public health research & practice

March 2015; Vol. 25(2):e2521518
doi: http://dx.doi.org/10.17061/phrp2521518
www.phrp.com.au

Research

Manual versus automated coding of free-text self-reported medication data in the 45 and Up Study: a validation study

Danijela Gnjidic[a,b,i], Sallie-Anne Pearson[a,c], Sarah N Hilmer[b,d], Jim Basilakis[e], Andrea L Schaffer[a], Fiona M Blyth[b,f,g] and Emily Banks[g,h], on behalf of the High Risk Prescribing Investigators

[a] Faculty of Pharmacy, University of Sydney, NSW, Australia
[b] Sydney Medical School, University of Sydney, NSW, Australia
[c] Sydney School of Public Health, University of Sydney, NSW, Australia
[d] Royal North Shore Hospital and Kolling Institute of Medical Research, Sydney, NSW, Australia
[e] School of Computing, Engineering and Mathematics, University of Western Sydney, NSW, Australia
[f] Centre for Education and Research on Ageing (CERA), Concord Hospital, Sydney, NSW, Australia
[g] The Sax Institute, Sydney, NSW, Australia
[h] National Centre for Epidemiology and Population Health, Australian National University, Canberra, ACT
[i] Corresponding author: danijela.gnjidic@sydney.edu.au

Article history

Publication date: March 2015

Citation: Gnjidic D, Pearson S, Hilmer S, Basilakis J, Schaffer AL, Blyth FM, Banks E. Manual versus automated coding of free-text self-reported medication data in the 45 and Up Study: a validation study. Public Health Res Pract. 2015;25(2):e2521518. doi: http://dx.doi.org/10.17061/phrp2521518

Key points

- Evidence on the validity of automated methods to code free-text medication data is limited
- Our study suggests that automated coding of free-text self-reported medication data, in particular prescription medications, shows very high levels of sensitivity compared with the 'gold standard' of manual expert coding
- These results have implications for large-scale studies using text-based self-reported medication data as a means of identifying medication exposure

Abstract

Background: Increasingly, automated methods are being used to code free-text medication data, but evidence on the validity of these methods is limited.

Aim: To examine the accuracy of automated coding of previously keyed in free-text medication data compared with manual coding of original handwritten free-text responses (the 'gold standard').

Methods: A random sample of 500 participants (475 with and 25 without medication data in the free-text box) enrolled in the 45 and Up Study was selected. Manual coding involved medication experts keying in free-text responses and coding using Anatomical Therapeutic Chemical (ATC) codes (i.e. chemical substance 7-digit level; chemical subgroup 5-digit; pharmacological subgroup 4-digit; therapeutic subgroup 3-digit). Using keyed-in free-text responses entered by non-experts, the automated approach coded entries using the Australian Medicines Terminology database and assigned corresponding ATC codes.

Results: Based on manual coding, 1377 free-text entries were recorded and, of these, 1282 medications were coded to ATCs manually. The sensitivity of automated coding compared with manual coding was 79% (*n* = 1014) for entries coded at the exact ATC level, and 81.6% (*n* = 1046), 83.0% (*n* = 1064) and 83.8% (*n* = 1074) at the 5, 4 and 3-digit ATC levels, respectively. The sensitivity of automated coding for blank responses was 100% compared with manual coding. Sensitivity of automated coding was highest for prescription medications and lowest for vitamins and supplements, compared with the manual approach. Positive predictive values for automated coding were above 95% for 34 of the 38 individual prescription medications examined.

Conclusions: Automated coding for free-text prescription medication data shows very high to excellent sensitivity and positive predictive values, indicating that automated methods can potentially be useful for large-scale, medication-related research.

Introduction

Self-report is a common source of medication exposure information in pharmacoepidemiological studies. Self-report has the advantage of potentially capturing information on all prescription, nonprescription, complementary and alternative medicines used, which is often not possible using other forms of ascertainment (such as pharmaceutical claims datasets, which may only capture medications subsidised by third-party payers). However, the reliability of self-report data depends on accurate recall, as well as data collection methods and the structure of the survey instrument administered to elicit information about medication use.[1-3]

Self-reported medication use is typically captured by asking participants to identify the medications they are taking from a checklist of commonly prescribed medications, and asking them to list medications that are not identified in the checklist in free-text or open-ended format. In large-scale studies that collect medication data, manual coding of free-text data by experts is often prohibitively resource intensive and potentially prone to error, particularly if the transcription of free-text information is done by individuals with limited content expertise. The use of software programs that can transcribe free-text information using an automated approach has the potential to save time by eliminating the transcription process without reducing accuracy.[4]

Despite the growing use of automated methods to code free-text medication data, evidence about the validity of these methods is limited. Research on the accuracy and validity of proprietary drug databases used to code medication information is scarce.[5,6] This validity is essential because accurate ascertainment of free-text medication data in pharmacoepidemiological studies requires reliable and valid coding methods. Misclassification of medication use can underestimate or overestimate the actual medication exposure.[7] Therefore, using data from the 45 and Up Study, the aim of this study was to compare the gold standard of medication-expert manual data entry and coding of self-reported medication data with non-expert data entry and automated coding.

Methods

Study population

This study used baseline questionnaire data from the Sax Institute's 45 and Up Study. Briefly, the 45 and Up Study includes 267 153 men and women aged 45 or over from New South Wales (NSW), randomly sampled from the Medicare Australia database.[8] Participants completed a self-administered postal questionnaire and provided data on sociodemographics, comorbidities and lifestyle (e.g. physical activity, smoking status, alcohol intake). Participants were recruited from February 2006 to April 2009, with an 18% response rate. This validation study forms part of a larger program of work that aims to determine the prevalence, risk factors, clinical consequences and costs of high-risk prescribing in older Australians.[9]

For this study, a random sample of 500 people was selected from the 45 and Up Study participants. Men and women across all age groups in the 45 and Up Study were included. To accurately capture the absence or presence of medication exposure, and to minimise the number of participants in the sample with no free-text data, we selected 5% of participants (n = 25) from those with no listed responses (according to summaries of data entered) and 95% (n = 475) from those with listed responses in the free-text data. This study was approved by the University of New South Wales Human Research Ethics Committee.

Medication data

In the 45 and Up Study, participants are asked to provide information about medication use using a comprehensive checklist of commonly used medicines and a free-text response box for medications that are not included in the checklist. Participants were asked, "Have you taken any medications, vitamins or supplements for most of the last four weeks?" The check box option included the 32 most common medications used in the Australian population at the time of the baseline survey.

In the 45 and Up Study, two questionnaire versions were used to ascertain medication use at baseline. Check boxes for three medications – citalopram, sertraline and venlafaxine were added to the version two questionnaire, which replaced the original version in October 2007.[10] For the current analysis, the second version of the questionnaire was administered to 417 participants, and the first version to 83 participants.

Manual coding of free-text medication data

All baseline questionnaires were scanned and saved electronically as PDFs. Expert manual data entry and coding of handwritten free-text data – hereafter referred to as ‘manual coding’ – involved review of the PDFs of participants’ original responses. A researcher trained in pharmacology (DG) did the manual coding. The free-text responses were recorded in an Excel spreadsheet as they were written on the questionnaire, irrespective of whether it had been written as a trade or generic medication name. When the free-text entry was unclear, consensus was reached by two medically qualified individuals (FB and EB). Medication names were then converted to generic names using the Australian registered product information for the medication.[11] The Excel file was exported to an SAS file using the SAS statistical package (SAS Institute Inc., Cary, NC), in which generic names were coded automatically to Anatomical Therapeutic Chemical (ATC) classification codes.[12] Medication generic names were mapped to chemical substance code level (full 7-digit ATC code), chemical subgroup code level (5-digit ATC code), pharmacological subgroup code level (4-digit ATC code) or therapeutic subgroup code level (3-digit ATC code).

Automated coding of free-text medication data

Non-expert automated coding – hereafter referred to as ‘automated coding’ – first involved data entry in Excel of participants’ original free-text responses by non-experts who had no specialised knowledge of medications. Data entry used a ‘key as you see’ method, whereby all the text contained in the text box was entered as it was written. Subsequent coding of medication terms from this text used the Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT), specifically the Australian Medicines Terminology (AMT) reference set.[13] This software extracts registered medications from free-text data fields and converts each medication to its generic component(s). A researcher with expertise and experience in applying SNOMED (JB) did the automated coding.[14] Once in generic form, the medication was coded to a second database, the ATC classification index.[12] Automated coding therefore relied on three separate databases for identifying terms in free-text medications: the AMT reference set[13], the ATC code index[12], and an electronic dictionary to differentiate between relevant medication terms and standard dictionary terms, as well as to correct for any misspelt terms based on a best-match algorithm. The automated coding approach was applied to free-text data in Excel. The output file was then converted to SAS format and compared with the SAS file from the manual entry.

Data analysis

The manual approach was considered the ‘gold standard’. The main comparisons were performed in relation to the individual medication ingredients and exact matches of ATC code records. Medications identified using manual and automated approaches were compared at the chemical substance level (full 7-digit ATC code), chemical subgroup level (5-digit ATC code), pharmacological subgroup level (4-digit ATC code) and therapeutic subgroup level (3-digit ATC code).

For the manual approach, medications that could not be assigned an ATC code (e.g. complementary and alternative supplements) were recorded. In addition, comparisons between individual medications within the most common therapeutic medications were made (ATC therapeutic class with 20 or more medications identified using manual approach). Sensitivity (i.e. the proportion of manual medication entries that were correctly identified using the automated method) was used as the main outcome measure in this analysis. Sensitivity and positive predictive values (PPVs) (i.e. the proportion of automated entries that were confirmed as correct using the manual approach) were then calculated at the individual medication level. SAS version 9.3 (SAS Institute, Cary, NC) was used for data analyses.

Results

The baseline characteristics of the 500 participants included in this study are shown in Table 1. The mean age of participants was 70.1 years, and 55.6% were female. Of 497 subjects who reported having had any medication in the past four weeks, 458 (92.2%) ticked at least one check box, and 39 (7.8%) did not tick any check box but listed something in the free-text box, according to existing non-expert entered study data. In this study population, the most common check-box medications were fish oil (34.0%), acetylsalicylic acid or aspirin (32.7%) and paracetamol (30.8%) (Table 2).

Table 1. Baseline characteristics of the study sample (*N* = 500)

Characteristic	Subcharacteristic	Study population
Age, mean (SD) years		70.1 (10.3)
Age groups, *n* (%)	45–59 years	84 (16.8)
	60–69 years	168 (33.6)
	70–79 years	147 (29.4)
	≥80 years	101 (20.2)
Sex, *n* (%)	Male	222 (44.4)
	Female	278 (55.6)
Country of birth, *n* (%)	Australia	353 (71.0)
	Country other than Australia	144 (29.0)
Highest educational qualification, *n* (%)	University degree	41 (8.5)
	Certificate/diploma	83 (17.2)
	Higher school/leaving certificate	42 (8.7)
	School, intermediate certificate/trade	223 (46.2)
	No certificate	94 (19.5)

(continued)

Table 1. Baseline characteristics of the study sample (*N* = 500) (*continued*)

Characteristic	Subcharacteristic	Study population
Household income, *n* (%)	<\$20 000	221 (47.4)
	\$20 000–<\$40 000	124 (26.6)
	\$40 000–<\$70 000	30 (6.4)
	≥\$70 000 or not stated	91 (19.5)
Marital status, *n* (%)	Married/de facto	299 (60.2)
	Not married[a]	198 (39.8)
Alcohol use, number of drinks per week, *n* (%)	None	220 (45.5)
	1–14	202 (41.7)
	≥15	62 (12.8)
Regular smoker, *n* (%)	Yes	237 (47.4)
	No	263 (52.6)
Needing assistance with daily tasks because of long-term illness or disability, *n* (%)	Yes	60 (12.7)
	No	411 (87.3)
Self-rated health, *n* (%)	Excellent	37 (7.7)
	Very good	119 (24.7)
	Good	197 (40.9)
	Fair	107 (22.2)
	Poor	22 (4.6)
Diabetes, *n* (%)	Yes	62 (12.4)
	No	438 (87.6)
Cardiovascular disease, *n* (%)[b]	Yes	164 (32.8)
	No	336 (67.2)
Any medication taken in the past four weeks according to check box		497 (99.4)

a Not married includes divorced, separated, single or widowed
b Diabetes, stroke, blood clots or heart disease
Note: Data were missing for the following variables: country of birth (*n* = 3), highest educational qualification (n=17), household income (*n* = 34), marital status (*n* = 3), alcohol use (*n* = 16), disability (*n* = 29), self-rated health (*n* = 18) and medications taken in the past four weeks (*n* = 3).

Using the manual approach, 25 blank entries and 1377 free-text entries were recorded. Of these, 1282 medications could be coded to ATC codes and 95 entries could not (i.e. 73 unique entries were identified that could not be coded to a medication name, and 22 were alternative supplements without ATC codes). Using the automated approach, 25 blank entries and 1204 free-text entries were recorded, of which 1128 records were coded to a medication name with corresponding ATC codes, and 51 unique entries were not coded to a medication name.

Table 2. Frequency of medications reported in check boxes (*N* = 500)

Medication type	Medication[a]	Frequency (%)
Cardiovascular medications	Atorvastatin	112 (22.4)
	Irbesartan, hydrochlorothiazide	62 (12.4)
	Simvastatin	56 (11.2)
	Perindopril, perindopril indapamide	39 (7.8)
	Atenolol	35 (7.0)
	Frusemide	34 (6.8)
	Amlodipine	22 (4.4)
	Ramipril	20 (4.0)
	Telmisartan	16 (3.2)
	Diltiazem	15 (3.0)
	Pravastatin	14 (2.8)
Gastrointestinal medications	Esomeprazole	56 (11.2)
	Pantoprazole	36 (7.2)
	Omeprazole	30 (6.0)
Psychoanaleptic medications	Citalopram[b]	10 (2.0)
	Sertraline[b]	9 (1.8)
	Venlafaxine[b]	8 (1.6)
Musculoskeletal medications	Allopurinol	31 (6.2)
	Alendronate	20 (4.0)
Multivitamins, complementary and alternative supplements	Fish oil	170 (34.0)
	Multivitamins and minerals[c]	120 (24.0)
	Glucosamine	119 (23.8)
	Calcium	56 (11.2)
	Omega 3	49 (9.3)
Other medications	Acetylsalicylic acid for heart or other reasons[d]	163 (32.7)
	Paracetamol	154 (30.8)
	Salbutamol	40 (8.0)
	Metformin	38 (7.6)
	Thyroxine	32 (6.4)
	Warfarin	31 (6.2)

a Medication brand names converted to generic names
b Data missing for participants who were administered version 1 baseline questionnaire (*n* = 83)
c Two separate check boxes for ‘multivitamins and minerals’ and ‘multivitamins alone’ combined
d Two separate check boxes for acetylsalicylic acid (aspirin) combined.
Note: Table sorted according to frequency within each therapeutic group or category.

The sensitivity of the automated approach for exact ATC codes was 79% (1014/1282) compared with manual coding (Table 3). Compared with the manual approach, the automated approach demonstrated 100% sensitivity for 25 blank free-text responses.

Table 3. Sensitivity of automated coding approach for blank free-text entries and medication entries to ATC classification compared with manual approach (gold standard)

Entry type	Manual approach	Correct using automated approach	Sensitivity (%)[a]
Blank entries	25	25	100.0
Exact ATC codes[b]	1282	1014	79.0
5-digit ATC codes	1282	1046	81.6
4-digit ATC codes	1282	1064	83.0
3-digit ATC codes	1282	1074	83.8

[a] Sensitivity of automated coding to chemical substance level (full 7-digit ATC code), chemical subgroup level (5-digit ATC code), pharmacological subgroup level (4-digit ATC code) or therapeutic subgroup level (3-digit ATC code) compared with manual coding (gold standard)

[b] 995 medication entries coded at 7-digit ATC code level and 19 coded at either 3, 4 or 5-digit ATC code level.

A disagreement between the manual and automated approaches was identified for 114 medication entries. Of these 114 entries, 32 were coded correctly at the 5-digit ATC code level, resulting in 81.6% (1046/1282) sensitivity of the automated approach at this level. The disagreement for this level largely occurred because manual coding identified all specific medication ingredients (e.g. calcium carbonate vs calcium, magnesium sulfate vs magnesium) that were then assigned the full 7-digit ATC code, while the automated approach entries were coded to the 5-digit ATC code. Another 18 entries were coded correctly at the 4-digit ATC level (e.g. ferrous fumarate vs iron, calcium carbonate combination vs calcium and magnesium), corresponding to 83.0% (1064/1282) sensitivity of the automated approach at this level. In addition, another 10 entries were coded correctly at the 3-digit ATC level (e.g. perindopril vs perindopril indapamide, irbesartan vs irbesartan hydrochlorothiazide), corresponding to 83.8% (1074/1282) sensitivity of the automated approach. Of the remaining 54 entries, 14 nonspecific free-text entries were coded using the automated but not manual approach (e.g. 'no aspirin (allergic)' response mapped to 'acetylsalicylic acid'), and disagreement for 40 entries occurred because of differences in entries interpreted and keyed in by non-experts and coded by the automated approach, versus entries keyed in by experts and coded by the manual approach (e.g. two entries were missed and not coded to medication names using the manual approach).

The manual approach identified 154 medication names with corresponding ATC codes that were not identified using the automated approach. These entries included free-text entries that were not keyed in or were keyed in incorrectly (*n* = 99) and consequently not coded by the automated approach, and free-text entries that were keyed in but not coded (*n* = 55) by the automated approach. Examples of keyed-in free-text entries not coded to medication names include some types of insulin and various minerals and vitamins.

Overall, compared with the manual approach, the automated approach demonstrated high to excellent sensitivity and PPVs for most of the common individual prescription medications. Overall, 34 of 38 prescription medications (rather than vitamins or supplements) had PPVs more than 95%, and the majority had PPVs of 100% (Table 4).

Discussion

To our knowledge, this is the first study to compare manual versus automated coding of free-text self-reported medication data. The findings of this study suggest very good sensitivity of the automated coding method for capturing the free-text self-reported prescription medication data compared with the expert coding. The sensitivity of automated coding was consistently high, with 79% of entries coded to exact ATC classification compared with manual coding, and increased with greater generality of the ATC level chosen, to 84% for the therapeutic subgroup. When individual medications within the most common therapeutic classes were compared, the automated approach demonstrated high sensitivity (>70%) and excellent PPVs for most of the individual medications compared with the manual approach.

In this study, the differences between manual and automated approaches occurred mostly because the automated approach did not identify all medication ingredients from original free-text entries keyed in by non-experts, or the free-text entries were not coded because they were keyed in incorrectly. The use of open-ended questions affects the accuracy of self-report[15], which in turn would affect the capability of the non-expert automated approach to correctly identify free-text entries. This may have been due to the misspelling of the medication name, or because the discernible spelling of medication names was such that the free-text entries could not be identified by the software. The manual approach by medication experts is more likely to correctly code less-specific free-text entries (e.g. 'high blood pressure medications' or 'HRT' for hormone replacement therapy) than the automated approach.

The sensitivity and PPVs of the automated approach for specific prescription medications were generally excellent when compared with the manual approach, with high sensitivity values and PPVs of more than 95% for 34 of the 38 individual prescription medications examined. Moderate or poor sensitivity of automated coding was demonstrated for a small proportion of therapeutic classes, including vitamins, mineral supplements and specific examples of prescription medications such as insulin (27.3%) and warfarin (33.3%). Poor sensitivity of the automated approach in relation to vitamins and

Table 4. Sensitivity and positive predictive values of automated coding for specific medication entries compared with manual coding (gold standard)

		Manual approach	Automated approach			
Therapeutic group (ATC class)[a]	Medication	Total number of entries[b]	Total number of entries	Number of correct automated entries	Sensitivity (%)	Positive predictive value (%)
Antacid agents (A02)	Rabeprazole	22	21	21	95.5	100.0
	Lansoprazole	14	15	14	100.0	93.3
Diabetes agents (A10)	Gliclazide	17	15	15	88.2	100.0
	Insulin	10	NP	NP	27.3	75.0
Vitamins (A11)	Ascorbic acid and other vitamins	18	NP	NP	0.0	0.0
	Alpha-tocopherol	12	10	8	66.7	80.0
Mineral supplements (A12)	Calcium carbonate	13	NP	NP	23.1	100
	Magnesium sulfate	7	NP	NP	0.0	0.0
Antithrombotic agents (B01)	Clopidogrel	25	23	23	92.0	100.0
	Warfarin	NP	NP	NP	33.3	100.0
Antianaemic agents (B03)	Folic acid	11	12	9	81.8	75.0
	Ferrous fumarate	NP	NP	NP	0.0	0.0
Cardiac agents (C01)	Isosorbide mononitrate	17	14	14	82.4	100.0
	Digoxin	16	11	11	68.8	100.0
Diuretics (C03)	Indapamide	22	18	18	81.8	100.0
	Spironolactone	10	9	9	90.0	100.0
Beta-blockers (C07)	Metoprolol	34	31	30	88.2	96.8
	Atenolol	13	10	10	76.9	100.0
Calcium channel blockers (C08)	Felodipine	16	13	13	81.3	100.0
	Lercanidipine	16	14	14	87.5	100.0
Angiotensin agents (C09)	Candesartan	31	30	30	96.8	100.0
	Enalapril	11	9	9	81.8	100.0
Lipid-lowering agents (C10)	Rosuvastatin	17	17	17	100.0	100.0
	Simvastatin	16	14	13	81.3	92.9
Hormones (G03)	Oestradiol	13	8	8	61.5	100.0
	Raloxifene	6	6	6	100.0	100.0
Corticosteroids (H02)	Prednisone	12	14	12	100.0	85.7
	Prednisolone	10	7	7	70.0	100.0
Anti-inflammatory agents (M01)	Meloxicam	17	16	16	94.1	100.0
	Celecoxib	14	13	13	92.9	100.0
Analgesics (N02)	Paracetamol	24	23	22	91.7	95.7
	Acetylsalicylic acid	19	16	15	78.9	93.8
Antiepileptic agents (N03)	Sodium valproate	11	11	11	100.0	100.0
	Carbamazepine	6	5	5	83.3	100.0
Psycholeptic agents (N05)	Diazepam	5	5	5	100.0	100.0
	Temazepam	5	5	5	100.0	100.0
Psychoanaleptic agents (N06)	Paroxetine	7	7	7	100.0	100.0
	Amitriptyline	6	6	6	100.0	100.0
Nasal agents (R01)	Fluticasone salmeterol	19	18	18	94.7	100.0
	Mometasone	NP	NP	NP	100.0	100.0

(continued)

Table 4. Sensitivity and positive predictive values of automated coding for specific medication entries compared with manual coding (gold standard) (*continued*)

		Manual approach	Automated approach			
Therapeutic group (ATC class)[a]	Medication	Total number of entries[b]	Total number of entries	Number of correct automated entries	Sensitivity (%)	Positive predictive value (%)
Obstructive airway disease agents (R03)	Budesonide eformoterol	11	11	11	100.0	100.0
	Tiotropium	11	10	10	90.9	100.0
Ophthalmologic agents (S01)	Latanoprost	8	7	7	87.5	100.0
	Timolol	6	5	5	83.3	100.0

NP = not publishable as cells contain less than five entries
a Medication classes with 20 or more entries identified using manual approach presented only
b Two most frequent medications within each therapeutic class identified using manual approach presented only
Note: Table sorted by ATC therapeutic class code.

supplements occurred because the AMT database had more comprehensive information on commercially available vitamin and mineral supplements than the ATC, and consequently had ingredients that were not successfully coded into ATC classifications. Poor accuracy of insulin coding was predominantly due to differences in classification of the variety of nonstandard insulin terms between the AMT and ATC databases. The availability of an AMT-to-ATC code mapping reference set would resolve the discrepancies between these datasets.[16]

Technological advances have resulted in major improvements in automated methods to capture and classify data. Increasingly sophisticated methods in text mining and natural language processing are being used to extract medication information from narrative clinical text such as electronic health records.[17-19] In a study assessing four commercial natural language processing engines for their ability to extract medication information, compared with the physician-derived manual gold standard, the medication extraction systems were successful at accurately capturing medication names.[17] However, further studies are required to assess the ability and accuracy of these automated methods to code extracted medication information from electronic health records and self-reported medication data captured in large-scale studies.[20]

The findings of this study indicate that automated coding of free-text self-reported medication information that has been captured through standard data entry is likely to be useful for future research into medications. The 45 and Up Study involves more than 267 000 participants, and manual expert coding of free-text data on medication use is not possible. A low-cost and feasible method for coding such medication use is important. As reflected by the excellent PPVs (>95% for the vast majority of individual prescription medications), the automated method can identify those who are most likely to have reported being exposed to specific medications, as well as a comparison group that is less likely to have been exposed. The 45 and Up Study links to national pharmaceutical claims data from the Pharmaceutical Benefits Scheme, which provides ongoing independent data on prescription medications dispensed to participants under the scheme.[9] However, these data do not necessarily capture all medications – for example, they do not include over-the-counter medications and medications obtained through private prescription. The check box and free-text self-report data therefore have the potential to add to the existing data framework. The findings here suggest that, at this stage, the automated capture of free-text data on vitamins, minerals and supplements is insufficiently accurate to be of major use. However, information on multivitamins and minerals in the 45 and Up Study can be specifically sought through check-box items that enquire about the self-reported use of these agents. The findings also indicate that automated coding of free-text data has the potential to contribute to other medical-related text-based data collections.

There are strengths and limitations to the current study. Validation studies of automated coding of self-reported medication information are useful for interpretation of results from large-scale pharmacoepidemiologic studies using this method. Additional strengths include that the comparison between manual and automated approaches was made for a range of medications and pharmacological classes, rather than concentrating on one medication class, and that an appropriate gold standard was used for comparison purposes. Although 500 participants is considered large and appropriate for validation purposes, it is likely that other unique free-text entries would have been identified if more study participants had been included.

Conclusion

In conclusion, our findings suggest that automated coding of free-text self-reported medication data, particularly prescription medications, shows very high levels of sensitivity compared with manual expert coding. These results have implications for other national and international large-scale studies using text-based self-reported medication data as a means of identifying medication exposure. However, our results also suggest that the automated approach used here would need further refinement before it could be used for classification of exposure to items such as vitamins, minerals and complementary medications.

Acknowledgements

This research was completed using data collected through the 45 and Up Study (www.saxinstitute.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council NSW, and partners the National Heart Foundation of Australia (NSW Division); NSW Ministry of Health; beyondblue; NSW Government Family & Community Services – Carers, Ageing and Disability Inclusion; and the Australian Red Cross Blood Service. We thank the many thousands of people participating in the 45 and Up Study. This study is funded by the National Health and Medical Research Council of Australia (NHMRC project grant number 1024450 and NHMRC Centre of Research Excellence in Medicines and Ageing 1060407). DG and EB are supported by the NHMRC.

Competing interests

Work by EB on this research, and other research projects, is funded by an NHMRC Senior Research Fellowship.

References

1. Boudreau DM, Daling JR, Malone KE, Gardner JS, Blough DK, Heckbert SR. A validation study of patient interview data and pharmacy records for antihypertensive, statin, and antidepressant medication use among older women. Am J Epidemiol. 2004;159(3):308–17.
2. Klungel OH, de Boer A, Paes AH, Herings RM, Seidell JC, Bakker A. Influence of question structure on the recall of self-reported drug use. J Clin Epidemiol. 2000;53(3):273–7.
3. Gama H, Correia S, Lunet N. Questionnaire design and the recall of pharmacological treatments: a systematic review. Pharmacoepidemiol Drug Saf. 2009;18(3):175–87.
4. Pahor M, Chrischilles EA, Guralnik JM, Brown SL, Wallace RB, Carbonin P. Drug data coding and analysis in epidemiologic studies. Eur J Epidemiol. 1994;10(4):405–11.
5. Qato DM, Schumm LP, Johnson M, Mihai A, Lindau ST. Medication data collection and coding in a home-based survey of older adults. J Gerontol B Psychol Sci Soc Sci. 2009;64 Suppl 1:i86–93.
6. Richesson RL. An informatics framework for the standardized collection and analysis of medication data in networked research. J Biomed Inform. 2014;52:4–10.
7. Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. Int J Epidemiol. 2005;34(3):680–7.
8. Banks E, Redman S, Jorm L, Armstrong B, Bauman A, Beard J, et al. Cohort profile: the 45 and up study. Int J Epidemiol. 2008;37(5):941–7.
9. Gnjidic D, Le Couteur DG, Pearson SA, McLachlan AJ, Viney R, Hilmer SN, et al. High risk prescribing in older adults: prevalence, clinical and economic implications and potential for intervention at the population level. BMC Public Health. 2013;13:115.
10. Sax Institute. Sydney: Sax Institute; 2015. Summary of modifications to the 45 and Up Study baseline questionnaire [cited 2015 Feb 26];p. 1–16. Available from: www.saxinstitute.org.au/wp-content/uploads/Questionnaire_modifications.pdf
11. Donohoo E, editor. MIMS Annual 2013. Sydney: CMP Medica Australia; 2013.
12. WHO Collaborating Centre for Drug Statistics Methodology, Norwegian Institute of Public Health. Oslo: WHO Collaborating Centre for Drug Statistics Methodology; 2014 Dec 16. ATC/DDD Index 2015 [cited 2015 Feb 26]; [database]. Available from: www.whocc.no/atcddd/
13. Côté RA and College of American Pathologists. SNOMED international: the systematized nomenclature of human and veterinary medicine. 3rd ed. Northfield, Il; Schaumberg, Il: College of American Pathologists; American Veterinary Medical Association; 1993.
14. Rahman SZ, Basilakis J, Rahmadi A, Lujic S, Musgrave I, Jorm L, et al. Use of serotonergic antidepressants and St John's wort in older Australians: a population-based cohort study. Australas Psychiatry. 2013;21(3)262–6.
15. Metlay JP, Hardy C, Strom BL. Agreement between patient self-report and a Veterans Affairs national pharmacy database for identifying recent exposures to antibiotics. Pharmacoepidemiol Drug Saf. 2003;12(1):9–15.

16. NEHTA: the personally controlled e-health record system. Canberra: National E-Health Transition Authority c2004–2015. 2012 survey results and development roadmap, Australian medicines terminology; 2012 Mar 30 [cited 2015 Feb 26]. Available from: www.nehta.gov.au/component/docman/doc_download/1796-nehta-amt-survey-results-and-roadmap-2012?Itemid

17. Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform. 2009;78(4):284–91.

18. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform. 2010;160(Pt 1):739–43.

19. Wilke RA, Xu H, Denny JC, Roden DM, Krauss RM, McCarty CA, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther 2011;89(3):379–86.

20. Pathak J, Murphy SP, Willaert BN, Kremers HM, Yawn BP, Rocca WA, et al. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. AMIA Annu Symp Proc. 2011;2011:1089–98.