Canadian Journal on Aging / La Revue canadienne du vieillissement

www.cambridge.org/cjg

Drug and Natural Health Product Data Collection and Curation in the Canadian Longitudinal Study on Aging

Benoit Cossette[1], Lauren Griffith[2], Patrick D. Emond[3], Dee Mangin[2], Lorraine Moss[3], Jennifer Boyko[3], Kathryn Nicholson[4], Jinhui Ma[2], Parminder Raina[2], Christina Wolfson[5], Susan Kirkland[6] and Lisa Dolovich[7]

[1]Department of Community Health Sciences, University of Sherbrooke, Sherbrooke, QC, Canada, [2]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada, [3]Canadian Longitudinal Study on Aging, Hamilton, ON, Canada, [4]Department of Epidemiology & Biostatistics, Western University, London, ON, Canada, [5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada, [6]Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada and [7]Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

Article

Cite this article: Cossette, B., Griffith, L., Emond, P.D., Mangin, D., Moss, L., Boyko, J., Nicholson, K., Ma, J., Raina, P., Wolfson, C., Kirkland, S., & Dolovich, L. (2024). Drug and Natural Health Product Data Collection and Curation in the Canadian Longitudinal Study on Aging. *Canadian Journal on Aging / La Revue canadienne du vieillissement* https://doi.org/10.1017/S0714980823000806

Received: 24 August 2023
Accepted: 07 December 2023

Mots-clés:
médicaments; produits de santé naturels; ELCV; aînés; cohorte longitudinale; vieillissement

Keywords:
medication; natural health product; CLSA; older adults; longitudinal cohort; aging

Corresponding author:
La correspondance et les demandes de tirésàpart doivent être adressées à : / Correspondence and requests for offprints should be sent to: Benoit Cossette, Research Center on Aging, CIUSSS de l'Estrie-CHUS, 1036, Belvedere South, Sherbrooke, QC J1H 4C4 (benoit.cossette@usherbrooke.ca).

The original version of this article was missing details in the Acknowledgements section. A notice detailing this has been published and the information added to the online PDF and HTML versions.

Résumé

Cette étude visait à développer un processus efficace de collecte et de recodage des données de tous les médicaments et produits de santé naturels (PSN) utilisés par les participants de l'Étude longitudinale canadienne sur le vieillissement (ELCV). Le processus séquentiel en trois étapes consistait à : 1) jumeler les médicaments colligés dans le cadre de l'étude avec les données de la Base de données sur les produits pharmaceutiques (BDPP) de Santé Canada, 2) recoder par algorithmes les médicaments et PSN non jumelés, et 3) recoder manuellement les médicaments et PSN non jumelés. Parmi les 30 097 participants de la cohorte globale de l'ELCV, 26 000 (86,4 %) utilisaient un médicament ou un PSN avec une moyenne de 5,3 (écart-type 3,8) médicaments ou PSN par participant-utilisateur pour un total de 137 366 médicaments ou PSN. Parmi ces médicaments ou PSN, 70 177 (51,1 %) ont été jumelés avec la BDPP de Santé Canada, 20 729 (15,1 %) ont été recodés par des algorithmes et 44 108 (32,1 %) ont été recodés manuellement. L'algorithme Direct a correctement classé 99,4 % des médicaments et 99,5 % des PSN. Nous avons développé un processus efficace en trois étapes pour la collecte et le recodage de médicaments et de PSN dans une cohorte longitudinale.

Abstract

This study aimed to develop an efficient data collection and curation process for all drugs and natural health products (NHPs) used by participants to the Canadian Longitudinal Study on Aging (CLSA). The three-step sequential process consisted of (a) mapping drug inputs collected through the CLSA to the Health Canada Drug Product Database (DPD), (b) algorithm recoding of unmapped drug and NHP inputs, and (c) manual recoding of unmapped drug and NHP inputs. Among the 30,097 CLSA comprehensive cohort participants, 26,000 (86.4%) were using a drug or an NHP with a mean of 5.3 (SD 3.8) inputs per participant user for a total of 137,366 inputs. Of those inputs, 70,177 (51.1%) were mapped to the Health Canada DPD, 20,729 (15.1%) were recoded by algorithms, and 44,108 (32.1%) were manually recoded. The Direct algorithm correctly classified 99.4 per cent of drug inputs and 99.5 per cent of NHP inputs. We developed an efficient three-step process for drug and NHP data collection and curation for use in a longitudinal cohort.

Background and context

Large databases of health information are an important resource to study the use and outcomes of health services including the use of medications (Cadarette & Wong, 2015; Metge et al., 2005; Murdoch & Detsky, 2013; Schneeweiss & Avorn, 2005; Zhan & Miller, 2003). Information on the prevalence, incidence, and duration of drug therapy is important in health research, health system planning, and assessment of appropriate prescribing for treatment patterns and burden (Galvin et al., 2014; Moriarty et al., 2015a, 2015b; Schneeweiss & Avorn, 2005). Moreover, as the global population of adults 65 years and older continues to grow, the need will also grow for timely and accurate information not only for prescribed medications but also for non-prescription medications and natural health product (NHP). Standardized coding and classification of medication data can improve the efficiency in data collection and curation processes, which are complex processes due to heterogeneous formats including generic names (e.g.,

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press

acetaminophen), trade names (e.g., Tylenol), and numeric drug identifiers (e.g., 02046040) (Nikiema et al., 2021; Richesson, 2014).

The mapping of medication data to standardized terminologies such as the RxNorm ontology (RxNorm, n.d.) has been proposed to allow efficient analysis and interpretation of drug data (Nikiema et al., 2021; Richesson, 2014). The performance of this mapping to standardized terminologies has been evaluated with medication data from hospital pharmacy systems (Hernandez et al., 2009; Waters et al., 2023), electronic health records (Zhou et al., 2012), drug adverse events database (Veronin et al., 2020), multi-site clinical trial (Lockery et al., 2019; Richesson et al., 2010), and longitudinal cohorts (Richesson et al., 2010). For prospective clinical studies, the ASPREE (Lockery et al., 2019) clinical trial in older adults and the 45 and Up study (Gnjidic et al., 2015) reported a method of structured medication data collection based on a list of common medications with the option of free-text data entry for other medications. Both studies used a structured process of automated and manual coding for the curation of the free-text data by medication experts (Gnjidic et al., 2015; Lockery et al., 2019). Systematic approaches for the curation of large free-text medication data have involved automated and manual approaches (Richesson, 2014; Veronin et al., 2020).

The Canadian Longitudinal Study on Aging (CLSA) is a population-based research platform established to better understand how biological, medical, psychological, and social determinants have an impact in maintaining health and in the development of disease and disability as people age (P. Raina et al., 2019; P. S. Raina et al., 2009). The complete documentation of all drugs and NHP used every 3 years over 20 years in a cohort of more than 30,000 participants requires efficient data mapping and curation processes. In this article, we describe a three-step process for the data entry and mapping of drug data to the Health Canada Drug Product Database (DPD) by CLSA interviewers, as well as the development and validation of a cleaning process of free-text/numeric drug and NHP inputs in a software algorithm approach followed by manual recoding.

Methods

Study population

The recruitment and baseline evaluations of the 51,338 CLSA participants aged 45–85 years at enrolment was completed in 2015 (P. Raina et al., 2019). The complete CLSA cohort is composed of the Tracking cohort of 21,241 participants who provide data via telephone interviews and the Comprehensive cohort of 30,097 participants who provide data via in-person home interviews and visits to a data-collection site. Comprehensive participants provided data in English and French on all regularly used drug and NHPs.

Drug and NHP data collection/mapping drug data to Health Canada database

In the first of a three-step process, Drug and NHP data were entered in the CLSA data collection software by interviewers who were trained to identify the relevant information from medication packaging (Figure 1). During an in-home visit, CLSA interviewers asked participants to present all regularly scheduled or taken medications (i.e., scheduled, once a day, every other day, taken occasionally, and as required), including prescription, non-prescription, over-the-counter, herbals, vitamins, or NHPs in all routes of administration. Information on study drugs and drugs commercialized in other countries than Canada was also collected. The interviewer entered either the generic name (e.g., atorvastatin), trade name (e.g., Lipitor), or drug identification number (DIN) (e.g., 02230711) in a type-to-search box that mapped the drug input to the Health Canada DPD and generated a list of corresponding generic or trade drug names. In the absence of adequate drug name correspondence, the name/DIN was entered as a free-text/numeric input. Since the type-to-search box was not mapped to the Health Canada Licensed Natural Health Products Database (LNHPD), NHP were entered as free-text/numeric inputs. The interviewer also recorded information about the dosage, frequency, duration, start date, and indications for use. Since dose and frequency data were gathered, strength was not collected.

Drugs authorized for sale by Health Canada are listed in the Health Canada DPD (Health Canada, n.d.a), which contains information notably on product name, list of active ingredients, DIN, and World Health Organization (WHO) anatomical therapeutic chemical (ATC) classification. NHP licensed by Health Canada are listed in the Health Canada LNHPD (Health Canada, n.d.b), which contains information notably on product name, product's medicinal ingredients, product's non-medicinal ingredients, and natural product number (NPN). The NHP database does not include ATC codes. Both databases are updated nightly.

Algorithm recoding

In a second step, sequential algorithms were applied to map free-text (drug or NHP names) or numeric (DINs or NPNs) inputs to the products of the Health Canada drug and NHP databases (Figure 1). Seven algorithms were developed in a software algorithm approach independent of the sample data (Table 1). The algorithms were run sequentially such that once an input was matched, it was no longer considered in the remaining algorithms. For a given input, the first algorithm attempted to map the input to the drug followed by the NHP database before moving on to the next algorithm. The Direct and Code algorithms were run first since they only ever matched a single input to a single drug or NHP, while the Word and Simple algorithms at times found multiple matches. In cases of multiple matches due to numerous dosage strengths, the input was matched to the suitable drug or NHP with the lowest DIN or NPN.

Work was conducted using SQL (database scripting language) and PHP (general programming language). The Health Canada databases and CLSA data were loaded into a secure MySQL database using SQL. Some pre-processing was conducted on these databases before using PHP to enhance performance, increase speed of matching, and make the computer algorithms more efficient. For instance, the Simple algorithm compared the unmapped inputs to drug and NHP names from the Health Canada databases by ignoring non-alpha-numeric characters. This was done by removing the non-alpha-numeric characters from both the unmapped inputs and the Health Canada databases names, then comparing the two. It would be slow to transform the drug names in this way every time a comparison is made. Instead, all drug names were electronically converted during this pre-process step once and used by the algorithm every time a match was searched for. Another example is a list that was made of all identical drug and NHP names. The final version of the algorithm sequence and variables from the Health Canada databases are presented in the Supplementary Material.

As part of an iterative algorithm improvement approach, two pharmacists (L.D. and B.C.) independently recoded 40 unmapped

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press

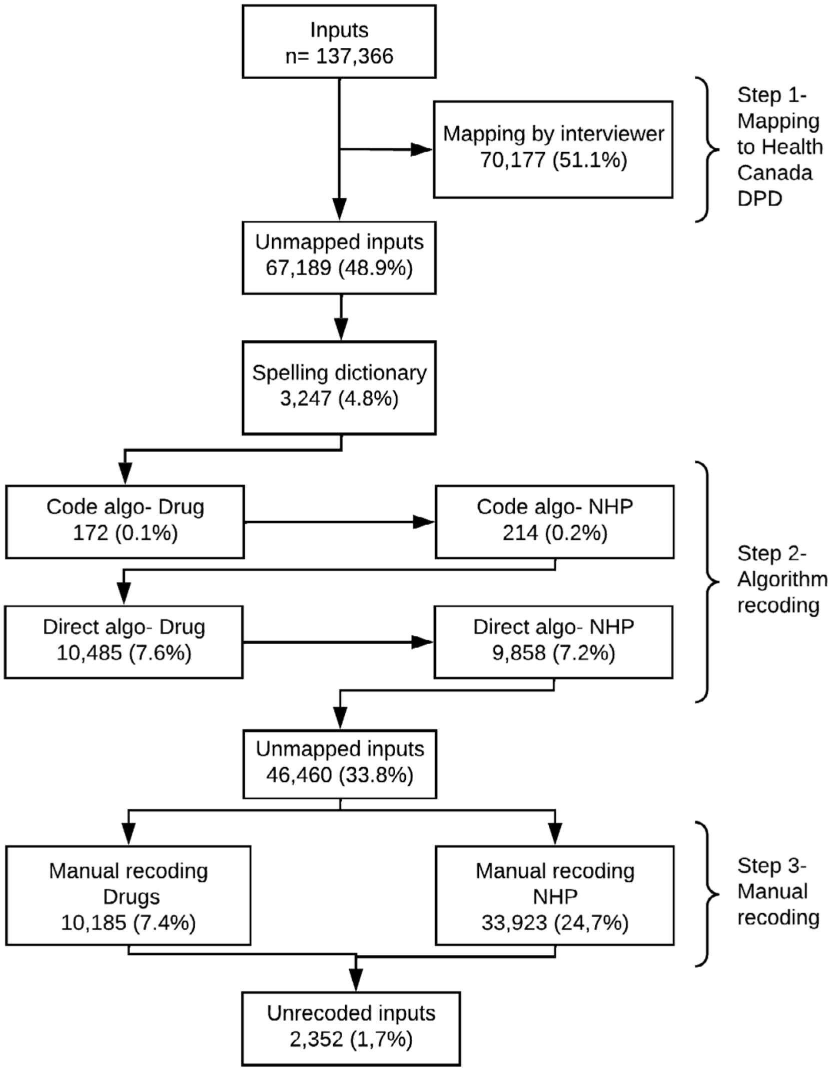

Figure 1. Input mapping by algorithm and manual processes in CLSA's baseline Comprehensive cohort participants.
Note: Algo = algorithm; DPD = Drug Product Database; NHP = natural health product.

drug and NHP inputs. The pharmacist-recoded inputs were compared to algorithm-recoded inputs during meetings of the research team, leading to algorithm refinement. This process of review – discussion – algorithm refinement was conducted three times for a total of 120 inputs, leading to two new algorithms: Predefined and No-units (Table 1). The greater complexity of recoding NHP inputs compared to drug inputs was identified early in this process and discussed throughout our work.

Manual recoding

In a third recoding step, following the application of the algorithms to the unmapped drug and NHP data, the remaining unmapped de-identified data were exported directly from the CLSA's database to an Excel file for manual recoding by three pharmacy technicians (Figure 1). The same group of recoders conducted the recoding and validation work. The recoders' work was supported by a set of decision rules (Supplementary Material) to assign selected NPNs for the most prevalent NHP inputs (e.g., NPN = 80083109 for calcium).

Spelling dictionary

As inputs were manually recoded, common misspellings were compiled into a dictionary and applied to future iterations of the computer algorithms. In the pre-processing stage, all inputs containing any of the misspelled words in the dictionary were replaced with the correct spelling before the algorithms were run (Figure 1).

Validation process

A validation sample of 100 Comprehensive cohort participants was randomly selected to evaluate the performance of the recoding algorithms and manual recoding. This sample included 352 free-text drug and NHP inputs for which a gold-standard recoded input was determined independently by two recoders with resolution of discrepancies by a pharmacist. A gold-standard recoded input could not be established for some inputs due to insufficient input information. Differing commercial products of the same generic drug or NHP were considered to be an agreement. After this first validation, the algorithms were further refined and validated in a second sample of 544 Comprehensive cohort participants with

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press

Table 1. Developed algorithms

Name	Description	Examples
Code	The input is compared to the DIN or NPN. A match is found when the input is identical to the DIN or NPN. There can only ever be one match.	The input '02275619' matches the DIN '02275619'. In comparison, the input '0227-5619' does not match the DIN '02275619'.
Direct	The input is compared to the drug or NHP's name. A match is found when the input is identical to the drug's or NHP's name (including all special characters and spaces). There can only ever be one match.	The input 'TYLENOL ALLERGY' matches the drug name 'TYLENOL ALLERGY'. In comparison, the input 'TYLENOL ALLERGY 100MG' does not match the drug name 'TYLENOL ALLERGY'.
Word	The input is compared to the drug's or NHP's name. A match is found when the drug's or NHP's name is found as a substring within the input. Spaces are considered such that only whole words can be matched. There may be multiple matches.	The input 'LARGE TYLENOL SUPER RELIEF 100MG' matches the drug name 'TYLENOL SUPER'. In comparison, the input 'LARGE TYLENOL SUPERIOR RELIEF 100MG' does not match the drug name 'TYLENOL SUPER'.
Simple	The input with all non-alpha-numeric characters removed is compared to the drug's or NHP's name with all non-alpha-numeric characters removed. A match is found when the two altered names are identical. There may be multiple matches.	The input 'TYLENOL-ALLERGY (50-MG)' (transformed into 'TYLENOLALLERGY50MG') matches the drug name 'TYLENOL ALLERGY 50MG' (transformed into 'TYLENOLALLERGY50MG'). In comparison, the input 'TYLENOL-ALLERGY (50-MG)' (transformed into 'TYLENOLALLERGY50MG') does not match the drug name 'TYLENOL ALLERGY' (transformed into 'TYLENOLALLERGY').
Reverse-word	This algorithm is identical to 'Word', but the input is searched as a substring within the drug or NHP.	
No-units	The input with all units of measurement removed.	The input 'ASPIRIN COATED CAPLETS 500MG' would have the units, 500MG, removed and become 'ASPIRIN COATED CAPLETS'.
Predefined	List of common drugs and NHPs established by our team. Inputs with predefined names would get coded first.	Aspirin, Vitamin B, Vitamin C, Vitamin D, multivitamin, etc.

Note: DIN = drug identification number; NHP = Natural Health Product; NPN = natural product number.

1,407 unmapped free-text drug and NHP inputs. In this second validation, the gold-standard recoded input was established by a single recoder based on the measured recoders consensus in the first validation.

Analysis

Manual recoding was considered the gold standard for free-text inputs. The proportion of algorithm-correctly recoded inputs was calculated as the number of algorithm-correctly recoded inputs, based on the gold standard, divided by the number of algorithm-recoded inputs. In the primary analysis, the denominator included only the inputs for which a gold standard could be established in order to distinguish between drug and NHP. In a sensitivity analysis, the denominator included all algorithm-recoded inputs, regardless of gold-standard coding, for a more conservative estimate that cannot differentiate between drug and NHP.

Ethics approval

The CLSA was approved by the Hamilton Integrated Research Ethics Board (approval number 10-423, for the Comprehensive cohort) at McMaster University and the research ethics boards of all collaborating institutions.

Results

Mapping and recoding of drug and NHP inputs

Among CLSA's 30,097 baseline Comprehensive cohort participants, 26,000 (86.4%) were using a drug or an NHP. Among drug or NHP users, a mean of 5.3 (SD 3.8) inputs per participant were documented for a total of 137,366 inputs. In the first of a three-step process, interviewers mapped 70,177 (51.1%) of the 137,366 inputs to a drug in the Health Canada DPD (Figure 1). Of the remaining 67,189 unmapped inputs (Figure 1), 3,247 (4.8%) were pre-processed by the spelling dictionary. In step 2, the Direct and Code algorithms recoded 10,657 (7.8%) drug and 10,072 (7.3%) NHP inputs. In step 3 (manual recoding), 10,185 (7.4%) drug and 33,923 (24.7%) NHP inputs out of the 46,460 (32.1%) remaining unmapped inputs were manually recoded (Figure 1). Insufficient input information resulted in an inability to code for 2,352 (1.7%) inputs (e.g., study drug and hypertension medication), made available to researchers as entered (Figure 1).

Algorithm and manual recoding validation

First validation sample

From the first validation sample, 352 free-text inputs were submitted to algorithm recoding and reviewed by two recoders (and pharmacist for non-consensus inputs) to establish a gold-standard recoded input. Of these 352 inputs, 12 free-text inputs were not recoded by the recoder nor the algorithms because of insufficient information. Of the remaining 340 inputs, 307 were recoded by the algorithms (Table 2). The Direct algorithm recoded the most (49.5%) inputs followed by the Word algorithm (22.5%). In the main analysis of the inputs for which a gold standard could be established, the Direct and Word algorithms correctly classified 97.9 per cent and 59.3 per cent of drugs and 96.2 per cent and 30.6 per cent of NHP inputs, respectively. In the sensitivity analysis of all algorithm-recoded inputs, the Direct and Word algorithms correctly classified 95.4 per cent and 39.1 per cent of inputs.

Of the 352 drug and NHP inputs, consensus was reached by both recoders for 294 (83.5%) inputs. Of these 352 inputs, the recoders agreed that there was insufficient information to recode 21 inputs, excluded from the following subgroup analysis. Of the remaining 329 inputs, consensus was reached by the recoders for 156 (89.7%) of the 174 drug inputs and for 116 (74.8%) of the

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press

Table 2. Validation of algorithm recoding with manual recoding (gold standard) – first validation sample

				Algorithm-correctly recoded inputs		
	Manual recoding (gold standard)			Primary analysis	Sensitivity analysis	
Algorithms	Not recoded	Drug	NHP	Drug	NHP	Drug or NHP
Direct	3	96	53	94	51	145
	(27.3%)	(65.3%)[a]	(35.6%)	(97.9%)[b]	(96.2%)[c]	(95.4%)[d]
No-units	0	0	2	–	1	1
	(0.0%)	(0.0%)	(1.3%)		(50.0%)	(50.0%)
Predefined	1	1	55	0	46	46
	(9.1%)	(0.7%)	(36.9%)	(0.0%)	(83.6%)	(80.7%)
Reverse-word	1	19	1	15	1	16
	(9.1%)	(12.9%)	(0.7%)	(78.9%)	(100%)	(76.2%)
Simple	0	4	2	4	2	6
	(0.0%)	(2.7%)	(1.3%)	(100%)	(100%)	(100%)
Word	6	27	36	16	11	27
	(54.5%)	(18.4%)	(24.2%)	(59.3%)	(30.6%)	(39.1%)
All	11	147	149	129	112	241
	(100%)	(100%)	(100%)	(87.8%)	(75.2%)	(78.5%)

Note: NHP = natural health product.

[a] Percent of the manually recoded drug inputs also recoded by the Direct algorithm out of the 147 manually recoded drug inputs.

[b] Percent of correctly recoded drug inputs by the Direct algorithm out of the 96 recoded drug inputs by the Direct algorithm.

[c] Percent of correctly recoded NHP inputs by the Direct algorithm out of the 53 recoded NHP inputs by the Direct algorithm.

[d] Percent of correctly recoded inputs by the Direct algorithm out of the 152 recoded inputs by the Direct algorithm.

155 NHP inputs. Based on these results, the second algorithms' validation was conducted with a gold standard established by a single recoder. The recoders' consensus was similar for algorithm-recoded inputs (83.4%) and non-algorithm-recoded inputs (84.4%).

Second validation sample

Of the 1,407 free-text inputs of the second validation sample, 27 were not recoded by the recoder nor the algorithms because of insufficient information. Of the remaining 1,380 inputs, 1,280 were recoded by the algorithms (Table 3). The Predefined algorithm recoded the most (44.8%) inputs followed by the Direct algorithm (29.0%). Modifications to the predefined algorithm for the coding of vitamins explains the increase in recoded inputs from the first to the second validation sample. In the main analysis of the inputs for which a gold standard could be established, the Direct and Pre-defined algorithms correctly classified 99.4 per cent and 86.4 per cent of drugs and 99.5 per cent and 78.2 per cent of NHP inputs, respectively. In the sensitivity analysis of all algorithm-recoded inputs, the Direct and Pre-defined algorithms correctly classified 94.6 per cent and 77.0 per cent of inputs. Following the second validation, the Code and Direct algorithms were selected for step-2 algorithm recoding of the unmapped free-text inputs of the baseline Comprehensive cohort participants.

Discussion

We described a three-step process for the mapping of drug and NHP data to Health Canada databases that included algorithm recoding of 15.1 per cent of all drug and NHP inputs with high confirmation against gold-standard manual recoding. The developed algorithms have and will continue to save significant manual recoding time considering the large volume of CLSA drug and NHP data collected every 3 years over 20 years. The three-step process will enable the medications data collected from CLSA participants to be curated more efficiently and released as part of the CLSA research data platform for use by researchers. The process has the potential to be tested and applied with other large studies.

In the first of the three-step process, CLSA in person interviewers mapped 51 per cent of 137,366 drug and NHP inputs to the Health Canada DPD. In CLSA, the mapping of all drugs and NHPs, a much more extensive and diverse data set, contrasts from the mapping to a selection of 2,025 common medications in the multinational, ASPREE clinical trial in older adults (Lockery et al., 2019) and to a list of the 32 most common medications used in the Australian population in the 45 and Up study (Gnjidic et al., 2015).

In the second mapping step, two (Code and Direct) of the seven developed algorithms were selected for algorithm recoding of unmapped drug and NHP inputs. The limited number of selected algorithms highlights the need for a validation process to identify the challenging inputs in a specific data set. In our final validation sample, the Direct algorithm correctly classified 99.4 per cent of drug and 99.5 per cent of NHP inputs among the inputs for which a gold standard could be established. Similar validations of drug mapping/recoding have been reported by other groups. In the 45 and Up study, the automated coding of drug terms first to generic names using the Systematized Nomenclature of Medicine – Clinical Terms followed by coding to the WHO – ATC classification achieved positive predictive values above 95 per cent and sensitivity of 79 per cent at the exact ATC level with higher sensitivity values for drugs than vitamins and supplements (Gnjidic et al., 2015). The cleaning of drug names in the Food

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press

Table 3. Validation of algorithm recoding with manual recoding (gold standard) – second validation sample

				Algorithm correctly recoded inputs		
	Manual recoding (gold standard)			Primary analysis	Sensitivity analysis	
Algorithms	Not recoded	Drug	NHP	Drug	NHP	Drug or NHP
Direct	18	171	182	170	181	351
	(26.1%)	(49.0%)[a]	(21.1%)	(99.4%)[b]	(99.5%)[c]	(94.6%)[d]
No-units	1	1	3	1	2	3
	(1.4%)	(0.3%)	(0.3%)	(100%)	(66.7%)	(60.0%)
Predefined	15	59	499	51	390	441
	(21.7%)	(16.9%)	(57.9%)	(86.4%)	(78.2%)	(77.0%)
Reverse-word	4	47	23	44	1	45
	(5.8%)	(13.5%)	(2.7%)	(93.6%)	(4.3%)	(60.8%)
Simple	3	1	14	1	13	14
	(4.3%)	(0.3%)	(1.6%)	(100%)	(92.9%)	(77.8%)
Word	28	70	141	64	101	165
	(40.6%)	(20.1%)	(16.4%)	(91.4%)	(71.6%)	(69.0%)
All	69	349	862	331	688	1019
	(5.4%)	(27.3%)	(67.3%)	(94.8%)	(79.8%)	(79.6%)

Note: NHP = natural health product.
[a] Percent of the manually recoded drug inputs also recoded by the Direct algorithm out of the 349 manually recoded drug inputs.
[b] Percent of correctly recoded drug inputs by the Direct algorithm out of the 171 recoded drug inputs by the Direct algorithm.
[c] Percent of correctly recoded NHP inputs by the Direct algorithm out of the 182 recoded NHP inputs by the Direct algorithm.
[d] Percent of correctly recoded inputs by the Direct algorithm out of the 371 recoded inputs by the Direct algorithm.

and Drug administration Adverse Event Reporting System (FAERS) database resulted in standardization of 95 per cent of drug name (Veronin et al., 2020). In another study on the FAERS database, drug name coverage of 93 per cent was achieved in the mapping to RxNorm standard code ingredients (Banda et al., 2016). With highly structured inpatient pharmacy data from the GEMINI database from seven Canadian hospitals over 8 years, the use of existing Rx-Norm functionality resulted in sensitivity greater than 98.5 per cent and an *F*-Measure above 90.0 per cent in the standardization of 13 selected drug classes (Waters et al., 2023).

In the third mapping step, 33.8 per cent of the remaining unmapped inputs were manually recoded with higher consensus for drug than NHP inputs. The mapping of the NHP inputs to Health Canada's LNHPD adequately documents the product name as recommended by the CONSORT statement on herbal interventions (Gagnier et al., 2006). It allows researchers using CLSA data to further detail the physical characteristics of the NHP such as the part of the plant used to produce the extract and the type of product used (e.g., fresh or dry) as suggested by the CONSORT statement. General NHP designations (e.g., multivitamins) were coded as per our decision rules.

Strengths and limitations

The main strength of our approach is the mapping/recoding of drug and NHP data to standardized information of Health Canada's Drug and NHP Databases. The availability of these regularly updated databases was essential to this project. This linkage included the WHO ATC categories for drugs, a derived variable particularly useful for researchers using CLSA data. Our sequential approach limited the manual recoding to 33.8 per cent drug and NHP inputs. The main limitation of our approach is in the initial free-text entry of all NHP inputs and the 74.8 per cent consensus during manual recoding. Also, our approach would need to be adapted for drug and NHP data collection in other countries because of varying names.

Making the CLSA drug and NHP data available to researchers

CLSA data are currently available to approved public sector researchers in Canada and elsewhere. The data application process is described on CLSA's website (http://www.clsa-elcv.ca), which also hosts the medication and NHP data support document providing a brief overview.

Ongoing developments

We continue to refine our collection and curation processes for medications data in the CLSA by exploring the linkage of the type-to-search box to Health Canada's LNHPD for the mapping of NHP information by CLSA interviewers. The multiple brand name extensions generating an important number of options that could increase interviewers' data collection time is a concern for NHP mapping. We are pursuing the refinement of the algorithms using new classification approaches and evaluating the integration of these refined algorithms to the type-to-search box to generate a list of possible matches to Health Canada's LNHPD and limit the need for manual recoding.

Conclusion

We created an efficient three-step sequential process for drug and NHP data collection and curation in a longitudinal cohort as shown by the mapping of half of the drug and NHP inputs by the interviewers and algorithm recoding of 15.1 per cent of inputs.

The accuracy of our approach was shown by the confirmation of algorithm coding compared to gold-standard manual recoding and recoders consensus for drug for the manual recoding process. Our approach has the potential to be applied by researchers using other large data sets requiring cleaning. We are pursuing the development of our approach for the data collection and mapping of NHP data to Health Canada's LNHPD and integrating the algorithms into the day-to-day working of the next set of follow-up data collection periods in the CLSA.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0714980823000806.

Acknowledgements. The authors would like to thank the participants who give their time to the Canadian Longitudinal Study on Aging. The authors would also like to thank Helga Weigelin and Bevonie Brown, who conducted the manual recoding, Jean-Philippe Turcotte and Claudie Rodrigue for the data analysis, and Joanne Ho, Carol Bassim, and Kasia Makara for their support in coordinating this work.

Financial support. Funding for the Canadian Longitudinal Study on Aging (CLSA) is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under grant reference LSA94473 and the Canada Foundation for Innovation, as well as the following provinces, Newfoundland, Nova Scotia, Quebec, Ontario, Manitoba, Alberta, and British Columbia. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The CLSA is led by Drs Parminder Raina, Christina Wolfson, and Susan Kirkland. B.C. is a Junior 1 Research Scholar from the Fonds de recherche du Québec – Santé. L.G. is supported by the McLaughlin Foundation Professorship in Population and Public Health. P.R. holds the Raymond and Margaret Labarge Chair in Optimal Aging and Knowledge Application for Optimal Aging, is the Director of the McMaster Institute for Research on Aging and the Labarge Centre for Mobility in Aging, and holds a Tier 1 Canada Research Chair in Geroscience.

References

Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., & Shah, N. H. (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific Data*, **3**, 160026. https://doi.org/10.1038/sdata.2016.26

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

Gagnier, J. J., Boon, H., Rochon, P., Moher, D., Barnes, J., Bombardier, C., & CONSORT Group. (2006). Recommendations for reporting randomized controlled trials of herbal interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, **59**(11), 1134–1149. https://doi.org/10.1016/j.jclinepi.2005.12.020

Galvin, R., Moriarty, F., Cousins, G., Cahir, C., Motterlini, N., Bradley, M., Hughes, C. M., Bennett, K., Smith, S. M., Fahey, T., & Kenny, R.-A. (2014). Prevalence of potentially inappropriate prescribing and prescribing omissions in older Irish adults: Findings from The Irish LongituDinal Study on Ageing study (TILDA). *European Journal of Clinical Pharmacology*, **70**(5), 599–606. https://doi.org/10.1007/s00228-014-1651-8

Gnjidic, D., Pearson, S.-A., Hilmer, S. N., Basilakis, J., Schaffer, A. L., Blyth, F. M., Banks, E., & High Risk Prescribing Investigators. (2015). Manual versus automated coding of free-text self-reported medication data in the 45 and Up Study: A validation study. *Public Health Research & Practice*, **25** (2), e2521518. https://doi.org/10.17061/phrp2521518

Health Canada. (n.d.a). *Health Canada Drug Product Database.* www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database.html.

Health Canada. (n.d.b). *Health Canada Licensed Natural Health Products Database.* https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/applications-submissions/product-licensing/licensed-natural-health-products-database.html.

Hernandez, P., Podchiyska, T., Weber, S., Ferris, T., & Lowe, H. (2009). Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse. *AMIA … Annual Symposium Proceedings. AMIA Symposium*, **2009**, 244–248.

Lockery, J. E., Rigby, J., Collyer, T. A., Stewart, A. C., Woods, R. L., McNeil, J. J., Reid, C. M., Ernst, M. E., & Group, on behalf of the ASPREE Investigator Group (2019). Optimising medication data collection in a large-scale clinical trial. *PLOS ONE*, **14**(12), e0226868. https://doi.org/10.1371/journal.pone.0226868

Metge, C., Grymonpre, R., Dahl, M., & Yogendran, M. (2005). Pharmaceutical use among older adults: Using administrative data to examine medication-related issues. *Canadian Journal on Aging = La Revue Canadienne Du Vieillissement*, **24**(Suppl 1), 81–95. https://doi.org/10.1353/cja.2005.0052

Moriarty, F., Bennett, K., Fahey, T., Kenny, R. A., & Cahir, C. (2015a). Longitudinal prevalence of potentially inappropriate medicines and potential prescribing omissions in a cohort of community-dwelling older people. *European Journal of Clinical Pharmacology*, **71**(4), 473–482. https://doi.org/10.1007/s00228-015-1815-1

Moriarty, F., Hardy, C., Bennett, K., Smith, S. M., & Fahey, T. (2015b). Trends and interaction of polypharmacy and potentially inappropriate prescribing in primary care over 15 years in Ireland: A repeated cross-sectional study. *BMJ Open*, **5**(9), e008656. https://doi.org/10.1136/bmjopen-2015-008656

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA*, **309**(13), 1351–1352. https://doi.org/10.1001/jama.2013.393

Nikiema, J. N., Liang, M. Q., Després, P., & Motulsky, A. (2021). OCRx: Canadian Drug Ontology. *Studies in Health Technology and Informatics*, **281**, 367–371. https://doi.org/10.3233/SHTI210182

Raina, P., Wolfson, C., Kirkland, S., Griffith, L. E., Balion, C., Cossette, B., Dionne, I., Hofer, S., Hogan, D., van den Heuvel, E. R., Liu-Ambrose, T., Menec, V., Mugford, G., Patterson, C., Payette, H., Richards, B., Shannon, H., Sheets, D., Taler, V., … Young, L. (2019). Cohort Profile: The Canadian Longitudinal Study on Aging (CLSA). *International Journal of Epidemiology*, **48**(6), 1752–1753j. https://doi.org/10.1093/ije/dyz173

Raina, P. S., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C. M., Hogan, D., Wister, A., Payette, H., Shannon, H., & Brazil, K. (2009). The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging = La Revue Canadienne Du Vieillissement*, **28**(3), 221–229. https://doi.org/10.1017/S0714980809990055

Richesson, R. L. (2014). An informatics framework for the standardized collection and analysis of medication data in networked research. *Journal of Biomedical Informatics*, **52**, 4–10. https://doi.org/10.1016/j.jbi.2014.01.002

Richesson, R. L., Smith, S. B., Malloy, J., & Krischer, J. P. (2010). Achieving standardized medication data in clinical research studies: Two approaches and applications for implementing RxNorm. *Journal of Medical Systems*, **34** (4), 651–657. https://doi.org/10.1007/s10916-009-9278-5

RxNorm. (n.d.). *RxNorm. National library of medicine.* www.nlm.nih.gov/research/umls/rxnorm/index.html.

Schneeweiss, S., & Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, **58**(4), 323–337. https://doi.org/10.1016/j.jclinepi.2004.10.012

Veronin, M. A., Schumaker, R. P., Dixit, R. R., Dhake, P., & Ogwo, M. (2020). A systematic approach to 'cleaning' of drug name records data in the FAERS database: A case report. *International Journal of Big Data Management*, **1**(2), 105–118. https://doi.org/10.1504/IJBDM.2020.112404

Waters, R., Malecki, S., Lail, S., Mak, D., Saha, S., Jung, H. Y., Imrit, M. A., Razak, F., & Verma, A. A. (2023). Automated identification of unstandardized medication data: A scalable and flexible data standardization pipeline using RxNorm on GEMINI multicenter hospital data. *JAMIA Open*, **6**(3), ooad062. https://doi.org/10.1093/jamiaopen/ooad062

Zhan, C., & Miller, M. R. (2003). Administrative data-based patient safety research: A critical review. *Quality & Safety in Health Care*, **12**(Suppl 2), ii58–ii63. https://doi.org/10.1136/qhc.12.suppl_2.ii58

Zhou, L., Plasek, J. M., Mahoney, L. M., Chang, F. Y., DiMaggio, D., & Rocha, R. A. (2012). Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. *Journal of Biomedical Informatics*, **45**(4), 626–633. https://doi.org/10.1016/j.jbi.2011.11.006

https://doi.org/10.1017/S0714980823000806 Published online by Cambridge University Press