International Journal of Population Data Science

Journal Website: www.ijpds.org

Generating synthetic data from administrative health records for drug safety and effectiveness studies

Olawale F. Ayilara[1,*], Robert W. Platt[2], Matt Dahl[3], Janie Coulombe[4], Pablo Gonzalez Ginestet[5], Dan Chateau[6], and Lisa M. Lix[1]

Submission History

Submitted:	13/06/2023
Accepted:	30/10/2023
Published:	27/11/2023

[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, Canada
[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada
[3]Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, Canada
[4]Department of Mathematics and Statistics, Université de Montréal, Montreal, Canada
[5]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden
[6]College of Health & Medicine, Australian National University, Canberra, Australia

Abstract

Introduction

Administrative health records (AHRs) are used to conduct population-based post-market drug safety and comparative effectiveness studies to inform healthcare decision making. However, the cost of data extraction, and the challenges associated with privacy and securing approvals can make it challenging for researchers to conduct methodological research in a timely manner using real data. Generating synthetic AHRs that reasonably represent the real-world data are beneficial for developing analytic methods and training analysts to rapidly implement study protocols. We generated synthetic AHRs using two methods and compared these synthetic AHRs to real-world AHRs. We described the challenges associated with using synthetic AHRs for real-world study.

Methods

The real-world AHRs comprised prescription drug records for individuals with healthcare insurance coverage in the Population Research Data Repository (PRDR) from Manitoba, Canada for the 10-year period from 2008 to 2017. Synthetic data were generated using the Observational Medical Dataset Simulator II (OSIM2) and a modification (ModOSIM). Synthetic and real-world data were described using frequencies and percentages. Agreement of prescription drug use measures in PRDR, OSIM2 and ModOSIM was estimated with the concordance coefficient.

Results

The PRDR cohort included 169,586,633 drug records and 1,395 drug types for 1,604,734 individuals. Synthetic data for 1,000,000 individuals were generated using OSIM2 and ModOSIM. Sex and age group distributions were similar in the real-world and synthetic AHRs. However, there were significant differences in the number of drug records and number of unique drugs per person for OSIM2 and ModOSIM when compared with PRDR. For the average number of days of drug use, concordance with the PRDR was 16% (95% confidence interval [CI]: 12%–19%) for OSIM2 and 88% (95% CI: 87%-90%) for ModOSIM.

Conclusions

ModOSIM data were more similar to PRDR than OSIM2 data on many measures. Synthetic AHRs consistent with those found in real-world settings can be generated using ModOSIM. Synthetic data will benefit rapid implementation of methodological studies and data analyst training.

Keywords

administrative health records; computer simulation; prescription drug records; validation

*Corresponding Author:
Email Address: olawale.ayilara@umanitoba.ca (Olawale F. Ayilara)

https://doi.org/10.23889/ijpds.v8i1.2176

Introduction

Administrative health records (AHRs), which are generated primarily for healthcare management and billing are also used for prescription drug safety and comparative effectiveness studies [1–4]. However, the cost of data extraction, and the challenges associated with securing ethics and data access approvals due to health privacy legislation, patient privacy and confidentiality issues can make it difficult for researchers to conduct methodological research in a timely manner using AHRs, and to train analysts to conduct drug safety and comparative effectiveness studies.

The development of analytic methods can benefit from the availability of synthetic (i.e., artificial) data, which do not require ethical approvals and data access permissions. Synthetic data are generated to preserve some of the statistical attributes of the original data sources without violating patient privacy and confidentiality issues. Availability of synthetic AHRs for developing new study designs and methods, and statistical programming codes to implement study protocols can facilitate timely completion of research.

Methods for generating synthetic data can broadly be classified as data-driven [5] and process driven [6]. Process-driven methods, including Monte Carlo and discrete-event simulations, generate synthetic data from computational or mathematical models of an underlying physical process. Data-driven methods, such as joint probability distribution, plasmode [7] and imputation-based methods [8, 9], generate synthetic data from generative models that use the observed data.

The Observational Medical Outcomes Partnership (OMOP) developed a simulation program to generate AHR data that incorporate the complex relationships among health conditions and prescription drug use [10]. The original version of the simulator, known as the Observational Medical Dataset Simulator (OSIM), employed a sequence of user-defined probability tables to model population demographic characteristics and prevalence distributions of prescription drugs and selected health conditions. However, the initial version of the simulator was inadequate to model certain characteristics of real-world data, such as the relationships among health conditions and drugs [11]. This limitation led to the development of OSIM2.

OSIM2 is an empirical simulation model of longitudinal patient data, which incorporates additional complexities observed in real-world AHR data. These include a depiction of the relationships between prescription drugs and health conditions. OSIM2 creates simulated data containing fictitious individuals with records of their health conditions and prescription drugs based on the characteristics of real-world data in the OMOP common data model (CDM). The OMOP CDM transforms data from different databases into a common format [10–12].

OSIM2 was initially applied to AHRs from the Regie de l'assurance maladie du Quebec (RAMQ) by the Canadian Network of Observational Drug Effect Studies (CNODES), a collaborating centre of the Drug Safety and Effectiveness Network (DSEN) [13]. The results of the simulation showed that OSIM2 adequately modeled the baseline characteristics of the RAMQ database. However, the model did not accurately capture the number of drug records and unique drugs of the target data set. This limitation led to the development of the modified OSIM2 (ModOSIM) simulation model. Specifically, ModOSIM aims to preserve the structure of OSIM2 with a number of modifications to the probability tables for simulating the drug stage, and a pre-processed format applied to the drug prescription information. This manuscript is the first study to our knowledge that aims to provide empirical evidence about the representativeness of synthetic data generated using the OSIM2 and ModOSIM models.

The overall purpose of this study was to generate AHRs using both the OSIM2 and ModOSIM models. Our objectives were to: (1) compare the representativeness of the synthetic data generated from both models to the real-world AHR data, and (2) describe the challenges associated with using synthetic data for real-world study.

Methods

Synthetic data generation methods

Observational medical dataset simulator II (OSIM2)

The OSIM2 simulator uses a Monte Carlo approach, whereby information for an individual is generated randomly from the empirical multinomial distributions of the analysis module. This module extracts individual information as probability distributions from the source data [11]. The individual simulation model process has four stages: (i) creating a simulated population with associated person-level demographic information and periods of observation; (ii) generating a set of underlying health conditions for each individual; (iii) assigning treatments (i.e., drug exposure) to the cohort members based on their underlying health conditions, and (iv) introducing associations between treatments and outcomes. A detailed description of these stages is described by Murray et al. (2011) [11].

Modified observational medical dataset simulator II (ModOSIM)

The ModOSIM model incorporates a number of modifications to the data pre-processing stage and the probability tables of the OSIM2 simulator. In the data pre-processing phase, ModOSIM dropped the OSIM2 grouping rule, which combines two or more prescription drug records of the same drug together into a "drug era" if the gap between the end date of the previous record and the start date of the following record is less than a 30-day window. Instead, ModOSIM used the database of the prescription drug information directly, which implies that each drug prescription record is a drug era in itself.

In addition, out of the four probability tables that govern the distribution of the different aspects of the drug stage in OSIM2, ModOSIM retained three tables and modified the fourth probability table. The three probability tables retained were: (i) a probability table for generating the total number of unique prescription drugs an individual should have for a given health condition, (ii) a probability table for generating the transition days from the given health condition to the first occurrence of each unique prescription drug, and (iii) a probability table for generating the number of re-occurrences for each unique prescription drug.

The fourth probability table, which generates the total number of days of each unique prescription drug exposure, from the initial to the final re-occurrence of that drug exposure was modified as follows: (i) the total number of days of exposure to the prescription drug was removed from the probability table, and, (ii) the total number of days of exposure associated with the first and last re-occurrence of the prescription drug was discarded. Instead, ModOSIM simulates the pair of number of days of exposure and the number of re-occurrences of that exposure for each unique prescription drug, until the sum of the latter is equal to the number of re-occurrences for each unique prescription drug of that drug exposure.

Illustrative example

In this section, we describe the real-world AHR data source, study variables and statistical analyses to compare the representativeness of the synthetic data and the challenges associated with using synthetic data for real-world study.

Real-World AHR data and study cohort

The Population Research Data Repository (PRDR) housed at the Manitoba Centre for Health Policy (MCHP), a research unit at the University of Manitoba in the province of Manitoba, Canada, was used as the real-world source of AHR data for this study. The province has a population of approximately 1.3 million, according to the Statistics Canada Census, and universal healthcare. The PRDR includes Drug Program Information Network (DPIN) records, hospital discharge abstracts, and physician billing claims for all individuals eligible to receive health services (Table 1). The population registry contains information for all residents registered under the Health Services Insurance Plan, including healthcare coverage start and end dates, demographic characteristics, and postal code of residence.

The DPIN is an electronic, online, point-of-sale database that contains accurate and comprehensive information about prescriptions filled by community pharmacies [14, 15]. Each approved drug is assigned a Drug Identification Number (DIN) by Health Canada; DINs can be linked to the World Health Organization's Anatomical Therapeutic Chemical (ATC) codes [16].

Hospital discharge abstracts contain records of demographic and clinical information of discharges from acute care facilities. Each abstract captures up to 25 diagnosis codes that use the World Health Organization's International Classification of Diseases (ICD), 10th revision, Canadian version (ICD-10-CA) [17], as well as up to 20 procedure codes.

Physician billing claims are submitted by fee-for-service physicians to the ministry of health for provider remuneration. Each claim includes a three-digit ICD-9-CM (Clinical Modification) diagnosis code that corresponds with the reason for the physician visit.

The study cohort comprised all individuals with health insurance coverage at any point between 1st April 2008 and 31st March , 2017. All prescription drug records, hospital discharge abstracts, and physician billing claims for this cohort were extracted from the PRDR.

Study variables from the real-world AHRs and synthetic data

Demographic information (i.e., age group, sex) were extracted from the PRDR population registry for the study cohort. Per-person measures of prescription drug use were calculated from DPIN data; measures including the number of prescription drug records, number of unique prescription drugs, and total number of days of prescription drug use, were produced. For the latter measure, all prescription drugs taken concurrently on the same day contributed to one day of use.

Selected health conditions were identified from diagnoses in hospital records and physician billing claims. Specifically, asthma, chronic obstructive pulmonary disease, diabetes mellitus, heart failure, myocardial infarction, ischemic stroke, cancer, and dementia were selected as example conditions [18–20].

To address our second objective, which describes some of the challenges associated with using synthetic data for real-world study, we used a study cohort described in one of the CNODES studies [21]. This study assessed the association between the risk of hospitalisation for community-acquired pneumonia (HCAP) and the use of proton pump inhibitors (PPI) in Canada. The diagnosis codes used to ascertain the health conditions are reported in Table 2.

Statistical analysis

Descriptive statistics, including means, standard deviations, and percentages, were used to describe the real-world data from the PRDR and the synthetic data produced using the OSIM2 and ModOSIM models for the study cohorts. Frequency distributions were used to compare demographic characteristics, number of unique diagnosis codes, number of unique ATC codes, selected health conditions, selected medication use, and number of unique drugs in each of the PRDR, OSIM2 and ModOSIM.

The concordance correlation coefficient with 95% confidence intervals (95% CIs) was estimated for the average number of days of drug use from the PRDR and each of the OSIM2 and ModOSIM models; this coefficient captures information on both precision and accuracy and evaluates the degree to which pairs of observations fall on the 45-degree line through the origin of a scatterplot of two variables [22]. Scatterplots of the average number of days of drug use for the three data sources were also produced. A 2x2 table to show the association between the use of PPIs and the risk of HCAP in the real-world data, OSIM2 and ModOSIM data was provided.

The metrics we considered in comparing the synthetic and the original data cover all the dimensions for evaluating administrative data quality [6, 23]. All analyses were conducted using SAS version 9.4 (SAS Institute, Cary, NC).

Results

The PRDR study cohort comprised of 169,586,633 drug records and 1,395 unique prescription drugs for 1,604,734 individuals with healthcare coverage at any point in a 10-year period between 1st April, 2008 and 31st March , 2017. For each of the OSIM2 and ModOSIM simulation models, synthetic data were generated for a total of 1,000,000 individuals.

Table 1: Attributes in synthetic data (OSIM2 and ModOSIM) compared with real-world data (AHRs)

Database	**Real-world data**		**OSIM2 and ModOSIM**	
	Attributes	**Role**	**Attributes**	**Limitations**
Health Insurance Registry	• De-identified PHIN • Sex • Age • Date of birth • Registry cancellation code • Health coverage start date • Health coverage end date	• To develop observation and person tables for the CDM. The observation table identifies the start and end dates of healthcare coverage, and the person table contains demographic information for each individual in the observation table.	• Identification number • Sex • Age • Date of birth • Health coverage start date • Health coverage end date	• The health coverage dates do not respect real-time ordering.
Hospital Discharge Abstracts	• De-identified PHIN • Date of hospital admission • Hospital separation date • Diagnosis code • Type of diagnosis • Transaction code • ICD9/10 • ICD9, 3 digits	• To create a condition era table for the CDM.	• Identification number • ICD9, 3 digits • Date	• Hospital discharge abstracts and physician billing claims formed a database for diagnosis. • The attribute "Date" in the synthetic data does not indicate dates of hospital admission or separation.
Physician Billing Claims	• De-identified PHIN • Diagnosis code • Number of services • Date of service	• To create a condition era table for the CDM.		
Drug Prescription Information Network	• De-identified PHIN • Drug identification number • Days of prescription drug supply • Metric quantity claim • Dispensing date • ATC code	• To build a drug era table for the CDM. This table contains all the prescription dispensation records for each individual in the observation table.	• Identification number • Duration of prescription drug supply • Date provided • ATC code	• Drug identification number is not available in the synthetic datasets.

PHIN: Personal Health Identification Number; ICD: International Classification of Diseases; ATC: Anatomical Therapeutic Chemical; CDM: Common Data Model.

Tables 3, 4 and 5 show the descriptive information for the demographics, prescription drug use measures and health conditions, respectively. Overall, the study cohort from the real-world AHR data were almost equally split in terms of male and female representation. The majority of the cohort was in the under 10 years age group (20.8%); 14.6% were 60 years or older. Over the 10-year period, slightly more than half (52.4%) of the cohort had less than 10 different prescription drugs; the

Table 2: Diagnosis codes used to define selected health conditions and medications in the real-world and synthetic data

Condition	Diagnosis codes		ATC Code
	ICD-9-CM	ICD-10-CA	
Diabetes Mellitus	250.x	E10.x–E14.x	
Cancer, excluding non-melanoma skin cancer	140.x–172.x 174.x	C00.x–C43.x C45.x–C97.x F00.0–F00.2, F00.9, F01.0-F01.3, F01.8, F01.9, F02.0–F02.4, F02.8, F03.x	
Asthma	493.x	J45 J46	
Heart Failure	402.01 402.11 402.91 404.01 404.03 404.11 404.13 404.91 404.93 428.x	I11.0 I13.0 I13.2 I50.x	
Myocardial Infarction	410.x	I21.x	
Ischemic Stroke	433.x 434.x	I63.x I64.x	
Dementia	290.0 331.0 331.2 797.x	F05.1 F06.5 F06.6 F06.8 F06.9 F09.x G30.0 G30.1 G30.8 G30.9 G31.0 G31.1 R45.x	
Drug			
Acetylsalicylic Acid			B01AC06, B01AC30, C08CA55, C10BX02, M03BC51, N02AA79
DPP-4 Inhibitors			A10BH01, A10BH03, A10BH04, A10BH05, A10BD07, A10BD09, A10BD10, A10BD11, A10BD13, A10BD19, A10BD21
Insulin			A10BD01, A10AB02, A10AB03, A10AB04, A10AB05, A10AB06, A10AB30, A10AC01, A10AC02, A10AC03, A10AC04, A10AC30, A10AD01, A10AD02, A10AD03, A10AD04, AD10AD05, A10AD06, A10AD30, A10AE01, A10AE02, A10AE03, A10AE04, A10AE05, A10AE06, A10AE30, A10AE54, A10AE56, A10AF01

Continued

Table 2: Continued

Condition	**Diagnosis codes**		**ATC Code**
	ICD-9-CM	**ICD-10-CA**	
SGLT2 Inhibitors			A10BK01, A10BK02, A10BK03, A10BK04, A10BD15, A10BD16, A10BD19, A10BD20, A10BD21
Statins			C10AA01, C10AA02, C10AA03, C10AA04, C10AA05, C10AA06, C10AA07, C10AA08, C10BA01, C10BA02, C10BA03, C10BA04, C10BA05, C10BA06, C10BX01, C10BX02, C10BX03, C10BX04, C10BX05, C10BX06, C10BX07, C10BX08, C10BX10, C10BX11, C10BX12, C10BX13, C10BX14, C10BX15

Note: ICD-9-CM is the International Classification of Diseases, 9th revision, Clinical Modification; ICD-10-CA is the International Classification of Diseases, 10th revision, Canadian version.

Table 3: Percentages of demographic characteristics for the real-world AHR data and the OSIM2 and ModOSIM synthetic data

Characteristic	**Real-world data (N = 1,604,734)**	**OSIM2 (N = 1,000,000)**	**ModOSIM (N = 1,000,000)**
Sex			
Female	49.7	49.9	50.1
Male	50.3	50.1	49.1
Age, years			
Less than 10	20.8	21.0	21.0
10–19	12.9	13.1	13.0
20–29	14.5	14.2	14.2
30–39	13.3	13.1	13.2
40–49	13.1	13.1	13.1
50–59	10.8	10.9	10.9
60–69	6.9	7.0	6.9
70–79	4.4	4.4	4.4
80–89	2.7	2.7	2.7
90+	0.6	0.6	0.6
Mean (SD)	32.7 (23.5)	32.7 (23.6)	32.7 (23.6)
Year of birth			
Before 1922	1.2	0.8	0.8
1922–1927	1.8	1.8	1.8
1928–1945	9.2	9.4	9.4
1946–1954	8.8	8.6	8.6
1955–1964	12.5	12.6	12.6
1965–1980	20.0	20.4	20.4
1981–1996	23.2	22.7	22.6
1997–2017	23.3	23.8	23.8

OSIM2: Observational Medical Dataset Simulator II; ModOSIM: Modified OSIM2.

average was 12.1. Three quarters of the cohort had less than 100 total drug records; the average number of drug records was 113.5. The percentages of individuals with diagnosed diabetes mellitus, cancer and dementia in the cohort were 10.8%, 7.1% and 2.8%, respectively.

The study cohorts from the OSIM2 and ModOSIM data sources had a similar distribution of males and females when compared to the study cohort defined from the real-world AHR data. Moreover, the percentages of males and females were similar in both OSIM2 and ModOSIM data sources. The

Table 4: Percentages for drug characteristics in real-world AHR data and OSIM2 and ModOSIM synthetic data

Characteristic	**Real-world data**	**OSIM2**	**ModOSIM**
Number of unique drugs per person, (%)			
Less than 5	26.7	40.2	40.8
5–9	25.6	29.4	30.3
10–14	17.2	15.3	15.4
15–19	11.2	7.0	7.0
20+	19.3	8.1	6.5
Mean (SD)	12.1 (11.0)	7.7 (7.1)	7.6 (6.5)
Number of drug records per person, (%)			
0–19	45.5	62.6	48.3
20–39	13.4	15.1	13.8
40–59	7.7	7.8	8.5
60–79	5.2	4.8	5.9
80–99	3.8	3.1	4.4
100+	24.6	6.7	19.0
Mean (SD)	113.5 (313.2)	28.1 (40.6)	60.5 (94.5)
Duration of drug use, (%)			
1–10	31.3	35.1	39.8
11–20	8.3	13.8	9.3
21–30	41.9	12.2	38.6
>30	18.4	38.9	12.3
Mean (SD)	29.4 (26.3)	41.3 (79.7)	24.1 (22.4)
Prevalence of prescription drug use, (%)			
Acetylsalicylic Acid	6.8	7.1	5.9
DPP-4 inhibitors	0.9	0.3	0.2
Insulin	2.3	2.8	2.4
SGLT2 inhibitors	0.6	0.1	0.1
Statins	14.2	7.7	6.7

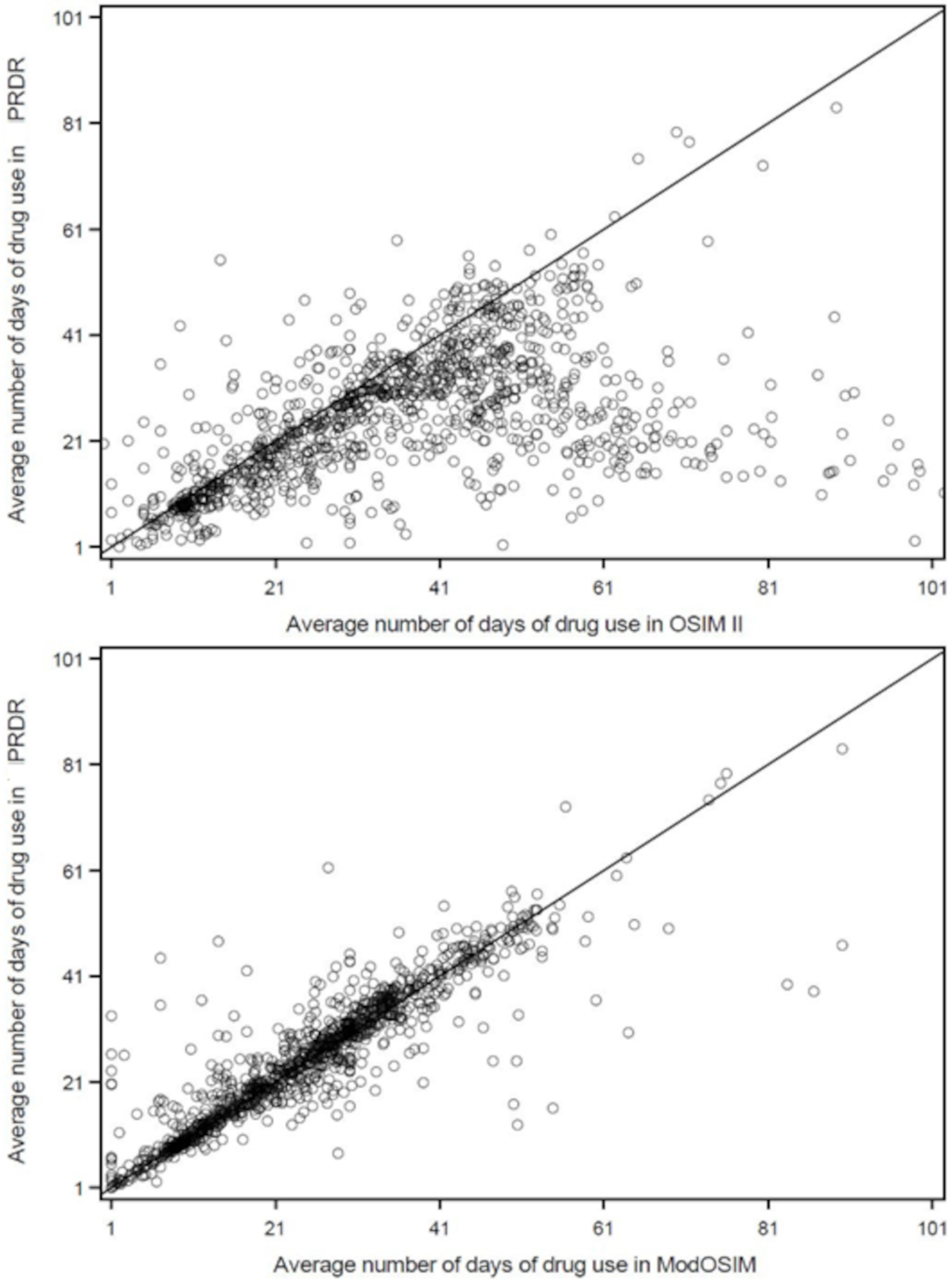
OSIM2: Observational Medical Dataset Simulator II; ModOSIM: Modified OSIM2.

Table 5: Percentages of health condition characteristics for the real-world AHR data and OSIM2 and ModOSIM synthetic data

Characteristic	**Real-world**	**OSIM2**	**ModOSIM**
Number of unique health conditions per person			
Less than 10	32.6	42.2	42.0
10–19	34.1	34.5	32.5
20–29	19.5	15.4	15.5
30–39	8.8	6.7	6.7
40–49	3.3	2.4	2.4
50–59	1.1	0.7	0.7
60+	0.5	0.2	0.2
Mean (SD)	16.5 (11.9)	13.9 (11.1)	13.9 (11.1)
Health condition			
Asthma	12.4	9.7	9.8
Chronic obstructive pulmonary disease (COPD)	14.3	14.1	14.1
Diabetes mellitus	10.8	9.0	9.0
Heart failure	3.6	3.9	3.9
Myocardial infarction	1.5	1.4	1.3
Ischemic stroke	2.6	3.1	3.0
Cancer (excluding non-melanoma skin cancer)	7.1	10.6	10.6
Dementia	2.8	3.3	3.3

OSIM2: Observational Medical Dataset Simulator II; ModOSIM: Modified OSIM2.

Figure 1: Scatter plots for the average number of days of prescription drug use in real-world administrative health data from the Population Research Data Repository (PRDR) and synthetic data from the OSIM2 (top) and ModOSIM (bottom) models



same was true for most age groups, with the exception of the 20-29 years and 50-59 years age groups. The percentages of individuals with diagnosed diabetes mellitus, cancer, and dementia, respectively, were similar in OSIM2 (9.2%, 10.6%, and 3.3%) and ModOSIM (9.0%, 10.6%, and 3.3%).

The numbers of drug records and unique drugs were different in OSIM2 and ModOSIM when compared with PRDR. However, the percentage of drug records in each category was similar for ModOSIM and PRDR. For OSIM2, the percentages of the numbers of drug records were most often higher than when compared to PRDR.

The estimated concordance correlation for the average number of days of drug use for PRDR and OSIM2 was 0.16 (95% CI: 0.12 - 0.19). For PRDR and ModOSIM the estimated concordance correlation was 0.88 (95% CI: 0.87 - 0.90). Scatter plots (Figure 1) revealed a strong linear relationship between the number of days for the PRDR and ModOSIM and a weak relationship for PRDR and OSIM2.

Information on the association between the number of prescription drugs and each of the demographic variables (i.e., sex and age group) in the real-world and synthetic data are presented in Figures 2 and 3, respectively. The frequency distribution of demographic variables and number of prescription drugs generated from ModOSIM was similar to the distributions from the real-world data. However, for the OSIM2 data there were substantial differences. In the ModOSIM cohort, two-thirds of individuals had fewer than 100 drug records, while more than 90% had fewer than 100 drug records in OSIM2 cohort, which is an overestimation of the 75.6% in the real-world data.

Table 6 reveals the association between the use of PPIs and the risk of HCAP in the real-world data, OSIM2 and ModOSIM data. The fraction (i.e., number of individual/total number) associated with the use of PPI were 2185/45215 (4.83%), 40/6595 (0.61%), and 18/6225 (0.30%) in the PRDR, OSIM2 and ModOSIM data, respectively. The number of individuals that use PPIs and at risk of HCAP was zero in the synthetic data.

Discussion

This study compared the attributes of real-world AHR data from one Canadian province to the data derived from two models for simulating AHR data. The OSIM2 model developed by OMOP was designed to overcome some of the limitations identified with the original simulation model, OSIM. The ModOSIM model was created by CNODES to address some of the limitations of OSIM2. The results of this study showed that synthetic data generated using ModOSIM model were more similar to real-world AHR data than OSIM2 on many attributes, which reflect the modifications incorporated in the ModOSIM simulator. The results of the association between the use of PPIs and the risk of HCAP from the synthetic data generated with OSIM2 and ModOSIM models showed a quasi-separation, which occurs when the outcome variable separates the exposure variable or a combination of exposure variables to a certain degree. Quasi-separation can result in large standard errors and infinite parameter estimates, making it difficult to draw meaningful conclusions from the statistical model. However, there are a number of techniques to address quasi-separation including penalized or regularization methods. This is one of the limitations associated with using synthetic dataset for real-world study.

Synthetic data are a valuable and important resource to gain hands-on experience in data exploration, transformation and validation. In observational studies, such as drug safety and comparative effectiveness studies, synthetic data can be used to train analysts in several areas. These include model development to address confounding using propensity scores models, and implementation of the sequence of steps required to complete a study protocol.

In Canada, each provinces/territories have its own health privacy legislation as well as its own process for data access approvals. No individual-level data are allowed to leave any of the provinces or territories; linked data can only be shared in aggregate form. Consequently, synthetic data are beneficial to: (i) test methods for conducting drug safety and effectiveness

Figure 2: Distribution of number of prescription drug records, stratified by sex, in real-world administrative health data from the Population Research Data Repository (PRDR) and synthetic data from the OSIM2 and ModOSIM models

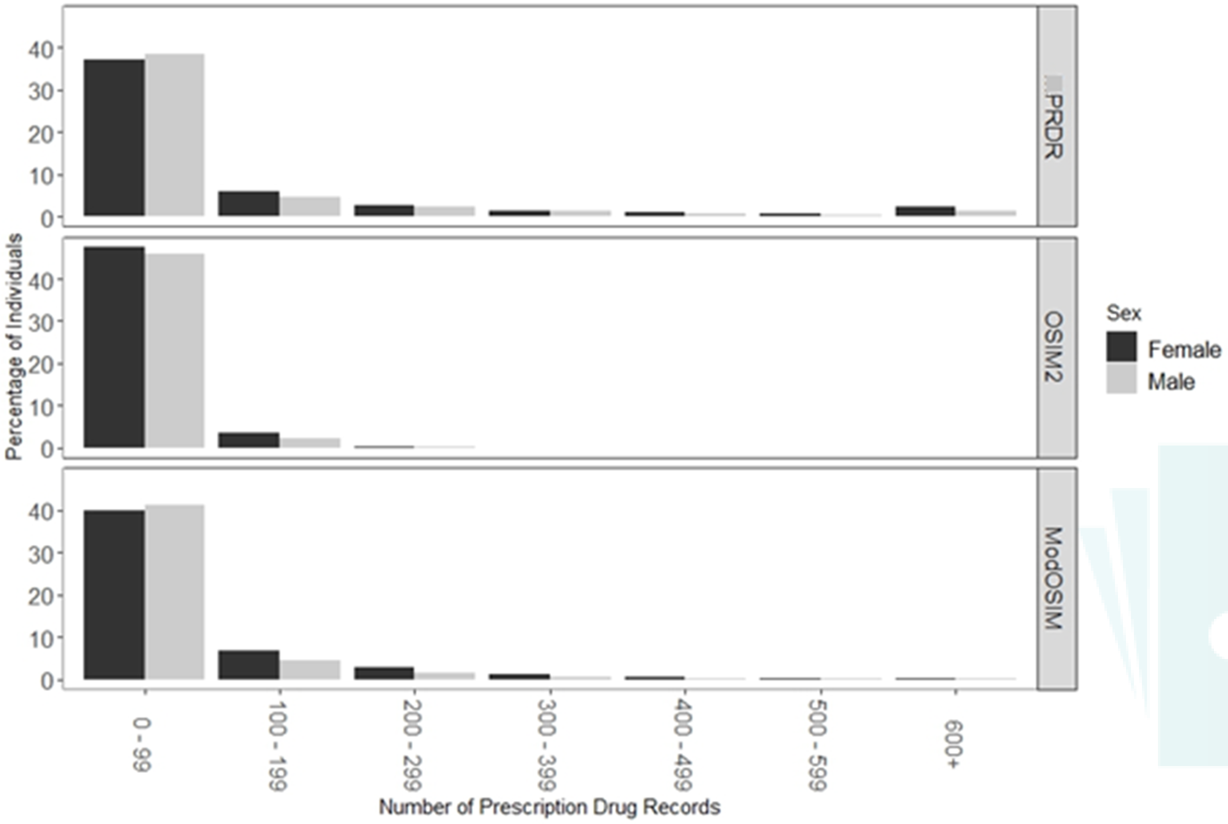


Figure 3: Distribution of number of prescription drug records, stratified by age group, in real-world administrative health data from the Population Research Data Repository (PRDR) and synthetic data from the OSIM2 and ModOSIM models

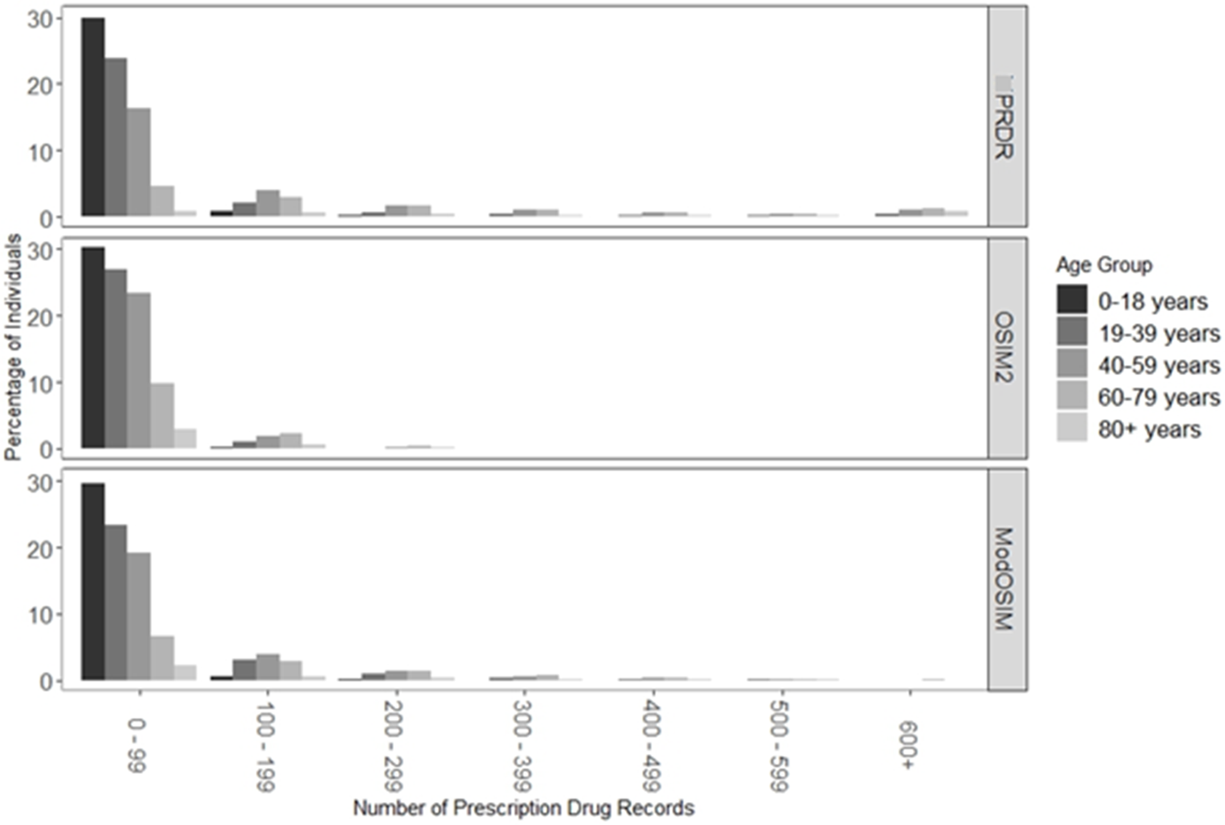


Table 6: Frequency of hospitalization for community-acquired pneumonia (HCAP) and the use of proton pump inhibitors (PPI) in the real-world PRDR, OSIM2, and ModOSIM synthetic data

Data source	Use of PPI	HCAP		Row total	Total
		Yes	No		
PRDR	Yes	20	2,165	2,185	45,215
	No	182	42,848	43,030	
OSIM2	Yes	0	40	40	6,595
	No	126	6,429	6,555	
ModOSIM	Yes	0	18	18	6,225
	No	99	6,108	6,207	

OSIM2: Observational Medical Dataset Simulator II; ModOSIM: Modified OSIM2.

studies and (ii) test program simultaneously by analysts in different provinces to ensure reproducibility and facilitate collaborative research.

There are, however, some limitations to the ModOSIM simulator. The longitudinal causal structure of the synthetic data is unlikely to follow the real-world time ordering of study variables [24]. This is because the model does not follow a causal framework. Other simulation methods that aim to preserve elements of the longitudinal causal structure include the Plasmode simulation approach, which relies on resampling real-world data [25], and simulation using marginal structural models or structural nested models with marginal parameters [26]. However, the latter simulator requires access to real-world data each time researchers wish to implement these models, in order to accurately model different sources of bias in the data; this can be challenging to achieve because of health privacy legislation. Future studies may consider generating synthetic data that follow the real-world time ordering of variables in real-world AHR data using general adversarial networks, which provide a promising framework for simulating complex distributions.

Conclusions

A key consideration when generating synthetic data is to ensure that they mirror many of the attributes of the source data. This study examined the representativeness of data generated from two different simulation models. Synthetic data generated using the ModOSIM model showed a greater similarity to real-world AHR data when compared to OSIM2 data on several measures. Synthetic AHRs consistent with those found in real-world settings can be generated using ModOSIM. Synthetic data will benefit rapid implementation of methodological studies and data analyst training. However, there is a need to further enlighten the research community on the limitations of using synthetic data for real-world studies.

Acknowledgements

The authors acknowledge the Manitoba Centre for Health Policy for use of data contained in the Population Research Data Repository (HIPC # 2018/2019–26). The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or other data providers is intended or should be inferred. Data used in this study are from the Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy, University of Manitoba and were derived from data provided by Manitoba Health.

Ethics approval and consent to participate

This study received ethical approval from the University of Manitoba Health Research Ethics Board. The Health Information Privacy Committee approved data access. Consent was not required from participants as all data were anonymized prior to their use.

Availability of data and material

The real-world data used in this article were derived from health data as a secondary source. The data were provided under specific data sharing agreements only for the approved use. The original source data are not owned by the researchers and as such cannot be provided to a public repository. The original data source and approval for use has been noted in the acknowledgments. Where necessary and with appropriate approvals, source data specific to this article or project may be reviewed with the consent of the original data providers, along with the required privacy and ethical review bodies.

Funding

This research was supported by funding from the Canadian Institutes of Health Research (Funding Reference Number DSE-146021).

Authors' contributions

All authors conceived the study and contributed to development of the analysis plan. OFA and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Cadarette, S. M., & Wong, L. (2015). An introduction to health care administrative data. *The Canadian Journal of Hospital Pharmacy*, *68*(3), 232–237. https://doi.org/10.4212/cjhp.v68i3.1457

2. Dormuth, C. R., Filion, K. B., Paterson, J. M., James, M. T., Teare, G. F., Raymond, C. B., . . . Lipscombe, L. (2014). Higher potency statins and the risk of new diabetes: multicentre, observational study of administrative databases. *BMJ*, 348. https://doi.org/10.1136/bmj.g3244

3. Dormuth, C. R., Hemmelgarn, B. R., Paterson, J. M., James, M. T., Teare, G. F., Raymond, C. B., . . . Ernst, P. (2013). Use of high potency statins and rates of admission for acute kidney injury: multicenter, retrospective observational analysis of administrative databases. *BMJ*, *346*(Mar18), f880–f880. https://doi.org/10.1136/bmj.f880

4. Azoulay, L., Filion, K. B., Platt, R. W., Dahl, M., Dormuth, C. R., Clemens, K. K., . . . Sketris, I. S. (2016). Association between incretin-based drugs and the risk of acute pancreatitis. *JAMA Internal Medicine*, *176*(10), 1464–1473. https://doi.org/10.1001/jamainternmed.2016.1522

5. Kokosi, T., De Stavola, B., Mitra, R., Frayling, L., Doherty, A., Dove, I., . . . Harron, K. (2022). An overview of synthetic administrative data for research. *International Journal of Population Data Science*. Swansea University. https://doi.org/10.23889/ijpds.v7i1.1727

6. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-00977-1

7. Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, *72*, 219–226. https://doi.org/10.1016/j.csda.2013.10.018

8. Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

9. Yucel, R. M., Zhao, E., Schenker, N., & Raghunathan, T. E. (2018). Sequential hierarchical regression imputation. *Journal of Survey Statistics and Methodology*, *6*(1), 1–22. https://doi.org/10.1093/jssam/smx004

10. Stang, P. E., Ryan, P. B., Racoosin, J. A., Overhage, J. M., Hartzema, A. G., Reich, C., . . . Woodcock, J.

(2010). Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of Internal Medicine*, *153*(9), 600–606. https://doi.org/10.7326/0003-4819-153-9-201011020-00010

11. Murray, R. E., Ryan, P. B., & Reisinger, S. J. (2011). Design and validation of a data simulation model for longitudinal healthcare data. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2011*, 1176–1185.

12. Reisinger, S. J., Ryan, P. B., O'Hara, D. J., Powell, G. E., Painter, J. L., Pattishall, E. N., & Morris, J. A. (2010). Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association*, *17*(6), 652–662. https://doi.org/10.1136/jamia.2009.002477

13. Suissa, S., Henry, D., Caetano, P., Dormuth, C. R., Ernst, P., Hemmelgarn, B., Canadian Network for Observational Drug Effect Studies (CNODES). (2012). CNODES: the Canadian Network for Observational Drug Effect Studies. *Open medicine : a peer-reviewed, independent, open-access journal*, *6*(4), e134–40.

14. Manitoba Health Services. (1994). *Pharmacy claims submission manual (DPIN pharmacy manual)*.

15. Kozyrskyj, A. L., & Mustard, C. A. (1998). Validation of an electronic, population-based prescription database. *Annals of Pharmacotherapy*, *32*(11), 1152–1157. https://doi.org/10.1345/aph.18117

16. World Health Organization Collaborating Centre for Drug Statistics Methodology. (1996). *Guidelines for ATC classification and DDD assignment*. Oslo, Norway.

17. Canadian Institute for Health Information. (2003). *International Statistical Classification of Diseases and Related Health Problems Tenth Revision, Canada [ICD-10-CA]*. Ottawa, Canada.

18. Alzheimer Society of Canada. (2016). *Prevalence and monetary costs of dementia in Canada: Population Health Expert Panel*. Toronto, Canada.

19. Canadian Cancer Statistics Advisory Committee. (2019). *Canadian Cancer Statistics 2019*.

20. Statistics Canada. (2018). *Diabetes, 2017*. Ottawa, Canada.

21. Filion, K. B., Chateau, D., Targownik, L. E., Gershon, A., Durand, M., Tamim, H., . . . Dormuth, C. R. (2014). Proton pump inhibitors and the risk of hospitalisation for community-acquired pneumonia: replicated cohort studies with meta-analysis. *Gut*, *63*(4), 552–558. https://doi.org/10.1136/gutjnl-2013-304738

22. Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–68.

23. Seeskin, Z. H., Ugarte, G., & Datta, A. R. (2019). Constructing a toolkit to evaluate quality of state and local administrative data. *International Journal of Population Data Science*, *4*(1). https://doi.org/10.23889/ijpds.v4i1.937

24. Gruber, S. (2015). A causal perspective on OSIM2 data generation, with implications for simulation study design and interpretation. *Journal of Causal Inference*, *3*(2), 177–187. https://doi.org/10.1515/jci-2014-0008

25. Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, *72*, 219–226. https://doi.org/10.1016/j.csda.2013.10.018

26. Young, J., Hernán, M., Picciotto, S., & Robins, J. (2019). Simulation from structural survival models under complex time-varying data structures.

List of Abbreviations

AHR:	Administrative Health Record
ATC:	Anatomical Therapeutic Chemical
CDM:	Common Data Model
CI:	Confidence Interval
CNODES:	Canadian Network for Observational Drug Effect Studies
DIN:	Drug Identification Number
DPIN:	Drug Program Information Network
DSEN:	Drug Safety and Effectiveness Network
ICD:	International Classification of Diseases
MCHP:	Manitoba Centre for Health Policy
ModOSIM:	Modified Observational Medical Dataset Simulator II
RAMQ:	Regie de l'assurance maladie du Quebec
OSIM:	Observational Medical Dataset Simulator
OSIM2:	Observational Medical Dataset Simulator II
OMOP:	Observational Medical Outcomes Partnership
PRDR:	Population Research Data Repository