Using Electronic Medical Record to Identify Patients With Dyslipidemia in Primary Care Settings: International Classification of Disease Code Matters From One Region to a National Database

Biomedical Informatics Insights
1–7

DOI: 10.1177/1178222616685880

Justin Oake[1,2], Erfan Aref-Eshghi[1,2], Marshall Godwin[1], Kayla Collins[3], Kris Aubrey-Bassler[1], Pauline Duke[1], Masoud Mahdavian[4] and Shabnam Asghari[1,2]

[1]Centre for Rural Health Studies, Faculty of Medicine, Memorial University of Newfoundland St. John's, NL, Canada.
[2]Primary Healthcare Research Unit, Department of Family Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada.
[3]Newfoundland and Labrador Centre for Health Information, St. John's, NL, Canada.
[4]Department of Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada.

ABSTRACT

OBJECTIVE: To assess the validity of the International Classification of Disease (ICD) codes for identifying patients with dyslipidemia in electronic medical record (EMR) data

METHODS: The EMRs of patients receiving primary care in St. John's, Newfoundland and Labrador (NL), Canada, were retrieved from the Canadian Primary Care Sentinel Surveillance Network database. International Classification of Disease codes were first compared with laboratory lipid data as an independent criterion standard, and next with a "comprehensive criterion standard," defined as any existence of abnormal lipid test, lipid-lowering medication record, or dyslipidemia ICD codes. The ability of ICD coding alone or combined with other components was evaluated against the two criterion standards using receiver operating characteristic (ROC) analysis, sensitivity, specificity, negative predictive value (NPV) and Kappa agreement. (No specificity was reported for the comparison of ICD codes against the comprehensive criterion standard as this naturally leads to 100% specificity.)

RESULTS: The ICD codes led to a poor outcome when compared with the serum lipid levels (sensitivity, 27%; specificity, 76%; PPV, 71%; NPV, 33%; Kappa, 0.02; area under the receiver operating characteristic curve (AUC), 0.51) or with the comprehensive criterion standard (sensitivity, 32%; NPV, 25%; Kappa, 0.15; AUC, 66%). International Classification of Disease codes combined with lipid-lowering medication data also resulted in low sensitivity (51.2%), NPV (32%), Kappa (0.28), and AUC (75%). The addition of laboratory lipid levels to ICD coding marginally improved the algorithm (sensitivity, 94%; NPV, 79%; Kappa, 0.85; AUC, 97%).

CONCLUSIONS: The use of ICD coding, either alone or in combination with laboratory data or lipid-lowering medication records, was not an accurate indicator in identifying dyslipidemia.

KEYWORDS: electronic medical records, algorithm, dyslipidemia, ICD codes, validation

RECEIVED: August 20, 2016. **ACCEPTED:** October 19, 2016

PEER REVIEW: Seven peer reviewers contributed to the peer review report. Reviewers' reports totaled 1355 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Newfoundland and Labrador Centre for Applied Health Research (NLCAHR).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Shabnam Asghari, Department of Family Medicine, Center for Rural Health Studies, Faculty of Medicine, Memorial University of Newfoundland, Agnes Cowan Hostel, Room 425, Health Sciences Centre, 300 Prince Philip Drive, St. John's, NL A1B 3V6, Canada. Email: shabnam.asghari@med.mun.ca

Introduction

Dyslipidemia is one of the most modifiable risk factors for cardiovascular disease (CVD), a significant chronic condition which imposes a substantial burden of morbidity and mortality; it is the leading cause of death worldwide.[1] As a result, dyslipidemia has been widely studied for projecting CVD population incidences, identifying CVD high-risk groups and evaluating prevention strategies for reducing individual and population risks. Accurate identification of dyslipidemia in the population is crucial to enhancing the ability to perform epidemiologic studies, including health systems planning, resource allocation, and pharmacoepidemiologic investigations to promote preventive and acute care programs related to CVDs.

Medico-administrative data, recorded according to the International Classification of Disease (ICD) coding system, have increasingly been used in large-scale studies in recent years due to higher accessibilities and lower costs compared with the population-based surveys. These data allow for the passive surveillance of the disease and are available at lower costs compared with active surveillance, particularly in Canada where a centralized government-based structure in health care exists. As the reliability of the findings from such studies depends on the accuracy of the medico-administrative billing data, studies have attempted to assess the reliability of such coding systems. Although the outcomes of these studies have

been varied according to the type of data used and the disease under study, inaccuracy of the ICD codes for the purpose of diagnosis of medical conditions is frequently reported.[2–4]

The majority of such studies have been performed on CVDs rather than their risk factors such as dyslipidemia and diabetes. The outcomes of these studies have questioned the sensitivity and specificity of medico-administrative record data in identifying the trends of stroke and CVDs,[5,6] although stroke coding has been found to be useful for high-level comparisons, particularly when compared with other diseases.[5] ICD codes have been particularly shown to have restricted potentials for patients with dyslipidemia. This has been shown by the result of few previous studies available using secondary data for lipid research. An algorithm developed by an American study reported that 62.3% of patients with dyslipidemia were not recorded by the ICD codes.[4] Another study in a large US medical insurance claims database found that only 15% of laboratory-defined patients had a dyslipidemia diagnosis.[7] In addition, some studies suggest more than one record of the ICD coding during 2 or 3 years to identify patients with a particular disease condition.[8–10] One disadvantage of ICD codes for CVDs is the inability to ascertain severity, which is the most important prognostic variable in the surveillance of CVDs. Coding for CVDs risk factors as an alternative, however, has the potential to tackle these limitations and enrich the utility of medico-administrative data for surveillance of CVDs; yet, they have been rarely examined using medico-administrative data, and a handful of such studies on dyslipidemia, the major modifiable risk factor to CVDs, have limitations from being performed using databases that contain no record of other potential identifying markers of dyslipidemia besides the ICD coding, such as the history of lipid-lowering medication use or laboratory lipid levels.[7]

The recent emergence of electronic medical records (EMRs), however, seems to have eliminated this barrier. Patients' records from a growing number of health providers are being collected in electronic format, which not only provides access to medico-administrative data (eg, ICD codes) but also contain information on medical histories, comorbidities, laboratory test results, and medication use.[11–13] The regular management of dyslipidemia is conducted using lipid-lowering medications and routine laboratory testing. The structured format of an EMR database would, therefore, be ideal for evaluating the accuracy of medico-administrative records compared with other diagnostic criteria. This study examines the degree to which the ICD codes alone, or in combination with lipid-lowering medication use or laboratory lipid levels, can predict a diagnosis of dyslipidemia relative to laboratory data or a more elaborate criterion standard. This investigation is conducted using the multidisease record surveillance system within the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), which contains the ICD codes and lipid-lowering medication records by primary care physicians as well as a link to laboratory data for every record. This strategy is particularly important because not all of the existing EMR databases have the entire components of criterion standard algorithms to allow for comparison.

Methods

This cross-sectional study was designed using the secondary analysis of data from EMRs of primary care clinics in St. John's, Newfoundland and Labrador (NL), Canada. Records of patients with complete lipid profiles undertaken during January 1, 2009, to December 31, 2010, were included. The records from the clinics in St. John's, NL, can be a good representative of the health records in the province, as more than 40% of the population of NL resides and commutes through the St. John's metropolitan region.

Study database

The multidisease record surveillance system within the CPCSSN is commonly used for chronic disease surveillance in primary care and for conducting primary care research.[14–16] This database contains the EMRs of family physicians which are abstracted quarterly and uploaded to a de-identified system to regional and central pan-Canadian databases. An electronic chart abstraction was performed using the EMRs of clinics in St. John's which form part of the NL component of the CPCSSN.[13] The data for this study come from three different sections of EMRs:

1. ICD coding for disease diagnosis which is an AutoFill section of EMR and is completed when the physician selects a disease diagnosis;
2. Laboratory results, which are electronically linked to the Laboratory Information System database. These data are independent of the EMR and are completed at the laboratory;
3. Medication prescriptions, which are entered into the EMR by physicians at every visit according to the medication prescribed during the visit.

Study population

The study population consisted of subjects from the NL component of the CPCSSN database aged 20 years or older. Among the total patients who received health care services during the study timeframe (January 1, 2009, to December 31, 2010), 4400 patients were identified as having had a complete lipid profile taken. Pregnant women were excluded from the analysis (Figure 1).

Algorithm development and evaluation

The algorithm validation/testing was performed in the following steps:

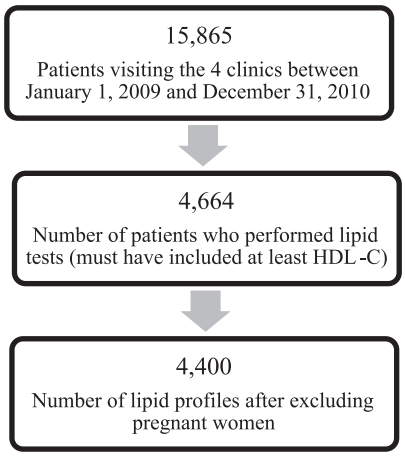

Figure 1. Study population flow chart.

1. To determine the performance of ICD coding in comparison with laboratory lipid measurements as an "independent criterion standard," ICD coding was compared with the lipid levels from laboratory data.
2. In the second step, a combination of the three criteria (laboratory lipid levels, ICD codes, and lipid-lowering medication use) was used to develop a "comprehensive criterion standard" algorithm to identify any record of patients with dyslipidemia in our database, as follows:
   (a) Any ICD of dyslipidemia recorded (ICD code v.9-272);
   (b) Any laboratory serum measurements of lipid levels deviating from the cutoffs defined by the Canadian guidelines for the diagnosis and management of dyslipidemias (Table 1);[17]
   (c) Any record of using lipid-modifying medications during the study period.

Use of lipid-modifying agents (HMG-CoA reductase inhibitors, fibrates, bile acid sequestrants, nicotinic acid, and other agents) was identified using the text record of the medication name and/or Anatomical Therapeutic Chemical (ATC) codes. Every group was assumed to have dyslipidemia independent of each other. For instance, the patients with normal lipid levels, but with a history of lipid modification use, were categorized as having dyslipidemia because the medication therapy is expected to alter the lipid levels.

Given that the local clinicians had determined that these three criteria would likely detect the significant majority of patients with dyslipidemia in the EMRs, we deemed that the existence of any one or several of these three criteria in an individual would be a criterion standard diagnosis of dyslipidemia. Furthermore, it is common in population screening studies to have results from one or more tests investigating the same condition, none of which can be considered the "criterion standard" alone.[18] In addition, the eMERGE network, a consortium of 5 US institutions linked to secure encrypted EMR data that are designed with the aim of identifying disease phenotypes from EMR, suggests the use of the above three criteria to detect the phenotype of low-density lipoprotein cholesterol (LDL-C) dyslipidemia from EMRs.[19,20]

Table 1. Healthy levels of serum lipids for Canadian adults.[17]

LIPID COMPONENT	NORMAL LEVELS
Total cholesterol (TC)	<5.2 mmol/L
Triglycerides (TG)	<1.7 mmol/L
Low-density lipoprotein cholesterol (LDL-C)	<3.4 mmol/L
High density lipoprotein cholesterol (HDL-C)	>1.0 mmol/L for men >1.3 mmol/L for women

The performance of ICD coding against this comprehensive criterion standard was then examined. Table 2 provides a detailed description of the three indicators, as well as the "criterion standard."

3. The combinations of ICD coding with medication use or laboratory lipid data were compared against the "comprehensive criterion standard."
4. The above analysis was repeated using the national CPCSSN data between 2010 and 2012 to assess the replicability of the findings.
5. In the end, the association of ICD coding with other factors associated with dyslipidemia, including age, sex, diabetes, hypertension, medication use, smoking, and body mass index (BMI), was examined to determine the factors with the most influence on the ICD coding. We assumed that individuals with different demographics and comorbidities may have variable ICD coding accuracy due to the difference in their management.

Statistical analysis

Analysis, using 2 × 2 table formats, was conducted to evaluate the variation in the diagnosis of dyslipidemia. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Kappa agreement, and the area under receiver operating characteristic curve (AUC) were calculated for every algorithm in comparison with the "criterion standard."

Sensitivity was defined as the proportion of patients identified by the testing algorithms who had dyslipidemia according to the "criterion standard." *Specificity* was defined as the proportion of patients excluded by the testing algorithms who did not have dyslipidemia according to the "criterion standard." *PPV* was defined as the proportion of patients with dyslipidemia identified by the testing algorithms that were also confirmed by the "criterion standard." *NPV* was defined similarly for patients who did not have dyslipidemia according to the testing algorithms (Table 3). The Kappa agreement was calculated between every testing algorithms and the "criterion standard." The Kappa values of 0 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, 0.81 to 0.90 and 0.91

Table 2. Number of patients with dyslipidemia and associated prevalence categorized by algorithm.

	DEFINITION	NO. OF CASES	APPARENT PREVALENCE (%)
Situation A An abnormal lipid level is reported in laboratory data	The most recent lipid profile (total cholesterol, high-density lipoprotein cholesterol [HDL-C], low-density lipoprotein cholesterol [LDL-C], and triglycerides) on an individual showed one component of the lipid profile was not in the normal range as recommended by the Canadian lipid guidelines: total cholesterol >5.2 mmol/L, HDL-C <1.0 mmol/L, LDL-C >3.4 mmol/L, and triglycerides >1.7 mmol/L (Statistics Canada, 2011; http://www.statcan.gc.ca/pub/82-625-x/2012001/article/11732-eng.htm)	3035	69.0
Situation B The individual is on a lipid-lowering drug	Any record of using a lipid-modifying agent including statins, fibrates, bile acid sequestrants, nicotinic acid and derivatives, and other lipid-modifying agents during the study period or an Anatomical Therapeutic Chemical Classification System (ATC) code C10 for these lipid-modifying agents, (WHO, 2012—within 2 years before the date the lipid tests were done; http://www.who.int/classifications/atcddd/en/)	1556	35.4
Situation C The individual has a diagnosis of abnormal lipids	There is a diagnosis of a "disorder of lipid metabolism" (ICD code 272) according to ICD code 272 in the EMR; (http://icd9cm.chrisendres.com/index.php?action=child&recordid=2055)	1147	26.1
Comprehensive criterion standard Any one or more of A, B, and C above	Patients were deemed to have dyslipidemia if they fitted into either one or more of the *Situations* above: (A) had one component of the lipid profile not in the normal range recommended by Canadian lipid guidelines; (B) there was record of using a lipid-modifying agent; (C) had an ICD code 272 diagnosis on record	3573	81.2

Table 3. Definitions for sensitivity, specificity, negative predictive value, and positive predictive value.

		CRITERION STANDARD	
		DYSLIPIDEMIA	HEALTHY LIPID
ICD code 272	Dyslipidemia	A (true-positive)	B (false-positive)
	Healthy Lipid	C (false-negative)	D (true-negative)

Sensitivity: $A/(A+C)\times 100$; specificity: $D/(D+B)\times 100$; positive predictive value: $A/(A+B)\times 100$; negative predictive value: $D/(D+C)\times 100$.

to 1.0 indicate poor, slight, fair, good, very good and excellent agreements, respectively.[21,22] A receiver operating characteristic (ROC) curve for each algorithm was measured against the "criterion standard." ROC curves were obtained by calculating the sensitivity and specificity of the test and plotting the sensitivity against 1-specificity. AUC of the ROC is a reflection of how reliable the test is in distinguishing between patients with disease and those without the disease.[23] The AUCs greater than 0.8 are considered to have high accuracy, whereas an AUC in the range of 0.7 to 0.9 indicates moderate accuracy, 0.5 to 0.7 indicates low accuracy, and 0.5 a chance result.[24] Prevalence was estimated according to the number of patients with dyslipidemia identified by each definition. A logistic regression analysis was performed to determine which factors influenced ICD coding for dyslipidemia. The significance of effects was evaluated at $\alpha = .05$.

All the analyses were conducted using Stata SE 11.2 (Stata Corp., College Station, Texas, USA).

Ethics

The Human Research Ethics Authority, Memorial University of Newfoundland, reviewed and approved the study protocol. All the data were de-identified before the analysis.

Results

The EMRs from a total of 4400 patients (mean age, 58.1 ± 14.8 years; 58.8% women) were included in the study. The population had a BMI of 31.1 ± 15.8, 42.3% of whom were present/former smokers. The prevalence of hypertension and diabetes was 33.5% and 15.2%, respectively. Among this population, 3573 patients had dyslipidemia during the study period according to the "comprehensive criterion standard" definition (prevalence of 81.2%). As shown in Table 2, among all patients, 69.0% were diagnosed with dyslipidemia according to the laboratory results (independent criterion standard), 26.1% had an ICD coding for dyslipidemia, and 35.4% had used one or more lipid-lowering medication. The overlap of these three components is shown as a Venn diagram in Figure 2.

The ICD codes resulted in a poor outcome when compared with the independent criterion standard (serum lipid levels). This analysis led to a sensitivity of 27.0%, specificity of 76.7%, PPV of 71.1%, NPV of 33.1%, a Kappa agreement of 0.02 and an AUC of 0.51.

In the second attempt, ICD coding was compared with the comprehensive criterion standard as shown in Table 4. The diagnostic data alone (ICD coding) led to the lowest sensitivity

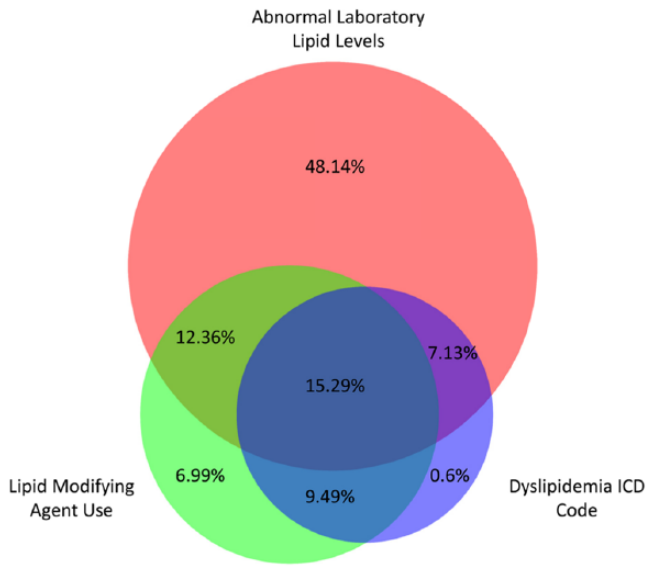

Figure 2. Venn diagram of the three components of the criterion standard algorithm.

(32.1%), NPV (25.4%), Kappa agreement (0.151), and AUC (66.1%) compared with the "comprehensive criterion standard."

The ICD coding data, combined with lipid-lowering medication data, also yielded a poor result. This algorithm led to a low sensitivity (51.2%), NPV (32.2%), Kappa agreement (0.283), and AUC (75.6%) (Table 4). The use of the laboratory results (Table 4) had the highest sensitivity (84.9%), NPV (60.6%), Kappa agreement (0.680), and AUC (92.5%) compared with ICD coding or prescription medication use on their own. Combining the laboratory results together with lipid-lowering medication data further increased the sensitivity (99.6%), NPV (98.1%), Kappa agreement (0.988), and AUC (99.8%) (Table 4).

To replicate our results, we assessed the repeatability of our findings using the Canada-wide records of 2010-2012 in a similar approach. This analysis also showed the lowest sensitivity (32.1%), NPV (26.0%), Kappa agreement (0.15), and AUC (0.66) for ICD coding compared with the "comprehensive criterion standard." The ICD coding data, combined with lipid-lowering medication data, also yielded a low sensitivity (51.2%), NPV (32.2%), Kappa agreement (0.28), and AUC (0.76).

Given that the ICD coding was such a poor predictor of dyslipidemia in the EMRs, an additional analysis was conducted using logistic regression to explore which demographic factors and comorbidities may influence the ICD coding for dyslipidemia (Table 5). Results from this analysis showed that patients prescribed lipid-lowering medication were very likely (odds ratio [OR], 9.75; 95% CI, 6.82-13.95; $P < .001$) to have the ICD codes for dyslipidemia.

Discussion

This study has demonstrated that using the ICD coding alone is an inaccurate indicator of dyslipidemia. The ICD coding data represented a substantial underestimation of dyslipidemia cases. The use of ICD codes in combination with data from laboratory results or lipid-lowering medication added insignificant marginal value to the respective algorithms. In addition, the ICD coding data alone yielded the most false-negatives. The ICD codes also performed poorly when compared with the serum lipid levels alone as a independent criterion standard. In addition, the case identification for dyslipidemia did not improve when we used one ICD code for a longer duration (during 2 years/during 3 years) or when we used more than one ICD code during 3 years to identify patients with dyslipidemia (data not shown).

Although the ICD codes are reported to be able to accurately identify patients with many medical conditions such as ischemic heart disease,[25] diabetes mellitus,[26] and preeclampsia,[27] their potential for patients with dyslipidemia is restricted. Our results are consistent with the few previous studies available using secondary data for lipid research. In support of our notion regarding the inaccuracy and unreliability of using the ICD coding data for lipid research, we learned that an American study created an algorithm for detecting dyslipidemia and diabetes. The algorithm identified 58.4% of patients with hyperlipidemia, 62.3% of whom were not recorded as having dyslipidemia by the ICD codes.[4] Another study in a large US medical insurance claims database found that only 15% of laboratory-defined patients had a dyslipidemia diagnosis.[11] In the province of Alberta, Canada, Kokotailo and Hill[28] showed that although the medico-administrative billing system is a good indicator of stroke and some of its risk factors including diabetes mellitus and hypertension, the identification of hyperlipidemia is not confidently made where the sensitivity was reported to be 57%. The exact reason for incomplete coding of dyslipidemia is unclear. Kokotailo et al[28] considered "a lack of perceived importance by health technologist coders, or a lack of time to code everything," as the putative reason.

The use of advanced technologies in disease coding may be a solution to this problem and to improve the accuracy of ICD codes. Natural language processing, a range of computational techniques for analyzing written or oral texts for the purpose of achieving human-like language processing, has been applied to the EMRs and have shown to improve the accuracy of case definition for inflammatory bowel disease,[29] venous thromboembolic disease,[30] and cancer.[31] Multimodal fusion/interaction are multiple modes of interaction with a system which provides several distinct tools for input and output of data. This technique has been implemented in different aspects of medical diagnosis including the processing of brain imaging[32] and magnetic resonance imaging (MRI)[33] data, as well as discriminative learning for Alzheimer's disease diagnosis.[34] This, however, has rarely been implemented in disease coding and EMR processing. The use of this method might have a potential for improving the disease coding in medical administrative data.

Consideration ought to be given to possible limitations when interpreting and applying these data. The possibility of information bias and data inaccuracy cannot be ignored, despite the fact that the direct link between the Laboratory Information

Table 4. Sensitivity, specificity, and predictive values of all combinations of situations A, B, and C compared with comprehensive criterion standard.

	SENSITIVITY	SPECIFICITY	POSITIVE PREDICTIVE VALUE	NEGATIVE PREDICTIVE VALUE	KAPPA VALUE	AUC (95% CI)
A	85%	100%	100%	60%	0.68	0.92 (0.92–0.93)
B	43%	100%	100%	29%	0.23	0.72 (0.71–0.72)
C	32%	100%	100%	25%	0.15	0.66 (0.65–0.67)
A and B	99%	100%	100%	98%	0.99	1.00 (0.99–1.00)
A and C	94%	100%	100%	79%	0.85	0.97 (0.96–0.97)
B and C	51%	100%	100%	32%	0.28	0.76 (0.75–0.76)

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval.

Table 5. Factors associated with ICD coding for dyslipidemia.

VARIABLE	ODDS RATIO (95% CONFIDENCE INTERVAL)
	ICD code for dyslipidaemia
Sex (male)	1.318 (0.943–1.843)
Aged 41-64 years	2.921 (1.060–8.050)*
Aged ⩾65 years	4.276 (1.521–12.021)*
Hypertension	1.502 (1.062–2.125)*
Diabetes	1.299 (0.847–1.990)
Former/current smoker	1.040 (0.876–1.233)
BMI ⩾30	1.044 (0.734–1.484)
Lipid-lowering medication user	9.754 (6.821–13.947)^

Abbreviations: BMI, body mass index; ICD, International Classification of Disease. Significant at *$P < .05$; ^$P < .001$.

System and EMRs should decrease the probability of data entry errors. Also, these results are based on ICD version 9.0. Newer versions of ICD, including ICD 10.0 and ICD 11.0, have been released. It is notable that the case definition for dyslipidemia does not change considerably between these versions, and thus, the findings of this study can be applied to newer versions of ICD.

One may question the representativeness of the study as this study focused only on the data from EMR clinics in St. John's, NL, Canada, to assess the validity of the ICD coding in identifying patients with dyslipidemia. We assessed the repeatability of our findings using a national dataset. A similar approach was performed using the data from CPCSSN across Canada between 2010 and 2012. This analysis showed the low accuracy of the ICD coding as is seen in this study.

In addition, our data only apply to primary care, and it may not be extended to hospital-based and specialized care where more severe and acute cases of CVDs and dyslipidemia exist.

Conclusions

Using secondary data to identify patients diagnosed with dyslipidemia could involve information on laboratory values, lipid-lowering medication data, or diagnostic data. Often, a given secondary database will have only one of these pieces of information. Results from laboratory data may only have levels of lipids, pharmacy data may only have prescription records, and provincial billing databases may only have diagnostic data. Databases that contain all three of these (lipid levels, medications, and diagnoses) can be used to understand how either one or any two of these pieces of information can predict whether dyslipidemia exists in an individual. The CPCSSN database contains all three of these types of information.

Although the ICD codes have typically been used for the diagnosis of many medical conditions in both research and practice, our research suggests that they are not an accurate indicator of patients with dyslipidemia. Therefore, caution ought to be taken into account when using the databases established according to the ICD codes for research involving dyslipidemia.

Author Contributions

SA and JO helped with the study design, data collection, and statistical analysis. SA, JO, EA-E, MG, PD, KC, KA-B, and MM helped with the interpretation of results. SA, JO, and EA-E helped with manuscript writing. EA-E, PD, KC, KA-B, MG, PD, and MM provided critical comments on the manuscript.

Acknowledgements

The authors are grateful to Emily Eaton, Andrea Pike, Adam Pike, and Scott Lee for their comments and contributions on earlier drafts of the manuscript.

REFERENCES

1. *Global status report on noncommunicable diseases 2010*. Geneva, Switzerland: World Health Organization; 2011.
2. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*. 2005;43:480–485.
3. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: iCD code accuracy. *Health Serv Res*. 2005;40:1620–1639.
4. Kandula S, Zeng-Treitler Q , Chen L, Salomon WL, Bray BE. A bootstrapping algorithm to improve cohort identification using structured data. *J Biomed Inform*. 2011;44:S63–S68.

5. Tirschwell DL, Longstreth WT Jr. Validating administrative data in stroke research. *Stroke*. 2002;33:2465–2470.
6. Mayo NE, Goldberg MS, Levy AR, Danys I, Korner-Bitensky N. Changing rates of stroke in the province of Quebec, Canada: 1981-1988. *Stroke*. 1991;22:590–595.
7. Li J, Motsko SP, Goehring EL, Vendiola R, Maneno M, Jones JK. Longitudinal study on pediatric dyslipidemia in population-based claims database. *Pharmacoepidemiol Drug Saf*. 2010;19:90–98.
8. Tu K, Mitiku T, Guo H, Lee DS, Tu JV. Myocardial infarction and the validation of physician billing and hospitalization data using electronic medical records. *Chronic Dis Can*. 2010;30:141–146.
9. Lix L, Yogendran M, Burchill C, et al. *Defining and Validating Chronic Diseases: an Administrative Data Approach*. Winnipeg, MB: Manitoba Centre for Health Policy; 2006.
10. Juurlink D, Preyra C, Croxford R, et al. *Canadian Institute for Health Information Discharge Abstract Database: a validation study*. Toronto, ON: Institute for Clinical Evaluative Sciences; 2006.
11. Birtwhistle R, Keshavjee K, Lambert-Lanning A, et al. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. *J Am Board Fam Med*. 2009;22:412–422.
12. Crawford AG, Cote C, Couto J, et al. Prevalence of obesity, type II diabetes mellitus, hyperlipidemia, and hypertension in the united states: findings from the GE centricity electronic medical record database. *Popul Health Manag*. 2010;13:151–161.
13. Young WB, Ryu H. Secondary data for policy studies: benefits and challenges. *Policy Polit Nurs Pract*. 2000;1:302–307.
14. Asghari S, Aref-Eshghi E, Hurley O, et al. Does the prevalence of dyslipidemias differ between Newfoundland and the rest of Canada? findings from the Electronic medical records of the Canadian Primary Care Sentinel Surveillance Network. *Front Cardiovasc Med*. 2015;2:1.
15. Aref-Eshghi E, Leung J, Godwin M, et al. Low density lipoprotein cholesterol control in Canadian high risk cardiovascular population: findings from Canadian Primary Care Sentinel Surveillance Network database. *Lipids Health Dis*. 2015;14:60.
16. Asghari S, Aref-Eshghi E, Godwin M, Duke P, Williamson T, Mahdavian M. Single and mixed dyslipidaemia in Canadian primary care settings: findings from the Canadian Primary Care Sentinel Surveillance Network database. *BMJ Open*. 2015;5:e007954.
17. Genest J, McPherson R, Frohlich J, et al. 2009 Canadian cardiovascular society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult - 2009 recommendations. *Can J Cardiol*. 2009;25:567–579.
18. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a criterion standard. *Am J Epidemiol*. 1995;141:263–272.
19. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4:13.
20. Rasmussen-Torvik LJ, Pacheco JA, Wilke RA, et al. High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin Transl Sci*. 2012;5:394–399.
21. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46:423–429.
22. Szklo M. *Epidemiology: Beyond the Basics*. 2nd ed. Burlington, MA: Jones & Bartlett Learning; 2007.
23. Vining DJ, Gladish GW. Receiver operating characteristic curves: a basic understanding. *Radiographics*. 1992;12:1147–1154.
24. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med*. 2003;29:1043–1051.
25. Cheng CL, Kao YH, Lin SJ, Lee CH, Lai ML. Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan. *Pharmacoepidemiol Drug Saf*. 2011;20:236–242.
26. Chen G, Khan N, Walker R, Quan H. Validating ICD coding algorithms for diabetes mellitus from administrative data. *Diabetes Res Clin Pract*. 2010;89:189–195.
27. Geller SE, Ahmed S, Brown ML, Cox SM, Rosenberg D, Kilpatrick SJ. International classification of diseases-9th revision coding for preeclampsia: how accurate is it? *Am J Obstet Gynecol*. 2004;190:1629–1633.
28. Kokotailo RA, Hill MD. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke*. 2005;36:1776–1781.
29. Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis*. 2013;19:1411–1420.
30. Hinz ER, Bastarache L, Denny JC. A natural language processing algorithm to define a venous thromboembolism phenotype. *AMIA Annu Symp Proc*. 2013;2013:975–983.
31. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Inform Assoc*. 2013;20:349–355.
32. Sui J, Adali T, Yu Q, Chen J, Calhoun VD. A review of multivariate methods for multimodal fusion of brain imaging data. *J Neurosci Methods*. 2012;204:68–81.
33. Meng X, Jiang R, Lin D, et al. Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data [published online ahead of print May 10, 2016]. *NeuroImage*. doi:10.1016/j.neuroimage.2016.05.026.
34. Lei B, Chen S, Ni D, Wang T. Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. *Front Aging Neurosci*. 2016;8:77.