

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Secondary use of routinely collected administrative health data for epidemiologic research: Answering research questions using data collected for a different purpose

Scott D. Emerson^{1,*}, Taylor McLinden², Paul Sereda¹, Amanda M. Yonkman¹, Jason Trigg¹, Sandra Peterson³, Robert S. Hogg^{1,2}, Kate A. Salters^{1,2,4}, Viviane D. Lima^{1,4}, and Rolando Barrios^{1,5}

Submission History

Submitted:	16/03/2024
Accepted:	11/09/2024
Published:	19/11/2024

¹British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

²Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

³Centre for Health Services and Policy Research, School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

⁴Department of Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

⁵School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

Abstract

The use of routinely collected administrative health data for research can provide unique insights to inform decision-making and, ultimately, support better public health outcomes. Yet, since these data are primarily collected to administer healthcare service delivery, challenges exist when using such data for secondary purposes, namely epidemiologic research. Many of these challenges stem from the researcher's lack of control over the quality and consistency of data collection, and – furthermore – a lessened understanding of the data being analyzed. That said, we assert that these challenges can be partly mitigated through careful, systematic use of these data in epidemiologic research. This article presents considerations derived from experiences analyzing administrative health data (e.g., healthcare practitioner billings, hospitalizations, and prescription medication data) in the Canadian province of British Columbia (population of over 5 million in 2024), though we believe the underlying principles generalize beyond this region. Key considerations were organized around four themes: 1) *Know the data and their primary use* (understand their scope and limitations); 2) *Understand classification and coding systems* (appreciate the nuances regarding classification systems, versions, how they are employed in the primary uses of the data, and querying the values); 3) *Transform data into meaningful forms* (process data and apply identification algorithms, when necessary); 4) *Recognize the importance of validity when defining analytic variables* (make meaningful inferences based on data/algorithms). Although this article is not an exhaustive list of all considerations, we believe that it will provide pragmatic insights for those interested in leveraging administrative health data for epidemiologic research.

Keywords

administrative health data; validity; Canada; epidemiology

*Corresponding Author:

Email Address: semerson@bccfe.ca (Scott D. Emerson)



Introduction

Routinely collected administrative health data (e.g., hospitalization records, prescription medication data) are produced primarily for administration purposes (e.g., billings from healthcare practitioners), rather than guided by *a priori* research questions. These data, however, can be repurposed to support endeavours such as epidemiologic research, program evaluation, and health system performance management [1–3].

Compared to primary sources of health research data, which are often collected with substantial researcher control and oversight (e.g., primary data collection directly within observational studies or clinical trials, through means such as participant questionnaires/surveys), the secondary use of administrative data offers several advantages. It is typically cheaper and more efficient to request to access existing administrative health data on a group of participants than it would be to collect primary data on the same number of participants. This is particularly the case in settings where longitudinal data are needed on large samples of people, within various geographic regions. Administrative data can reflect information from virtually all documented healthcare interactions by residents of a region, administrative group, or program – and can do so over many years [1, 2]. By their very nature, these data are also generally representative of the real world in terms of healthcare use patterns including challenges such as untimely follow-up, and other suboptimal health system realities [1, 4, 5]. The secondary use of administrative data for research presents a unique scenario – the researcher knows which data were collected, so must ask themselves: can my research question, with the appropriate partnerships, collaborations, and expertise, be validly answered using these data? This contrasts with primary data collection scenarios, where *a priori* research questions are typically posed before the data are collected (e.g., survey items are included based on *a priori* research goals, and the researcher does not know what the responses will be). In the administrative data scenario, the data have already been collected. As such, the focus is on understanding the strengths and limitations of such information and, after careful consideration, repurposing the administrative data for addressing suitable research questions.

Administrative health data as a source of information for research can be more desirable in some contexts than others. The real-world reality of healthcare – characterized by factors such as missed appointments, suboptimal continuity of care, and persons not accessing healthcare care for certain conditions – can bias inferences based on healthcare data. For example, if one wanted to document the prevalence of depression, one could use administrative health data to measure recorded depression-related healthcare interactions (e.g., based on International Classifications of Diseases [ICD] diagnostic coding). However, because depression can go undiagnosed, it may be under-captured in the healthcare data; administrative data alone may underestimate the prevalence of depression in a given sample of people [6]. Thus, a question that can be more validly addressed with administrative data would concern the prevalence of depression-related healthcare interactions, which would implicitly acknowledge that depression is underdiagnosed (or undiagnosed) in many settings. Nevertheless, group differences in diagnosis as well as healthcare access or care seeking behaviours have also been

documented. For instance, lower socio-economic level groups and some ethno-cultural or immigrant groups may be less likely to access healthcare for mental health and substance use for various reasons [7, 8]. Additionally, observed trends may also be biased by changes over time in the provision of healthcare services, policy, and data coding standards, such as in 2011 in the United States where the number of diagnostic codes permitted in a hospitalization record increased from 9 to 25. This resulted in a higher number of comorbidities being recorded in the healthcare data [9]. In summary, an overarching challenge related to repurposing administrative data for research is the lack of control over the timing, content, and/or format of data collection [10, 11]: one can solely see the events that resulted in a recorded healthcare system interaction. This reality can evoke concerns about data quality, misclassification (e.g., when defining analytic variables using algorithms), and the usefulness of these data for research [12–14].

We believe that the concerns listed above are valid, yet we assert that many concerns can be partly mitigated via the pragmatic use of administrative data for research. In this article, we present four key considerations derived from experiences analyzing administrative health data (see Table 1), chiefly in the context of the Canadian province of British Columbia (BC). Many of the noted principles, however, generalize elsewhere and in several instances represent what are likely to be universal considerations. Although important discussions on using administrative data for research exist [10, 15, 16], we believe this article provides a novel contribution by discussing issues from the perspective of a researcher engaging in epidemiologic research. From the outset, we wish to emphasize the importance of multidisciplinary, team-based approaches to applying these considerations, as having various partners and collaborators with diverse areas of expertise, research backgrounds, and lived experiences [17] is vital to ensuring a more informed approach to epidemiologic research.

Consideration 1: Know the data and their primary use

Given the absence of researcher control over data collection processes for administrative data, and since research questions are devised *after* data are collected, it is valuable to understand data collection mechanisms and how their relation to the primary use of the data can introduce challenges when repurposing for research. Such an understanding also requires an awareness of common misconceptions surrounding use of administrative data in research [15]. As administrative health data are collected for non-research purposes, a better understanding of how and why they are collected (i.e., their *intended* purposes, such as billing, planning, or administrative aims), as well as which populations are included [15] can help inform the validity, limitations, and generalizability of their use for research. To help further such understanding, one may consider asking of the data they intend to analyze: *Why, when, how, and for whom were they collected?*

To illustrate one example about gaps in knowing one's data, take the instance of the healthcare practitioner and laboratory services billings dataset in BC (the Medical Services Plan [MSP] Payment Information File). Published

Table 1: Summary of considerations when using administrative health data for epidemiologic research

Consideration	Key points
Know the data and their primary use	– Understand their scope – Appreciate their limitations, and formulate research questions accordingly
Understand classification and coding systems	– Nuances of diagnostic code systems, versions, and the code values – Consider how version changes over time or between databases may impact patterns
Transform data into meaningful forms	– Value of pre-processing/cleaning – Case-finding and other identification algorithms can help provide less biased classifications of cases, events, characteristics
Recognize the importance of validity when defining analytic variables	– Various factors can influence the validity of the approaches used to identify cases/events/characteristics used in analyses – Useful to quantify the extent to which algorithms yield meaningful, appropriate inferences for one's intended context

research papers, spanning various domains across many time periods, occasionally describe this dataset as an “outpatient physician” billings dataset despite the coverage being broader – it also contains billings from non-physician healthcare practitioners (including nurse practitioners) and medical laboratory staff, and includes interactions from settings beyond solely outpatient care, such as services provided by certain practitioners in inpatient, emergency department, and medical laboratory diagnostic facilities (16). This exemplifies a data definition/description issue that may not be well publicly documented but will be known by researchers/analysts familiar with the data. Accessible, living documentation and metadata that clearly describe data and related quality issues is thus vital to support informed research using administrative data.

A better understanding of the overall data lifecycle – including the systems, technologies, and steps involved from data entry through to processing and finalizing the records as they appear in one's research database is also useful to better understand potential systematic quality issues with administrative data. For instance, text in fields may be truncated due to limits in data collection systems, special characters may appear as implausible values, or decimal places may be removed from codes that rely on such values to correctly identify categories. For example, in Canada's largest province, Ontario, only diagnostic code values of up to three digits/characters are enterable in the healthcare practitioner billings database [18]. In BC, decimal points are stripped from ICD diagnostic codes in common datasets (e.g., the MSP Payment Information File) and leading zeroes can sometimes be added (by billing systems), which can introduce uncertainty in ascertaining the original/intended value (e.g., ICD-9 values 250: *diabetes*, and 0250: *meliodosis* may be conflated if leading zeroes are introduced) [19]. Awareness of such data infrastructure contexts can thus help ensure appropriate querying of the datasets representing the healthcare interactions of large samples of people.

To improve understanding of the data, one small step can be creating simple tabulations and visualizations of data fields over time, as well as stratifications. Doing so can provide clues of data changes warranting further investigation and enable cross-referencing with other sources. For example, data artefacts can occur whereby an apparent finding or change in some pattern (e.g., a sudden increasing/decreasing trend in prevalence of a condition) is the product of a data collection or coverage issue. Examples that can create such false findings, for example, include changes to the diagnostic classification system over time (e.g., ICD-9 to ICD-10-CA in Canada's Discharge Abstract Database [DAD]), the introduction of newer fields to capture information previously unavailable, or changes to the categories captured in existing fields. For instance, the number of service locations in BC's healthcare practitioner and laboratory services billings database have expanded markedly over time from initially two categories (“Hospital” or “Physician's Office”) to numerous settings including inpatient hospitalization, emergency department, outpatient office within a hospital, and healthcare practitioner office (Figure 1) [20]. If unaccounted for, these changes can pose substantial challenges when repurposing administrative data for epidemiologic research and introduce artefacts of spurious changes. Visual and tabular explorations of administrative data can help broaden one's knowledge and understanding of potential issues requiring attention.

While such descriptive examination of data over time and across strata can provide some clues for further examination, this approach is often, by itself, insufficient; data-driven explorations and investigations should be complemented by a deeper understanding of system and contextual factors that may impact the observed data patterns. Administrative data are typically linked with other datasets; for example, people's demographic information may be linked to prescription medication data and hospitalizations. People may be missing from the linkage, such as those whose personal healthcare number, name, birthdate, and/or sex, are missing or invalid, or

Figure 1: Longitudinal availability of service location categories for healthcare practitioner and laboratory services billings in British Columbia, Canada (within the Medical Services Plan [MSP] Payment Information File)

Service Location	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
O - Physician's office													Retired Oct 2006														
A - Practitioner's office - in community																											
T - Practitioner's office - in publicly administered facility																											
H - Hospital													Retired Oct 2006														
E - Hospital - emergency room (unscheduled patient)																											
I - Hospital inpatient																											
P - Hospital - outpatient																											
G - Hospital - day care (surgery)																											
R - Patient's private home																											
S - Community location, e.g. school, hall													Retired Oct 2006														
Z - Other (i.e., accident site or in an ambulance)																											
D - Diagnostic facility																											
F - Private medical/surgical facility																											
C - Residential care/assisted living residence																											
M - Mental health centre																											

Note. In 2006, service location became mandatory in MSP Payment Information Plan billings. In 2021, numerous additional codes were added (in particular, the 'A – Practitioner's office - in community' category has been split into numerous new codes).

when they do not consent to have their data linked [21]. More fulsome discussion of linkage/non-linkage and its impacts on data omissions can be found elsewhere, but is an important point to consider with respect to who is, and is not, included in an analytical research dataset [15, 22]. More broadly, one should remain aware of how social determinants of health can shape various aspects of the available data – throughout various stages of the process from linkage through to the presence and patterns of healthcare use observed. Health system literacy, structural barriers, discrimination, and group-based differences in healthcare use and access are all factors that can shape the data available for a research study [7, 8]. For instance, in BC one indicator of low socio-economic status is receipt of a subsidy for health insurance. One-quarter of eligible (low-income) households did not, however, receive this subsidy, since it had an opt-in policy [23]. This is but one example of how social determinants of health – in this case healthcare system knowledge and barriers – could shape the type of information present in administrative data and thus impact research findings.

Despite an individual researcher often having to work with what already exists in the administrative datasets, some pitfalls can be mitigated by having a data-informed understanding of the data being used. Diligent consultation of existing documentation by data providers [24], researchers/analysts familiar with the data, healthcare practitioners whose services are populated in the datasets, and publications or concept dictionary repositories demonstrating application and use of data [19, 25] are all worthwhile avenues to pursue to improve understanding of this type of data source. Such knowledge can help guide which analyses, derivations, and data uses are appropriate under which circumstances.

Consideration 2: Understand classification and coding systems

Related to the first consideration regarding knowing the data, one specific aspect pertains to standardized classification and coding systems; due to the complexity and centrality of these systems, we believe a dedicated consideration is warranted. These systems assign groupings or labels to characterize events

such as diagnoses, reasons for visits, and surgical interventions. Hospitalization data typically include information about healthcare interventions that occur during a person's stay at a medical facility (e.g., typically, but not exclusive to, a hospital), including surgical operations. Currently, the classification system used in Canadian hospitalization data comprises ~18,000 unique values [26], which demonstrate a hierarchical format such that the first digit indicates a broad intervention type (e.g., therapeutic, obstetric), and subsequent characters delineate greater nuance (e.g., anatomic area, technique/device used). Another key classification system pertains to characterizing the reasons for healthcare use. Classification of diagnoses and factors affecting healthcare use is often conducted via the World Health Organization (WHO)'s ICD system [27]. ICD diagnostic codes "form the backbone structure of disease classification worldwide" [28]. Understanding the structure of this system, its versioning over time (and country-specific modifications, like the -CA, Canadian modification), is fundamental when leveraging any datasets containing ICD codes. The ICD codes form a nested hierarchical system that thematizes diseases and events into groupings, and subgroupings (see Figure 2 for an illustration). For instance, the ICD-10 (tenth revision) code J45 – *Asthma*, is nested within the subgrouping: J45-J47 – *Chronic lower respiratory diseases*, which is further nested within the grouping: J – *Diseases of the respiratory system*. The ICD system has undergone multiple revisions/updates since its inception. Currently, ICD-9 (ninth revision) and ICD-10 (tenth revision), and variants therein (e.g., ICD-10-CA), appear in many healthcare database systems globally [27]. Despite much overlap in fundamental goals, important changes exist between these revisions: ICD-10 codes start with a letter, whereas most ICD-9 codes are numeric; the former has approximately 5 times more codes than ICD-9 [29]. The level of detail has increased with newer versions. For instance, at the broader, 3-digit level, substance use related ICD-9 codes (292 – *Drug psychoses*, 304 – *Drug dependence*, 305 – *Nondependent abuse of drugs*) do not specify the substance involved (except alcohol: 291 – *Alcoholic psychoses*, 303 – *Alcohol dependence syndrome*), whereas comparable codes in ICD-10 generally denote the substance involved (e.g., *Mental and behavioural disorders due to use of... alcohol* (F10), *opioids* (F11),

cocaine (F14), *hallucinogens* (F16), *tobacco* (F17)) [30, 31]. As mentioned, modifications also appear, such as ICD-10-CM (“clinical modification version”) in the United States, Canada’s -CA (Canadian version), and Australia’s -AM (Australian modification) [32]. Such modifications introduce additional codes, groupings, and complexities to address country-specific needs [29, 32, 33]. Hence, knowing the classification systems for core information pieces in the data – and versioning changes – is becoming ever more important as international studies (involving administrative health data from multiple countries) continue to be published.

Appropriately querying ICD diagnostic codes

There are many factors to consider when analyzing data that contain ICD codes. First, the researcher should know which ICD versions exist in a database to ensure queries are appropriate. It is also important to understand version changes across time and geography. Some ICD codes may only exist – at all or to a certain level of detail – in particular regions [27]. Temporally, version changes can impact the patterns observed. For instance, BC hospitalization data in the DAD database used ICD-9 and ICD-9-CM (United States version) until April 2001 and thereafter switched to ICD-10-CA [33]. Some conditions were absent or specified in less detail prior to this switch (perhaps because less was known clinically at the time of the previous version), and it has been shown that the transition may have introduced some relevant data-related artefacts at the point of transition [33]. For example, the transition to ICD-10-CA led to apparent (but likely spurious) increases in occurrence of secondary diagnoses (those present at admission) in BC [33], and differences in cause-specific mortality rates in Canada [34]. Hence, understanding potential time-varying artefacts introduced as the classification systems change over time is important to avoid misinterpreting trends.

Another factor to note is inconsistency in diagnostic coding data in health databases containing ICD codes. In some cases, unofficial mixing of code versions can occur. For instance, ICD-9 is the official version for BC’s MSP database, but occasionally some ICD-9-CM codes appear (e.g., ICD-9 lacks codes specific to asthma exacerbation but ICD-9-CM has one: 493.02 – *Extrinsic asthma with (acute) exacerbation*). Although speculation, this may be why this particular ICD-9-CM began to appear in what was “officially” an ICD-9-based database. Hence, it could be useful to query both possible versions to ensure better capture. Relatedly, due to data entry errors or structural aspects of the datasets, unofficial values, simply introduced by typographical errors, may be included after codes (e.g., “J45” may appear as “J45a”) [19]. For these reasons, a useful general rule can be to query codes beginning with the values of interest, rather than the exact value; this is occasionally denoted as, for example, “J45*”. Regardless, this should be investigated at the exploratory analysis stage and the strategy for querying data (based on diagnostic coding) be clearly articulated in the methods section of a paper. The exact value may be queried in cases where solely that exact diagnostic code is of interest – and one wants to avoid potentially including

subtypes within. This choice is context-specific, indeed some case-finding algorithms state which codes should be queried ‘exactly’ or ‘starting with’ [35]. A further consideration is that it may be more pragmatic to query a higher/broader level of diagnostic code (sometimes called a ‘parent code’) as it may be that healthcare practitioners use ICD values at the 3-digit level rather than 4- or 5-digit level, as seen in the BC context for the healthcare practitioner and laboratory services billings [19] as well as some Canadian regions where only 3-digit level ICD codes are entered in similar databases [18]. Also, inclusion of unofficial, non-ICD values may occur such as in BC’s healthcare practitioner and laboratory services billings dataset [36, 37]. For example, a case-finding algorithm in BC for mood and anxiety disorders includes a region-specific diagnostic code “50B – *Anxiety/Depression*” not present in the official WHO ICD system [38]. Similarly, in Canada a subset of urgent care-related ICD-10-CA codes are included in the pan-Canadian emergency visit database – The Canadian Emergency Department Diagnosis Shortlist (CED-DxS) – with some modifications to the descriptions and structure of the diagnostic code names [39]. Therefore, knowing which diagnostic coding systems exist in one’s data is key to performing appropriate queries and transparently described what was in fact done in the methods section of a published study.

It is important to carefully select how diagnostic code fields should be queried. For example multiple diagnostic codes can be included within a single healthcare record (e.g., in a hospitalization record) [40]. One should consider whether all diagnostic code fields warrant being queried (to ensure a broader/more sensitive capture) or if only certain codes should be (to ensure consistency with an algorithm or rationale). For many conditions, established code sets may be available based on literature reviews, landmark studies, or other documentation [41, 42]. These may represent useful starting points for researchers attempting to ascertain which diagnostic codes are relevant for their goals. Although this section has focused primarily on ICD diagnostic codes, we believe the principles of broadening one’s understanding of the classification/coding systems and diligently considering how to query such values can generalize to other diagnostic and coding/classification systems (in other Canadian datasets) and in other administrative health datasets around the world.

Consideration 3: Transform data into meaningful forms

The raw nature of administrative data can pose challenges for individuals attempting to use these data for a research project. These pitfalls include data quality issues, a lack of adequate corroboration (e.g., diagnoses are sometimes preliminary, pending further confirmation), multiple billings/services associated with a single unique healthcare encounter, and mistakes during data entry (including, but not limited to, typographical errors), all of which can result in misclassification bias [43]. Hence, raw administrative data often require processing and cleaning before they are research-ready [44, 45].

Figure 2: Example of the nested hierarchical structure of International Classification of Diseases (ICD) diagnostic codes, 10th revision



Based on ICD data from: <https://github.com/holtzy/Visualizing-the-ICD10-Classification/> [73].

Effective data management and cleaning methods can allow more meaningful use of data for epidemiologic research. As an example, one method is processing data to ensure redundant data are not included in analysis (e.g., cancelation/no-show encounters, or healthcare practitioner encounters without direct patient contact). Depending on the analytic use of these data, one may desire to retain these data (as no-shows may incur billing or penalty fees) or one may opt to omit them (as they do not reflect patient-practitioner encounters) [46]. Other data processing considerations include how to handle apparently erroneous, ambiguous, or impossible events. Values that should follow a certain schema such as fixed categories or approved diagnostic codes but do not follow such systems should also be accounted for. As an example, the Canadian Primary Care Sentinel Surveillance Network applies various coding algorithms to healthcare practitioner billing records including the diagnostic codes, transforming the data into a standardized 'coded' version [18]. Another instance is discordant indication of a person's status (e.g., one source indicating a person had a death date, other sources indicating healthcare contact thereafter). This information can be examined that can be examined and reconciled appropriately as part of the data processing stage (e.g., discordance may indicate linkage errors). Some frameworks, examples, and recommendations for reporting data quality, cleaning data, and quality assurance exist [18, 47–49]. One example is the Manitoba Centre for Health Policy Data Quality Framework, containing five dimensions: accuracy, internal validity, external validity, timeliness, and interpretability [48]. Appropriately cleaning and processing data can therefore be an important step in ensuring administrative datasets are better prepared for analysis.

In addition to processing/cleaning administrative data, the dataset structure itself can also be a challenge needing addressing in order to appropriately identify events or characteristics. One example is the enumeration of hospitalization readmissions based on raw data from the standardized hospitalization database in Canada, which captures interhospital transfers in separate records. This could lead to misclassification of interhospital transfers as

readmissions, thus inflating this number [50]. An algorithm can be applied that combines planned transfers into a hospitalization episode of care, thus avoiding misclassification of planned transfers as unplanned readmissions (see Figure 3) [50, 51].

A related example of benefits conferred by transforming data is ascertainment of emergency department (ED) visits in BC, where they are captured in multiple partially-overlapping databases [52]. One can leverage ED-specific indications from three databases – remaining mindful of duplication across these sources [52]. These examples indicate scenarios where it is more appropriate to apply algorithms or some criteria to use a converted/transformed format of the data, rather than the raw format of the data. Thus, processing, cleaning, and applying algorithms/transformations to one's administrative data can provide a more effective, research-ready database and help mitigate potential information biases.

Consideration 4: Recognize the importance of validity when defining analytic variables

The building blocks of much administrative health data research are derived analytic variables, used to ascertain events, disease cases, or characteristics. Identification algorithms are definitions/criteria used to identify diseases, events, or characteristics – oftentimes classifying if a person has a disease/condition based on a pattern of healthcare use [53–55]. One example for identifying people with asthma would be requiring one asthma-related hospitalization or two physician encounters within a 24-month period [56]. This is typically how one defines variables that can be used to, say, define a study sample, define exposures, confounders, effect modifiers, and/or outcomes in a given study [43, 57]. The validity of inferences therefore depends on the appropriate use of identification algorithms for a given research question. Validity is often assumed but not always evidenced. Rather than being a static property of a measurement (i.e., a piece

Figure 3: Demonstration of how interhospital transfers can appear in raw hospital data and the concept of a hospitalization episode of care

An interhospital transfer produces multiple rows

Raw data structure

ID	Admit date	Discharge date	Hospital	Hospital FROM	Hospital TO
X	June 1	June 3	A		B
X	June 3	June 5	B	A	

2 hospitalization records

Episode concept

ID	Episode start	Episode end
X	June 1	June 5

1 hospitalization episode of care
(containing 2 records)

Note: This is a fictitious example of an interhospital transfer between two hospitals, based on data structure in the standard Canadian hospitalization record database.

of 'data'), validity can be more holistically viewed as an interaction between the measurement, the person, and the context where the data occur [58]. Indeed, the validity of a given assumption or inference can vary by characteristics of the healthcare provider, the person receiving the care, geography, and many other elements including time [59]. This is but one hypothetical example that can compound across thousands of healthcare practitioners, millions of patients, and tens of millions of healthcare interactions, within administrative health datasets. Since researchers need to define various concepts using administrative data, the derivation of these concepts is based on algorithms using information from healthcare system interactions (such as diagnostic codes). Hence, the validity of the inferences based on such algorithms depends on the underlying data generating mechanism (e.g., healthcare interactions, quality of data documented therein).

Threats to validity include issues of data quality and coverage, which may vary depending on the source. For instance, within most Canadian regions, virtually all acute inpatient hospitalizations will be captured via a standardized hospitalization database [60]. By contrast, in BC, healthcare practitioner billings (which include outpatient interactions) appear in two broad streams rather than a single source: fee-for-service (FFS) and alternate payment plan (APP). FFS billings are those provided in a model whereby there is a payment associated with each fee item, and each encounter. Thus, FFS-related healthcare practitioner encounters are, relative to APP records, well-captured in BC. Conversely, in the APP model, payment can take various forms including salaried and capitation. Unlike with FFS where there is a fee per each service provided, APP can, for example, reflect a fixed payment amount based on roster size and complexity. While the APP encounters that do appear in BC's main healthcare practitioner billings database are entered

for record-keeping purposes rather than for reimbursement of services (termed 'shadow billings') [61], not all APP practitioners have their encounters captured in this database, and coverage (of patient-practitioner encounters in the APP model) can vary by specialty [61]. This scenario exemplifies how healthcare administration payment models can impact coverage and quality of data that are generated through administering healthcare, thus potentially threatening the validity of inferences based on secondary uses of these data for epidemiologic research. Relatedly, misclassification of certain measures may occur if data availability varies across certain characteristics under comparison. For example, in Alberta, Canada, unspecified stroke diagnostic codes were more frequently used in rural hospitals than urban hospitals (perhaps due to lack of specialty equipment/expertise) – so ascertainment of an exposure or outcome based on these codes could result in misclassifying certain types of stroke [62]. In another example of geographic patterning of diagnosis and healthcare use data in Alberta, time to breast cancer diagnosis was found to vary substantially by health region of residence [63]; it was hypothesized that such regional variation could be due to lack of specialist healthcare resources such as radiologists in more rural regions. Therefore, it is important to understand whether varying healthcare coverage, over time, or across different geographies, may have an impact on validity when applying case-finding algorithms to define variables for health research.

Validity in the context of case-finding algorithms

A common challenge faced by researchers using administrative data for epidemiologic research is identifying if a person has

a condition, without a verified indication/reference standard. For example, one may be interested in identifying a prevalent asthma case in the absence of some confirmed indication (e.g., no confirmatory laboratory result). In this instance, one could apply a case-finding algorithm – a type of identification algorithm, with criteria often bound by a time constraint [64, 65] – to administrative health data to more validly ascertain a person as having a condition. Multiple events, such as physician encounters (usually in outpatient settings), reflect corroboration/confirmation of an initial diagnosis with subsequent confirmation(s) of that same diagnosis (hence, in the example, two asthma-related encounters within a 24-month period) [56]. Thus, case-finding algorithms – if supported by validity evidence and applied appropriately – can help produce a less biased ascertainment of people with a condition than using raw data. One caveat to note is that the types and quantity of diagnostic/intervention codes applied by service providers may be influenced by financial incentives, a phenomenon termed “code creep”, or “upcoding” particularly in the context of hospital care [66]. In particular, incentives may impact patterning of diagnostic codes evident in the data [67], and those diagnostic codes are often what are relied upon when using administrative data for research purposes.

Another key validation consideration for case-finding algorithms relates to amount of time to lookback when querying healthcare records. When applying algorithms to query a database, one must decide which period of time will be queried in the data. For instance, for ascertaining asthma prevalence one may search ‘X’ years prior to a given date. Persons who have lived in a region (e.g., the province of BC) for a longer time will naturally have more health data available, and therefore have a longer ‘lookback’ window than those with fewer years of data to query. Those with a longer available follow-up time will likely have more healthcare encounters and hence will be more likely to be ascertained as a prevalent case than persons with only a minimal follow-up time [68], a product of surveillance bias [11]. To help mitigate this bias, a fixed lookback window may be applied when ascertaining conditions to ensure all persons in an analysis have an equal level of data to query (e.g., a set window of 5 years). Further elaboration on impacts of lookback window length on prevalence and incidence is available elsewhere [68]. Careful consideration of the length of time to query when defining variables from administrative data is therefore an important validity consideration.

Evaluating validity evidence

Where possible, validated and context-relevant approaches should be implemented, given that diagnostic code versions, data availability, and healthcare use patterns can vary by region [19, 32]. Using case-finding algorithms that are applicable to the population of interest – for example in the same country, sub-national region, and clinical population, would be ideal. Validity evidence for case-finding algorithms is typically determined by agreement between algorithm-based classification of a case vs non-case and classification from a reference (“gold”) standard (e.g., confirmatory laboratory result for a clinical condition) [43, 57]. Two common metrics

yielded by this analysis are sensitivity (the proportion of ‘true’ cases identified by the algorithm) and specificity (the proportion of ‘true’ non-cases identified by the algorithm) [43]. Several other metrics can also be computed, including positive predictive values (PPV; the proportion of algorithm-classified cases that are ‘true’ cases) and negative predictive values (NPV; the proportion of algorithm-classified non-cases that are ‘true’ non-cases) [69]. Such predictive values are termed ‘prevalence-dependent’ metrics as they yield biased values unless the validation sub-sample has a disease prevalence comparable to the population in which the algorithm is intended to be applied [69]. Typically, there is a trade-off in algorithm performance – an inverse relationship between specificity and sensitivity. The research question should guide which metric(s) to prioritize. Prioritizing sensitivity is useful when ensuring a broad capture of a disease is important, whereas prioritizing specificity may be apt when the goal is minimizing false positives (see [64] for example). Although in this section we have focused on validation in the context of case-finding algorithms, it can be useful to note that validity extends beyond this specific aspect of research leveraging administrative health data. It is crucial to ensure valid approaches are employed throughout all aspects from data linkage [21, 22, 70], through to data processing [47, 48], to the statistical analysis of the data (including efforts to mitigate biases) [71, 72], and to the drafting, publication, and knowledge translation of the findings.

Conclusion

We have highlighted numerous considerations garnered through our experiences working with administrative health data in a Canadian province, considerations which we believe can help researchers better answer their research questions. The strengths and limitations of administrative data render it crucial to formulate research questions (with appropriate partnerships, collaborations, and expertise) that are suitable this paper, and beyond, can help guide such formulations. The set of topics discussed is of course not an exhaustive list, and readers should examine the extent to which these considerations may apply (or not) to their own contexts and datasets. We hope the information shared herein will be valuable for newcomers and experienced researchers alike who wish to ask reasonable questions of administrative health data repurposed for research.

Statement on conflicts of interest

None declared.

Ethics statement

As this paper is entirely conceptual, not based on any data involving individuals, no associated ethics approval was required/applicable.

Data availability statement

As this paper is entirely conceptual, not based on any data involving individuals, there are no data to be made available.

References

- Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. *Annu Rev Public Health* [Internet]. 2011 [cited 2023 Nov 10];32:91–108. <https://doi.org/10.1146/annurev-publhealth-031210-100700>
- Quan H, Smith M, Bartlett-Esquilant G, Johansen H, Tu K, Lix L, et al. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. *Can J Cardiol* [Internet]. 2012;28:152–4. <https://doi.org/10.1016/j.cjca.2012.01.005>
- McGrail KM, Jones K, Akbari A, Bennett TD, Boyd A, Carinci F, et al. A Position Statement on Population Data Science: The Science of Data about People. *Int J Popul Data Sci* [Internet]. 2018 [cited 2023 Nov 10];3:415. <https://doi.org/10.23889/ijpds.v3i1.415>
- Rockhold FW, Goldstein BA. Pragmatic Randomized Trials Using Claims or Electronic Health Record Data. In: Piantadosi S, Meinert CL, editors. *Princ Pract Clin Trials* [Internet]. Cham: Springer International Publishing; 2022 [cited 2023 Nov 10]. p. 2307–17. https://doi.org/10.1007/978-3-319-52636-2_270
- Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, et al. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. *BMJ Health Care Inform* [Internet]. 2020 [cited 2023 Nov 10];27:e100161. <https://doi.org/10.1136/bmjhci-2020-100161>
- Edwards J, Thind A, Stranges S, Chiu M, Anderson KK. Concordance between health administrative data and survey-derived diagnoses for mood and anxiety disorders. *Acta Psychiatr Scand* [Internet]. 2020;141:385–95. <https://doi.org/10.1111/acps.13143>
- Urbanoski K, Inglis D, Veldhuizen S. *Service Use and Unmet Needs for Substance Use and Mental Disorders in Canada*. Can J Psychiatry [Internet]. SAGE PublicationsSage CA: Los Angeles, CA; 2017 [cited 2024 Jul 24]; Available from: <https://journals.sagepub.com/doi/10.1177/0706743717714467>.
- Chiu M, Amartey A, Wang X, Kurdyak P. Ethnic differences in mental health status and service utilization: a population-based study in Ontario, Canada. *Can J Psychiatry* [Internet]. Sage Publications Sage CA: Los Angeles, CA; 2018;63:481–91. <https://doi.org/10.1177/0706743717741061>
- Ody C, Msall L, Dafny LS, Grabowski DC, Cutler DM. Decreases In Readmissions Credited To Medicare's Program To Reduce Hospital Readmissions Have Been Overstated. *Health Aff Proj Hope* [Internet]. 2019;38:36–43. <https://doi.org/10.1377/hlthaff.2018.05178>
- Sarrazin MSV, Rosenthal GE. Finding pure and simple truths with administrative data. *JAMA* [Internet]. 2012;307:1433–5. <https://doi.org/10.1001/jama.2012.404>
- Haut ER, Pronovost PJ, Schneider EB. Limitations of administrative databases. *JAMA* [Internet]. 2012;307:2589; author reply 2589–2590. <https://doi.org/10.1001/jama.2012.6626>
- Ioannidis JPA. Are mortality differences detected by administrative data reliable and actionable? *JAMA* [Internet]. 2013;309:1410–1. <https://doi.org/10.1001/jama.2013.3150>
- McGuckin T, Crick K, Myroniuk TW, Setchell B, Yeung RO, Campbell-Scherer D. Understanding challenges of using routinely collected health data to address clinical care gaps: a case study in Alberta, Canada. *BMJ Open Qual* [Internet]. 2022;11:e001491. <https://doi.org/10.1136/bmjopen-2021-001491>
- Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. *BMC Health Serv Res* [Internet]. 2017;17:766. <https://doi.org/10.1186/s12913-017-2697-y>
- Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. *Int J Popul Data Sci* [Internet]. 2023;8:2115. <https://doi.org/10.23889/ijpds.v8i1.2115>
- Iezzoni LI. Assessing Quality Using Administrative Data. *Ann Intern Med* [Internet]. American College of Physicians; 1997 [cited 2024 Jul 24];127:666–74. Available from: https://www.acpjournals.org/doi/abs/10.7326/0003-4819-127-8_part_2-199710151-00048.
- McKenna S, Nelson E, Maguire A. Data-driven research with historically excluded groups: Towards a model for co-production and democratisation. *Int J Popul Data Sci* [Internet]. 2023 [cited 2024 Aug 14];8. Available from: <https://ijpds.org/article/view/2261>.
- Morken R, Salman A, Herman C, Shah R, Wong S, Barber D. CPCSSN Data Quality: An opportunity for enhancing Canadian primary care data [Internet]. Canadian Primary Care Sentinel Surveillance Network; 2023. Available from: <https://cpcssn.ca/wp-content/uploads/2023/08/2023-CPCSSN-Report-Data-Quality-FINAL-June2111.pdf>.
- Hu W. Diagnostic codes in MSP claim data. Medical Services Plan - Program Monitoring and Information Management Branch, Resource Management Division; 1996.

20. British Columbia Ministry of Health. MSP service locations codes [Internet]. Available from: <https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/msp-service-locations-codes.pdf>.
21. Sakshaug JW. Measuring and Controlling for Non-Consent Bias in Linked Survey and Administrative Data. *Adm Rec Surv Methodol* [Internet]. John Wiley & Sons, Ltd; 2021 [cited 2024 Jul 19]. p. 155–78. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119272076.ch7>.
22. Harron K, Goldstein H, Dibben C. Methodological Developments in Data Linkage [Internet]. [cited 2024 Jul 19]. Available from: <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119072454>.
23. Warburton RN. Takeup of Income-Tested Health-Care Premium Subsidies: Evidence and Remedies for British Columbia. *Can Tax J* [Internet]. 2005;53:1. Available from: <https://heinonline.org/HOL/Page?handle=hein.journals/cdntj53&id=17&div=&collection=>.
24. Population Data BC. Population Data BC - Services For Researchers [Internet]. Available from: <https://www.popdata.bc.ca/researchers>.
25. Smith M, Turner K, Bond R, Kawakami T, Roos LL. The Concept Dictionary and Glossary at MCHP: Tools and Techniques to Support a Population Research Data Repository. *Int J Popul Data Sci* [Internet]. 2019 [cited 2023 Nov 10];4. <https://doi.org/10.23889/ijpds.v4i1.1124>
26. Canadian Institute for Health Information (CIHI). A Guide for Users of the Canadian Classification of Health Interventions (CCI) [Internet]. 2022. Available from: <https://www.cihi.ca/sites/default/files/document/guide-for-users-of-cci-manual-en.pdf>.
27. Jetté N, Quan H, Hemmelgarn B, Drosler S, Maass C, Moskal L, et al. The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Med Care* [Internet]. 2010;48:1105–10. <https://doi.org/10.1097/MLR.0b013e3181ef9d3e>.
28. Yu AXY, Holodinsky JK, Zerna C, Svenson LW, Jetté N, Quan H, et al. Use and Utility of Administrative Health Data for Stroke Research and Surveillance. *Stroke* [Internet]. 2016;47:1946–52. <https://doi.org/10.1161/STROKEAHA.116.012390>
29. Cartwright DJ. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Adv Wound Care* [Internet]. 2013 [cited 2023 Nov 10];2:588–92. <https://doi.org/10.1089/wound.2013.0478>
30. British Columbia Ministry of Health. Diagnostic Code Descriptions (ICD9) - Mental Disorders [Internet]. Available from: https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/diag-codes_mental.pdf.
31. Canadian Institute for Health Information. Hospital Stays for Harm Caused by Substance Use Appendices to Indicator Library [Internet]. 2022. Available from: <https://www.cihi.ca/sites/default/files/document/appendix-hospital-stays-for-harm-caused-by-substance-use-en-web.pdf>.
32. Hirsch JA, Nicola G, McGinty G, Liu RW, Barr RM, Chittle MD, et al. ICD-10: History and Context. *AJNR Am J Neuroradiol* [Internet]. 2016 [cited 2023 Nov 10];37:596–9. <https://doi.org/10.3174/ajnr.A4696>
33. Walker RL, Hennessy DA, Johansen H, Sambell C, Lix L, Quan H. Implementation of ICD-10 in Canada: how has it impacted coded hospital discharge data? *BMC Health Serv Res* [Internet]. 2012 [cited 2023 Nov 10];12:149. <https://doi.org/10.1186/1472-6963-12-149>
34. Statistics Canada. Comparability of ICD-10 and ICD-9 for Mortality Statistics in Canada [Internet]. 2005. Available from: <https://www150.statcan.gc.ca/n1/en/catalogue/84-548-X>.
35. Janjua NZ, Islam N, Kuo M, Yu A, Wong S, Butt ZA, et al. Identifying injection drug use and estimating population size of people who inject drugs using healthcare administrative datasets. *Int J Drug Policy* [Internet]. 2018;55:31–9. <https://doi.org/10.1016/j.drugpo.2018.02.001>
36. Complex Care Planning and Management Fees [Internet]. General Practice Services Committee; 2022. Available from: <https://fpscbc.ca/sites/default/files/GPSC-Complex-Care-Billing-Guide.pdf>.
37. British Columbia Ministry of Health. Additional Diagnostic codes (MSP) [Internet]. Available from: <https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/additional-diag-codes.pdf>
38. BC Centre for Disease Control. British Columbia Chronic Disease Registries (BCCDR) Case Definitions - Mood and Anxiety Disorders [Internet]. 2022. Available from: <http://www.bccdc.ca/resource-gallery/Documents/Chronic-Disease-Dashboard/mood-anxiety-disorders.pdf>.
39. Unger B, Afilalo M, Boivin JF, Bullard M, Grafstein E, Schull M, et al. Development of the Canadian Emergency Department Diagnosis Shortlist. *CJEM* [Internet]. 2010;12:311–9. <https://doi.org/10.1017/s1481803500012392>
40. Statistics Canada. Multiple causes of death. *Health Rep* [Internet]. 1997;9. Available from: <https://www150.statcan.gc.ca/n1/en/pub/82-003-x/1997002/article/3235-eng.pdf?st=HleRq6zJ>.
41. Lix LM, Ayles J, Bartholomew S, Cooke CA, Ellison J, Emond V, et al. The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance. *Int J Popul Data Sci* [Internet]. [cited 2024 Aug 13];3:433. <https://doi.org/10.23889/ijpds.v3i3.433>

42. Fiest KM, Jette N, Quan H, St Germaine-Smith C, Metcalfe A, Patten SB, et al. Systematic review and assessment of validated case definitions for depression in administrative data. *BMC Psychiatry* [Internet]. 2014;14:289. <https://doi.org/10.1186/s12888-014-0289-5>
43. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttman A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* [Internet]. 2011 [cited 2023 Nov 10];64:821–9. <https://doi.org/10.1016/j.jclinepi.2010.10.006>
44. Grath-Lone LM, Jay MA, Blackburn R, Gordon E, Zylbersztejn A, Wijlaars L, et al. What makes administrative data “research-ready”? A systematic review and thematic analysis of published literature. *Int J Popul Data Sci* [Internet]. [cited 2023 Nov 10];7:1718. <https://doi.org/10.23889/ijpds.v7i1.1718>
45. Tasker RC. Why everyone should care about “Computable Phenotypes.” *Pediatr Crit Care Med J Soc Crit Care Med World Fed Pediatr Intensive Crit Care Soc* [Internet]. 2017 [cited 2023 Nov 10];18:489–90. <https://doi.org/10.1097/PCC.0000000000001115>
46. Peterson S, Lavergne R, Morgan J, McGrail K. Defining type of contact in Medical Services Plan data in British Columbia [Internet]. UBC Centre for Health Services and Policy Research; 2021. Available from: <https://chspr.sites.olt.ubc.ca/files/2022/07/Type-of-Contact-2021.pdf>.
47. Tran DT, Havard A, Jorm LR. Data cleaning and management protocols for linked perinatal research data: a good practice example from the Smoking MUMS (Maternal Use of Medications and Safety) Study. *BMC Med Res Methodol* [Internet]. 2017 [cited 2024 Jul 26];17:97. <https://doi.org/10.1186/s12874-017-0385-6>
48. Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, et al. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *J Am Med Inform Assoc JAMIA* [Internet]. 2018;25:224–9. <https://doi.org/10.1093/jamia/ocx078>
49. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med* [Internet]. Public Library of Science; 2015 [cited 2023 Nov 10];12:e1001885. <https://doi.org/10.1371/journal.pmed.1001885>
50. Peng M, Li B, Southern DA, Eastwood CA, Quan H. Constructing Episodes of Inpatient Care: How to Define Hospital Transfer in Hospital Administrative Health Data? *Med Care* [Internet]. 2017;55:74–8. <https://doi.org/10.1097/MLR.0000000000000624>
51. Emerson S, McLinden T, Sereda P, Yonkman A, Trigg J, Barrios R, et al. Identifying hospitalization episodes of care among people with and without HIV in British Columbia, Canada. *Int J Popul Data Sci* [Internet]. 2024 [cited 2024 Oct 20];9. <https://doi.org/10.23889/ijpds.v9i5.2549>
52. Peterson S, Wickham M, Lavergne R, Beaumier J, Ahuja M, Mooney D, et al. Methods to comprehensively identify emergency department visits using administrative data in British Columbia [Internet]. UBC Centre for Health Services and Policy Research; 2021 [cited 2023 Nov 10]. Available from: <https://chspr.sites.olt.ubc.ca/files/2021/02/CHSPR-ED-Report-2021.pdf>.
53. Leung KM, Hasan AG, Rees KS, Parker RG, Legorreta AP. Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. *J Clin Epidemiol* [Internet]. 1999;52:57–64. [https://doi.org/10.1016/s0895-4356\(98\)00143-7](https://doi.org/10.1016/s0895-4356(98)00143-7)
54. Kim JY, Lee K-J, Kang J, Kim BJ, Han M-K, Kim S-E, et al. Development of stroke identification algorithm for claims data using the multicenter stroke registry database. *PloS One* [Internet]. 2020;15:e0228997. <https://doi.org/10.1371/journal.pone.0228997>
55. Benchimol EI, Guttman A, Mack DR, Nguyen GC, Marshall JK, Gregor JC, et al. Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada. *J Clin Epidemiol* [Internet]. 2014;67:887–96. <https://doi.org/10.1016/j.jclinepi.2014.02.019>
56. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying patients with physician-diagnosed asthma in health administrative databases. *Can Respir J* [Internet]. 2009;16:183–8. <https://doi.org/10.1155/2009/963098>
57. Ehrenstein V, Petersen I, Smeeth L, Jick SS, Benchimol EI, Ludvigsson JF, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. *Clin Epidemiol* [Internet]. 2016;8:49–51. <https://doi.org/10.2147/CLEP.S104448>
58. Zumbo BD. Trending away from routine procedures, toward an Ecologically Informed In Vivo View of Validation Practices. *Meas Interdiscip Res Perspect* [Internet]. Routledge; 2017 [cited 2023 Nov 10];15:137–9. <https://doi.org/10.1080/15366367.2017.1404367>
59. Shiff NJ, Jama S, Boden C, Lix LM. Validation of administrative health data for the pediatric population: a scoping review. *BMC Health Serv Res* [Internet]. 2014 [cited 2023 Nov 10];14:236. <https://doi.org/10.1186/1472-6963-14-236>
60. Butler AL, Smith M, Jones W, Adair CE, Vigod SN, Lesage A, et al. Multi-province epidemiological research using linked administrative data: a case study from Canada. *Int J Popul Data Sci* [Internet]. 2018 [cited 2022 Oct 11];3:443. <https://doi.org/10.23889/ijpds.v3i3.443>
61. Cunningham CT, Cai P, Topps D, Svenson LW, Jetté N, Quan H. Mining rich health data from Canadian physician

- claims: features and face validity. *BMC Res Notes* [Internet]. 2014;7:682. <https://doi.org/10.1186/1756-0500-7-682>
62. Yiannakoulis N, Svenson LW, Hill MD, Schopflocher DP, James RC, Wielgosz AT, et al. Regional comparisons of inpatient and outpatient patterns of cerebrovascular disease diagnosis in the province of Alberta. *Chronic Dis Can*. 2003;24:9–16.
 63. Yuan Y, Li M, Yang J, Elliot T, Dabbs K, Dickinson JA, et al. Factors related to breast cancer detection mode and time to diagnosis in Alberta, Canada: a population-based retrospective cohort study. *BMC Health Serv Res* [Internet]. 2016 [cited 2024 Jul 31];16:65. <https://doi.org/10.1186/s12913-016-1303-z>
 64. Emerson SD, McLinden T, Sereda P, Lima VD, Hogg RS, Kooij KW, et al. Identification of people with low prevalence diseases in administrative healthcare records: A case study of HIV in British Columbia, Canada. *PloS One* [Internet]. 2023;18:e0290777. <https://doi.org/10.1371/journal.pone.0290777>
 65. Widdifield J, Ivers NM, Young J, Green D, Jaakkimainen L, Butt DA, et al. Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario, Canada. *Mult Scler Houndmills Basingstoke Engl* [Internet]. 2015;21:1045–54. <https://doi.org/10.1177/1352458514556303>
 66. Miller A. Medical fraud north of the 49th. *CMAJ Can Med Assoc J* [Internet]. 2013 [cited 2023 Nov 10];185:E31–3. <https://doi.org/10.1503/cmaj.109-4358>
 67. Simborg DW. DRG Creep. *N Engl J Med* [Internet]. Massachusetts Medical Society; 1981 [cited 2023 Nov 10];304:1602–4. <https://doi.org/10.1056/NEJM198106253042611>
 68. Nanditha NGA, Dong X, McLinden T, Sereda P, Kopec J, Hogg RS, et al. The impact of lookback windows on the prevalence and incidence of chronic diseases among people living with HIV: an exploration in administrative health data in Canada. *BMC Med Res Methodol* [Internet]. 2022;22:1. <https://doi.org/10.1186/s12874-021-01448-x>
 69. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008;56:45–50.
 70. Setoguchi S, Zhu Y, Jalbert JJ, Williams LA, Chen C-Y. Validity of deterministic record linkage using multiple indirect personal identifiers: linking a large registry to claims data. *Circ Cardiovasc Qual Outcomes* [Internet]. 2014;7:475–80. <https://doi.org/10.1161/CIRCOUTCOMES.113.000294>
 71. Matthay EC, Glymour MM. A Graphical Catalog of Threats to Validity: Linking Social Science with Epidemiology. *Epidemiology* [Internet]. 2020 [cited 2024 Jul 26];31:376. <https://doi.org/10.1097/EDE.0000000000001161>
 72. Shaw RJ, Harron KL, Pescarini JM, Pinto Junior EP, Allik M, Siroky AN, et al. Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses. *Eur J Epidemiol* [Internet]. 2022 [cited 2024 Jul 29];37:1215–24. <https://doi.org/10.1007/s10654-022-00934-w>
 73. Yan Holtz. Visualizing the ICD10 classification [Internet]. 2018. Available from: <https://github.com/holtzy/Visualizing-the-ICD10-Classification>

Abbreviations

APP:	Alternate payment plan
BC:	British Columbia
CED-DxS:	Canadian Emergency Department Diagnosis Shortlist
ED:	Emergency department
FFS:	Fee-for-service
ICD:	International Classifications of Diseases
MSP:	Medical Services Plan Payment Information File
WHO:	World Health Organization

