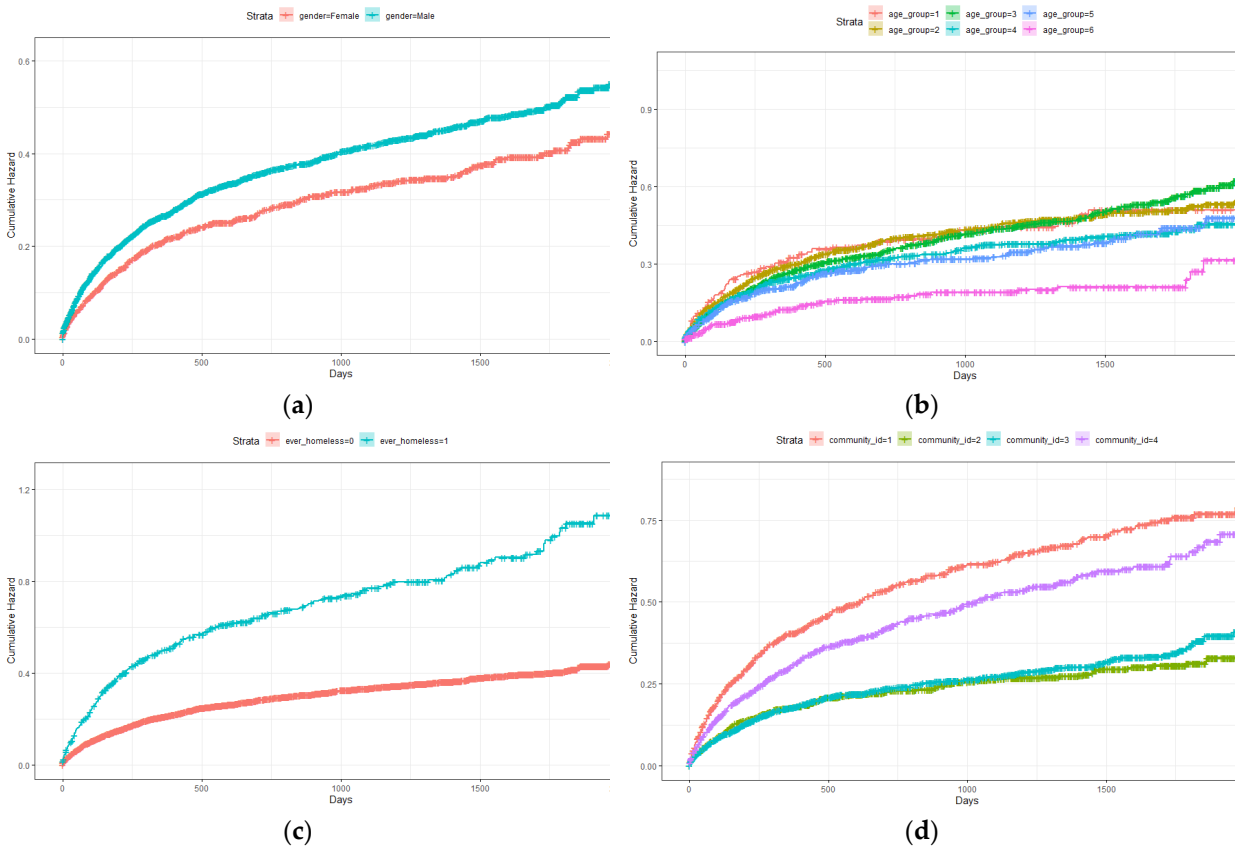


Figure 7. (**a**) Comparing the survival probabilities over time against second overdose for male and female patients. (**b**) Comparing the survival probabilities over time against second overdose among different age groups. (**c**) Comparing the survival probabilities over time against second overdose among different homeless groups. (**d**) Comparing the hazard probabilities over time against second overdose among different community groups.

In this analysis, incorporating all relevant covariates, it becomes evident that certain factors emerge as particularly influential in shaping the likelihood of repeated overdose. Specifically, male sex carries a statistically significant aHR of 1.30 (95% CI 1.15, 1.46), indicating that male individuals are 1.30 times more likely to experience a second overdose compared to their female counterparts, after accounting for others.

Overall, these hazard ratios provide valuable insights into the differential risks associated with demographic and contextual factors, emphasizing the importance of tailored interventions targeting vulnerable subpopulations to mitigate the burden of opioid overdose recurrence. Figure 6 shows the survival probability of patients experiencing a second overdose over time, stratified by gender, age group, and community ID, and illustrates these associations. The hazard rate of 1.34 for male gender indicates a 34% increase in the likelihood of a second overdose within two thousand days, while the hazard rate of 1.65 for individuals who have ever experienced homelessness signifies a 65% increase. Notably, there was a significant decrease in the hazard rate among age groups two and three, with over a 50% reduction observed among individuals aged over 60 years. Cumulative

proportion plots further illustrate these trends, demonstrating the impact of gender, age, community ID, and homelessness status on the likelihood of experiencing a second overdose. In a comprehensive model encompassing all covariates, male sex (aHR = 1.30; 95% CI 1.15, 1.46), homelessness (aHR = 1.86; 95% CI 1.63, 2.11), and community ID were identified as significant factors associated with a second overdose, Figure 7.

In this analysis, incorporating all relevant covariates, it becomes evident that certain factors emerge as particularly influential in shaping the likelihood of repeated overdose. Specifically, male sex carries a statistically significant aHR of 1.30 (95% CI1.15, 1.46), indicating that male individuals are 1.30 times more likely to experience a second overdose compared to their female counterparts, after accounting for other variables. This also measures the level of vulnerability of males to repeated overdose events and emphasizes the importance of gender-sensitive interventions in addressing this disparity.

Similarly, homelessness status emerges as a significant predictor of second overdose risk, with an aHR of 1.86 (95% CI 1.63, 2.11). This suggests that individuals with a history of homelessness are nearly twice as likely to experience a second overdose compared to those who have not experienced homelessness, even after controlling for other factors. This also highlights the profound impact of housing instability on overdose risk and underscores the urgent need for targeted interventions to support individuals experiencing homelessness in managing their substance use disorders.

5. Discussion

Treatment homogeneity may not work in all circumstances. For a cohort of patients that are suffering from an organ-bound illness such as diabetes or kidney disease, treatment homogeneity may be the appropriate approach to use. However, for a cohort of patients suffering from opioid overdose, treatment homogeneity may not work: Healthcare programs benefit certain opioid overdose patients but not others. This is mostly due to the fact that a cohort of opioid patients is heterogeneous with regard to a variety of factors, and as such, it requires a different approach than the traditional taxonomic diagnostic approach to care management.

In this study, we used a graph community detection to systematically partition the OD cohort based on their pattern of engagement with the health service system across the continuum of care. Additionally, we used diagnosis profiles to fine-tune the partition and facilitate the labeling of the groups from the community detection. Our result has shown that there are four distinct communities of patients that constitute the OD cohort. Working with team members with clinical and service system operation backgrounds, the communities were labeled as follows: (1) high risk and reciprocally engaged with the service system; (2) relatively disengaged with the service system; (3) complex health problems and heavy users of Med/Surg services; and (4) high risk with psychiatric issues and unilaterally engaged with the service system. Additionally, their age profile demonstrates a younger population for Communities one, two, and four, with an average age varying between 35 and 36 years. However, Community three is older with an average age of 46 years, and a considerable number of patients are over 60 years.

Focusing on the generated communities of patients and applying survival analysis has shown that the risk profile among these communities is not the same. Two of the communities, including Communities one and four, have a risk twice as great as Communities two and three in experiencing a second overdose. The high-risk Community one has a 51% chance of experiencing a second overdose compared to Community two.

In this study, the combination of the emergency department and acute care (hospital admission) is not used as a proxy for full cross-continuum service utilization. Instead, access to all comprehensive services, including secondary and tertiary services provided by the host organization, is considered and brought into focus. Hence, services such as rehab recovery and harm reduction are brought into focus and are used to distinguish characteristics between the patients constituting the opioid cohort. This made it possible to fine-tune the clustering of patients. Using access to the emergency department and acute care as a proxy for full cross-continuum service utilization would not make such a fine-tuning of the communities possible to enable a more targeted response to their respective needs.

Moreover, the behaviors of individuals are an important determinant of the prevalence of a disease, treatment adherence, as well as health outcome [43]. Opioid overdoses are conditioned by patients' behavior [27]. Bringing individuals' behaviors into focus can be analytically challenging. However, patterns of service utilization reflect the behavior of individuals in relation to the behavior of the system. Hence, PSUs across the continuum of care can be used to bring proximal determinants and behavioral determinants of health into focus. This allows a fine-tuning of clusters, making it possible to distinguish characteristics between high-risk communities, as an example. Although both communities (Communities one and four) are considered high-risk, their behavior vis-à-vis the service system is dissimilar. As a result, a potentially useful approach to reach out and support the "high-risk and reciprocally engaged community" is going to be different from one considered for the "high-risk with psychiatric issues and unilaterally engaged with the service system community." If one looks at the emergency department and acute care (hospital admission) only, it would be impossible to bring into focus the proximal determinants of the health profile of individuals into the analysis.

Other factors outside the use of patterns of service engagement as the basis for partitioning the OD cohort were used. Overall, the result has shown that the identification of male sex, homelessness status, and community ID as significant factors associated with second overdose risk. This underscores the complex interplay of individual, social, and environmental factors in shaping substance use outcomes. By understanding and addressing these factors comprehensively, healthcare providers and policymakers can develop more effective strategies to prevent overdose recurrence and improve the long-term health outcomes of individuals affected by substance use disorders.

Given the data used for this analysis, not all factors that can influence predisposition for a second opioid overdose were included in the analysis. These include patients' distal determinants of health and social determinants that were not collected by the host organization. Moreover, the cohort used for the analysis only captures patients whose overdose was reported and recorded by the host organization. Any unreported overdose that took place

in the community was not included in the analysis. Finally, due to incomplete/inconsistent collection of demographic data at source, as well as strict privacy limitations, information on race or ethnicity was not available for this study. These factors limit the findings of this study. Additionally, the findings of this study are limited to the host organization and hence not immediately generalizable/transferable to other jurisdictions. This is another limitation of the findings from this study. However, the methods outlined in this study are generalizable. to other healthcare jurisdictions.

6. Conclusions

This paper has provided a methodology that can help inform a treatment heterogeneity approach that is likely to be more efficient for a cohort made of diverse individuals, such as an opioid overdose cohort. By grouping opioid overdose patients into different communities informed by their PSUs for the healthcare services across the continuum of care, it is providing an opportunity for the healthcare service system to apply a more targeted approach to care that is likely to be more efficient for each of the communities constituting the overdose cohort.

Using PSUs, the findings from the paper demonstrated that the overdose cohort is not homogeneous with respect to the determinant of risk. This conclusion corroborates results reported in other studies, including [17]. In addition to previous studies findings, the number of groups constituting the various communities that make up the overdose cohort was determined and labeled based on their healthcare service engagement across the continuum of care and clinical characteristics. Finally, the risk for a subsequent overdose was quantified for each of the communities constituting the opioid overdose cohort. Providing such information to a healthcare organization will equip the organization with required information to provide a more differentiated package of services to different fractions of the overdose-at-risk population that are distinguishable on the basis of their proximal determinants of risk profiles, specifically, patterns of interacting with the service system. The intent is better evidence-informed efforts to prevent opioid overdoses.

Author Contributions: Conceptualization, J.B., K.O., Y.S., K.M., A.R., and A.K.; methodology, J.B., K.O., and Y.S.; software, K.O., H.S., and Y.S.; validation, J.B., K.M., and E.C.; formal analysis, J.B., K.O., Y.S., and K.M.; investigation, J.B., K.O., and Y.S.; resources, J.B. and K.M.; data curation, K.O. and Y.S.; writing—original draft preparation, J.B., K.O., and Y.S.; writing—review and editing, J.B., K.O., Y.S., G.D., K.M., A.R., E.C., and A.K.; visualization, J.B., K.O., Y.S., and K.M.; supervision, J.B., K.M., A.R., and A.K.; project administration, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: A certificate of approval was provided by the University of Victoria Research Ethics Board (REB), following British Columbia, Canada Ethics Harmonization Guidelines. The REB number is H21-02817 (approved on 17 June 2022).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are unavailable because of privacy or ethical restrictions. Requests to access the datasets require a

certificate of approval by the University of Victoria Research Ethics Board, following British Columbia, Canada Ethics Harmonization Guidelines.

Conflicts of Interest: The authors declare no conflicts of interest.

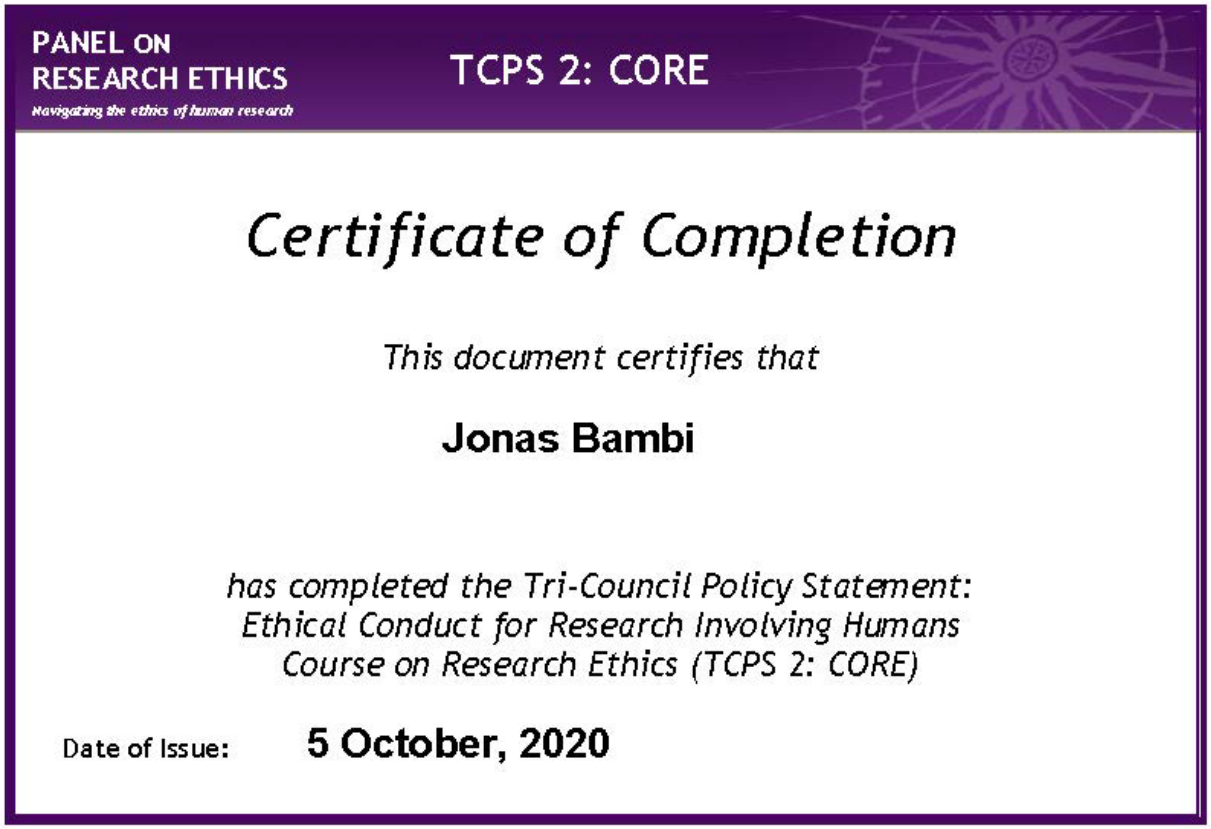
References

46. Dalgleish, T.; Black, M.; Johnston, D.; Bevan, A. Transdiagnostic approaches to mental health problems: Current status and future directions. *J. Consult. Clin. Psychol.* **2020**, *88*, 179–195. https://doi.org/10.1037/ccp0000482. (In English)
47. Kawachi, I.; Berkman, L. Social cohesion, social capital, and health. *Soc. Epidemiol.* **2000**, *174*, 290–319.
48. Chandler, R.K.; Villani, J.; Clarke, T.; McCance-Katz, E.F.; Volkow, N.D. Addressing opioid overdose deaths: The vision for the HEALing communities study. *Drug Alcohol Depend.* **2020**, *217*, 108329.
49. Larochelle, M.R.; Slavova, S.; Root, E.D.; Feaster, D.J.; Ward, P.J.; Selk, S.C.; Knott, C.; Villani, J.; Samet, J.H. Disparities in opioid overdose death trends by race/ethnicity, 2018–2019, from the HEALing communities study. *Am. J. Public Health* **2021**, *111*, 1851–1854.
50. Drake, J.; Charles, C.; Bourgeois, J.W.; Daniel, E.S.; Kwende, M. Exploring the impact of the opioid epidemic in Black and Hispanic communities in the United States. *Drug Sci. Policy Law* **2020**, *6*, 2050324520940428.
51. Altekruse, S.F.; Cosgrove, C.M.; Altekruse, W.C.; Jenkins, R.A.; Blanco, C. Socioeconomic risk factors for fatal opioid overdoses in the United States: Findings from the Mortality Disparities in American Communities Study (MDAC). *PLoS ONE* **2020**, *15*, e0227966.
52. Mitra, A.; Ahsan, H.; Li, W.; Liu, W.; Kerns, R.D.; Tsai, J.; Becker, W.; Smelson, D.A.; Yu, H. Risk factors associated with nonfatal opioid overdose leading to intensive care unit admission: A cross-sectional study. *JMIR Med. Inform.* **2021**, *9*, e32851.
53. Bohnert, A.S.B.; Ilgen, M.A.; Ignacio, R.V.; McCarthy, J.F.; Valenstein, M.; Blow, F.C. Risk of death from accidental overdose associated with psychiatric and substance use disorders. *Am. J. Psychiatry* **2012**, *169*, 64–70.
54. Yule, A.M.; Carrellas, N.W.; DiSalvo, M.; Lyons, R.M.; McKowen, J.W.; Nargiso, J.E.; Bergman, B.G.; Kelly, J.F.; Wilens, T.E. Risk factors for overdose in young people who received substance use disorder treatment. *Am. J. Addict.* **2019**, *28*, 382–389.
55. Karmali, R.N.; Ray, G.T.; Rubinstein, A.L.; Sterling, S.A.; Weisner, C.M.; Campbell, C.I. The role of substance use disorders in experiencing a repeat opioid overdose, and substance use treatment patterns among patients with a non-fatal opioid overdose. *Drug Alcohol Depend.* **2020**, *209*, 107923.
56. van Draanen, J.; Tsang, C.; Mitra, S.; Phuong, V.; Murakami, A.; Karamouzian, M.; Richardson, L. Mental disorder and opioid overdose: A systematic review. *Soc. Psychiatry Psychiatr. Epidemiol.* **2020**, *57*, 647–671.
57. Yamamoto, A.; Needleman, J.; Gelberg, L.; Kominski, G.; Shoptaw, S.; Tsugawa, Y. Association between homelessness and opioid overdose and opioid-related hospital admissions/emergency department visits. *Soc. Sci. Med.* **2019**, *242*, 112585.
58. Gjersing, L.; Amundsen, E. Increasing trend in accidental pharmaceutical opioid overdose deaths and diverging overdose death correlates following the opioid prescription policy liberalization in Norway 2010–2018. *Int. J. Drug Policy* **2022**, *108*, 103785.
59. Mueller, S.R.; Glanz, J.M.; Nguyen, A.P.; Stowell, M.; Koester, S.; Rinehart, D.J.; Binswanger, I.A. Restrictive opioid prescribing policies and evolving risk environments: A qualitative study of the perspectives of patients who experienced an accidental opioid overdose. *Int. J. Drug Policy* **2021**, *92*, 103077.
60. Yarborough, B.J.H.; Stumbo, S.P.; Janoff, S.L.; Yarborough, M.T.; McCarty, D.; Chilcoat, H.D.; Coplan, P.M.; Green, C.A. Understanding opioid overdose characteristics involving prescription and illicit opioids: A mixed methods analysis. *Drug Alcohol Depend.* **2016**, *167*, 49–56.
61. Bahji, A.; Bajaj, N. Opioids on trial: A systematic review of interventions for the treatment and prevention of opioid overdose. *Can. J. Addict.* **2018**, *9*, 26–33.
62. Chu, K.; Carriere, G.; Garner, R.; Bosa, K.; Hennessy, D.; Sanmartin, C. Exploring the intersectionality of characteristics among those who experienced opioid overdoses: A cluster analysis. *Health Rep.* **2023**, *34*, 3–14.
63. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233–234. https://doi.org/10.1038/nmeth.4642.
64. Clark, T.G.; Bradburn, M.J.; Love, S.B.; Altman, D.G. Survival Analysis Part I: Basic concepts and first analyses. *Br. J. Cancer* **2003**, *89*, 232–238.
65. Ohno-Machado, L. Modeling medical prognosis: Survival analysis techniques. *J. Biomed. Inform.* **2001**, *34*, 428–439.

66. Agarwal, N.; Biswas, B.; Singh, C.; Nair, R.; Mounica, G.; Jha, A.R.; Das, K.M. Early Determinants of Length of Hospital Stay: A Case-Control Survival Analysis among COVID-19 Patients Admitted in a Tertiary Healthcare Facility of East India. *J. Prim. Care Community Health* **2021**, *12*, 21501327211054281.
67. Ho, S.-H. Survival analysis of living arrangements and health care utilization in terms of total mortality among the middle-aged and elderly in Taiwan. *J. Nurs. Res.* **2008**, *16*, 160–168.
68. Bambi, J.; Santoso, Y.; Sadri, H.; Moselle, K.; Rudnick, A.; Robertson, S.; Chang, E.; Kuo, A.; Howie, J.; Dong, G.Y.; et al. A Methodological Approach to Extracting Patterns of Service Utilization from a Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems. *BioMedInformatics* **2024**, *4*, 946–965.
69. Bambi, J.; Sadri, H.; Moselle, K.; Chang, E.; Santoso, Y.; Howie, J.; Rudnick, A.; Elliott, L.T.; Kuo, A. Approaches to Generating Patterns of Service Utilization for Patients with Complex Conditions: Graph Community Detection vs Natural Language Processing Clustering. *BioMedInformatics* **2024**, *4*, 1884–1900.
70. Bambi, J.; Santoso, Y.; Moselle, K.; Robertson, S.; Rudnick, A.; Chang, E.; Kuo, A. Analyzing Patterns of Service Utilization Using Graph Topology to Understand the Dynamic of the Engagement of Patients with Complex Problems with Health Services. *BioMedInformatics* **2024**, *4*, 1071–1084.
71. Bambi, J.; Dong, G.Y.; Santoso, Y.; Moselle, K.; Dugas, S.; Olobatuyi, K.; Rudnick, A.; Chang, E.; Kuo, A. Patterns of Service Utilization across the Full Continuum of Care: Using Patient Journeys to Assess Disparities in Access to Health Services. *Knowledge* **2024**, *4*, 252–264.
72. Park, T.W.; Lin, L.A.; Hosanagar, A.; Kogowski, A.; Paige, K.; Bohnert, A.S.B. Understanding risk factors for opioid overdose in clinical populations to inform treatment and policy. *J. Addict. Med.* **2016**, *10*, 369–381.
73. Tseregounis, I.E.; Henry, S.G. Assessing opioid overdose risk: A review of clinical prediction models utilizing patient-level data. *Transl. Res.* **2012**, *234*, 74–87. https://doi.org/10.1016/j.trsl.2021.03.012.
74. Suffoletto, B.; Zeigler, A. Risk and protective factors for repeated overdose after opioid overdose survival. *Drug Alcohol Depend.* **2020**, *209*, 107890. https://doi.org/10.1016/j.drugalcdep.2020.107890.
75. Government of British Columbia. "Health Authorities " Government of British Columbia. Available online: https://www2.gov.bc.ca/gov/content/health/about-bc-s-health-care-system/partners/health-authorities (accessed on 20 June 2024).
76. Koval, A.; Moselle, K. Clinical Context Coding Scheme—Describing Utilisation of Services of Island Health between 2007–2017. In Proceedings of the Conference of the International Population Data Linkage Association, Banf, AB, Canada, 12–14 September 2018.
77. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008.
78. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer: Berlin/Heidelberg, Germany, 2003.
79. Li, Y.; Rakesh, V.; Reddy, C.K. Project success prediction in crowdfunding environments. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 247–256.
80. Dunn, O.J.; Clark, V.A. *Basic Statistics: A Primer for the Biomedical Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
81. Zupan, B.; Demšar, J.; Kattan, M.W.; Beck, J.R.; Bratko, I. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artif. Intell. Med.* **2000**, *20*, 59–75.
82. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481.
83. Lee, E.T.; Wang, J. *Statistical Methods for Survival Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
84. Allison, P.D. *Survival Analysis Using SAS: A Practical Guide*; Sas Institute: Cary, NC, USA, 2010.
85. Bohnert, A.S.; Valenstein, M.; Bair, M.J.; Ganoczy, D.; McCarthy, J.F.; Ilgen, M.A.; Blow, F.C. Association between opioid prescribing patterns and opioid overdose-related deaths. *JAMA* **2011**, *305*, 1315–1321.
86. Paulozzi, L.J.; Kilbourne, E.M.; Shah, N.G.; Nolte, K.B.; Desai, H.A.; Landen, M.G.; Harvey, W.; Loring, L.D. A history of being prescribed controlled substances and risk of drug overdose death. *Pain Med.* **2012**, *13*, 87–95.
87. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical: Vienna, Austria, 2008.
88. Horwitz, R.I.; Horwitz, S.M. Adherence to treatment and health outcomes. *Arch. Intern. Med.* **1993**, *153*, 1863–1868.

Appendices

A. Ethics Review and Approval Letter



PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE

Certificate of Completion

This document certifies that

Jonas Bambi

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

Date of Issue: **5 October, 2020**

Island Health

Certificate of Ethical Approval: Amendments for Harmonized Minimal Risk Behavioural Study

Research Ethics and Compliance Office
Island Health
2400 Arbutus Road
Victoria, BC, V8N 1V7
Tel: 250-519-6726

Also reviewed and approved by:

- **University of Victoria**

Principal Investigator:	**Primary Appointment:**	**Board of Record REB Number:**	**REB Number: H21-02817**
Alex MH Kuo			PAA #: H21-02817-A002

Study Title:
Understanding Patterns of Service Utilization for the Underserved Vulnerable Population with Complex Healthcare Issues to Optimize Service Delivery

Approval Date: May 23, 2023 **Expiry Date: May 15, 2024**

Research Team Members: Kenneth Moselle, Abraham Rudnick, Alex MH Kuo

Sponsoring Agencies: N/A

Documents included in this approval:

Document Name	Version	Date
Protocol: HREB Protocol for Caring for the Vulnerable Population with Complex Healthcare Issues Study_V3.0. May 16 2023.	V3	May 16, 2023

This ethics approval applies to research ethics issues only and does not include provision for any administrative approvals required from individual institutions before research activities can commence.

The Board of Record (as noted above) has reviewed and approved this study in accordance with the requirements of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2, 2018).

The "Board of Record" is the Research Ethics Board delegated by the participating REBs involved in a harmonized study to facilitate the ethics review and approval process.

The application for ethical review and the document(s) listed above have been reviewed and the procedures were found to be acceptable on ethical grounds for research involving human subjects.

This study has been approved either by the Board of Record's full REB or by an authorized delegated reviewer.

B. Operational Review and Approval Letter

Institutional Approval Certificate

Vancouver Island Health Authority
Research Ethics & Compliance Office
Queen Alexandra Centre for Children's Health
Main Building Room 205 - 2400 Arbutus Road, Victoria, BC V8N 1V7

island health

INSTITUTIONAL APPROVAL TO CONDUCT A RESEARCH PROJECT

Study Number: H21-02817
Study Title: Understanding Patterns of Service Utilization for "Medically Hard-to-House" Populations to Optimize Service Delivery

Institutional Approval Date: 21 June 2023
Certificate of Ethical Approval Date: 17 June 2022

Principal Investigator: Alex Kuo
Island Health Position: N/A

Supervisor: N/A
Department: N/A

Island Health Collaborator: Jonas Bambi
Project Team Members: Dr. Ken Moselle; Dr. Abraham Rudnick

Sponsoring Agencies: N/A
Funding Title: N/A

This is to inform you that your research project may now be initiated as of the Institutional Approval date above and is approved based on the following:

1. The Certificate of Approval dated above issued by the Health Research Ethics Board on behalf of Island Health.
2. All Island Health Operational Review approvals are received.

The Institutional Approval to Conduct a Research Project will remain in effect as long as the Health Research Ethics Board approval is renewed annually and all amendments submitted are approved as required throughout the duration of this project. The Institutional Approval to Conduct a Research Project will expire upon the HREB receipt and acknowledgement of the study closure report.

This Institutional Approval to Conduct Research does not represent approval to use a new intervention or product within Island Health. Please discuss with the appropriate Director/Executive Director regarding making any changes to practice.

All necessary ethical, hospital department/facilities approvals and institutional agreements/contracts are now in place and you have permission to begin your research. *
Cindy Trytten
Director, Research
Island Health

***Island Health Departments, Unit, and Programs may require a copy of this certificate prior to granting access to their services for this project.**

C·A·R·E
Our Values
Courage: to do the right thing – to change, innovate and grow.
Aspire: to the highest degree of quality and safety.
Respect: to value each individual and bring trust to every relationship.
Empathy: to give the kind of care we would want for our loved ones.

v.21Feb2023

C. Contribution Narrative and Publication Status

Manuscript	Members	Conceptualization	Methodology	Project Administration	Resources	Supervision	Data Curation	Software	Formal Analysis	Validation	Visualization	Investigation	Writing - Original Draft	Writing -Review and Editing	Rewriting and Addressing Reviewers Concerns	Status
Manuscript 1: A Methodological Approach to Extracting Patterns of Service Utilization from a Cross-Continuum High Dimensional Healthcare Dataset to Support Care Delivery Optimization for Patients with Complex Problems Utilization	Jonas Bambi	X	X	X	X	X	X	X	X	X	X	X	X	X	X	Accepted
	Yudi Santoso		X					X	x	X				X		
	Hanieh Sadri		X					X	X					x		
	Ken Moselle	X	X		X	X			X	X	X	X		X	X	
	Abraham Rudnick	X	X			X			X	X				X		
	Stan Robertson						X	X						X		
	Ernie Chang							X						X		
	Alex Kuo	X				X								X		
	Joseph Howie							X						X		
	Gracia Dong													X		
	Kahinde Olobatuyi													X		
	Mahdi Hajiabadi							X						X		
	Ashlin Richardson							X						X		
Manuscript 2: Approaches to Extracting Patterns of Service Utilization for	Jonas Bambi	X	X	X	X	X	X	X	X	X	X	X	X	X	X	Accepted
	Hanieh Sadri						X	X						X		
	Ken Moselle	X			X	X			X					X		
	Ernie Chang		X						X					X		

Patients with Complex Conditions: Graph Community Detection vs NLP Clustering	Yudi Santoso							X						X		
	Joseph Howie							X						X		
	Abraham Rudnick	X				X								X		
	Lloyd Elliot													X		
	Alex Kuo	X				X								X		
Manuscript 3: Analyzing PSUs Graph Topology to Understand the Dynamic of the Engagement of Patients with Complex Problems with Health Services	Jonas Bambi	X	X	X	X	X			X	X	X	X	X	X	X	Accepted
	Yudi Santoso	X	X					X	X	X	X	X	X	X		
	Ken Moselle	X	X		X	X			X	X	X	X		X	X	
	Stan Robertson				X		X	X						X		
	Abraham Rudnick	X				X								X		
	Ernie Chang							X		X				X		
	Alex Kuo	X				X								X		
Manuscript 4: Disparities in Access to Services, as Evident in Patients Journeys: Illustrating a Nuanced Approach in Assessing Healthcare Equity Using PSUs Across the Full Continuum of Care	Jonas Bambi	X	X	X	X	X			X	X	X	X	X	X	X	Accepted
	Gracia Dong	X	X				X	X	X		X	X	X	X	X	
	Yudi Santoso	X	X				X	X	X		X	X	X	X		
	Ken Moselle	X			X	X			X	X				X		
	Sophie Dugas													X		
	Kehinde Olobatuyi													X		
	Abraham Rudnick	X				X								X		
	Ernie Chang									X				X		
	Alex Kuo	X				X								X		
Manuscript 5: Use of PSUs and Hierarchical Survival Analysis in	Jonas Bambi	X	X	X	X	X			X	X	X	X	X	X	X	Accepted
	Kehinde Olobatuyi	X	X				X	X	X		X	X	X	X	X	

Planning and Providing Care for Overdose Patients and Predicting the Time-to-Second Overdose	Yudi Santoso	X	X				X	X	X		X	X	X	X		
	Hanieh Hadri													X		
	Ken Moselle	X			X	X			X	X	X			X		
	Abraham Rudnick	X				X								X		
	Gracia Dong													X		
	Ernie Chang									X				X		
	Alex Kuo	X				X								X		

Contributor Roles Taxonomy

Conceptualization	Ideas; formulation or evolution of overarching research goals and aims.
Data curation	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.
Formal analysis	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.
Investigation	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
Methodology	Development or design of methodology; creation of models.
Project administration	Management and coordination responsibility for the research activities planning and execution, to ensure that all the activities are aligned with overall research goals and aims
Resources	Provision of research materials, including data, computing resources, and other analysis tools.

Software	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
Supervision	Oversight and leadership responsibility for the research activity planning and execution, including mentorship internal and external to the core team.
Validation	Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
Visualization	Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.
Writing - original draft	Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
Writing - review & editing	Critical review, commentary or revision of manuscript pre publication stages.
Rewriting and Addressing Reviewers Concerns	Addressing reviewers' feedback and rewriting/restructuring the manuscript if required