


Research and Applications

Automated identification of unstandardized medication data: a scalable and flexible data standardization pipeline using RxNorm on GEMINI multicenter hospital data

Riley Waters¹, Sarah Malecki², Sharan Lail¹, Denise Mak¹, Sudipta Saha¹, Hae Young Jung¹, Mohammed Arshad Imrit¹, Fahad Razak^{1,2,3,†}, and Amol A. Verma ^{1,2,3,†,*}

¹St. Michael's Hospital, Unity Health Toronto, Toronto, Ontario, Canada, ²Department of Medicine, University of Toronto, Toronto, Ontario, Canada, and ³Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada

*Corresponding Author: Amol A. Verma, Li Ka Shing Knowledge Institute, St Michael's Hospital, 30 Bond Street, 14-077B, Toronto, ON M5B 1W8, Canada; amol.verma@mail.utoronto.ca

†Co-senior authors

ABSTRACT

Objective: Patient data repositories often assemble medication data from multiple sources, necessitating standardization prior to analysis. We implemented and evaluated a medication standardization procedure for use with a wide range of pharmacy data inputs across all drug categories, which supports research queries at multiple levels of granularity.

Methods: The GEMINI-RxNorm system automates the use of multiple RxNorm tools in tandem with other datasets to identify drug concepts from pharmacy orders. GEMINI-RxNorm was used to process 2 090 155 pharmacy orders from 245 258 hospitalizations between 2010 and 2017 at 7 hospitals in Ontario, Canada. The GEMINI-RxNorm system matches drug-identifying information from pharmacy data (including free-text fields) to RxNorm concept identifiers. A user interface allows researchers to search for drug terms and returns the relevant original pharmacy data through the matched RxNorm concepts. Users can then manually validate the predicted matches and discard false positives. We designed the system to maximize recall (sensitivity) and enable excellent precision (positive predictive value) with efficient manual validation. We compared the performance of this system to manual coding (by a physician and pharmacist) of 13 medication classes.

Results: Manual coding was performed for 1 948 817 pharmacy orders and GEMINI-RxNorm successfully returned 1 941 389 (99.6%) orders. Recall was greater than 0.985 in all 13 drug classes, and the F1-score and precision remained above 0.90 in all drug classes, facilitating efficient manual review to achieve 100% precision. GEMINI-RxNorm saved time substantially compared with manual standardization, reducing the time taken to review a pharmacy order row from an estimated 30 to 5 s and reducing the number of rows needed to be reviewed by up to 99.99%.

Discussion and Conclusion: GEMINI-RxNorm presents a novel combination of RxNorm tools and other datasets to enable accurate, efficient, flexible, and scalable standardization of pharmacy data. By facilitating efficient manual validation, the GEMINI-RxNorm system can allow researchers to achieve near-perfect accuracy in medication data standardization.

LAY SUMMARY

Medication data are very useful for research and quality measurement applications but are frequently stored in different forms across healthcare organizations. RxNorm is a publicly available platform that includes various tools to connect medication data and enable standardization across different vocabularies. In this article, we describe the GEMINI-RxNorm system, which automates the use of multiple RxNorm tools in tandem with other datasets to identify drug concepts from pharmacy orders. The system is designed primarily for the purposes of data querying and exploration to support research and other analyses. GEMINI-RxNorm was used to process 2 090 155 pharmacy orders from 245 258 hospitalizations between 2010 and 2017 at 7 hospitals in Ontario, Canada. The system was able to accurately identify medication orders across 13 different drug classes, and facilitate efficient manual review to confirm the automated matches, reducing the number of rows of medication data that needed to be reviewed manually by 99.99%. GEMINI-RxNorm enables accurate, efficient, flexible, and scalable standardization of pharmacy data. By facilitating efficient manual validation, the GEMINI-RxNorm system can allow researchers to achieve near-perfect accuracy in medication data standardization.

Key words: RxNorm, data, pharmacy, standardization, medication

BACKGROUND AND SIGNIFICANCE

Patient data repositories are growing in size and complexity and have increasing importance in a wide range of research applications.^{1,2} Medication data are an essential component of these data repositories and require extensive and accurate standardization to enable reliable and reproducible research.

This presents a challenge for large medication databases aggregated from multiple sources that may use varying medication vocabularies, data formats, storage systems, or data quality protocols.

Existing standardization methods for medication data are often inflexible, providing limited support for situations

Received: 20 September 2022. Revised: 18 July 2023. Editorial Decision: 21 July 2023. Accepted: 24 July 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

where standardized medication identifiers have low coverage or inconsistent formats.^{3–6} They typically do not account for free-text drug fields that may contain abbreviations, spelling variations, or additional information that obscures simple parsing.^{3–6} Point-in-time manual mapping operations may be time-consuming or result in human error.

Mapping operations are typically performed with the aid of medication ontologies that record normalized drug descriptions along with their interactions and properties.⁷ RxNorm, maintained by the National Library of Medicine,⁸ is one such platform that is often used to connect and convert data sources between drug vocabularies.⁹ Its publicly available Application Programming Interface (API)¹⁰ includes tools to: link drug concepts with their related concepts, search drug classifications, convert drug identifiers, and approximately match strings to concepts using a custom natural language processing (NLP) method.¹¹ These tools have been utilized previously to evaluate standardized representations of ambulatory care e-prescriptions⁴ and medication lists for clinical decision support.¹² However, there remains unexplored potential in creating a holistic procedure, using multiple tools in tandem, that works on all drug categories and supports research query inputs at any level of granularity using any drug vocabulary input.

Our objective was to define, implement, and evaluate an automated pipeline to standardize medication data collected from multiple healthcare organizations, primarily for the purposes of data querying and exploration to support research and other analyses. The system makes use of existing RxNorm functionality and other available datasets and was evaluated for use on the GEMINI database—a large, real-world clinical dataset that aggregates data from hospitals in Ontario, Canada.¹³ This study describes a highly flexible, scalable, and accurate approach to drug data standardization that can be replicated in most drug databases internationally with minimal modifications, significantly reducing the need for manual mapping. In this study, we use the term “framework” to refer to the theoretical flow of steps that may be replicated in other medication databases while “tool” refers to the code implementation of the steps and “system” refers to both of these in tandem.

METHODS

Overall approach

The GEMINI-RxNorm system consists of 2 independent modules that can be used together to standardize medication data. The Matching Module extracts all potential RxNorm Concept Unique Identifier (CUI) matches for each unique input containing drug-identifying information in the data. Each CUI identifies a specific RxNorm “concept”—a term denoting a commonly accepted definition of a drug given to RxNorm by a source vocabulary.⁸ These concepts have a wide range of granularity. For example, “insulin, isophane” (ingredient), “Humulin” (brand name), and “3 ML insulin isophane, human 70 UNT/ML/insulin, regular, human 30 UNT/ML Pen Injector” (Semantic Clinical Drug/Generic Pack) are each unique concepts with their own CUI, which are considered related by RxNorm. The Matching Module runs with each new ingestion of unprocessed data and stores matches in a cache for efficient querying.

The Query Module provides a user interface for researcher queries, determines which RxNorm concepts to search for, and performs back-matching to return the original pharmacy data that relates to those concepts. The Query Module also enables a final review by a subject matter expert prior to use of data. Both modules were implemented in R and utilized RxMix.¹⁰ RxMix is an interface that allows users to combine multiple APIs around RxNorm, enabling them to perform actions such as searching for RxCUI given a generic name.

We validated the performance of the GEMINI-RxNorm system by comparing its outputs to comprehensive manual mappings assembled by a physician and pharmacist. The initial mapping of the validation data was first performed by the physician who manually assigned medication orders to classes. All of the results were reviewed by both the physician and pharmacist to establish the gold standard. They were then independently sent result files from the GEMINI-RxNorm System where discrepancies were discussed and resolved together by consensus. Because this process was iterative, until both the physician and pharmacist were satisfied with medication categorization, we did not measure inter-rater reliability.

Setting

The GEMINI database collects administrative and clinical data for hospital admissions at multiple sites in Ontario.^{13,14} GEMINI data undergo a series of data quality checks as previously described.¹⁴ These data quality checks are primarily designed to ensure completeness and plausibility of data, and do not influence the actual content of the rows of pharmacy data. Therefore, although the GEMINI data quality checks are essential for ensuring the overall quality of the data, they would not affect the specific matches identified by the GEMINI-RxNorm system. This evaluation used inpatient pharmacy data for patients admitted to or discharged from the general medicine inpatient service of 7 hospital sites.¹³ It covers physician medication orders from 245 258 unique admissions between April 2010 and October 2017.

The GEMINI pharmacy data include generic names, brand names, Drug Identification Numbers (DIN, unique identifiers assigned by Health Canada to all drug products),¹⁵ National Drug Codes (NDC, unique identifiers assigned to drugs by the U.S. Food and Drug Administration),¹⁶ internal hospital identification codes, route, dose, frequency, and other administrative and prescription information. However, raw data coverage and quality varies greatly because hospitals differ in how they store, extract, and manipulate data prior to GEMINI receiving it. Some hospitals may only record generic names or allow free-text entries with additional prescription instructions. Others may not record NDC or may change their data formats and internal codes over time. NDC is often present in Canadian medication databases as a byproduct of the data input systems used by Canadian hospitals that are closely connected to those in the United States. Similarly, brand names are often available in medication databases and can provide an additional method to map the drugs. The GEMINI-RxNorm system is designed for maximum flexibility, utilizing any available drug-identifying information (key identifier fields) in the North American context regardless of data format or field coverage.

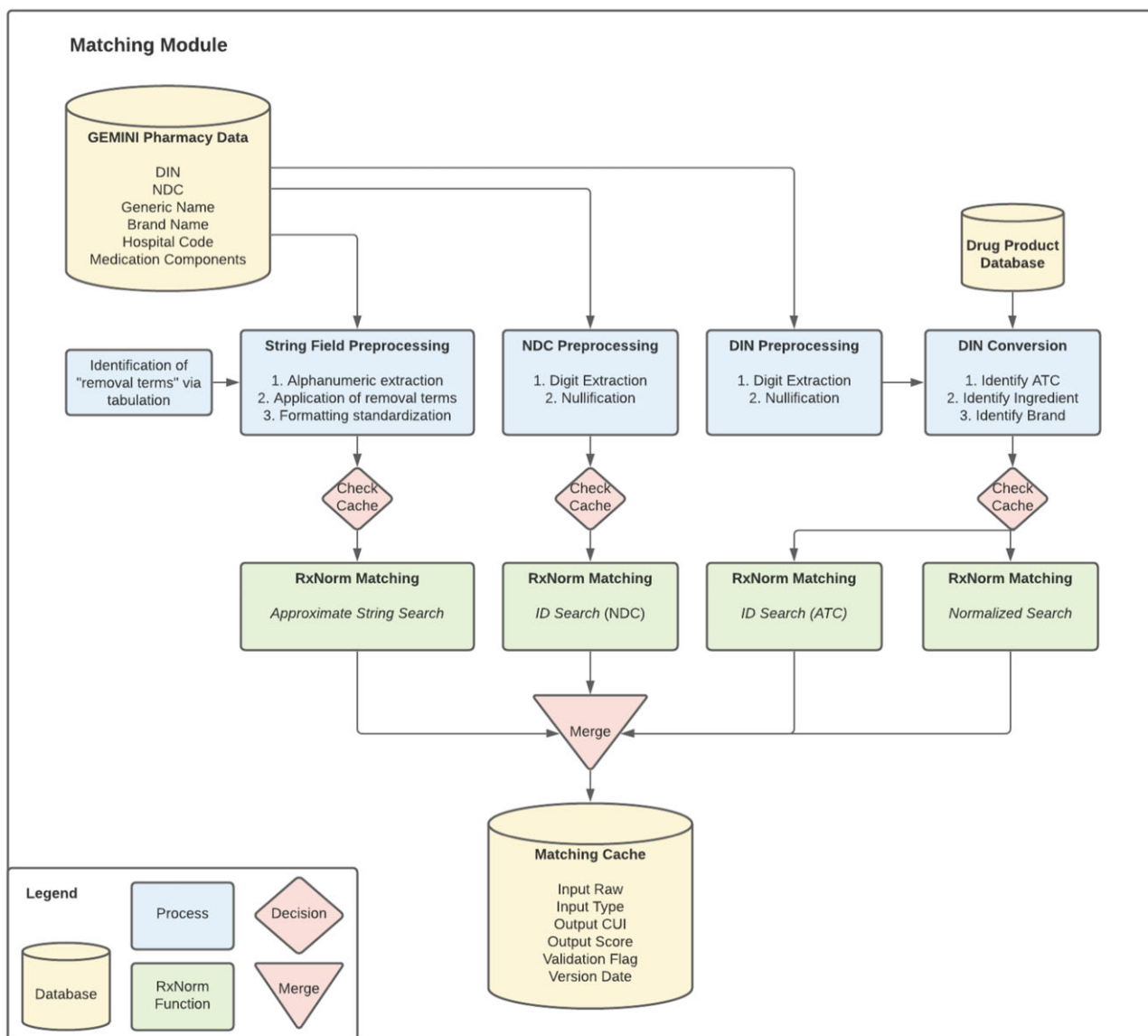


Figure 1. Matching module. The Matching Module matches pharmacy order data to potential RxNorm drug concepts. Pharmacy order data are preprocessed according to their data types (String, NDC, or DIN). A store of outputs called the Matching Cache is checked to see if the data have been encountered before, and if not, an appropriate RxNorm tool is used to determine the matching drug concept identifiers (CUI). All outputs are stored in the Matching Cache along with additional information such as the score indicating how well the data match each CUI.

Table 1. Key identifier fields used in the GEMINI database

Key identifier field	Data type
Generic name	Text
Brand name	Text
Drug identification number (DIN)	Identifier
National drug code (NDC)	Identifier
Internal hospital code	Text
Medication components	Text

Notes: This table lists the fields that often contain drug identifying information (key identifier fields) in the GEMINI database. Other medication repositories may contain different key identifier fields, but GEMINI-RxNorm will support any number as long as they can be categorized into text or identifier data types.

Matching module

Step 1. Preprocessing

The matching module (Figure 1) began by extracting all raw data from the key identifier fields (Table 1). Each field was

then preprocessed in a different way. NDC data were searched for 10 or 11 digit sequences, ignoring separations by symbols or additional lettering. Sequences above 11 digits were nullified to avoid non-NDC identifiers that were mistakenly entered in this field. Sequences below 10 digits were padded with leading zeros to account for entries that had leading zeros stripped. Similarly, DIN data were searched for 8 digit sequences after removing symbols and letters. Any sequence below 8 digits was padded with leading zeros as DINs are often entered without them in the GEMINI data.

The remaining string fields were preprocessed to only retain strings relevant to RxNorm concepts. Alphanumeric sequences were extracted and symbols discarded excluding those commonly found in RxNorm concept names (“.”, “%”, “-”, “/”). A list of “removal terms” was then identified by tabulating all individual words in the string fields and manually flagging high-frequency terms that were not relevant to drug identification (eg, “[NF]” indicating non-formulary, and

“[NP]” indicating “not preferred”). This is an optional preprocessing step to improve match scores, but it only needs to be done once and can be applied to all string data. Additional custom formatting steps were undertaken such as collapsing whitespace. To retain flexibility for future string formats, no other preprocessing was done. RxNorm Approximate String Search is designed to handle specific wording often found in pharmacy data including dosage and units.¹¹

Step 2. Concept matching

Processed data from fields that potentially contained drug-identifying information such as names or IDs, referred to as “key identifier fields” (Table 1), were matched to potential RxNorm CUIs using different tools provided by the RxNorm API. NDC data were inputted directly into the RxNorm *ID Search* function. All string fields were sent through the *Approximate String Search* function. This function supports all RxNorm concept types including generic and brand names.

As RxNorm only supports US drug vocabularies, we converted the processed DINs into identifiers that are supported: Anatomical Therapeutic Chemical (ATC) code, ingredient name, and brand name. This was accomplished by assembling the required fields from the most current Drug Product Database provided by Health Canada¹⁷ then searching for the identifiers related to the processed DINs. ATC matches were inputted into RxNorm’s *ID Search* while ingredient and brand name matches were inputted into the *Normalized String Search* function. Important to note is that ATC searches lose route information when converted to CUI. However, the system attempts to find the route using raw string processing and NDC matching alongside the DIN conversion when possible. This provides additional avenues to include the route-specific RxCUI in the matching cache. When validating the GEMINI-RxNorm system, incorrect routes were considered to be false positive matches. For example, Ciprofloxacin ear drops were considered a false positive when searching for systemic antibiotics.

Step 3. Building the matching cache

A cache of each raw input and its CUI outputs were saved as a database with fields to store the Rx-Norm-defined match score, manual validation flag, and the date the cache was built. This ensured replicability in matching, enabled manual adjustment of matching results, and eliminated the need for reprocessing the entire raw data each time a query was made. The output of the *Approximate String Search* can yield multiple potential CUIs along with a match score out of 100 that indicates how closely the input matches the concept. Matches made using the DIN or NDC inputs were assigned a match score of 100 because these were explicit conversions. DIN conversions that lost ATC information were still assigned a score of 100 as the system is confident that they are associated with the resulting generalized CUI. All of these potential outputs and their match scores were saved.

Query module

Step 1. Finding concepts related to a query

The query module (Figure 2) is an interface for researchers to identify pharmacy data of interest by allowing them to search using any RxNorm supported classification (eg, ATC classifications). For example, a researcher may want to extract all pharmacy records related to diabetes drugs. The user enters

keyword(s) of interest (eg, “diabetes”) and the Query Module returns a list of valid concept names using RxNorm’s *Normalized String Search*, *Spelling Suggestions*, *ID Search*, *All Classes*, and *Class Members* functions.¹⁰ After the user confirms all valid concepts of interest, the module uses RxNorm’s *Get Related by Type* function to discover all other directly related concepts. This step ensures that all brand, generic names, and dosage variations of the selected drugs are also searched. For example, a search for “metformin” would also return rows where the only identifying information is the metformin brand name “Glucophage.” Users are able to search for specific routes of a drug when selecting from the returned list of related RxNorm concepts in the query module. The system will return entries with matching route and where no route was found and those are left for manual review.

Step 2. Matching concepts to the pharmacy data

With the final concept list constructed, the corresponding CUIs are searched in the matching cache to return the GEMINI pharmacy data variables that match them. These data variables are then back-matched in the GEMINI pharmacy data repository, depending on their field type, to return the rows of original GEMINI pharmacy data that match the search (ie, orders that included diabetes drugs). The results cache may be further limited to only return matches above a specified match score, between certain dates, or belonging to certain patient encounters or hospital sites.

Step 3. Output flagging

Data users may want to manually check standardized medication data for errors before conducting analyses. It is much less labor intensive to check for “false positive” results (ie, incorrect matches) than “false negatives” (ie, missed matches), because the latter requires manually searching the entire dataset whereas the former only requires checking the suggested matches for correctness. Thus, the GEMINI-RxNorm system was designed to maximize recall (sensitivity), allowing researchers to focus only on manually flagging false positive results. Other applications could easily adjust the system to optimize for other balances of precision and recall if a human in the loop is not desired.

To allow a person to easily identify false positive results, outputted rows are condensed into unique combinations of key identifier fields and their predicted CUI matches. In our case, a pharmacist manually reviewed the matches and removed false matches. Any pharmacist-flagged false matches were removed from the matching cache so that they would not be made on future runs of the system.

Validation of the GEMINI-RxNorm system

To establish a gold standard of drug mapping, key identifier fields covering 1 948 817 total pharmacy orders for a subset of commonly used medication classes (Table 2) were manually mapped and validated by a physician (3rd year internal medicine resident) and pharmacist. The physician performed line-by-line manual coding using a master file of pharmacy data for each hospital site to retrieve the medications belonging to each drug category. We used a combination of pre-established drug categories based on existing ATC groupings and several custom-developed categories to demonstrate the flexibility of our approach across a wide range of potential applications (Table 2). Variables used to code drugs into categories included brand name, generic name, DIN, and route

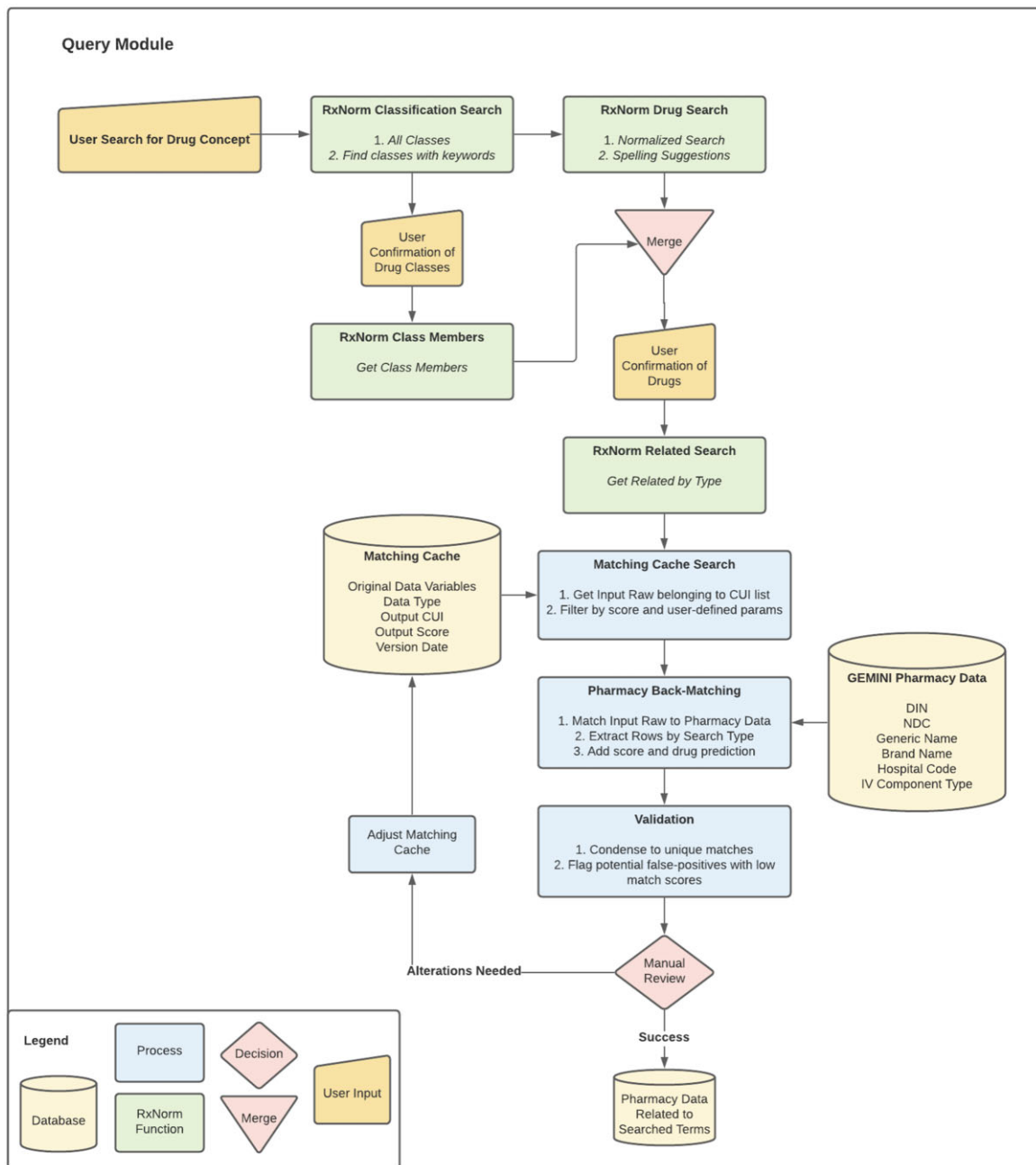


Figure 2. Query module. The Query Module is a researcher interface for retrieving pharmacy orders after the Matching Cache has been created by the Matching Module. The user inputs a list of drug keywords related to the pharmacy orders they want returned (eg, generic names, brand names, ATC codes). RxNorm is then used to determine the list of concept identifiers (CUI) that relate to the user input. The pharmacy orders with potential matches to the CUI list are then identified through a process of back-matching. Outputs are condensed and validated by manual review. False positives may be altered in the Matching Cache to improve precision in future queries.

information where available. This annotation process was based on exact match requirements to the chosen drug categories. Unique combinations of variables were displayed as rows and lines for mapping. The number of occurrences of each unique combination was also provided to offer additional context with respect to the most common medication orders. Categories of interest were defined by clinical (physician and pharmacist) expert opinion, and chosen because of their helpfulness in answering specific research questions.

The GEMINI-RxNorm system was run once on the same subset of manually coded data. It queried the data for the same drug names as the manual mappings and had no restrictions on the minimum match score required to make a match.

No manual revisions were made to the GEMINI-RxNorm outputs to ensure that only the automated standardization procedure would be evaluated. We then calculated the F1-score, precision, and recall of the outputs by comparing to the gold standard. Analyses were performed using R version 4.0.1. Research ethics board approval for this study was obtained from all participating sites.

RESULTS

The GEMINI-RxNorm system performed matching on 2 090 155 medication orders, occurring during 245 258 hospital admissions at 7 hospitals between April 1, 2010 and

Table 2. Drug classifications used for validation

Drug classification	Drugs included
Insulin	Anatomical Therapeutic Chemical (ATC) code: A10A
Non-insulin glucose lowering drugs	ATC: A10B
Vitamin K antagonists	ATC: B01AA
P2Y12 inhibitors	Clopidogrel, ticagrelor, prasugrel
Ace inhibitors/Angiotensin II receptor blockers	ATC: C09A, C09C
Benzodiazepines	ATC: N05BA
Direct-acting oral anticoagulants	ATC: B01AF
Antipsychotics	ATC: N05A
Dementia medications	Donepezil, memantine, galantamine, rivastigmine
Furosemide	furosemide (Lasix)
Corticosteroids	Dexamethasone, hydrocortisone, methylprednisolone, prednisolone, prednisone, deflazacort, cortisone
Puffers (limited to inhalation route)	Albuterol, salbutamol, levalbuterol, terbutaline, Proventil, ProAir, AccuNeb, pirbuterol, ventolin, formoterol, salmeterol, indacaterol, arformoterol, olodaterol, orciprenaline, ipratropium, umeclidinium, glycopyrrolate, glycopyrronium, tiotropium, aclidinium, budesonide, fluticasone, beclomethasone, ciclesonide, mometasone, albuterol/ipratropium, albuterol/ipratropium, umeclidinium/vilanterol, olodaterol/tiotropium, glycopyrronium/indacaterol, formoterol/glycopyrronium, aclidinium/formoterol, tiotropium/olodaterol, fluticasone/salmeterol, fluticasone/vilanterol, budesonide/formoterol, beclomethasone, salmeterol, formoterol/mometasone, beclomethasone/formoterol/glycopyrronium, fluticasone/umeclidinium/vilanterol, budesonide/glycopyrronium/formoterol, ibudilast, montelukast, pranlukast, zafirlukast, roflumilast, aminophylline, doxofylline, theophylline
Antibiotics	Ceftriaxone, cefotaxime, cefepime, cefdinir, cefditoren, cefpodoxime, ceftaroline, azithromycin, clarithromycin, erythromycin, streptomycin, levofloxacin, moxifloxacin, ciprofloxacin, gemifloxacin, doxycycline, amoxicillin-clavulanic acid, ampicillin-sulbactam, ticarcillin-clavulanate, piperacillin-tazobactam, meropenem, imipenem, imipenem+cilastatin, ertapenem, vancomycin (excluding oral, rectal, ophthalmic routes), aztreonam, colistin, gentamicin, septrin, trimethoprim-sulfamethoxazole, cefazolin, cefprozil, cefuroxime, cephalixin, penicillin G, amoxicillin, ticarcillin, flucloxacillin, ampicillin, piperacillin

Notes: This table defines the input drugs used in each query. Combination products were listed separately. Items that list an ATC code indicate that the entire list of generic drugs as specified by ATC were used as input.

October 31, 2017. The dataset included 27 473 unique medication name entries and 22 824 unique DIN numbers. The system categorized the medication orders into 29 249 distinct RxNorm CUIs.

Validation

The pharmacy-order level results (Table 3) reflect the real-world performance of the system when retrieving individual GEMINI pharmacy data for the given queries. In total, GEMINI-RxNorm successfully returned 1 941 389 of the 1 948 817 (99.62%) manually identified orders. The recall (sensitivity) of the GEMINI-RxNorm system was above 98.5% for all medication classes and the F1-score was above 0.95 in all drug categories except steroids (0.92) and antibiotics (0.90). With minimal manual review to discard false positives, precision of 100.0% can be achieved. The majority of false positives were caused by orders that included drugs with a similar string-name to a drug concept related to a queried drug. For example, “Ciprallex” (escitalopram) orders were assigned a 50% match score to the brand name “Ciprodex” leading these orders to be incorrectly matched to the antibiotic ciprofloxacin. Cases such as these could be avoided by setting a minimum match score above 50% but doing so could potentially lower system’s recall (Figure 3). Some medications in source data from hospitals are described using only brand names, and therefore including brand name identifiers that match to all the related RxNorm concepts increases the sensitivity of the system. There was a marked improvement in precision at match scores of 50% with relatively little tradeoff in recall. Medication orders that the system could not match

were orders where the only drug-identifying information was a Canada-specific drug name such as “Gravol” which RxNorm cannot recognize.

The validation of the GEMINI-RxNorm system was based on exact match requirements to a specific medication grouping and route of medication delivery. Incorrect routes were considered to be false positive matches, for example, Ciprofloxacin ear drops were considered a false positive when searching for systemic antibiotics. In these cases, the limitation of losing route information affected the precision but not the recall.

Time savings

To illustrate the time saved by the GEMINI-RxNorm system compared with manual medication mapping, we estimated that the time required to manually map a single row of medication data to a drug category was 30 s whereas the time required to manually verify suggested medication category matches through GEMINI-RxNorm was 5 s per row. To identify insulin medications, a manual reviewer might need to check up to all 2 090 155 pharmacy orders whereas after application of GEMINI-RxNorm, manual review was only required for 662 consolidated rows of data.

DISCUSSION

This article describes the development, implementation, and extensive validation of GEMINI-RxNorm, a medication data standardization and exploration system that uses a novel combination of RxNorm tools and external datasets. It is

Table 3. Validation results

Drug class	Number of orders	F1-score	Precision	Recall
Insulin	TP: 141 349 FP: 1910 FN: 1580	0.988	98.7	98.9
Non-insulin glucose lowering drugs	TP: 106 463 FP: 9480 FN: 39	0.957	91.8	>99.9
Vitamin K antagonists	TP: 100 023 FP: 4632 FN: 0	0.977	95.6	100
P2Y12 inhibitors	TP: 42 720 FP: 0 FN: 0	1.00	100	100
Ace inhibitors/Angiotensin II receptor blockers	TP: 133 398 FP: 6548 FN: 86	0.976	95.3	99.9
Benzodiazepines	TP: 167 577 FP: 2268 FN: 1465	0.989	98.7	99.1
Direct-acting oral anticoagulants	TP: 22 970 FP: 82 FN: 0	0.998	99.6	100
Antipsychotics	TP: 149 213 FP: 128 FN: 287	0.999	99.9	99.8
Dementia meds	TP: 16 262 FP: 0 FN: 0	1.00	100	100
Furosemide	TP: 220 422 FP: 14 FN: 52	>0.99	>99.9	>99.9
Steroids	TP: 147 470 FP: 26 286 FN: 0	0.918	84.9	100
Puffers	TP: 242 622 FP: 22 232 FN: 3261	0.950	91.6	98.7
Antibiotics	TP: 450 900 FP: 97 906 FN: 658	0.902	82.2	99.9

Notes: Results of the validation for the 13 drug classification queries. For each query, the pharmacy orders returned by GEMINI-RxNorm (with no limitations on match scores) were compared with the gold standard manual mappings. TP, true positive; FP, false positive; FN, false negative.

primarily designed to permit the querying and exploration of unstandardized data to enable research and other analyses. By creating a separate cache which matches row-level data in the underlying dataset to standardized RxNorm concepts, the tool also serves a data standardization function, but does so without altering the underlying data elements.

GEMINI-RxNorm demonstrates a flexible approach to medication standardization that can support a variety of input types, data formats, quality, and coverage that may be found when aggregating raw medication orders from multiple sources. The system was compared with manual expert mappings of 13 drug classes from 7 Canadian hospital sites over 8 years. It was found to have recall greater than 98.5% and an F1-score above 0.90 in all classes. GEMINI-RxNorm enabled substantial time savings compared with full manual standardization, reducing the time taken to review a pharmacy order row from an estimated 30 to 5 s and reducing the number of rows needed to be reviewed by up to 99.99%, from 2 090 155 rows to 662. Our experience suggests that the GEMINI-RxNorm system can be used independently to

efficiently extract and standardize pharmacy data with a high degree of accuracy. With minimal additional manual validation, researchers can achieve nearly perfect accuracy of standardized medication data in multisite patient data repositories.

Data standardization in clinical research repositories is crucial as it can have major impacts on research outcomes and policy decisions. Zhou et al¹⁸ describe an automated method to map the Partners Master Drug Dictionary (MDD) to RxNorm concepts using the MTERMS NLP tool. Similarly, Jiang et al¹⁹ standardized medication information in clinical text using MedEx. These studies focus on point-in-time mappings with single-source inputs. Less has been published on defining scalable frameworks for use in growing research repositories. Other drug data standardization efforts have involved mapping specific drug categories to RxNorm concepts. Klevens et al⁵ categorized outpatient antibiotic prescription claims and Dhavle et al⁴ evaluated 49 997 ambulatory e-prescription claims using RxNorm. These studies had a limited range of data inputs and were mapped using NDCs provided in their starting data without support for free-text parsing. We have been unable to find any studies that combine and evaluate several RxNorm tools in tandem with the intention of mapping all drug categories and supporting the continuous standardization that is required by ongoing data collection in growing research repositories. We describe the implementation of RxNorm in a health system outside of the United States, and by defining processes to link Canadian DINs to RxNorm concepts, the GEMINI-RxNorm framework can be easily extended for other international applications such as the central vocabulary service *Athena* (<https://athena.ohdsi.org/>). The GEMINI-RxNorm system could also benefit from the Canadian Drug Ontology OCRx⁷ by enabling it to process French data inputs. Although optimized for use within the GEMINI dataset, the GEMINI-RxNorm system is flexible and can be implemented in medication repositories with different types of input data as it is designed to extract drug-identifying concepts from a wide range of fields.

Some limitations remain in the GEMINI-RxNorm system. The RxNorm functions that we used do not support non-US drug concepts and therefore, it was not possible to match some non-US medication brand names to RxNorm concepts. This represents an opportunity for future work involving custom NLP tools or efforts to expand RxNorm to international medications such as with OHDSI's RxNorm Extension. Additionally, GEMINI-RxNorm allows users to search using classifications such as ATC, but doing so will not allow for specific medication routes to be searched. GEMINI-RxNorm was able to identify drugs even when abbreviated or obscured among additional information, but it can misidentify similar-sounding drugs and does not take into account a complete view of a drug order that a human may find. For example, some furosemide orders included the string "Hold Lasix for Today," indicating that Lasix was not ordered. However, GEMINI-RxNorm returned this row in the furosemide query as it only saw the word "Lasix." Our experience indicates that manual data validation is still necessary to resolve these cases, and thus we designed a module to facilitate this process.

The tool does not achieve complete interoperability as it cannot enable analysis on data more granular than the standardized results. For example, it does not support querying by medication indication due to limitations in converting

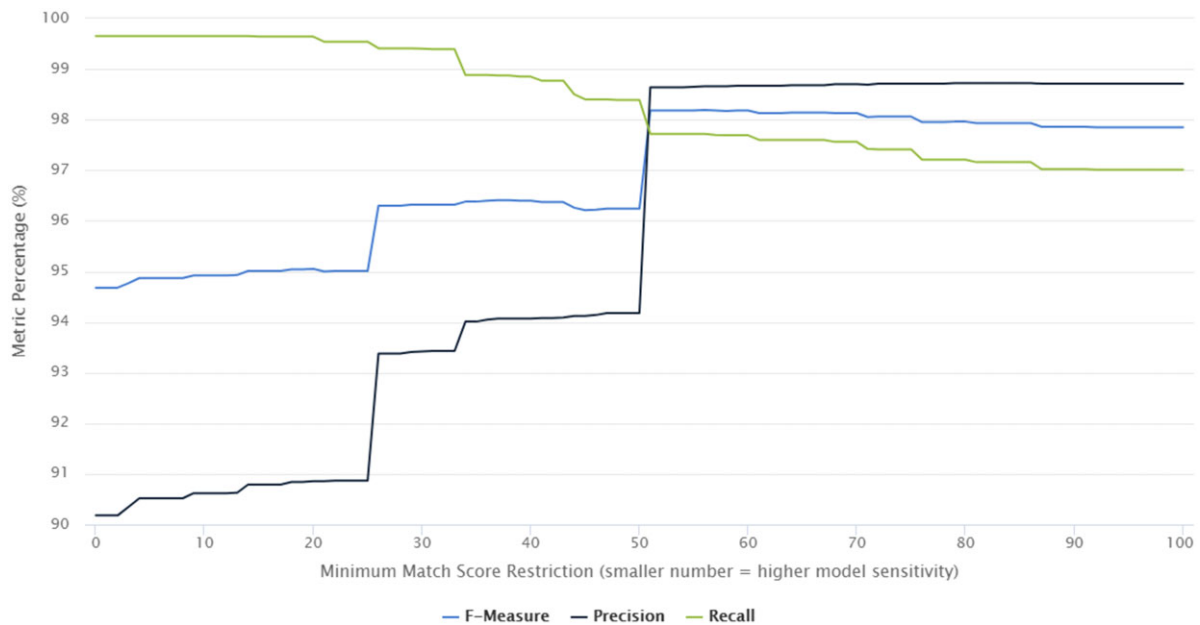


Figure 3. Gemini-RxNorm performance at different match score levels. This chart displays the performance metrics of GEMINI-RxNorm when run using different “minimum match scores.” The match score is a number between 0 and 100 that RxNorm uses to indicate how closely a free-text string matches an RxNorm concept. The x-axis in this chart represents the minimum match score needed for GEMINI-RxNorm to return the match. The y-axis represents the percentage metrics of *F*-measure (F1-score), precision, and recall obtained by comparing the pharmacy orders returned by GEMINI-RxNorm to the 1 948 817 gold standard manually mapped pharmacy orders.

identifiers. Starting with automated standardization, maximizing the system’s recall, and then condensing the outputs for manual review can minimize the manual workload while maintaining data quality. Any false positives flagged by the reviewer can then be removed from the matching database so that future queries do not make the same mismatch. The drug classes we validated were chosen to represent a wide range of medications commonly used in research, but many drug classifications were not validated. Manual annotation of medication data was performed by a physician and pharmacist with strong clinical background and subject matter expertise, but who did not have specific training in biomedical informatics. The non-ATC medication categories that we used (eg, “antibiotics” in Table 2) were not meant to be generalizable or comprehensive but to highlight how a user could retrieve a custom list of medications using our system. We believe that the GEMINI-RxNorm tool is likely to perform with the same excellent recall/sensitivity for most common medications, given its consistent performance across the wide range of classes that were validated. Finally, the estimated time savings associated with GEMINI-RxNorm is based on the reduced number of rows of medication data that require manual review, and is imprecise. The real mapping procedure for the validation data occurred over multiple sessions and days without being timed, which limited how accurately we were able to estimate time savings.

CONCLUSION

The GEMINI-RxNorm system is a comprehensive, flexible, scalable, and highly accurate automated pipeline for drug standardization in multisite patient data repositories. Extensive manual validation demonstrates consistently excellent recall and very good precision for medications across a wide range of medication classes. Thus, with limited additional

manual validation, the GEMINI-RxNorm system can allow researchers to achieve near-perfect accuracy in medication data standardization.

CODE AVAILABILITY

The GEMINI Rxnorm function is publicly available at <https://github.com/GEMINI-Medicine/RxNorm>.

FUNDING

This work was supported by the Digital Research Alliance of Canada through the “Data Champions Grant” and the “NDRIO-Portage COVID-19 Data Curation Funding.”

AUTHOR CONTRIBUTIONS

RW, FR, and AAV conceptualized the study. RW, DM, SS, HYJ, and MAI performed data collection and analysis. SM and SL performed manual validation and provided clinical subject matter expertise. All authors critically revised the manuscript for important intellectual content.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

DATA AVAILABILITY

Data from this manuscript can be accessed upon request to the corresponding author, to the extent that is possible in compliance with local research ethics board requirements and data sharing agreements. The GEMINI Rxnorm function is publicly available at <https://github.com/GEMINI-Medicine/RxNorm>

REFERENCES

1. Campion TR, Craven CK, Dorr DA, *et al.* Understanding enterprise data warehouses to support clinical and translational research. *J Am Med Inform Assoc* 2020; 27 (9): 1352–8.
2. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc* 2012; 19 (e1): e119–24.
3. Klevens RM, Caten E, Olesen SW, DeMaria A, Troppy S, Grad YH. Outpatient antibiotic prescribing in Massachusetts, 2011–2015. *Open Forum Infect Dis* 2019; 6 (5): ofz169.
4. Dhavle AA, Ward-Charlerie S, Rupp MT, Kilbourne J, Amin VP, Ruiz J. Evaluating the implementation of RxNorm in ambulatory electronic prescriptions. *J Am Med Inform Assoc* 2016; 23 (e1): e99–107.
5. Hoopes M, Angier H, Raynor LA, *et al.* Development of an algorithm to link electronic health record prescriptions with pharmacy dispense claims. *J Am Med Inform Assoc* 2018; 25 (10): 1322–30.
6. McDonough CW, Smith SM, Cooper-DeHoff RM, Hogan WR. Optimizing antihypertensive medication classification in electronic health record-based data: Classification system development and methodological comparison. *JMIR Med Inform* 2020; 8 (2): e14777.
7. Nikiema JN, Liang MQ, Després P, Motulsky A. *OCRx: Canadian Drug Ontology*. IOS Press; 2021. <https://ebooks.iospress.nl/doi/10.3233/SHTI210182>. Accessed August 2, 2023.
8. RxNorm Technical Documentation [Internet]. National Library of Medicine. Available from: <https://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html>
9. Warnekar PP, Bouhaddou O, Parrish F, *et al.* Use of RxNorm to exchange codified drug allergy information between Department of Veterans Affairs (VA) and Department of Defense (DoD). In: Annual Symposium proceedings, 2007: 781–5; AMIA.
10. RxNorm API [Internet]. National Library of Medicine. Available from: <https://rxnav.nlm.nih.gov/RxNormAPIs.html>. Accessed August 2, 2023.
11. Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. *AMIA. Annu Symp Proceedings AMIA Symp* 2011; 2011: 1117–26.
12. Freimuth RR, Wix K, Zhu Q, Siska M, Chute CG. Evaluation of RxNorm for medication clinical decision support. *AMIA. Annu Symp Proceedings AMIA Symp* 2014; 2014: 554–63.
13. Verma AA, Guo Y, Kwan JL, *et al.* Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open* 2017; 5 (4): E842–9.
14. Verma AA, Pasricha SV, Jung HY, *et al.* Assessing the quality of clinical and administrative data extracted from hospitals: The General Medicine Inpatient Initiative (GEMINI) experience. *J Am Med Inform Assoc* 2021; 28 (3): 578–87.
15. Drug Identification Number (DIN) [Internet]. Government of Canada. Available from: <https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/fact-sheets/drug-identification-number.html#shr-pg0>.
16. National Drug Code Directory [Internet]. U.S. Food & Drug Administration. Available from: <https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory>.
17. Drug Product Database: Access the database [Internet]. Government of Canada. Available from: <https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database.html>.
18. Zhou L, Plasek JM, Mahoney LM, Chang FY, DiMaggio D, Rocha RA. Mapping partners master drug dictionary to RxNorm using an NLP-based approach. *J Biomed Inform* 2012; 45 (4): 626–33.
19. Jiang M, Wu Y, Shah A, Priyanka P, Denny JC, Xu H. Extracting and standardizing medication information in clinical text—the MedEx-UIMA system. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 37–42.