Identification of Patients With Diabetes From the Text of Physician Notes in the Electronic Medical Record

ALEXANDER TURCHIN, MD[1,2]
ISAAC S. KOHANE, MD, PHD[2]
MERRI L. PENDERGRASS, MD, PHD[1]

In this report, we describe a software tool that rapidly and reliably identifies a diagnosis of diabetes documented in physician notes in the electronic medical record.

Diabetes care in the U.S. is suboptimal: 20–40% of diabetic patients have inadequate glycemic or blood pressure control or do not have annual eye or foot examinations (1–3). Effective public health surveillance is mandatory to address this problem. It is crucial for the assessment of prevalence of diabetes, its economic and social costs, and evaluation of the dynamics of disease care measures and outcomes (4,5).

Disease surveillance can be significantly hampered by the difficulty in identifying the target population (6). A number of approaches have been used to identify patients with diabetes, including death certificates (7,8), billing data (9,10), and surveys (11,12). Each of these methods has its own shortcomings, and sensitivity remains relatively low. Consequently, manual chart review remains the gold standard for the identification of individuals diagnosed with a particular disease. This is a labor-intensive process that is not scalable to the level needed in public health surveillance.

Because most elements of the patient chart are increasingly available in digital format, there have been a number of attempts to identify diagnoses from the text of physician notes (13–15). However, low sensitivity and specificity remain a problem. We therefore have designed a software tool, DITTO (Diabetes Identification Through Textual element Occurrences), that accurately and rapidly identifies patients with diabetes through analysis of the texts of physician notes.

RESEARCH DESIGN AND METHODS

— Data were obtained from the Research Patient Data Registry, a database containing clinical (laboratory results, physician notes, and radiology reports) and administrative (billing and encounter data) records on all patients treated at Massachusetts General Hospital and Brigham and Women's Hospital. Billing codes and outpatient physician notes of 7,203 adult patients who were seen in four primary care practices in 2002–2003 and who had either at least one billing ICD-9 code of 250.xx, one serum glucose >199 mg/dl, or one measurement of HbA_{1c} (A1C) were retrieved for analysis. Although these selection criteria identified a population with a high likelihood of diabetes diagnosis, only about a third of the patients actually had diabetes (further described in RESULTS).

DITTO is a program written in Perl language that takes one or more text files containing patient notes as input. The entire text of physician notes was analyzed by DITTO for the presence of words, word roots, or groups of words (word tags) that potentially indicated the presence of a diagnosis of diabetes. These included two terms naming diabetes ("diabet," "IDDM"—also capturing "NIDDM") and 32 names of medications exclusively used to treat diabetes. Metformin and all insulins/insulin analogs except glargine were excluded because in the practices being studied, they are commonly used to treat polycystic ovarian syndrome and gestational diabetes. Sentences with diabetes word tags were ignored if they also contained 1 of 31 negative qualifiers (e.g., "insipidus," "family history," "work up for," "not"), indicating that the patient did not have diabetes. Patients with at least two sentences with diabetes word tags but without negative qualifiers were considered to have diabetes. No data other than the text of the notes were used in the analysis.

The ability of DITTO to identify patients with diabetes was compared with *1*) billing codes and *2*) manual chart review. At least two codes of diabetes over 2 years were required to make the diagnosis of diabetes from billing data (16,17). One hundred fifty patient records randomly selected for manual review were examined independently by two investigators who were also blinded to the conclusions of the note text and billing codes analyses. When there was a discrepancy between the two reviewers (10 of 150 records), the charts were reexamined jointly and agreement was made on 100% of the charts. McNemar's test was used to estimate statistical significance of the difference between text and billing code analysis (18). κ statistic (19) was used to evaluate the agreement between DITTO and manual chart review.

RESULTS

— DITTO processed 182,345 physician notes over 40 min. The estimated processing speed was 2.7×10^5 notes/h. Of the 7,023 records that were analyzed, billing data identified 2,007 and DITTO identified 2,982 diabetic patients.

In the manual chart review, κ statistic

From the [1]Division of Endocrinology, Brigham and Women's Hospital, Boston, Massachusetts; and the [2]Medical Informatics Program, Children's Hospital, Boston, Massachusetts.

Address correspondence and reprint requests to Alexander Turchin, MD, Division of Endocrinology, Brigham and Women's Hospital, 221 Longwood Ave., Boston, MA 02115. E-mail: aturchin@partners.org.

Received for publication 28 December 2004 and accepted in revised form 18 March 2005.

Abbreviations: DITTO, Diabetes Identification Through Textual element Occurrences.

A table elsewhere in this issue shows conventional and Système International (SI) units and conversion factors for many substances.

Table 1—*Comparison of DITTO and billing data analysis to manual chart review*

	DITTO$^+$	Billing data$^+$	DITTO$^-$	Billing data$^-$
Chart review$^+$	50	40	2	12
Chart review$^-$	2	2	96	96
Sensitivity	96.2%	76.9%	—	—
Specificity	—	—	98.0%	98.0%

Chart review$^+$, number of records identified by manual chart review as having the diagnosis of diabetes; chart review$^-$, number of records identified by chart review as not having the diagnosis of diabetes; DITTO$^+$, number of records identified by DITTO as having the diagnosis of diabetes; DITTO$^-$, number of records identified by DITTO as not having the diagnosis of diabetes.

for agreement between the two reviewers was 0.87 ($P = 0.05$ that $\kappa \geq 0.8$). Of the 150 records randomly selected from 7,023 patient records, manual review identified 52 patients as having a documented diagnosis of diabetes. Billing data analysis detected 40, and DITTO detected 50 of these patients (Table 1).

κ statistic for agreement between DITTO and manual chart review was 0.94 ($P < 0.001$ that $\kappa \geq 0.8$). Using manual chart review as the gold standard, sensitivity of DITTO was 96.2% and specificity 98.0%. Compared with DITTO, billing code analysis had substantially lower sensitivity (76.9%) but similarly high specificity (98.0%), consistent with findings of other investigators (9,10). This difference between the two methods was statistically significant ($P = 0.02$).

CONCLUSIONS — We have designed DITTO, a software tool that identifies the diagnosis of diabetes documented in the chart using analysis of the text of physician notes. DITTO is more sensitive and at least as specific as the best of the previously reported methods. It is very fast and can process over a quarter of a million patient notes (~10,000 individual patient records) per hour. It requires minimal customization (mostly related to the format of the patient medical record numbers and separators between the notes in the text file) for adaptation to a different health care organization. It requires a minimal set of data (e.g., no insurance medication claims) that can be commonly obtained in many health care facilities.

Identification of patients with a particular diagnosis is a problem of great importance for public health care at every level: national, regional, and individual health care facilities. Our tool is an important advance in this field, and we plan to continue to develop this concept further to improve its performance, comprehensiveness, and functionality.

References

1. Saaddine JB, Engelgau MM, Beckles GL, Gregg EW, Thompson TJ, Narayan KM: A diabetes report card for the United States: quality of care in the 1990s. *Ann Intern Med* 136:565–574, 2002
2. Jencks SF, Cuerdon T, Burwen DR, Fleming B, Houck PM, Kussmaul AE, Nilasena DS, Ordin DL, Arday DR: Quality of medical care delivered to Medicare beneficiaries: a profile at state and national levels. *JAMA* 284:1670–1676, 2000
3. Jencks SF, Huff ED, Cuerdon T: Change in the quality of care delivered to Medicare beneficiaries, 1998–1999 to 2000–2001. *JAMA* 289:305–312, 2003
4. Thacker SB, Stroup DF, Rothenberg RB: Public health surveillance for chronic conditions: a scientific basis for decisions. *Stat Med* 14:629–641, 1995
5. Brownson RC, Bright FS: Chronic disease control in public health practice: looking back and moving forward. *Public Health Rep* 119:230–238, 2004
6. Saydah SH, Geiss LS, Tierney E, Benjamin SM, Engelgau M, Brancati F: Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Ann Epidemiol* 14:507–516, 2004
7. Sasaki A, Horiuchi N, Hasegawa K, Uehara M: The proportion of death certificates of diabetic patients that mentioned diabetes in Osaka District, Japan. *Diabetes Res Clin Pract* 20:241–246, 1993
8. Vauzelle-Kervroedan F, Delcourt C, Forhan A, Jougla E, Hatton F, Papoz L: Analysis of mortality in French diabetic patients from death certificates: a comparative study. *Diabetes Metab* 25:404–411, 1999
9. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM: Identifying persons with diabetes using Medicare claims data. *Am J Med Qual* 14:270–277, 1999
10. Kashner TM: Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. *Med Care* 36:1324–1336, 1998
11. Haapanen N, Miilunpalo S, Pasanen M, Oja P, Vuori I: Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. *Am J Epidemiol* 145:762–769, 1997
12. Heliovaara M, Aromaa A, Klaukka T, Knekt P, Joukamaa M, Impivaara O: Reliability and validity of interview data on chronic diseases: the Mini-Finland Health Survey. *J Clin Epidemiol* 46:181–191, 1993
13. Cooper GF, Miller RA: An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. *J Am Med Inform Assoc* 5:62–75, 1998
14. Berrios DC, Kehler A, Fagan LM: Knowledge requirements for automated inference of medical textbook markup. *Proc AMIA Symp* 676–680, 1999
15. Friedman C, Shagina L, Lussier Y, Hripcsak G: Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 11:392–402, 2004
16. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR: Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? *Am J Med Qual* 19:201–206, 2004
17. National Committee on Quality Assurance: *HEDIS 2005: Technical Specifications*. Vol. 2. Washington, D.C., NCQA, 2005
18. Hawass NE: Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *Br J Radiol* 70:360–366, 1997
19. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46, 1960