Deriving comorbidities from medical records using natural language processing

Hojjat Salmasian,[1] Daniel E Freedberg,[2] Carol Friedman[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001889).

[1]Department of Biomedical Informatics, Columbia University, New York, USA
[2]Division of Digestive and Liver Diseases, Columbia University Medical Center, New York, USA

Correspondence to
Dr Hojjat Salmasian, Department of Biomedical Informatics, Columbia University, 622 West 168th Street VC-5, New York, NY 10032, USA; hojjat@dbmi.columbia.edu

Received 6 April 2013
Revised 1 October 2013
Accepted 22 October 2013
Published Online First 31 October 2013

ABSTRACT

Extracting comorbidity information is crucial for phenotypic studies because of the confounding effect of comorbidities. We developed an automated method that accurately determines comorbidities from electronic medical records. Using a modified version of the Charlson comorbidity index (CCI), two physicians created a reference standard of comorbidities by manual review of 100 admission notes. We processed the notes using the MedLEE natural language processing system, and wrote queries to extract comorbidities automatically from its structured output. Interrater agreement for the reference set was very high (97.7%). Our method yielded an F1 score of 0.761 and the summed CCI score was not different from the reference standard (p=0.329, power 80.4%). In comparison, obtaining comorbidities from claims data yielded an F1 score of 0.741, due to lower sensitivity (66.1%). Because CCI has previously been validated as a predictor of mortality and readmission, our method could allow automated prediction of these outcomes.

INTRODUCTION

Extracting comorbidity information is crucial for identifying phenotype–genotype associations, studying clinical outcomes because of the confounding effect of comorbidities,[1–3] and patient cohort identification because it helps identify cohorts with certain outcomes.[4 5] As composite representations of burden of disease and patient complexity, comorbidity measures have also been related to health resource use, mortality and readmission.[6 7]

Numerous methods exist to measure comorbidities in clinical research including the Charlson comorbidity index (CCI), cumulative illness rating scale, index of coexisting disease, and Kaplan index.[8] Among these, various versions of the CCI are most extensively studied.[8 9] The index encompasses 19 medical conditions, each weighted between 1 and 6 based on the relative risk of their association with 1-year mortality. The summed CCI score, calculated by adding the individual scores for each item, has proved to be an accurate predictor of mortality, disability, readmission and length of stay.[10–12]

For most applications, the CCI score is calculated by manual record review or using claims data, typically coded using the International Classification of Diseases, 9th Version (ICD-9).[13–15] The former approach is costly and the latter introduces biases due to coding errors, heterogeneous coding conventions, and the granularity of the coding system.[16] In addition, previous research has suggested that manually extracting comorbidity information from medical records is superior to the use of ICD-9 codes[17] and claims data are not available until after discharge time.

Most of the information needed for comorbidities must be extracted from narrative reports, particularly from the 'history of present illness' and 'past medical history' sections. Previous efforts in automating the extraction of comorbidity information from narratives are limited to specific domains[18] or rudimentary methodologies such as keyword search.[19] However, it should be possible to utilize natural language processing (NLP) to build a generalizable method for extracting comorbidities from electronic health records (EHR). Therefore, we aim to develop an effective automated and generalizable NLP-based method that derives comorbidities from narrative records.

METHODS

We used a 15-item modified version of the CCI (table 1), which we will hereafter refer to as 'the index'. It is different from the original Charlson index[9] in that it combines mild, moderate and severe liver disease into one category, combines diabetes mellitus with and without complications into one category, and excludes metastatic solid tumors. To formulate the index, we performed a preliminarily manual review of randomly selected notes and included in the index only those categories that could be accurately captured. We found that the notes frequently lacked sufficient information to distinguish the severity of liver disease or diabetes, and were not explicit regarding tumor staging. Thus, in order to have a reliable gold standard comparator, we collapsed or excluded the relevant categories.

We randomly selected 100 admission notes for patients admitted at the New York-Presbyterian Hospital during 2009–12, each corresponding to a unique patient. Two physicians annotated the notes to record the presence or absence of each of the 15 items in the index, creating a 'reference standard'. We measured agreement between two coders using a subset of 30 notes. We then processed all the notes using the MedLEE NLP system, which handles negations, interprets the level of certainty associated with the concepts in the notes, and normalizes the concepts to concept unique identifiers (CUI) in the unified medical language system.[20] We excluded findings attributed to patient's family and only included 'diseases and conditions' by excluding medications, laboratory results, procedures and measurements, resulting in a table containing ongoing conditions of each patient represented as CUI. We then used a randomly selected subset of 30 subjects for training and the remaining 70 for testing.

To cite: Salmasian H, Freedberg DE, Friedman C. *J Am Med Inform Assoc* 2013;**20**:e239–e242.

Table 1 Items included in the modified Charlson comorbidity index, and their relative scores

	Item	Points	POS_{train}	TP	FN	FP	TN
1	Myocardial infarction	1	8	5	2	8	55
2	Congestive heart failure	1	10	14	4	3	49
3	Peripheral vascular disease	1	3	6	1	6	57
4	Cerebrovascular disease	1	9	10	2	8	50
5	Dementia	1	3	11	1	0	58
6	Chronic pulmonary disease	1	11	5	9	0	56
7	Connective tissue disease	1	3	2	0	2	66
8	Peptic ulcer disease	1	3	2	2	2	64
9	Liver disease (mild, moderate or severe)	1	3	4	1	1	64
10	Diabetes mellitus (with/without complications)	1	12	23	3	0	44
11	Hemiplegia	2	2	3	1	2	64
12	Renal disease	2	10	20	1	4	45
13	Leukemia	2	2	1	1	0	68
14	Lymphoma	2	2	1	1	1	67
15	AIDS/HIV	6	1	1	0	2	67

Fourth column (POS_{train}) indicates total number of positive cases in the training set (N = 30), and columns 5–8 show the number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) assignments by the automated method in the test set (N = 70), in comparison to manual review. The total number of positive cases in the test set can be easily calculated by adding the values for TP and FN in each row.

Using knowledge we developed queries to determine the comorbidities. Each query contained a list of CUI representing the pertinent item in the score (for an example, see table 2). Queries were developed by consensus between the first two authors. For each disease or condition in the index, we included CUI representing that disease or condition (eg, 'diabetes mellitus') as well as those that imply the problem exists (eg, 'diabetic foot'). We manually evaluated and refined the queries using the training set. Finally, we applied these queries to the test set to evaluate the accuracy of the automated method against the reference set. Concepts included in each query can be found in the supplementary data (available online only).

In addition, we compared our method with an approach that determined comorbidities using the ICD-9 codes recorded at discharge. First, the accuracy of each method was calculated based on the proportion of patient–condition pairs that were correctly classified. For example, if a patient was correctly classified as having diabetes mellitus this patient–condition pair was counted towards true positives, and when diabetes mellitus was wrongly ruled out it was counted towards false negatives, and so on. We calculated the sensitivity, specificity and F1 score to measure performance, and also compared the overall agreement of each approach with the reference set using Cohen's κ. In addition, we compared the summed CCI score of each method versus that of the reference standard using a standard t test. All the analyses were performed on the test set (N=70). Our study was designed with 80% power to detect a minimum difference in mean CCI scores of 0.5 between manual and automated methods in the test set. To select a minimum difference of 0.5 we referred to previous studies suggesting that differences in CCI scores of at least 1.0 were necessary for clinically important differences in outcomes.[21 22]

Table 2 The list of concepts included in the query for diabetes mellitus, and their preferred names in the unified medical language system

CUI	Preferred name
C0011849	Diabetes mellitus
C0011854	Diabetes mellitus, insulin-dependent
C0011860	Diabetes mellitus, non-insulin-dependent
C0206172	Diabetic foot
C1456868	Diabetic foot ulcer
C0011880	Diabetic ketoacidosis
C0011881	Diabetic nephropathy
C0011884	Diabetic retinopathy
C0011882	Diabetic neuropathies
C1720297	Disorder associated with type I diabetes mellitus
C2919365	Macroalbuminuric diabetic nephropathy

The concepts are represented using their unique identifiers (CUI) in the unified medical language system.

RESULTS

Overall, 53% of patients were women, and the average age was 63.6 years (range 1–98 years, SD 23.5 years). Based on manual review, each of these patients had on average 2.12 of the index conditions reported in their admission note (range 0–5, SD 1.51). Interrater agreement for manual review was high (97.8%) with the Cohen's κ equal to 0.923 (95% CI 0.876 to 0.970). Reasons for disagreement included miscoding and different interpretation of uncertain results (eg, 'possible CHF' was coded as positive by one coder and negative by the other).

Within the test set, the NLP-based method yielded a sensitivity of 81.2%, a specificity of 95.3%, an F1 score of 0.761, and high agreement with the reference set (Cohen's κ 0.723). Among the 15 index items, the best results were obtained for diabetes (sensitivity 100% and specificity 85%), and the poorest for chronic pulmonary diseases (sensitivity 36% and specificity 100%). The latter was a relatively uncommon diagnosis, and low sensitivity was due to a missing CUI ('chronic obstructive airway disease'), which was not included in our query and did not occur in our training data but occurred in the test data. The summed CCI score was not significantly different in the automated versus manual calculations (mean CCI score of 2.9 for automated method vs 2.6 for the manual approach, p=0.329) (figure 1).

Compared with the reference standard for the test set, comorbidity data extracted using the ICD-9 codes recorded at the time of discharge had a sensitivity and specificity of 66.1% and 97.6%, respectively (F1 score 0.741). While the NLP-based method was more sensitive and slightly more accurate than the claims-based model, the summed CCI score was not significantly different when calculated using the claims data versus the reference set (p=0.109) or the NLP-based method (p=0.251).

DISCUSSION

We were able to calculate a modified version of the CCI with high accuracy using a combination of NLP and tailored queries. Our method was able to calculate this summed score with minimal error, which can potentially allow automated predictions of patients' outcomes including mortality, disability, readmission and length of stay.[6 7]

Automating the CCI score using EHR was studied previously. Singh *et al*[19] used a keyword search to find terms pertinent to each item on the CCI score, and then used an Excel formula to calculate the score. While they achieved high accuracy compared to manual data extraction, their method of correcting for negations was basic and their success could also be attributed partly to

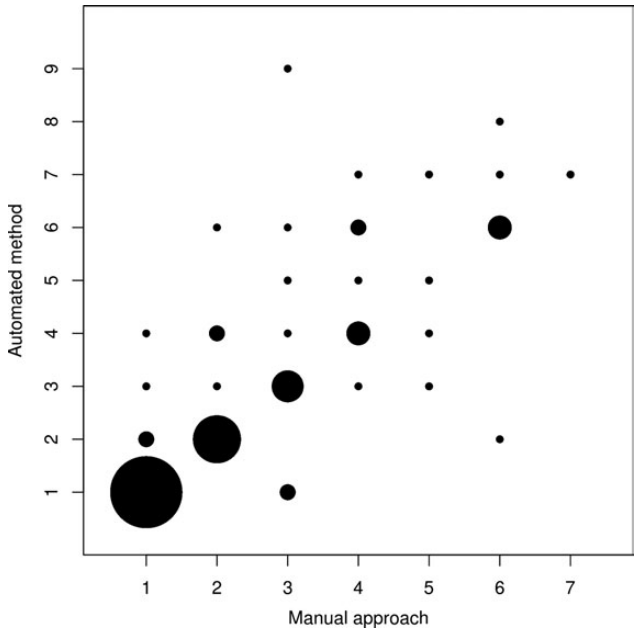

Figure 1 Scatter plot of summed Charlson comorbidity index scores calculated for the test set (N=70), using the automated method versus those calculated by manual chart review. The size of circles is in proportion to the number of patients for whom the respective scores were calculated by each method.

the availability of a specific query builder for their EHR system. In comparison, our method does not depend on a specific EHR system and uses NLP to deal with negations, temporality and certainty. In another study, Chuang *et al*[18] used an approach that utilized the MedLEE NLP system for processing of the notes; however, their method did not gain a high accuracy on the test set and was evaluated only regarding a specific domain (pneumonia patients). However, their reference standard was based on ICD-9 codes, which may not be an ideal choice of reference standard (see below). In comparison, our method achieved a higher accuracy, probably because we used a different gold standard (manual chart review) and a more current version of the MedLEE NLP system, with improvements in disambiguating abbreviations and normalizing the terms to CUI.

Our findings indicate that obtaining comorbidity information from claims data is less effective, primarily because of a lower sensitivity. This can be due to situations in which chronic conditions (such as a history of myocardial infarction) were in the note but not in the claims data. Under-reporting of chronic comorbidities in the claims data has been demonstrated in previous research.[23 24] These are necessary for the CCI score and the higher sensitivity of the NLP method provides a more complete list of comorbidities using medical records. Previous research has also shown that claims data is subject to coding errors and granularity issues.[25 26] In addition, deriving comorbidities from claims data is only possible retrospectively, because these data become available after the patient's discharge, hindering the utility of claims data for comorbidity information in applications such as reducing readmission, when it is ideal to identify patients with higher chances of readmission before discharge so that they could be managed appropriately. Nevertheless, because both NLP-based and claims-based approaches showed high accuracy in capturing comorbidity information, we recommend using them as complementary sources of information whenever applicable.

Our study has limitations. While our sample size was statistically large enough, the small size of the training set for query development contributed to the low sensitivity of at least one of the queries (ie, chronic pulmonary diseases). Because we randomly divided the notes into training and test sets, the distribution of positives and negatives among the different conditions was not balanced in the two sets. For some conditions (eg, HIV) there were more positive cases in the training set than expected, and for others (eg, dementia) there were fewer. However, the majority of the terms in the queries were mainly determined using clinical knowledge and consensus and therefore the choice of terms for inclusion in the queries minimally depended on the number of training cases. While we achieved good results using a small training set, it is likely that we could achieve higher sensitivity and better overall performance by using a larger sample size. However, caution is needed to ensure that the larger sample size would not lead to 'statistically significant' differences that are not of clinical value. In addition, we only used admission notes in the current study, but intend to expand our model to use other types of notes in the future. Finally, we used a modified version of the CCI. This is not the first study to use a modified version of the CCI, as multiple previous studies have found that various modifications of the CCI predict outcomes across different populations.[27–29] Nevertheless, direct conclusions about the utility of our index in clinical studies cannot be inferred from the current finding, and the modification could have affected the performance of our method. Future research will include exploring ways to improve further the accuracy of our approach, demonstrating the utility of this approach in clinical studies, and reducing confounding in studies of patient cohort identification.

CONCLUSION

We were able to derive comorbidities automatically from narrative admission notes with high accuracy. Our method has a higher sensitivity and accuracy than determining comorbidities from claims data, and has an additional advantage of utility in prospective studies that need to identify phenotypes from medical records and correct for the confounding effects of comorbidities.

Contributors All authors collaborated in the study design. HS and CF collaborated in acquisition of the data, and HS and DEF collaborated on analysis of the results. All authors contributed to the preparation of the final manuscript, and approved it before submission.

Funding This work has been supported by the National Library of Medicine grants R01 LM010016, R01 LM010016-0S1, R01 LM010016-0S2, R01 LM008635, and 5 T15 LM007079. DEF was partly supported by National Institutes of Health training grant T32 DK083256-04.

Competing interests CF consults for a company that licenses the NLP software MedLEE. The remaining authors declare no other competing interests.

Ethics approval Ethics approval was obtained from the Institutional Review Board of Columbia University.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The authors agree to share the list of concepts used for each of the index items as an online supplement to the manuscript.

REFERENCES

1 Smoller JW, Lunetta KL, Robins J. Implications of comorbidity and ascertainment bias for identifying disease genes. *Am J Med Genet* 2000;96:817–22.
2 Schneeweiss S, Seeger JD, Maclure M, *et al*. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol* 2001;154:854–64.
3 Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. *Int J Epidemiol* 2000;29:891–8.

4 Della Porta MG, Malcovati L, Strupp C, *et al*. Risk stratification based on both disease status and extra-hematologic comorbidities in patients with myelodysplastic syndrome. *Haematologica* 2011;96:441–9.
5 Condon JR, You J, McDonnell J. Performance of comorbidity indices in measuring outcomes after acute myocardial infarction in Australian indigenous and non-indigenous patients. *Intern Med J* 2012;42:e165–73.
6 Elixhauser A, Steiner C, Harris DR, *et al*. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
7 Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011;306:1688–98.
8 De Groot V, Beckerman H, Lankhorst GJ, *et al*. How to measure comorbidity. a critical review of available methods. *J Clin Epidemiol* 2003;56:221–9.
9 Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
10 D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the Charlson comorbidity index with administrative data bases. *J Clin Epidemiol* 1996;49:1429–33.
11 Ghali WA, Hall RE, Rosen AK, *et al*. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. *J Clin Epidemiol* 1996;49:273–8.
12 Librero J, Peiró S, Ordiñana R. Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. *J Clin Epidemiol* 1999;52:171–9.
13 Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–19.
14 Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. *J Clin Epidemiol* 1993;46:1075–9; discussion 1081–90.
15 Quan H, Sundararajan V, Halfon P, *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
16 Dong Y, Chang C, Shau W, *et al*. Development and validation of a pharmacy-based comorbidity measure in a population-based automated health care database. *Pharmacotherapy* 2013;33:126–36.
17 Kieszak SM, Flanders WD, Kosinski a S, *et al*. A comparison of the Charlson comorbidity index derived from medical record data and administrative billing data. *J Clin Epidemiol* 1999;52:137–42.
18 Chuang J-H, Friedman C, Hripcsak G. A comparison of the Charlson comorbidities derived from medical language processing and administrative data. *AMIA Annu Symp Proc* 2002;160–4.
19 Singh A, Kuo Y-F, Goodwin JS. Many patients who undergo surgery for colorectal cancer receive surveillance colonoscopies earlier than recommended by guidelines. *Clin Gastroenterol Hepatol* 2013;11:65–72.e1.
20 Friedman C, Shagina L, Lussier Y, *et al*. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;11:392–402.
21 Núñez JE, Núñez E, Fácila L, *et al*. [Prognostic value of Charlson comorbidity index at 30 days and 1 year after acute myocardial infarction]. *Rev Esp Cardiol* 2004;57:842–9.
22 Goldstein LB, Samsa GP, Matchar DB, *et al*. Charlson Index comorbidity adjustment for ischemic stroke outcome studies. *Stroke* 2004;35:1941–5.
23 Kern EFO, Maney M, Miller DR, *et al*. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res* 2006;41:564–80.
24 Romano PS, Roos LL, Luft HS, *et al*. A comparison of administrative versus clinical data: coronary artery bypass surgery as an example. Ischemic Heart Disease Patient Outcomes Research Team. *J Clin Epidemiol* 1994;47:249–60.
25 Rawson NS, D'Arcy C. Assessing the validity of diagnostic information in administrative health care utilization data: experience in Saskatchewan. *Pharmacoepidemiol Drug Saf* 1998;7:389–98.
26 Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol* 2004;57:131–41.
27 Volk ML, Hernandez JC, Lok AS, *et al*. Modified Charlson comorbidity index for predicting survival after liver transplantation. *Liver Transpl* 2007;13:1515–20.
28 Habbous S, Chu KP, Harland LTG, *et al*. Validation of a one-page patient-reported Charlson comorbidity index questionnaire for upper aerodigestive tract cancer patients. *Oral Oncol* 2013;49:407–12.
29 Hemmelgarn BR, Manns BJ, Quan H, *et al*. Adapting the Charlson Comorbidity Index for use in patients with ESRD. *Am J Kidney Dis* 2003;42:125–32.