Journal of Biomedical Informatics 79 (2018) 41–47

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Exploration of association rule mining for coding consistency and completeness assessment in inpatient administrative health data

Mingkai Peng[a,*], Vijaya Sundararajan[b], Tyler Williamson[a], Evan P. Minty[c], Tony C. Smith[d], Chelsea T.A. Doktorchik[a], Hude Quan[a]

[a] Department of Community Health Sciences, University of Calgary, Calgary, Canada
[b] Department of Medicine, St. Vincent's Hospital, University of Melbourne, Melbourne, Australia
[c] Cumming School of Medicine, University of Calgary, Calgary, Canada
[d] Department of Computer Science, University of Waikato, Hamilton, New Zealand

ARTICLE INFO

Keywords:
Coding completeness
Coding inconsistency
Association rule mining
Inpatient administrative health data
Diagnosis code
International classification of disease

ABSTRACT

Objective: Data quality assessment is a challenging facet for research using coded administrative health data. Current assessment approaches are time and resource intensive. We explored whether association rule mining (ARM) can be used to develop rules for assessing data quality.

Materials and methods: We extracted 2013 and 2014 records from the hospital discharge abstract database (DAD) for patients between the ages of 55 and 65 from five acute care hospitals in Alberta, Canada. The ARM was conducted using the 2013 DAD to extract rules with support ≥0.0019 and confidence ≥0.5 using the bootstrap technique, and tested in the 2014 DAD. The rules were compared against the method of coding frequency and assessed for their ability to detect error introduced by two kinds of data manipulation: random permutation and random deletion.

Results: The association rules generally had clear clinical meanings. Comparing 2014 data to 2013 data (both original), there were 3 rules with a confidence difference > 0.1, while coding frequency difference of codes in the right hand of rules was less than 0.004. After random permutation of 50% of codes in the 2014 data, average rule confidence dropped from 0.72 to 0.27 while coding frequency remained unchanged. Rule confidence decreased with the increase of coding deletion, as expected. Rule confidence was more sensitive to code deletion compared to coding frequency, with slope of change ranging from 1.7 to 184.9 with a median of 9.1.

Conclusion: The ARM is a promising technique to assess data quality. It offers a systematic way to derive coding association rules hidden in data, and potentially provides a sensitive and efficient method of assessing data quality compared to standard methods.

1. Introduction

Coded administrative health data contain rich health information and are widely used in population and health services research, disease surveillance and in comparative effectiveness research [1,2]. Such routinely collected health information is called "clinically coded" because diseases documented within the medical record are assigned unique codes using the International Classification of Diseases (ICD), now in its 10th Revision (ICD-10) [3]. ICD-10 has been used by many countries throughout the world for coding cause of death and for hospital morbidities since 1994 as mandated by World Health Organization [4]. These diagnosis codes are used during all phases of analysis, including sample selection, the identification of exposures/interventions, the identification of covariates (comorbidities), and the labeling of outcomes [5].

In the case of hospital discharge records, a clinical coder abstract patients' health information into ICD codes. This is largely done for reimbursement or administrative purposes. The coding abstraction may contain some errors or inconsistencies due to incompleteness or ambiguity inherent in providers' documentation, and/or inappropriate assignment of ICD codes by the clinical coder [6]. Health systems and researchers have sought to understand the quality of coded data, typically hospital discharge data, using coding audits and validation studies [6–8]. Coding incompleteness and inconsistency have been identified as major issues in coded health data. Validation studies have found coded diagnoses to have high positive predictive values (PPV) but low or moderate sensitivities to define specific conditions [7]. Further, studies have found high coding variability between different coders. As useful as these audits and validation studies are, they are expensive and time

* Corresponding author.
E-mail address: mpeng@ucalgary.ca (M. Peng).

https://doi.org/10.1016/j.jbi.2018.02.001
Received 16 October 2017; Received in revised form 23 January 2018; Accepted 4 February 2018
Available online 06 February 2018
1532-0464/

consuming. Many of them, including systematic coding audits, do not have sample sizes sufficient to assess data quality across large numbers of diagnostic codes, or relationships between these codes. Thus, despite these efforts to evaluate data quality, research results can still be confounded by differential misclassifications. For example, in a study on coding quality of hypertension, diabetes, obesity, and depression, the coding validity (e.g., sensitivity, PPV) depended on the presence of the other conditions [9,10], demonstrating the importance of context-dependent data quality assessment.

The problems inherent in observational data necessitate the development and use of quality assessment methodologies, to determine the suitability of data for given research tasks. Data quality has been conceptualized as a multi-dimensional issue including completeness, correctness, concordance, plausibility, and currency [11]. Data quality assessment methods include comparisons with a gold standard dataset or other data sources through linkage, compatibility or agreement for two or more elements within the same data source, presence of expected data elements, or comparison between observed statistics and expected values. For coded administrative data, it is very expensive to build a comprehensive gold standard dataset, and the major clinical information are often only recorded by diagnosis codes. Assessment of validity for individual codes is difficult as the main clinical information is often embedded within clinical notes in EMR or hard copy medical records. Checking for the presence or absence of data is relatively easy as databases often employ various constraints and format checks to ensure the integrity of certain data elements. Meanwhile, various data quality frameworks or guidelines have been developed to discuss the theory of data quality assessment (e.g. conformance, completeness and plausibility), or share best practices for data quality assessment [12–14]. Such frameworks and guidelines provide important insights about which aspects to assess data quality, but are limited in terms of how to assess it. An open source data quality tool was recently published with a set of data quality rules (e.g. patient age > 0; drug quantity < 600) to generate a list of errors and warnings [15]. Those rules are very simple, and often enforced by applying various selection or exclusion criteria during data extraction process.

To clean data, quality rules are needed to detect inconsistent and anomalous entries. In database theory, function dependency, inclusion dependencies, conditional function dependencies, denial constraints, and equality-generating dependencies have examined the dependency between attributes, which often only hold true for a portion of the data [16]. For example, a woman with prostate cancer violates the integrity constraint and can be used as a data quality rule. However, this kind of rule has limited functionality as it can only cover a few sex-specific conditions. Rule development also relies on domain experts to conceptualize and design rules, which is a time-consuming and manual process.

Association rule mining (ARM) has been proposed as a cost-efficient way to develop rules for data error correction [17–19]. In epidemiological studies, many disease associations have been identified, which provide evidence to support the development of coding association rules. In coded health data, coding associations can be identified using the ARM. The identified associations are contextual and probabilistic in nature, reflecting probabilistic dependence between diagnosis codes in the coded health data. The change of coding associations can be potentially used as an indicator of data quality change over time or sites. The expected disease association is expected to be reproducible with reasonable strength in valid datasets.

In this study, we explored the use of ARM in data quality assessment of administrative health data from Alberta, Canada by comparing the performance of association rules to the currently used method of analyzing code frequency. Furthermore, we tested the performance of association rules in two simulated datasets derived from an original dataset: randomly permutated data to break the coding associations, and randomly deleted data to assess the impact of coding incompleteness on the sensitivity of association rules to detect departures in data quality.

2. Methods

2.1. Association rule mining

ARM is the process of finding clinical and interesting associations or patterns hidden in data [23]. An association rule is an expression of $X \rightarrow Y$, where X and Y are disjoint and nonempty code sets. Code sets of X and Y are the left-hand side (LHS) and right-hand side (RHS) of the rule, respectively. The strength of an association rule can be measured in terms of support and confidence. Support, denoted as P(X,Y), reflects the proportion of records with both X and Y and is used as a measure of importance of rules. Support of the LHS, P(X), reflects the proportion of records with X and indicates the coverage of the rule. Support of the RHS, P(Y), is the proportion of records with the code set of Y. Confidence, P (Y|X), determines how frequently the code set of Y appears in records with the code set of X and measures the reliability of inference made by the rule.

The first step in applying this approach is to identify all rules with support and confidence greater than a set of pre-defined threshold values. The setting of threshold values is an iterative process and depends on the context and proposed applications. If support and confidence are set to a pair of small values, a large number of rules may be identified with the limitation that many of the rules are uninformative, unreliable, and redundant. The latter is at times unavoidable, as rule mining methods often exploit the downward closure property of frequency patterns. It is important to use appropriate strategies that limit or eliminate redundant association rules. We identified and excluded the nested rules. Two rules are considered nested if they have the same RHS, and the LHS of one rule is a subset of the LHS of the other. For example, two rules of $X \rightarrow Y$ and $\{X, Z\} \rightarrow Y$ are nested. In this case, only the first rule $X \rightarrow Y$ would be kept, as it is simpler and has higher coverage.

2.2. Data source

We used the Alberta hospital discharge abstract database (DAD) for rule development and evaluation. The Alberta DAD follows the Canadian Coding Standard created by the Canadian Institution of Health Information (CIHI) and captures all patients discharged from hospitals in Alberta [24]. To create the DAD records, coders review the hospital chart and assign up to 25 diagnosis codes to each discharge using ICD-10, Canada (ICD-10-CA). To ensure international generalizability of the association rules, we mapped the ICD-10-CA codes back to the original version of ICD-10 created by the World Health Organization (WHO).

We used the 2013 DAD data for rule mining and the 2014 DAD data for assessment and evaluation. In this study, we only focused on patients in the age group of 55–65 from the top 5 high volume hospitals in Alberta. All 5 hospitals are major acute care hospitals and account for 50% of total hospital discharges for 55–65 year-old patients in Alberta.

2.3. Development of association rules

After testing a few iterations of different support and confidence thresholds to balance the coverage and reliability of rules, we set the support P(X,Y) at ≥ 0.0019. This ensured that at least 50 DAD discharge records (out of a dataset of approximately 26,000 records) would have codes of both X and Y, and a confidence of ≥ 0.5, to find rules where the coding probability of Y is greater than 50% conditional on the code set of X.

The process of rule mining was bootstrapped to improve rule quality and generalizability. We randomly sampled the training data with replacement 100 times and applied the ARM algorithm on each sample, keeping only rules identified in all 100 replicates for further analysis. Rule mining was conducted using the Apriori Algorithm from the package of *arules* in R [22]. Nested rules were identified and removed from the analysis.

2.4. Comparing coding frequency to ARM

Coding frequency is a common way to assess data quality. We compared the differences in coding frequency as well as the differences in rule confidence between 2013 and 2014. The change of coding frequency and change of rule confidence on the same diagnosis code were compared.

2.5. Assessing the sensitivity of association rules for detecting changes in simulated data quality

It is assumed that high quality data should capture the meaningful associations with the reasonable degree of strength. We introduced two types of data manipulations on the 2014 data. The first data manipulation was random permutation, which would reduce the strength of association between diagnosis codes. Random permutation could result in coding inconsistency and impair the relational conformance in the data. In the random permutation process, we shuffled a set of codes to remove them from one set of records and added them back to the other records. The proportion of permutated codes ranged from 0 to 100% with increments of 10%. Random permutation keeps the coding frequency unchanged while changing the positions of codes in the original data. The confidence of rules was recalculated after permutation.

The second type of data manipulation was random deletion. This simulates the issues of coding incompleteness as the completeness of coding are impaired after deletion. We randomly deleted 5–25% of the diagnosis codes, in an increment of 5% for each version. Deletion of the main diagnosis codes (being most responsible for the patients' stay in hospital) was prevented to ensure each record had at least one diagnosis code. Deletion is expected to change the coding frequency as well as coding association. We compared the differences in rule confidence and coding frequency between different versions of deleted data. We further assessed the number of rules being violated and the magnitude of change in rule confidence. We used the Fisher-exact test to compare the confidence of the rules in 2013 and 2014 DAD. To correct for multiplicity, we implemented the Benjamini-Hochberg adjustment to control for the false discovery rate (FDR) [20]. To reduce the type I error, we assumed a rule is violated if the adjusted p-value is blow 0.01.

3. Results

3.1. Characteristics of Alberta DAD

In total, there were 26,378 and 26,665 DAD records for patients 55–65 years of age in 2013 and 2014, respectively (Table 1). More than half of the admission records had two to five ICD diagnosis codes. The number of ICD codes in the records was similar between the two years. The volume of admissions for the five hospitals ranged from approximately 3400–7200, and was consistent between 2013 and 2014. There were approximately 4200 unique ICD codes used in these records. The distribution of ICD codes was highly right-skewed, with around 500 codes (12% unique codes) accounting for 80% of the total number of documented ICD codes. Around 56% of ICD codes were coded less than six times. The most frequently used ICD code was I10.0 (benign hypertension), which is a common condition in people aged 55–65. The four most frequently used ICD codes are all cardiovascular-related conditions, which, for this age group, is the leading cause of hospitalization in Canada.

3.2. Association rules identified from the Alberta DAD

ARM identified 421 rules; from those, 86 non-redundant rules were kept. Table 2 lists the top 10 association rules sorted by support of LHS, P(X). The confidence of rules ranges from 0.55 to 0.90. The association rules reflect various kinds of associations between codes. For example, rule 1 suggests that patients with a diagnosis of atherosclerotic heart disease often have hypertension, as high blood pressure is associated with atherosclerosis. Rule 5 reflects the association between respiratory diseases. Rules 2 and 6 capture the association of an abnormal finding and diagnosis of conditions. Presence of abnormal cardiovascular function is needed for the diagnosis of cardiovascular conditions. Rules 3, 4 and 7 reflect the relationship between etiology and manifestation codes, which is a coding convention specified in the coding standard. For example, rule 7 reflects the fact that urinary tract infection is often caused by *E. coli*. Rules 8 and 9 reveal the relationship between a risk factor, health service use and disease. All the rules can be in Appendix.

Table 1
Characteristics of 2013 and 2014 hospital discharge abstracted data (DAD) in the age group of 55–65.

	DAD in 2013	DAD in 2014
Records levels		
Total number of records	26,378	26,665
Male, N (%)	14,924 (56.5)	14,992 (56.2)
# of ICD-10 codes in each record, N (%)		
1	3966 (15.0)	3974 (14.9)
2–5	13,845 (52.5)	13,667 (51.3)
6–10	6203 (23.5)	6457 (24.2)
11–25	2364 (9.0)	2567 (9.6)
Hospital		
1	7225 (27.4)	7227 (27.1)
2	4384 (16.6)	4647 (17.4)
3	6269 (23.8)	6251 (23.4)
4	5053 (19.2)	5129 (19.2)
5	3447 (13.1)	3411 (12.8)
Coding field levels		
Total number of ICD codes	131,335	136,254
Total number of unique ICD codes	4201	4235
Top 5 frequent ICD-10 codes, N (%)		
I10.0: benign hypertension	8156 (6.2)	8215 (6.0)
I25.1: atherosclerotic heart disease	2881 (2.2)	2937 (2.2)
E11.9: Type 2 DM without complications	2012 (1.5)	1912 (1.4)
E11.5: Type 2 DM with circulatory complications	1871 (1.4)	1960 (1.4)
Z72.0: Tobacco use	1642 (1.3)	1774 (1.3)

DAD: Hospital discharge abstracted database; ICD: international classification of disease; DM: diabetes mellitus.

3.3. Support and confidence of rules in 2013 and original 2014 records

Fig. 1 presents the change of support and confidence in 2013 and 2014. For all rules, the change of coding proportion of RHS codes ranged from −0.003 to 0.003. There were 3 rules with confidence differences greater than 0.1, and 22 rules with absolute differences greater than 0.05. Large differences in coding frequency between certain subgroups (indicated by the vertical axis in Fig. 1) were identified for codes with minimal difference in overall coding frequency (as indicated by horizontal axis). For example, the proportion of the diagnosis code I25.1 in 2013 and original 2014 records was 0.1071 and 0.1078 respectively (difference was around 0.00065). However, the rule {E78.9, Z72.0} → {I25.1} had a confidence of 0.62 in 2013 and 0.51 in 2014, a difference of 0.11. This indicates that smokers {Z72.0: Tobacco use} with disorders of lipid metabolism {E78.9} are less likely to have atherosclerotic heart disease {I25.1} coded in original 2014 data than 2013 data.

3.4. Assessing the sensitivity of association rules to detect random permutation of 2014 data

As expected, we observed decreases of rule confidence as a higher proportion of codes were permutated (Fig. 2(a)). This was consistent across all association rules. It should be noted that the coding frequency in random permutation is unchanged as the permutation only changes

Table 2
Top 10 association rules sorted by the coverage of rules: support of X – P(X).

#	Association rules	Support P(X,Y)	Confidence P(Y|X)	Support of LHS, P(X)	Support of RHS, P(Y)
1	{I25.1: atherosclerotic heart disease} → {I10.0: benign hypertension}	0.0686	0.64	0.1071	0.3090
2	{R94.3: Abnormal results of cardiovascular function studies} → {I25.1: atherosclerotic heart disease}	0.0435	0.81	0.0541	0.1071
3	{E11.2: Type 2 diabetes mellitus with kidney complications} → {N08.3: Glomerular disorders in diabetes mellitus}	0.0243	0.67	0.0365	0.0270
4	{N08.3: Glomerular disorders in diabetes mellitus} → {E11.2: Type 2 diabetes mellitus with kidney complications}	0.0243	0.90	0.0270	0.0365
5	{J44.0: Chronic obstructive pulmonary disease with acute lower respiratory infection} → {J18.9: Pneumonia, unspecified}	0.0102	0.57	0.0179	0.0285
6	{E11.5: Type 2 diabetes mellitus with circulatory complications, R94.3: Abnormal results of cardiovascular function studies} → {I21.4: Acute subendocardial myocardial infarction}	0.0093	0.59	0.0157	0.0263
7	{B96.2: *E. coli* as the cause of disease classified to other chapter} → {N39.0: urinary tract infection, site not specified}	0.0122	0.82	0.0149	0.0371
8	{I25.1: atherosclerotic heart disease; Z54.0: Convalescence following surgery} → {Z95.5: Presence of coronary angioplasty implant and graft}	0.0089	0.64	0.0138	0.0285
9	{I20.8: other forms of angina pectoris} → {Z54.0: Convalescence following surgery}	0.0076	0.55	0.0115	0.0266
10	{E11.4: Type 2 diabetes mellitus with neurological complications} → {G63.2: Diabetic polyneuropathy}	0.0038	0.67	0.0056	0.0041

The arrow → indicates the direction of rules, which is from X to Y; Support, P (X,Y): the proportion of patients with both X and Y and is used as a measure of importance of rules. Confidence, P (Y|X): frequency with which code set of Y appears in the patients with the code set of X and measures the reliability of inference made by a rule. Support of LHS, P (X): the proportion of patients with X. Support of RHS, P (Y): the proportion of patients with the code set of Y.

Fig. 1. Comparison of support and confidence of rules with values from 2014 minus 2013.

Fig. 2. Confidence change of rules after random permutation (a) and random deletion (b) in 2014 data.

the position of codes in the dataset. As indicated by the black line on Fig. 2(a), the average confidence dropped from 0.72 to 0.27 when 50% of codes were permutated.

3.5. Assessing the sensitivity of association rules to detect random deletion of 2014 data

Fig. 2(b) presents the confidence change after code deletion for all the rules. The rule confidence decreased as more codes were deleted. In Fig. 2(b), the slope of each line indicates the ratio of confidence change of a rule, to support change of the code, on the right-hand side of the rule. The slopes ranged from 1.7 to 184.9 with median of 9.1. A slight change of coding frequency results in a big change of rule confidence. Therefore, rule confidence is more sensitive to data quality change than coding frequency.

Confidence change after random deletion was calculated and assessed using statistical testing. We randomly deleted 5% to 25% of

Table 3
Rules performance in the 2014 DAD data with different percentages of codes deletion.

		Hospital				
	Overall	1	2	3	4	5
Volumes of admissions	26,665	7227	4647	6251	5129	3411
Dataset with 0% of codes deleted						
# of rules applied	86	70	32	76	68	31
# of rules with absolute confidence difference > 0.2	0 (0.0)	7 (10.0)	7 (21.9)	2 (2.6)	8 (11.8)	4 (12.9)
# of rules with adjusted P value < 0.01	5 (5.8)	10 (14.3)	8 (25.0)	10 (13.2)	13 (19.1)	7 (22.6)
Dataset with 5% of codes deleted						
# of rules applied	86	68	30	74	68	31
# of rules with absolute confidence difference > 0.2	0 (0.0)	7 (10.3)	9 (30.0)	4 (5.4)	8 (11.8)	6 (19.4)
# of rules with adjusted P value < 0.01	9 (10.5)	12 (17.6)	10 (33.3)	9 (12.2)	14 (20.6)	8 (25.8)
Dataset with 10% of codes deleted						
# of rules applied	86	64	27	73	67	31
# of rules with absolute confidence difference > 0.2	1 (1.2)	7 (10.9)	9 (33.3)	4 (5.5)	11 (16.4)	9 (29.0)
# of rules with adjusted P value < 0.01	23 (26.7)	15 (23.4)	11 (40.7)	8 (11.0)	14 (20.9)	12 (38.7)
Dataset with 15% of codes deleted						
# of rules applied	86	63	26	72	63	28
# of rules with absolute confidence difference > 0.2	4 (4.7)	11 (17.5)	10 (38.5)	7 (9.7)	11 (17.5)	8 (28.6)
# of rules with adjusted P value < 0.01	45 (52.3)	22 (34.9)	13 (50.0)	11 (15.3)	18 (28.6)	11 (39.3)
Dataset with 20% of codes deleted						
# of rules applied	86	63	26	71	62	27
# of rules with absolute confidence difference > 0.2	7 (8.1)	15 (23.8)	12 (46.2)	10 (14.1)	16 (25.8)	12 (44.4)
# of rules with adjusted P value < 0.01	60 (69.8)	28 (44.4)	15 (57.7)	19 (26.8)	30 (48.4)	15 (55.6)
Dataset with 25% of codes deleted						
# of rules applied	86	61	26	70	59	24
# of rules with absolute confidence difference > 0.2	16 (18.6)	21 (34.4)	13 (50.0)	18 (25.7)	22 (37.3)	12 (50.0)
# of rules with adjusted P value < 0.01	67 (77.9)	32 (52.5)	16 (61.5)	29 (41.4)	35 (59.3)	15 (62.5)

A rule is applied if data has more than 20 records with the codes in left hand side of rules.

codes in 2014 data and recalculated the confidence after code deletion. We defined a rule as successfully applied if the number of admissions with LHS codes was ≥ 20, because estimates of rule confidence can be inaccurate if the sample sizes are too small. We expect fewer rules to be applied when fewer codes are assigned to individual discharge records. The number of rules with statistically significant changes in confidence between 2013 data and manipulated 2014 data increased with increasing coding deletion (Table 3). For example, in the dataset with 25% of the codes deleted, we identified 16 of 86 rules with absolute confidence change > 0.2, and 67 of 86 rules with adjusted p-values < 0.01 for the 2014 data while only 5 rules with adjusted p-value < 0.01 in the original 2014 data.

We also applied the rules at hospital level and compared the rule confidence between 2014 and the overall confidence in 2013. On the original hospital data (without deletion), there were around 2–8 rules with confidence differences greater than 0.2 at hospital levels and 7 to 13 rules with a statistically significant confidence change compared with overall confidence in 2013. Therefore, hospital level variations did exist in our data. The rules with large differences for a particular hospital provide a direction for data quality improvement. As expected, fewer rules can be applicable if a hospital has a smaller volume of discharges. Additionally, the number of rules with a confidence difference greater than 0.2 increased as the proportion of codes deleted increased. For example, after 25% code deletion in the 2014 data, hospitals had around 25.7–50.0% of applied rules with confidence difference greater than 0.2.

4. Discussion

Our study explored the feasibility to use ARM to develop data quality rules for coding quality assessment in administrative health data. We decreased the degree of coding associations through introduction of random permutation and random deletion in original data. As expected, after random permutation, confidence of rules dropped dramatically although the coding frequency was unchanged. Rule confidence was more sensitive to data quality change than coding frequency, as indicated by the slope of change. Association rules provide an efficient way to identify the difference in quality between datasets. However, further investigations are still needed to confirm whether change in rule confidence is a data quality issue, as the confidence changes could also come with the advancement of health prevention or treatment, introduction of new health policy, or change of coding practices.

4.1. Why is data mining coding association useful for data quality assessment?

ARM provides a systematic way to summarize the relationships between data elements using statistical measures, and an automated way to derive association rules with reasonable clinical meanings. The method can be easily implemented and transferable to other types of data (e.g. electronic medical records). Association rules can be used to assess the completeness of data by checking whether the observed proportion of patients with a diagnosis code and known characteristics differ from an expected proportion. For example, in a study on disease X, the association rules containing the code for disease X at the RHS can be applied to check the coding quality of that disease. If the confidence of the rules in one dataset is lower than in another dataset, this may indicate potential issues of false negatives for disease X in the first dataset (given that other settings of data collection between the two datasets were similar).

Checking the enforcement of expected associations in a database has been used in literature to assess data quality. In drug safety surveillance, reference drug-outcome sets with negative (i.e. no effect) and positive (i.e. increased effect) controls are used to check whether a research dataset is valid, by checking reproducibility of expected drug-outcome associations [21]. Faulconer and de Lusignan compared the prevalence of chronic obstructive pulmonary disease (COPD) in study

data to the literature-reported prevalence in different age and sex specific groups [25]. Kahn et al. proposed a conceptual framework for assessing the quality of electronic health data with attribute dependency rules [26]. In epidemiological studies, differential misclassification where coding completeness differs between groups may result in a biased estimation either towards or away from the null value [27]. Association rules could have the potential to check whether differential misclassification exists in a subpopulation within the data.

4.2. How can ARM for data quality assessment be applied in practice?

ARM can be used to assess the temporal consistency of coding within one site as well as coding consistency across different sites. Data from one year or site may be used as a basis for the development of association rules and applied the derived rules for comparison over time or sites. For example, observational studies often require the combination of multiple disparate data sources. Such studies are susceptible to heterogeneous data quality originating from different data capture processes, different source populations with varied patient demographics, or from different health care systems [28]. We can apply ARM on one site and test the learned rules in other sites to check the confidence change of rules between sites. This approach to data quality assessment is flexible and cost-effective, and does not require comparison to external data sources or specific audits.

Previously, data quality assessment relied comparison to a “gold standard” data source. Some examples of “gold standards” include chart review data, registry data, or information supplied by patients or physicians [11]. Notably, not all of these external reference data sources can be considered true gold standards, and many are simply not available [29]. It is possible to conduct rule mining on a “gold standard” or high-quality dataset to develop benchmark association rules, and subsequently applied to other datasets to assess coding quality. For example, data linkage is a useful way to enhance the data quality as it provides more complete data on individual subjects [30]. We can apply the ARM on the linked data and use the learned rules as reference standards to check the data quality.

Violation of an association rule in a record is not necessarily a sign of error, but large differences in the confidence of a rule between data sets call for an investigation of data quality. The variability of confidence could reflect true differences in clinical practices, such as workflow, or variation in data capture and treatment strategies [26,31]. Data holders need to interpret observed variability to determine if the difference can be explained by other factors, such as coding guidelines, coder training, or experience. Once the issues of data variability is clarified, the appropriate study design can be adopted to minimize pitfalls in the dataset, and results can be interpreted with increased awareness. Not only can association rules help to suggest variability in the data, but they also provide a granular assessment of data quality including the identification of differential misclassification, false-positives, and false-negatives in a study [32]. Furthermore, they also provide an opportunity to be part of an ongoing data quality improvement program to supplement or replace widely used and expensive audit programs.

4.3. Future development of ARM for data quality assessment

ARM is an unsupervised data mining method that generates associations without any prioritization. It can return duplicated or unusable rules without support from clinical patterns. In this study, we identified nested rules and retained only the most unique rules with the same RHS. This is an effective method for selecting rules with high coverage. Further advancements in pruning methods will improve the ability to cull redundant rules. Incorporation of clinical perspectives could provide additional insight into the appropriateness and applicability of identified association rules. The rules can also be divided into different clinical categories (cardiovascular disease, oncology, *etc.*) to conduct task-dependent data quality assessments for different research questions. In future work, we will collaborate with clinical experts and systematically develop data quality rules.

4.4. Limitations

Our study demonstrates the advantages and benefits of ARM for the assessment of coded hospital data quality. However, the following limitations should be noted. First, our study demonstrates the feasibility of association rules for data quality assessment using simulations, and does not provide any validated rules for use in other studies. We also only focused on one age group; however, the method can be generalized to other age groups. Second, although we pruned the number of rules based only on statistical measures, we believe that incorporation of clinical judgement would be beneficial in the process of refining initially developed association rules. This could be conducted through a physician panel review process.

5. Conclusion

Data quality assessment is one of the most challenging problems researchers face when using routinely collected administrative health data. The ARM is a promising technique to systematically assess data quality as demonstrated in simulated data with artificially introduced error patterns. Further work comparing association rules versus other data quality assessment methods is required to confirm that it is indeed an effective way to assess data quality.

Conflict of interest

None.

Acknowledgements

This work was supported by the Canadian Institute of Health Research (CIHR).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2018.02.001.

References

[1] W.R. Hersh, M.G. Weiner, P.J. Embi, J.R. Logan, P.R. Payne, E.V. Bernstam, H.P. Lehmann, G. Hripcsak, T.H. Hartzog, J.J. Cimino, et al., Caveats for the use of operational electronic health record data in comparative effectiveness research, Med. Care 51 (8 Suppl 3) (2013) S30–S37.
[2] C. Robitaille, S. Dai, C. Waters, L. Loukine, C. Bancej, S. Quach, J. Ellison, N. Campbell, K. Tu, K. Reimer, et al., Diagnosed hypertension in Canada: incidence, prevalence and associated mortality, CMAJ 184 (1) (2012) E49–E56.
[3] N. Jette, H.D. Quan, B. Hemmelgarn, S. Drosler, C. Maass, L. Moskal, W. Paoin, V. Sundararajan, S. Gao, R. Jakob, et al., The development, evolution, and modifications of ICD-10 challenges to the international comparability of morbidity data, Med. Care 48 (12) (2010) 1105–1110.
[4] C. De Coster, H. Quan, A. Finlayson, M. Gao, P. Halfon, K.H. Humphries, H. Johansen, L.M. Lix, J.C. Luthi, J. Ma, et al., Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium, BMC Health Serv. Res. 6 (2006) 77.
[5] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.C. Luthi, L.D. Saunders, C.A. Beck, T.E. Feasby, W.A. Ghali, Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, Med. Care 43 (11) (2005) 1130–1139.
[6] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, Measuring diagnoses: ICD code accuracy, Health Serv. Res. 40 (5) (2005) 1620–1639.
[7] H. Quan, B. Li, L.D. Saunders, G.A. Parsons, C.I. Nilsson, A. Alibhai, W.A. Ghali, I. Investigators, Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database, Health Serv. Res. 43 (4) (2008) 1424–1441.
[8] T. Henderson, J. Shepheard, V. Sundararajan, Quality of diagnosis and procedure coding in ICD-10 administrative data, Med. Care 44 (11) (2006) 1011–1019.
[9] C. van Walraven, P.C. Austin, A. Jennings, H. Quan, A.J. Forster, A modification of

the Elixhauser comorbidity measures into a point system for hospital death using administrative data, Med. Care 47 (6) (2009) 626–633.
[10] M. Peng, D.A. Southern, T. Williamson, H. Quan, Under-coding of secondary conditions in coded hospital health data: Impact of co-existing conditions, death status and number of codes in a record, Health Informat. J. (2016).
[11] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.
[12] A.P. Reimer, A. Milinovich, E.A. Madigan, Data quality assessment framework to assess electronic medical record data for use in research, Int. J. Med. Inform. 90 (2016) 40–47.
[13] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data, eGEMs 4 (1) (2016).
[14] M.G. Kahn, M.A. Raebel, J.M. Glanz, K. Riedlinger, J.F. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research, Med. Care 50 (2012).
[15] V. Huser, F.J. DeFalco, M. Schuemie, P.B. Ryan, N. Shang, M. Velez, R.W. Park, R.D. Boyce, J. Duke, R. Khare, Multisite evaluation of a data quality tool for patient-level clinical data sets, eGEMs 4 (1) (2016).
[16] W.F. Fan, Data quality: from theory to practice, Sigmod. Rec. 44 (3) (2015) 7–18.
[17] P. Alpar, S. Winkelsträter, Assessment of data quality in accounting data with association rules, Expert Syst. Appl. 41 (5) (2014) 2259–2268.
[18] J. Hipp, U. Güntzer, U. Grimmer, Data quality mining-making a virute of necessity, DMKD (2001).
[19] F. Chiang, R. Miller, J: Discovering data quality rules, Proceed. VLDB Endowment 1 (1) (2008) 1166–1177.
[20] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining, first ed., Pearson Addison Wesley, Boston, 2005.
[21] Canadian Institute for Health Information: Canadian Coding Standards for ICD-10-CA and CCI for 2015. In. Ottawa: CIHI, 2015.
[22] M. Hahsler, B. Grün, K. Hornik, Arules – a computational environment for mining association rules and frequent item sets, J. Stat. Softw. 14 (15) (2005).
[23] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate - a practical and powerful approach to multiple testing, J. Roy. Stat. Soc. B Met. 57 (1) (1995) 289–300.
[24] P.B. Ryan, M.J. Schuemie, E. Welebob, J. Duke, S. Valentine, A.G. Hartzema, Defining a reference set to support methodological research in drug safety, Drug. Saf. 36 (Suppl 1) (2013) S33–S47.
[25] E.R. Faulconer, S. de Lusignan, An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar, Inform. Prim. Care 12 (4) (2004) 243–254.
[26] M.G. Kahn, M.A. Raebel, J.M. Glanz, K. Riedlinger, J.F. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research, Med. Care 50 (Suppl) (2012) S21–S29.
[27] A.M. Jurek, S. Greenland, G. Maldonado, T.R. Church, Proper interpretation of non-differential misclassification effects: expectations vs observations, Int. J. Epidemiol. 34 (3) (2005) 680–687.
[28] D. Madigan, P.B. Ryan, M. Schuemie, P.E. Stang, J.M. Overhage, A.G. Hartzema, M.A. Suchard, W. DuMouchel, J.A. Berlin, Evaluating the impact of database heterogeneity on observational study results, Am. J. Epidemiol. 178 (4) (2013) 645–651.
[29] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C.H. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inform. 46 (5) (2013) 830–836.
[30] P.D. Faris, W.A. Ghali, R. Brant, C.M. Norris, P.D. Galbraith, M.L. Knudtson, A. Investigators, Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses, J. Clin. Epidemiol. 55 (2) (2002) 184–191.
[31] S. Schneeweiss, J. Avorn, A review of uses of health care utilization databases for epidemiologic research on therapeutics, J. Clin. Epidemiol. 58 (4) (2005) 323–337.
[32] S. Greenland, J.M. Robins, Confounding and misclassification, Am. J. Epidemiol. 122 (3) (1985) 495–506.