Generating realistic simulated administrative health data for population-based drug safety and effectiveness research and training

Olawale Ayilara

University of Manitoba

Robert W Platt

McGill University

Matthew Dahl

Manitoba Centre for Health Policy

Janie Coulombe

McGill University

Pablo Gonzalez Ginestet

Karolinska Institutet

Dan Chateau

University of Manitoba Faculty of Health Sciences

Lisa M Lix (✉ lisa.lix@umanitoba.ca)

<https://orcid.org/0000-0001-8685-3212>

Research article

Keywords: administrative health records, computer simulation, prescription drug records, 69 validation

Posted Date: December 9th, 2019

DOI: <https://doi.org/10.21203/rs.2.18325/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Generating realistic simulated administrative health data for population-based drug safety and effectiveness research and training

Olawale F. Ayilara, Robert W. Platt, Matt Dahl, Janie Coulombe, Pablo Gonzalez Ginestet, Dan Chateau, Lisa M. Lix

Author Addresses:

OFA: Department of Community Health Sciences, University of Manitoba, Winnipeg, CANADA; olawale.ayilara@umanitoba.ca

MD: Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, CANADA; matthew_dahl@cpe.umanitoba.ca

JC: Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, CANADA; janie.coulombe@mail.mcgill.ca

PGG: Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, SWEDEN; pablo.gonzalez.ginestet@ki.se

DC: Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, CANADA; dan_chateau@cpe.umanitoba.ca

RWP: Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, CANADA; robert.platt@mcgill.ca

LML: Department of Community Health Sciences, University of Manitoba, Winnipeg, CANADA; lisa.lix@umanitoba.ca

Corresponding Author:

Lisa Lix, PhD

Department of Community Health Sciences, University of Manitoba

S113-750 Bannatyne Avenue

Winnipeg, Manitoba, CANADA

R3E 0W3

Phone: 204-789-3573

e-mail: lisa.lix@umanitoba.ca

Abstract Word Count: 340

Manuscript Word Count: 2530

Abstract

Background: Administrative health records (AHRs), which are generated primarily for management and billing purposes, are now widely used in drug safety and comparative effectiveness studies. The development of analytic methods for multi-site studies can benefit from the availability of simulated data, which do not require ethical approvals and data access permissions. We simulated AHRs using both the Observational Medical Dataset Simulator II (OSIM2) proposed by the Observational Medical Outcomes Partnership, and a modified OSIM (ModOSIM) method developed by the Canadian Network for Observational Drug Effect Studies (CNODES). Our objective was to compare the simulated data to real-world AHR data to assess the representativeness of the simulated data.

Methods: The real-world data comprised prescription drug records for all individuals with healthcare coverage at any point in a 10-year period (2008 – 2017) from the Manitoba Population Research Data Repository (MPRDR) in the province of Manitoba, Canada. OSIM2 and ModOSIM, which are empirical simulation models for longitudinal patient data, were used to simulate AHRs. The data were described using frequencies and percentages. We estimated agreement of prescription drug use measures in MPRDR, OSIM2 and ModOSIM using the concordance coefficient.

Results: The MPRDR cohort included 169,586,633 drug records and 1,395 drug types for 1,604,734 individuals. Data for 50,000 individuals were simulated using OSIM2 and ModOSIM. Sex and age group distributions were similar in the real-world and simulated data. There were significant differences in the total number of drug records and number of unique drugs for OSIM2 and ModOSIM when compared with MPRDR; the median number of unique drugs in

61 MPRDR, OSIM2 and ModOSIM was 9.0, 6.0 and 10.0, respectively. For average number of days
62 of drug use, concordance was 16% (95% confidence interval [CI]: 12% – 19%) for MPRDR and
63 OSIM2 and 88% (95% CI: 87%-90%) for MPRDR and ModOSIM.

64 **Conclusions:** ModOSIM data were more similar to MPRDR than OSIM2 data on many measures
65 of prescription drug use. Simulated AHRs that are consistent with those found in real-world
66 settings can be generated using ModOSIM; these simulated data will benefit methodological
67 studies and data analyst training.

68

69 **Keywords:** administrative health records, computer simulation, prescription drug records,
70 validation

Generating realistic simulated administrative health records for population-based drug safety and effectiveness research and training

Background

Administrative health records (AHRs), which are generated primarily for healthcare system management and remuneration are also increasingly being used for prescription drug safety and comparative effectiveness research (1–4). However, the cost of data extraction, and the challenges associated with securing ethics and data access approvals due to patient privacy and confidentiality issues and health privacy legislation can make it difficult for researchers to conduct methodological research in a timely manner using AHRs and to train analysts with real-world data. To address these challenges, the Observational Medical Outcomes Partnership (OMOP) developed a simulation program to generate realistic AHR data that incorporate the complex relationships among health conditions and prescription drug use (5). 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 (6). 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 (7). 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 (5,6,8) .

In Canada, OSIM2 was applied to AHRs from the Regie de l'assurance maladie du Quebec (RAMQ) by the Canadian Network of Observational Drug Effect Studies, a collaborating centre of the Drug Safety and Effectiveness Network (DSEN) (9). 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 preserves 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.

The overall purpose of this study was to simulate AHRs using both the OSIM2 and ModOSIM model. Our objective was to compare the simulated data to real-world AHR data to assess the representativeness of the former.

Methods

Real-World AHR Data and Study Cohort

The Manitoba Population Research Data Repository (MPRDR) 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 MPRDR includes Drug Program Information Network (DPIN) records, hospital discharge abstracts, and physician billing claims for all individuals

eligible to receive health services in Manitoba (Table 1). The population registry contains information for all Manitoba residents registered with the Manitoba 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 in Manitoba (10,11). 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 (12).

Hospital discharge abstracts contain records of demographic and clinical information of discharges from acute care facilities in the province. 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) (13), 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 April 1, 2008 and March 31, 2017. All prescription drug records, hospital discharge abstracts, and physician billing claims for this cohort were extracted from the MPRDR.

Ethics approval for this research was received from the Manitoba Health Research Ethics Board and permission to access the study data was provided by the Manitoba Health Information Privacy Committee.

Observational Medical Dataset Simulator II (OSIM2)

The OSIM2 simulator uses an individual Monte Carlo approach, whereby information for an individual is generated using a random probability at each decision point that selects a value from the empiric multinomial distributions generated with the analysis module. This module extracts individual information as probability distributions from the source data (6). 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) injecting associations between treatments and outcomes. A detailed description of these stages are described by Murray et al. (2011).

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 that contain the prescription drug information directly, which implies that each drug 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, which are: (i) a probability table for generating the total number of unique prescription drug 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.

ModOSIM modifies 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 as follows: (i) the total of number days of exposure to the prescription drug was removed from the probability table, and, (ii) the structural equation for drawing 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.

Study Variables from the Real-World AHRs and Simulated Data

Demographic information (i.e., age group, sex) were extracted from the MPRDR population registry for the study cohort. Per-person measures of prescription drug use were calculated from DPIN data. Annual and overall measures were produced; these included the number of prescription drug records, number of unique prescription drugs, and total number of days of prescription drug use. 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, diabetes, cancer and dementia were selected as example

conditions (14–16). The diagnosis codes used to ascertain the selected 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 MPRDR and the simulated data produced using the OSIM2 and ModOSIM models for the study cohorts. Frequency distributions were used to compare demographic characteristics, health conditions, number of unique drugs, and number of drug records in each of the MPRDR, 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 MPRDR 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 (17).

Scatterplots of the average of number of days of drug use for the three data sources were also produced. All analyses were conducted using SAS version 9.4 (SAS Institute, Cary, NC).

Results

The MPRDR study cohort was 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 April 1, 2008 and March 31, 2017. For each of the OSIM2 and ModOSIM simulation models, data were generated for a total of 50,000 individuals. We selected this cohort size for the simulated data because it could be produced with a reasonable amount of computation time and was sufficiently large for comparative evaluations of the study measures.

Table 3 contains descriptive information for the demographic, health condition, and prescription drug use measures in the real-world AHR data and the simulated data generated using the OSIM2 and ModOSIM models. 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. The percentage of individuals with diagnosed diabetes, cancer and dementia in the cohort were 11.2%, 7.3% and 3.2% respectively. Over this 10-year period, slightly more than half (52.4%) of the cohort had less than 10 different prescription drugs; the median was 9.0. Three quarters of the cohort had less than 100 total drug records; the median number of records was 25.0.

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 same was true for most age groups, with the exception of the 20-29 years and 50-59 years age groups. The percentage of individuals with diagnosed diabetes, cancer and dementia, respectively, were similar in OSIM2 (8.8%, 10.8%, and 3.3%) and ModOSIM (9.0%, 10.8%, and 3.3%).

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

The estimated concordance correlation for the average number of days of drug use for MPRDR and OSIM2 was 0.16 (95% CI: 0.12 - 0.19). For MPRDR 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 MPRDR and ModOSIM and a weak relationship for MPRDR 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 simulated 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 distribution for the real-world data. However, for the OSIM2 data there were substantial differences. In the ModOSIM cohort, two-thirds of individuals (31.2% males; 35.6% females) had fewer than 100 drug records, while more than 90% (46.5% males; 47.0% females) had fewer than 100 drug records in OSIM2 cohort, which is an overestimation of the 75.6% (38.5% males; 37.1% females) in the real-world data. The average number of unique drugs stratified by health condition in the real-world data and the two simulated datasets are presented in Figure 4. The average number of unique drugs for individuals diagnosed with diabetes, cancer and dementia in MPRDR cohort were 17.1, 16.0 and 18.3 respectively. The average number of unique drugs for the health conditions (diabetes, cancer, and dementia) in ModOSIM (18.2, 18.6, 22.0) was closer to MPRDR than OSIM2 (10.5, 10.8, 12.8).

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 (7). The ModOSIM model was created by CNODES to address some of the

limitations of OSIM2. The results of this study showed that data simulated 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.

Simulated data are a valuable and important resource for conducting methodological studies, such as comparing different methods for developing propensity scores models to address confounding in observational studies. As well, the data can be used for the training of analysts who conduct observational studies. In particular, new analysts can benefit from the data generated using ModOSIM by using it to implement the sequence of steps required to interpret and program a drug safety and effectiveness protocol.

In Canada, each province and territory has 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. Hence, simulated data are beneficial to test methods for conducting multi-site drug safety and effectiveness studies.

While the ModOSIM simulator has many benefits, it also has some limitations. The longitudinal causal structure of the simulated data may not respect real-world time ordering of study variables (18). 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 from real-world data (19), and simulation using marginal structural models or structural nested models with marginal parameters (20). However, the latter simulator requires access to real-world data each time

268 researchers wish to run these models, in order to accurately model different sources of bias in
269 the data; this can be challenging to achieve because of health privacy legislation.

270 **Conclusions**

271 A key consideration when developing simulated datasets is to ensure that they mirror
272 many of the attributes of the source data. In this study, we examined the representativeness of
273 data derived from two different simulation models to real-world data. These simulated data
274 provide an opportunity to test new methods for conducting drug safety and effectiveness
275 studies and to enable new learners to develop hands-on skills in working with complex health
276 data.

277 List of Abbreviations

- 278 AHR: Administrative Health Record
- 279 OSIM: Observational Medical Dataset Simulator
- 280 OSIM2: Observational Medical Dataset Simulator II
- 281 OMOP: Observational Medical Outcomes Partnership
- 282 CDM: Common Data Model
- 283 RAMQ: Regie de l'assurance maladie du Quebec
- 284 CNODES: Canadian Network for Observational Drug Effect Studies
- 285 DSEN: Drug Safety and Effectiveness Network
- 286 ModOSIM: Modified Observational Medical Dataset Simulator II
- 287 MPRDR: Manitoba Population Research Data Repository
- 288 MCHP: Manitoba Centre for Health Policy
- 289 DPIN: Drug Program Information Network
- 290 DIN: Drug Identification Number
- 291 ATC: Anatomical Therapeutic Chemical
- 292 ICD: International Classification of Diseases
- 293 CI: Confidence Intervals

294 **Declarations**

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

299 ***Consent for publication.*** Not applicable.

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

308 ***Competing interests.*** The authors declare that they have no competing interests.

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

311 ***Authors' contributions.*** All authors conceived the study and contributed to development of the
312 analysis plan. OFA and LML conducted the analysis and prepared the draft manuscript. All
313 authors reviewed and approved the final version of the manuscript.

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

References

1. Dormuth CR, Hemmelgarn BR, Paterson JM, James MT, Teare GF, Raymond CB, et al. Use of high potency statins and rates of admission for acute kidney injury: multicenter, retrospective observational analysis of administrative databases. *BMJ*. 2013;346(7900):1–10.
2. Dormuth CR, Filion KB, Paterson JM, James MT, Teare GF, Raymond CB, et al. Higher potency statins and the risk of new diabetes: multicentre, observational study of administrative databases. *BMJ*. 2014;348(May):1–9.
3. Cadarette SM, Wong L. An introduction to health care administrative data. *Can J Hosp Pharm*. 2015;68(3):232–7.
4. Azoulay L, Filion KB, Platt RW, Dahl M, Dormuth CR, Clemens KK, et al. Association between incretin-based drugs and the risk of acute pancreatitis. *JAMA Intern Med*. 2016;176(10):1464–73.
5. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med*. 2010;153(9):600–6.
6. Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. *AMIA Annu Symp Proc Arch*. 2011;1176–85.
7. Observational Medical Outcomes Partnership. Data dictionary for the enhanced observational medical dataset simulator (OSIM 2). United BioSource Corp. 2011;1–39.

- 341 8. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development
342 and evaluation of a common data model enabling active drug safety surveillance using
343 disparate healthcare databases. *J Am Med Inform Assoc*. BMJ Group; 2010;17(6):652–62.
- 344 9. Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: the
345 Canadian network for observational drug effect studies. *Open Med*. Canada;
346 2012;6(4):e134-40.
- 347 10. Manitoba Health Services. Pharmacy claims submission manual (DPIN Pharmacy Manual)
348 [Internet]. 1994. Available from:
349 https://www.gov.mb.ca/health/pharmacare/profdocs/dpin_manual.pdf. Accessed 15
350 Nov 2018.
- 351 11. Kozyrskyj AL, Mustard CA. Validation of an electronic, population-based prescription
352 database. *Ann Pharmacother*. 1998;32(11):1152–7.
- 353 12. World Health Organization Collaborating Centre for Drug Statistics Methodology.
354 Guidelines for ATC classification and DDD assignment [Internet]. Oslo, Norway; 1996.
355 Available from: <https://www.who.int/classifications/atcddd/en/>. Accessed 10 Aug 2019.
- 356 13. Canadian Institute for Health Information. International statistical classification of
357 diseases and related health problems tenth revision, Canada [ICD-10-CA]. Ottawa, ON;
358 2003.
- 359 14. Statistics Canada. Diabetes, 2017 [Internet]. Ottawa, ON; 2018. Available from:
360 <https://www150.statcan.gc.ca/n1/en/pub/82-625-x/2018001/article/54982->

eng.pdf?st=ICN8JLi5. Accessed 21 Sep 2019.

15. Canadian Cancer Statistics Advisory Committee. Canadian cancer statistics 2019 [Internet]. Available from: cancer.ca/Canadian-Cancer-Statistics-2019-EN. Accessed 15 Oct 2019.
16. Alzheimer Society of Canada. Prevalence and monetary costs of dementia in Canada: population health expert panel. Alzheimer Society of Canada. Toronto, Canada; 2016.
17. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. United States; 1989 Mar;45(1):255–68.
18. Gruber S. A Causal Perspective on OSIM2 Data generation, with implications for simulation study design and interpretation. *J Causal Inference*. 2015;3(2):177–87.
19. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. Netherlands; 2014 Apr;72:219–26.
20. Young J, Hernán M, Picciotto S, Robins J. "Simulation from structural survival models under complex time-varying data structures." 2019.
21. Overhage JM, Ryan PB, C.G. R, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inf Assoc*. 2012;19(1):54–60.

Table 1: Manitoba Population Research Data Repository (MPRDR) databases and attributes used in the simulation study

Database	Attributes	Purpose in the Simulation
Manitoba Health Insurance Registry	Healthcare coverage start and end dates, postal code, birthdate, de-identified personal health identification number (PHIN), sex	To develop observation and person tables for the Common Data Model [CDM] (Overhage, Ryan, Hartzema, & Stang, 2012). 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.
Drug Prescription Information Network	drug identification number (DIN), dispensing date, days of prescription drug supply and quantity, de-identified PHIN	To build a drug era table for the CDM. This table contain all the prescription dispensation records for each individual in the observation table.
Hospital Discharge Abstracts	diagnosis codes, dates of hospital admission and separation, procedure codes, de-identified PHIN	To create a condition era table for the CDM.
Physician Billing Claims	diagnosis code, date of service, de-identified PHIN	To create a condition era table for the CDM.

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

Condition	Diagnosis codes	
	ICD-9-CM	ICD-10-CA
Diabetes	250.x	E10.x - E14.x
Cancer, excluding non-melanoma skin cancer	140.x - 172.x	C00.x - C43.x
	174.x - 209.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
Dementia	290.0	F05.1
	331.0	F06.5
	331.2	F06.6
	797.x	F06.8
		F06.9
		F09.x
		G30.0
		G30.1
		G30.8
		G30.9
		G31.0
		G31.1
		R54.x

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: Cohort characteristics for Manitoba Population Research Data Repository (MPRDR) real-world data and OSIM2 and ModOSIM simulated data

Characteristic	MPRDR	OSIM2	ModOSIM
Sex, (%)			
Female	49.7	49.9	50.0
Male	50.3	50.1	50.0
Age in years, (%)			
Less than 10	20.8	20.7	21.1
10-19	12.9	13.0	13.0
20-29	14.5	13.9	13.8
30-39	13.3	13.0	13.1
40-49	13.1	13.4	13.2
50-59	10.8	11.2	11.1
60-69	6.9	7.0	6.9
70-79	4.4	4.4	4.4
80-89	2.7	2.8	2.8
90+	0.6	0.6	0.6
Number of unique drugs per person, (%)			
Less than 10	52.4	69.4	49.0
10-19	28.4	22.5	28.7
20-29	11.7	6.4	11.0
30-39	4.6	1.6	6.2
40-49	1.8	0.1	3.3
50-59	0.7	0.0	1.4
60+	0.4	0.0	0.4
Number of drug records per person, (%)			
0 - 99	75.6	93.5	66.9
100 - 199	10.4	5.7	13.7
200 - 299	4.8	0.8	6.6
300 - 399	2.7	0.1	3.6
400 - 499	1.7	0.0	2.5
500 - 599	1.1	0.0	1.8
600+	3.8	0.0	4.9
Health conditions, (%)			
Diabetes	11.2	8.8	9.0
Cancer	7.3	10.8	10.8
Dementia	3.2	3.3	3.3

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

Figure 1: Average number of days of prescription drug use for real-world data from the Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 and ModOSIM

Figure 2: Number of prescription drug records, stratified by sex, for real-world data from the Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 and ModOSIM

Figure 3: Number of prescription drug records, stratified by age group, for real-world data from the Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 and ModOSIM

Figure 4: Prescription drug use measures, stratified by health condition, for real-world data from the Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 and ModOSIM

Figures

Figure 1

Average number of days of prescription drug use for real-world data from the 392 Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 393 and ModOSIM

Figure 2

Number of prescription drug records, stratified by sex, for real-world data from the 397 Manitoba Population Research Data Repository (MPRDR) and simulated data from OSIM2 398 and ModOSIM

Figure 3

Number of prescription drug records, stratified by age group, for real-world data 401 from the Manitoba Population Research Data Repository (MPRDR) and simulated data from 402 OSIM2 and ModOSIM

Figure 4

Prescription drug use measures, stratified by health condition, for real-world data 407 from the Manitoba Population Research Data Repository (MPRDR) and simulated data from 408 OSIM2 and ModOSIM