International Journal of Population Data Science

Coding reliability and agreement of International Classification of Disease, 10th revision (ICD-10) codes in emergency department data

Peng, M[1], Eastwood, C[1], Boxill, A[1], Jolley, RJ[1], Rutherford, L[2], Carlson, K[2], Dean, S[2], and Quan, H[1]

Abstract

Introduction
Administrative health data from emergency departments play important roles in understanding health needs of the public and reasons for health care resource use. International Classification of Disease (ICD) diagnostic codes have been widely used to code reasons of clinical encounters for administrative purposes in emergency departments.

Objective
The purpose of the study is to examine the coding agreement and reliability of ICD diagnosis codes in emergency department records through auditing the routinely collected data.

Methods
We randomly sampled 1 percent of records (n=1636) between October and December 2013 from 11 emergency departments in Alberta, Canada. Auditors were employed to review the same chart and independently assign main diagnosis codes. We assessed coding agreement and reliability through comparison of codes assigned by auditors and hospital coders using proportion of agreement and Cohen's kappa. Error analysis was conducted to review diagnosis codes with disagreement and categorized them into six groups.

Results
Overall, the agreement was 86.5% and 82.2% at 3 and 4 digits levels respectively, and reliability was 0.86 and 0.82 respectively. Variations of agreement and reliability were identified across different emergency departments. The major two categories of coding discrepancy were the use of different codes for same condition (23.6%) and the use of codes at different levels of specificity (20.9%).

Conclusions
Diagnosis codes in emergency departments show high agreement and reliability, although there are variations of coding quality across different hospitals. Stricter coding guidelines regarding the use of unspecified codes are needed to enhance coding consistency.

Keywords
coding agreement; coding reliability; ICD codes; emergency department

Introduction

Routinely collected administrative health data are increasingly used in population based health services and policy research(1). Administrative health data are created when clinical information is coded using International Classification of Disease (ICD) system through reabstraction of clinical documentation by coders or physicians(2, 3). Validation studies have been conducted to assess validity of ICD codes in identifying clinical conditions through comparison with chart-reviewed data. Validity of ICD codes varies across different conditions and health care systems and depends on how data is collected(4).

Code assignment is an onerous process with many potential sources of errors(2). Coding personnel generate data, and are key determinants of ICD code quality(3). Therefore, knowledge on coding reliability and agreement among coders provide important information about data quality of administrative health data. Canadian Institute of Health Information (CIHI) conducts routine reabstraction studies to assess coding quality on hospital discharge abstract data through auditing. Selected charts are re-coded by CIHI coders and agreement of codes between CIHI and hospital coders are compared(5). Similar studies on hospital administrative data were also conducted in Australia and UK(6, 7). Emergency and ambulatory

*Corresponding Author:
  Email Address: mpeng@ucalgary.ca (M Peng)

care are some of the largest-volume patient activities, making these settings key components of the continuum of health services(8). Emergency care is different from inpatient care in terms of reasons for visits and length of patient-physician interaction. Coding reliability and agreement in administrative health data from emergency care is still unknown.

The aim of this study was to assess the coding reliability and agreement of ICD-10 codes in emergency departments. We audited selected charts from emergency departments, which were previously coded by certified health information coding specialists.

Method

Data source

Following national guidelines developed by CIHI, clinical information from emergency visits are collected in Alberta. There are up to ten ICD-10th revision, Canada (ICD-10-CA) diagnosis codes in each record. For each record, the first diagnosis code is the main problem, which is deemed to be the most clinically significant reason for the client's visit and that requires evaluation and/or treatment or management. The main problem can be a diagnosis, symptom, sign, abnormal test result, or reason for encounter(9). This study focused on the assessment of the main diagnosis code.

We randomly sampled 1 percent (n=1636) of the total visits from 11 emergency departments between October and December in 2013. Hospital coders code the charts based on the Canadian coding guideline. Auditors were employed as external coders to re-code the sampled records following the same coding guideline. One auditor was a coding coordinator who conducted the audit of all records from Calgary. In other Alberta hospitals, the initial audit was performed by students enrolled in health information management (HIM) programs as part of their mandatory practicum for their certification program and were subsequently validated by experienced coding coordinators. The HIM program is an accredited program in Canada to train for a career in managing personal health information.

Analysis

Following the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (10), we used reliability and agreement to assess the reproducibility of ICD-10-CA coding in emergency departments. Agreement is whether ICD-10-CA codes assigned by auditor and hospital coders are identical for the same record and is reported as the proportion of agreement. The proportion of agreement was calculated by dividing the number of records with identical codes by the total number of records. Reliability is the ability of a coding method/system to differentiate among subjects (e.g. patients who visited emergency departments in this study) and is closely related to setting where method/systems are applied. We used Cohen's kappa to assess the reliability. Cohen's kappa is calculated by dividing the difference between observed agreement and expected agreement by the difference between 1 and the expected agreement. The value of Cohens' kappa ranges from -1

to +1. The agreement and reliability were assessed at 3-digits and 4-digits level of ICD-10-CA codes.

We also assessed the agreement between auditor and the hospital's coder on the top 20 most frequently used diagnosis codes. Sensitivity and positive predictive value (PPV) were calculated with the auditor's codes as the reference. Sensitivity indicates the proportion of records with a diagnosis code assigned by the auditor in which hospital coders also assigned the same codes. PPV measures the proportion of records with a diagnosis code assigned by hospital coders in which auditors also assigned the same code.

Error analysis was conducted on the records with different diagnosis codes. We invited one clinical nurse and one coding specialist to review the diagnosis code and categorized the codes with disagreements into 6 different categories(7). First, the nurse and coding specialist independently reviewed the codes and assign the codes into one of the 6 categories of discrepancy. Records assigned into different categories were discussed between the nurse and coding specialist and final category was assigned after consensus was reached. Detailed descriptions and examples for each category can be found in the results section. The proportion of records in each category was presented. All the analysis was conducted in R software and 95% confidence intervals were provided for all the statistics.

Results

Eight HIM students and 2 coding coordinators were involved in the audit of 1636 records. Overall, the agreement was 86.5% and 82.2% at 3 and 4-digits levels respectively, and reliability was 0.86 and 0.82 respectively (Table 1). At the hospital level, agreement ranged from 68.4% to 95.2% at the 3-digits level and from 57.9% to 93.4% at the 4-digits level; reliability ranged from 0.68 to 0.95 at the 3-digits level and from 0.57 to 0.93 at the 4-digits level. There were 3 out of 11 hospitals with agreement below 70% at 3-digits level and below 65% at the 4-digits level. There were no significant differences in agreement and reliability between teaching and non-teaching hospitals. Edmonton had significantly lower agreement and reliability at 3 and 4-digits level than Calgary and other regions.

Table 2 presents the agreement on the top 20 diagnosis codes at 3 digits level, which accounted for 35% of sampled records. Hospital coders and auditors had similar coding frequency for the top 20 diagnosis, in which 17 codes has coding frequency difference less than 3. Using auditor codes as the reference, 13 of 20 diagnosis codes had both sensitivity and PPV above 0.9, although only 1 code showed perfect agreement. Three out of the 20 codes had sensitivity below 0.9 while 2 out of the 20 codes had PPV below 0.9. Codes J18 (pneumonia, organism unspecified) and S01 (open wound of head) had both sensitivity and PPV below 0.9.

Results of error analysis on the records with different diagnoses are shown in Table 3. The most common category of discrepancy was that coders assigned different codes for the same condition (23.6%). The issue with coding specificity is the second most common category of discrepancy (20.9%) as one coder assigned more specific codes than the other coder. There were around 16.5% with completely different codes.

Table 1: Overall agreement and reliability for coding at 3 and 4 digit levels

		3-digits level		4-digits level	
Category	N	Agreement (%)	Reliability (Cohen's kappa)	Agreement (%)	Reliability (Cohen's kappa)
Type of hospital					
Teaching	1249	86.9 (84.9, 88.7)	0.87 (0.85, 0.89)	83.1 (80.9, 85.1)	0.83 (0.81, 0.85)
Non-teaching	387	85.0 (81.0, 88.3)	0.85 (0.81, 0.88)	79.1 (74.6, 82.9)	0.79 (0.75, 0.83)
Location					
Calgary	806	93.9 (92.0, 95.4)	0.94 (0.92, 0.96)	91.1 (88.8, 92.9)	0.91 (0.89, 0.93)
Edmonton	443	74.3 (69.9, 78.2)	0.74 (0.70, 0.78)	68.6 (64.0, 72.9)	0.68 (0.64, 0.73)
Others	387	85.0 (81.0, 88.3)	0.85 (0.81, 0.88)	79.1 (74.6, 82.9)	0.79 (0.75, 0.83)
Hospitals					
1	166	95.2 (90.4, 97.7)	0.95 (0.92, 0.98)	93.4 (88.2, 96.5)	0.93 (0.89, 0.97)
2	128	84.4 (76.7, 90.0)	0.84 (0.78, 0.90)	76.6 (68.1, 83.4)	0.76 (0.69, 0.84)
3	182	94.0 (89.2, 96.8)	0.94 (0.90, 0.97)	90.7 (85.2, 94.3)	0.91 (0.86, 0.95)
4	95	68.4 (58.0, 77.4)	0.68 (0.58, 0.77)	57.9 (47.3, 67.8)	0.57 (0.48, 0.67)
5	185	93.5 (88.7, 96.5)	0.93 (0.90, 0.97)	90.3 (84.8, 94.0)	0.90 (0.86, 0.94)
6	164	95.1 (90.3, 97.7)	0.95 (0.92, 0.98)	93.3 (88.0, 96.4)	0.93 (0.89, 0.97)
7	190	93.7 (89.0, 96.5)	0.94 (0.90, 0.97)	92.6 (87.7, 95.8)	0.93 (0.89, 0.96)
8	156	69.9 (61.9, 76.8)	0.69 (0.62, 0.77)	61.5 (53.4, 69.1)	0.61 (0.54, 0.69)
9	83	92.8 (84.4, 97.0)	0.93 (0.87, 0.98)	85.5 (75.7, 92.0)	0.85 (0.78, 0.93)
10	134	84.3 (76.8, 89.8)	0.84 (0.78, 0.90)	81.3 (73.5, 87.3)	0.81 (0.74, 0.88)
11	153	69.9 (61.9, 76.9)	0.70 (0.62, 0.77)	64.7 (56.5, 72.1)	0.64 (0.57, 0.72)
Overall	1636	86.5 (84.7, 88.1)	0.86 (0.85, 0.88)	82.2 (80.2, 84.0)	0.82 (0.80, 0.84)

Discussion

Overall, there was high agreement and reliability of coding for emergency visits, although variability of coding was observed across different emergency departments. Coders and auditors showed high agreement with each other on individual codes based on assessment of the top 20 diagnosis codes. The main categories of disagreement between coders and auditors were different codes for the same condition and the issue of coding specificity.

What is the coding quality in emergency departments compared to other settings? The proportion of agreement on the main diagnosis at the 3-digit level in Canadian hospital data was around 77%(11), which is lower than emergency department data. There are 14 different diagnosis types in Canadian hospital inpatient data to reflect clinical significance and timing of diagnosis. If diagnosis type in hospital setting is ignored, the agreement in hospital setting data was around 85.4%, which is very close to the proportion of agreement (86.5%) in the emergency department data. Compared with Canadian emergency department data, Australian hospital data has similar agreement at the 3-digit (around 86 %) and 4-digit levels (around 80.3%) on main or principal diagnosis(6). At the individual code level, emergency department data had higher reproducibility than coding in hospital settings in Canada, as the sensitivity (median (interquartile range): 0.82 (0.71 to 0.89)) and PPV (0.82 (0.74 to 0.89)) for 50 most frequent main diagnosis codes is lower than the sensitivity and PPV for the 20 most frequent main diagnosis codes in emergency departments. Inpatient hospital care is more complex and has more clinical information documented during the patient's stay, which increases the difficulty for coders to translate clinical notes into standardized diagnosis codes and diagnosis types. However, high coding reproducibility does not directly imply high validity of coding in the emergency department. For example, a validity study on emergency department diagnosis codes for identifying acute heart failure (AHF) conducted in Edmonton, Alberta showed moderate sensitivity and high PPV but low specificity (0.50 (95% confidence interval: 0.398, 0.601)), with the chart review as the reference within a cohort with suspected AHF(12). In another study, for hyperkalemia and hyponatremia, ICD-10 codes had very high specificity (>0.99) but very low sensitivity (0.15 for hyperkalemia and 0.075 for hyponatremia) (13). Therefore, it is cautioned when using these ICD diagnosis codes for research purposes, as further studies on coding validity of emergency department data are still needed.

How to improve the coding quality in emergency departments? The issue of different codes for similar conditions is partly due to the use of "not otherwise specified (NOS) or unspecified" and "not elsewhere classified (NEC)" in code description. The NOS code description was created to meet the challenge that a specific diagnosis is often hard to achieve in practice, and is therefore not documented by physicians, while the NEC is created to classify residual categories or categories not explicitly specified in other codes. This gives the room for flexibility in coding which in turn can result in coding discrepancies. For example, one study found almost half of food allergy patients were found with unspecified allergy codes(14). The issue of coding specificity is another main reason for discrepancy. This could be related to lack of specificity in clinical documentation as coders reported vague terms or ambiguity

Table 2: Agreement on top 20 frequently used diagnosis codes

ICD-10 codes at three digits levels	# of records Hospital	Audit	Sensitivity[*] (95% CI)	PPV[*] (95% CI)
Symptom and sign:				
R07: Pain in throat and chest	61	66	0.92 (0.83, 0.97)	1.00 (0.94, 1.00)
R10: Abdominal and pelvic pain	84	84	0.95 (0.88, 0.99)	0.95 (0.88, 0.99)
R11: Nausea and vomiting	26	27	0.93 (0.76, 0.99)	0.96 (0.80, 1.00)
R50: Fever of unknown origin	22	22	0.95 (0.77, 1.00)	0.95 (0.77, 1.00)
R51: Headache	15	15	0.93 (0.68, 1.00)	0.93 (0.68, 1.00)
Disease of the respiratory system:				
J02: Acute pharyngitis	24	24	1.00 (0.86, 1.00)	1.00 (0.86, 1.00)
J05: Acute obstructive laryngitis (croup) and epiglottitis	31	30	1.00 (0.88, 1.00)	0.97 (0.83, 1.00)
J06: Acute upper respiratory infections of multiple and unspecified sites	34	36	0.89 (0.74, 0.97)	0.94 (0.80, 0.99)
J18: Pneumonia, organism unspecified	20	17	0.88 (0.64, 0.99)	0.75 (0.51, 0.91)
J45: Asthma	22	21	1.00 (0.84, 1.00)	0.95 (0.77, 1.00)
Infectious and parasitic diseases:				
A09: Diarrhea and gastroenteritis of presumed infectious origin	33	31	0.97 (0.83, 1.00)	0.91 (0.76, 0.98)
B34: Viral infection of unspecified site	25	27	0.93 (0.76, 0.99)	1.00 (0.86, 1.00)
Injury, poisoning and other sequence of external causes:				
S01: Open wound of head	15	17	0.82 (0.57, 0.96)	0.93 (0.68, 1.00)
S61: Open wound of wrist and hand	24	23	0.96 (0.78, 1.00)	0.92 (0.73, 0.99)
S82: Fracture of lower leg, including ankle	15	15	0.87 (0.60, 0.98)	0.87 (0.60, 0.98)
S93: Dislocation, sprain and strain of joints and ligaments at ankle and foot level	15	16	0.94 (0.70, 1.00)	1.00 (0.78, 1.00)
Other Problems:				
F10: Use of alcohol	36	34	0.91 (0.76, 0.98)	0.86 (0.71, 0.95)
M54: Dorsalgia	31	34	0.85 (0.69, 0.95)	0.94 (0.79, 0.99)
N39: other disorder of urinary system, including urinary tract infection (unspecified site)	30	32	0.94 (0.79, 0.99)	1.00 (0.88, 1.00)
L03: Cellulitis	19	17	1.00 (0.80, 1.00)	0.89 (0.67, 0.99)

[*]Sensitivity and PPV are calculated using the auditors' codes as the reference.

Table 3: Proportion of cases in each category of discrepancy on ICD codes showed disagreement between hospital coder and auditor.

Category	Categories of discrepancy	N (%)
1	One coder recorded a symptom while the other coder recorded a diagnosis related to the symptom. e.g. N76.0: acute vaginitis and N93.9: Abnormal uterine and vaginal bleeding, unspecified	49 (16.5)
2	Coders recorded codes for similar but not identical condition; often one coder is more specific than the other. E.g. T78.4: allergy, unspecified and T78.1: Other adverse food reactions, not elsewhere classified	62 (20.9)
3	Coders recorded obviously different codes for similar conditions. E.g. T09.5: injury of unspecified muscle and tendon of trunk, NOS and M54.5: low back pain	35 (11.8)
4	Coders recorded codes for conditions which were not similar but obviously related conditions. E.g. I49.5: sick sinus syndrome and I47.1: supraventricular tachycardia.	32 (10.8)
5	Coders recorded completely different conditions. E.g. F43.0: acute stress reaction and T48.5: poisoning by anti-common-cold drugs	49 (16.5)
6	The coders recorded different codes for the same conditions. E.g. R07.4: chest pain unspecified and R07.3: other chest pain	70 (23.6)

NOS: not otherwise specified

in physician's documentation(3). Therefore, improvement of clinical documentation is necessary to improve coding specificity. There were also cases with different codes for similar conditions or completely different codes. This could be due to different understandings of clinical documentation by different coders or possibly codes without clear definitions in the coding standards. Precise code definitions, as well as continuous education for coders is necessary to ensure current clinical knowledge, for the best possible coding consistency(15).

What is the generalizability of our results to administrative data from other provinces and countries? Alberta follows the same coding guidelines and practices as other provinces in Canada. Therefore, overall emergency department data quality should be comparable across Canada. However, high variability of coding quality in different hospitals should be expected as different hospitals might have different clinical practice patterns, charting requirements, and coding management. Coding quality is closely associated with the data collection and use of data. Canada has trained health information professionals to assign the codes while some countries uses physicians, medical clerks, or nurses to assign the codes. Use of administrative health data to classify disease related groups for purpose of payment could also affect how codes are assigned, as more specific codes are often required for billing purpose. In summary, within Canada, generalizability is possible as the provinces use similar coding standards. Greater variation from our findings would be expected in countries with very different documentation and coding practices.

This study has the following limitations. First, our study only focused on one province and at one time point. Spatial and temporal variation of data quality cannot be assessed. Second, we used health information management students as the auditors to recode the chart. This might increase the level of discrepancy due to their lack of coding experience. However, codes assigned by students were also reviewed by an experienced coder and found that they were likely to agree with students if there was a coding discrepancy. Third, this study only assessed coding reproducibility. Validity in emergency department coding still needs further investigations.

Conclusion

Alberta emergency department administrative health data had high agreement and reliability of ICD code assignments. The main reasons of coding discrepancies between coders were due to use of codes at different levels of specificity and use of unspecified codes in coding practice.

Conflict of Interest Statement

The authors declare that they have no conflicts of interest.

References

1. Khan Y, Glazier RH, Moineddin R, Schull MJ. A Population-based Study of the Association Between Socioeconomic Status and Emergency Department Utilization in Ontario, Canada. Acad Emerg Med. 2011;18(8):836-43. https://doi.org/10.1111/j.1553-2712.2011.01127.x

2. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005;40(5):1620-39. https://doi.org/10.1111/j.1475-6773.2005.00444.x

3. Tang KL, Lucyk K, Quan H. Coder perspectives on physician-related barriers to producing high-quality administrative data: a qualitative study. CMAJ open. 2017;5(3):E617. https://doi.org/10.9778/cmajo.20170036

4. Quan H, Li B, Saunders LD, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008;43(4):1424-41. https://doi.org/10.1111/j.1475-6773.2007.00822.x

5. Canadian Institute for Health Information. CIHI Data Quality Study of the 2009–2010 Discharge Abstract Database. Ottawa: CIHI: 2012.

6. Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. Med Care. 2006;44(11):1011-9. https://doi.org/10.1097/01.mlr.0000228018.48783.34

7. Dixon J, Sanderson C, Elliott P, Walls P, Jones J, Petticrew M. Assessment of the reproducibility of clinical coding in routinely collected hospital activity-data: a study in two hospitals. Journal of Public Health. 1998;20(1):63-9. https://doi.org/10.1093/oxfordjournals.pubmed.a024721

8. Canadian Institute for Health Information. Understanding Emergency Department Wait Times. Ottawa: CIHI: 2005.

9. Canadian Institute for Health Information. Canadian Coding Standards for ICD-10-CA and CCI for 2015. Ottawa: CIHI: 2015.

10. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002

11. Juurlink D PC, Croxford R, Chong A, Austin P, Tu J, Laupacis A. Canadian Institute for Health Information Discharge Abstract Database: A Validation Study. Toronto: Institute for Clinical Evaluative Sciences: 2006.

12. Frolova N, Bakal JA, McAlister FA, Rowe BH, Quan HD, Kaul P, et al. Assessing the Use of International Classification of Diseases-10th Revision Codes From the Emergency Department for the Identification of Acute Heart Failure. Jacc-Heart Fail. 2015;3(5):386-91. https://doi.org/10.1016/j.jchf.2014.11.010

13. Fleet JL, Shariff SZ, Gandhi S, Weir MA, Jain AK, Garg AX. Validity of the International Classification of Diseases 10th revision code for hyperkalaemia in elderly patients at presentation to an emergency department and at hospital admission. BMJ open. 2012;2(6):e002011. https://doi.org/10.1136/bmjopen-2012-002011

14. Clark S, Gaeta TJ, Kamarthi GS, Camargo CA. ICD-9-CM coding of emergency department visits for food and insect sting allergy. Ann Epidemiol. 2006;16(9):696-700. https://doi.org/10.1016/j.annepidem.2005.12.003

15. Lorenzoni L, Da Cas R, Aparo UL. The quality of abstracting medical information from the medical record: the impact of training programmes. Int J Qual Health Care. 1999;11(3):209-13. https://doi.org/10.1093/intqhc/11.3.209