Chapter 28
Improving Patient Cohort Identification Using Natural Language Processing

Raymond Francis Sarmiento and Franck Dernoncourt

Learning Objectives
To compare and evaluate the performance of the structured data extraction method and the natural language processing (NLP) method when identifying patient cohorts using the Medical Information Mart for Intensive Care (MIMIC-III) database.

1. To identify a specific patient cohort from the MIMIC-III database by searching the structured data tables using ICD-9 diagnosis and procedure codes.
2. To identify a specific patient cohort from the MIMIC-III database by searching the unstructured, free text data contained in the clinical notes using a clinical NLP tool that leverages negation detection and the Unified Medical Language System (UMLS) to find synonymous medical terms.
3. To evaluate the performance of the structured data extraction method and the NLP method when used for patient cohort identification.

28.1 Introduction

An active area of research in the biomedical informatics community involves developing techniques to identify patient cohorts for clinical trials and research studies that involve the secondary use of data from electronic health records (EHR) systems. The widening scale of EHR databases, that contain both structured and unstructured information, has been beneficial to clinical researchers in this regard. It has helped investigators identify individuals who may be eligible for

The two authors contributed equally to this work.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_28

clinical trials as well as conduct retrospective studies to potentially validate the results of prospective clinical studies at a fraction of the cost and time [1]. It has also helped clinicians to identify patients at a higher risk of developing chronic disease, especially those who could benefit from early treatment [2].

Several studies have investigated the accuracy of structured administrative data such as the World Health Organization's (WHO) International Classification of Diseases, Ninth Revision (ICD-9) billing codes when identifying patient cohorts [3–11]. Extracting structured information using ICD-9 codes has been shown to have good recall, precision, and specificity [3, 4] when identifying distinct patient populations. However, for large clinical databases, information extraction can be time-consuming, costly, and impractical when conducted across several data sources [12] and applied to large cohorts [13].

Using structured queries to extract information from an EHR database allows one to retrieve data easily and in a more time-efficient manner. Structured EHR data is generally useful, but may also contain incomplete and/or inaccurate information especially when each data element is viewed in isolation. For example [14], to justify ordering a particular laboratory or radiology test, clinicians often assign a patient with a diagnosis code for a condition that the patient is suspected to have. But even when the test results point to the patient not having the suspected condition, the diagnosis code often remains in the patient's medical record. When the diagnosis code is then viewed without context (i.e., without the benefit of understanding the nuances of the case as provided in the patient's clinical narrative), this becomes problematic because it prohibits the ability of investigators to accurately identify patient cohorts and to utilize the full statistical potential of the available populations. Compared to narratives from clinical notes, relying solely on structured data such as diagnostic codes can be unreliable because they may not be able to provide information on the overall clinical context. However, automated examination of a large volume of clinical notes requires the use of natural language processing (NLP). The domain of study for the automated analysis of unstructured text data is referred to as NLP, and it has already been used with some success in the domain of medicine. In this chapter, we will be focusing on how NLP can be used to extract information from unstructured data for cohort identification.

NLP is a field of computer science and linguistics that aims to understand human (natural) languages and facilitate more effective interactions between humans and machines [13, 15]. In the clinical domain, NLP has been utilized to extract relevant information such as laboratory results, medications, and diagnoses from de-identified medical patient record narratives in order to identify patient cohorts that fit eligibility criteria for clinical research studies [16]. When compared to human chart review of medical records, NLP yields faster results [17–20]. NLP techniques have also been used to identify possible lung cancer patients based on their radiology reports [21] and extract disease characteristics for prostate cancer patients [22].

We considered chronic conditions where both a disease diagnosis and an intervention diagnosis were likely to be found together in an attempt to better highlight the differences between structured and unstructured retrieval techniques, especially given the limited number of studies that have looked at interventions or treatment procedures, rather than illness or disease, as outcomes [14]. The diabetic population was of particular interest for this NLP task because the numerous cardiovascular, ophthalmological, and renal complications associated with diabetes mellitus eventually require treatment interventions or procedures, such as hemodialysis in this case. Moreover, clinical notes frequently contain medical abbreviations and acronyms, and the use of NLP techniques can help in capturing and viewing these information correctly in medical records. Therefore, in this case study, we attempted to determine whether the use of NLP on the unstructured clinical notes of this population would help improve structured data extraction. We identified a cohort of critically ill diabetic patients suffering from end-stage renal failure who underwent hemodialysis using the Medical Information Mart for Intensive Care (MIMIC-III) database [23].

28.2 Methods

28.2.1 Study Dataset and Pre-processing

All data from this study were extracted from the publicly available MIMIC-III database. MIMIC-III contains de-identified [24] data, per Health Insurance Portability and Accountability Act (HIPAA) privacy rules [25], on over 58,000 hospital admissions in the intensive care units (ICU) at Beth Israel Deaconess Medical Center from June 2001 to October 2012 [26]. Aside from being publicly accessible, we chose MIMIC-III because it contains detailed EHR data on critically ill patients who are likely to have multiple chronic conditions, including those with complications from chronic diseases that would require life-saving treatment interventions.

We excluded all patients in the database who were under the age of 18; diagnosed with diabetes insipidus only and not diabetes mellitus; underwent peritoneal dialysis only and not hemodialysis; or those diagnosed with transient conditions such as gestational diabetes or steroid-induced diabetes without any medical history of diabetes mellitus. We also excluded patients who had received hemodialysis prior to their hospital admission but did not receive it during admission. From the remaining subjects, we included those who were diagnosed with diabetes mellitus and those who had undergone hemodialysis during their ICU admission. We extracted data from two primary sources: the structured MIMIC-III tables (discharge diagnoses and procedures) and unstructured clinical notes.

28.2.2 Structured Data Extraction from MIMIC-III Tables

Using the ICD-9 diagnosis codes from the discharge diagnoses table and ICD-9 procedure codes from the procedures table, we searched a publicly available ICD-9 [27] database to find illness diagnosis and procedure codes related to diabetes and hemodialysis as shown in Table 28.1. We used structured query language (SQL) to find patients in each of the structured data tables based on specific ICD-9 codes.

Table 28.1 ICD-9 codes and descriptions indicating a patient was diagnosed with diabetes mellitus and who potentially underwent hemodialysis from structured data tables in MIMIC-III

Structured data table	ICD-9 code and description
Diabetes mellitus	
Discharge diagnosis codes	249 secondary diabetes mellitus (includes the following codes: 249, 249.0, 249.00, 249.01, 249.1, 249.10, 249.11, 249.2, 249.20, 249.21, 249.3, 249.30, 249.31, 249.4, 249.40, 249.41, 249.5, 249.50, 249.51, 249.6, 249.60, 249.61, 249.7, 249.70, 249.71, 249.8, 249.80, 249.81, 249.9, 249.90, 249.91)
	250 diabetes mellitus (includes the following codes: 250, 250.0, 250.00, 250.01, 250.02, 250.03, 250.1, 250.10, 250.11, 250.12, 250.13, 250.2, 250.20, 250.21, 250.22, 250.23, 250.3, 250.30, 250.31, 250.32, 250.33, 250.4, 250.40, 250.41, 250.42, 250.43, 250.5, 250.50, 250.51, 250.52, 250.53, 250.6, 250.60, 250.61, 250.62, 250.63, 250.7, 250.70, 250.71, 250.72, 250.73, 250.8, 250.80, 250.81, 250.82, 250.83, 250.9, 250.90, 250.91, 250.92, 250.93)
Hemodialysis	
Discharge diagnosis codes	585.6 end stage renal disease (requiring chronic dialysis)
	996.1 mechanical complication of other vascular device, implant, and graft
	996.73 other complications due to renal dialysis device, implant, and graft
	E879.1 kidney dialysis as the cause of abnormal reaction of patient, or of later complication, without mention of misadventure at time of procedure
	V45.1 postsurgical renal dialysis status
	V56.0 encounter for extracorporeal dialysis
	V56.1 fitting and adjustment of extracorporeal dialysis catheter
Procedure codes	38.95 venous catheterization for renal dialysis
	39.27 arteriovenostomy for renal dialysis
	39.42 revision of arteriovenous shunt for renal dialysis
	39.43 removal of arteriovenous shunt for renal dialysis
	39.95 hemodialysis

28.2.3 Unstructured Data Extraction from Clinical Notes

The unstructured clinical notes include discharge summaries (n = 52,746), nursing progress notes (n = 812,128), physician notes (n = 430,629), electrocardiogram (ECG) reports (n = 209,058), echocardiogram reports (n = 45,794), and radiology reports (n = 896,478). We excluded clinical notes that were related to any imaging results (ECG_Report, Echo_Report, and Radiology_Report). We extracted notes from MIMIC-III with the following data elements: patient identification number (SUBJECT_ID), hospital admission identification number (HADM_IDs), intensive care unit stay identification number (ICUSTAY_ID), note type, note date/time, and note text.

We used an SQL query to extract pertinent information from all patients' notes that will be helpful in identifying a patient as someone belonging to the cohort, then wrote a Python script to filter the notes by looking for keywords and implementing heuristics in order to refine our search results. As part of our search strategy, we removed the family history sections when searching the clinical notes and ensured that the search for clinical acronyms did not retrieve those that were part of another word. For example, our filters did not retrieve those where "DM" appeared as part of another words such as in 'a**dm**ission' or 'a**dm**it'. Finally, we used cTAKES [28, 29] version 3.2 with access to Unified Medical Language System (UMLS) [30] concepts to use the negation detection annotator when searching the note text. The negation detection feature in cTAKES works by trying to detect which entities in the text are negated. Examples of negation words that may be found in the clinical notes include 'not', 'no', 'never', 'hold', 'refuse', 'declined'. For example, in this case study, if "DM" or "HD" is consistently negated when searching the clinical notes, then the patient should not be considered part of the cohort.

The Metathesaurus [31] in UMLS contains health and biomedical vocabularies, ontologies, and standard terminologies, including ICD. Each term is assigned to one or more concepts in UMLS. Different terms from different vocabularies or ontologies that have similar meanings and assigned with the same concept unique identifier (CUI) are considered UMLS synonyms [32]. In order to identify diabetes mellitus patients who underwent hemodialysis during their ICU stay, we scanned the clinical notes containing the terms "diabetes mellitus" and "hemodialysis". We used the UMLS Metathesaurus to obtain synonyms for these terms because using only these two terms will restrict our search results.

cTAKES is an open-source natural language processing system that extracts information from clinical free-text stored in electronic medical records. It accepts either plain text or clinical document architecture (CDA)-compliant extensible markup language (XML) documents and consists of several annotators such as attributes extractor (assertion annotator), clinical document pipeline, chunker, constituency parser, context dependent tokenizer, dependency parser and semantic role labeler, negation detection, document preprocessor, relation extractor, and dictionary lookup, among others [33]. When performing named entity recognition

or concept identification, each named entity is mapped to a specific terminology concept through the cTAKES dictionary lookup component [28], which uses the UMLS as a dictionary.

We refined our query parameters iteratively and searched the clinical notes containing our final query parameters based on UMLS synonyms to diabetes and hemodialysis. These were as follows: (A) include documents that contained any of the following terms: diabetes, diabetes mellitus, DM; (B) include documents that contained any of the following terms: hemodialysis, haemodialysis, kidney dialysis, renal dialysis, extracorporeal dialysis, on HD, HD today, tunneled HD, continue HD, cont HD; (C) finalize the set of documents to be run in cTAKES by only including documents that contained at least one of the terms from group A and at least one of the terms from group B; and (D) exclude documents by using the negation detection annotator in cTAKES to detect negations such as avoid, refuse, never, declined, etc. that appear near any of the terms listed in groups A and B.

28.2.4 Analysis

We manually reviewed all the notes for all patients identified by the structured data extraction method and/or the clinical NLP method as those potentially to have a diagnosis of diabetes mellitus and who had undergone hemodialysis during their ICU stay in order to create a validation database that contains the positively identified patients in the population of MIMIC-III patients. We used this validation database in evaluating the precision and recall of both the structured data extraction method and the clinical NLP method. We compared the results from both methods to the validation database in order to determine the true positives, false positives, recall, and precision. We defined these parameters using the following equation: recall = TP/(TP + FN), where TP = true positives and FN = false negatives; and precision = TP/(TP + FP), where FP = false positives. In this case study, we defined recall as the proportion of diabetic patients who have undergone hemodialysis in the validation database who were identified as such. We defined precision as the proportion of patients identified as diabetic and having undergone hemodialysis whose diagnoses were both confirmed by the validation database.

28.3 Results

In the structured data extraction method using SQL as illustrated in Fig. 28.1, we found 10,494 patients diagnosed with diabetes mellitus using ICD-9 codes; 1216 patients who underwent hemodialysis using ICD-9 diagnosis and procedure codes; and 1691 patients who underwent hemodialysis when searching the structured data tables using the string '%hemodial%'. Figure 28.2 shows the number of patients

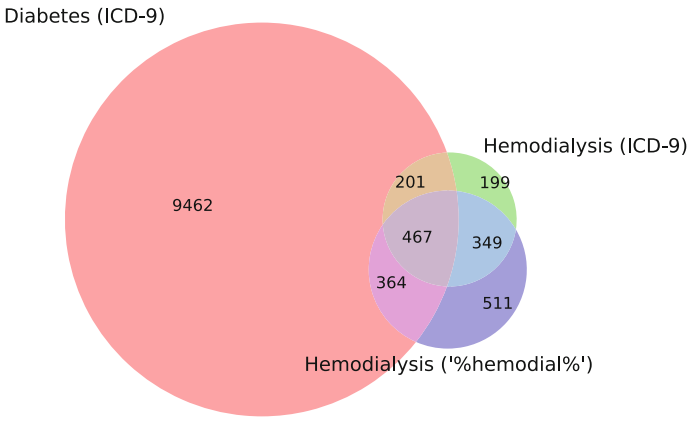

Fig. 28.1 Patients identified by structured data extraction, clockwise from *left* diagnosed with diabetes mellitus using ICD-9 diagnosis codes, underwent hemodialysis using ICD-9 discharge diagnosis and procedure codes, and underwent hemodialysis using the string '%hemodial%'

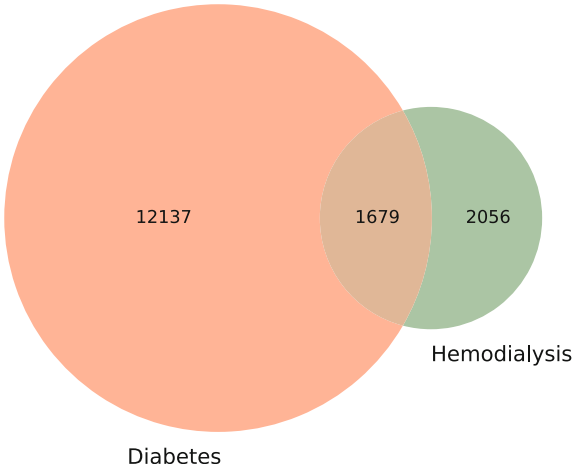

Fig. 28.2 Patients identified by clinical NLP method, from *left* diagnosed with diabetes, diagnosed with diabetes and who underwent hemodialysis, and who underwent hemodialysis

identified using the clinical NLP method: 13,816 patients diagnosed with diabetes mellitus and 3735 patients identified as having undergone hemodialysis during their ICU stay.

There were 1879 patients in the validation database consisting of 1847 (98.3 %) confirmed diabetic patients who had undergone hemodialysis. We identified 1032 (54.9 % of 1879) patients when using SQL only and 1679 (89.4 % of 1879) when using cTAKES. Of these, 832 (44.3 % of 1879) were found by both approaches as illustrated in Fig. 28.3.

Table 28.2 shows the results of the two methods used to identify patient cohorts compared to the validation database. The clinical NLP method had better precision compared to the structured data extraction method. The clinical NLP method also

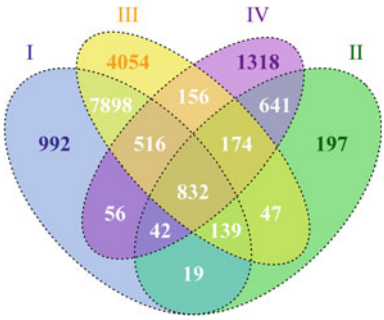

Fig. 28.3 Patients identified by structured data extraction and clinical NLP methods: *I*—diabetes patients found using SQL; *II*—patients who underwent hemodialysis found using SQL; *III*—diabetic patients found using cTAKES and; *IV*—patients who underwent hemodialysis found using cTAKES

Table 28.2 Precision of identifying patient cohorts using structured data extraction and clinical NLP compared to the validation database

Validation database (n = 1879)	Structured data extraction method, positive (n = 1032)	Clinical NLP method, positive (n = 1679)
Positive	TP = 1013	TP = 1666
Negative	FP = 19	FP = 13
Precision	98.2 %	99.2 %

identified fewer FP (0.8 % of 1679) compared to the structured data extraction method (1.8 % of 1032).

In this case study, the recall value could not be computed. But because recall is calculated by dividing TP by the sum of TP and FN, and the denominator for both methods is the same, we can use the TP count as a proxy to determine which method showed a higher recall. Based on the results, we found that more TPs were identified using NLP compared to the structured data approach. Hence, the clinical NLP method yielded a higher recall than the structured data extraction method.

We also analyzed the clinical notes for the 19 patients identified as FP using the structured data extraction method. We found that 14 patients were incorrectly identified as diabetic patients, 3 patients were incorrectly identified as having undergone hemodialysis, and 2 patients were not diabetic nor did they undergo hemodialysis during their ICU stay. In the 13 patients identified as FP when using the clinical NLP method, we also analyzed the clinical notes and found that 5 did not undergo hemodialysis during their ICU stay, 2 had initially undergone hemodialysis but was stopped due to complications, and 6 did not have diabetes (3 did not have any history of diabetes, 1 had initially been presumed to have diabetes according to the patient's family but was not the case, 1 had gestational diabetes without prior history of diabetes mellitus, and 1 was given insulin several times during the patient's ICU stay but was not previously diagnosed with diabetes nor was a diagnosis of new-onset diabetes indicated in any of the notes).

28.4 Discussion

Both the structured data extraction method and the clinical NLP method achieved high precision in identifying diabetic patients who underwent hemodialysis during their ICU stay. However, the clinical NLP method exhibited better precision and higher recall in a more time-saving and efficient way compared to the structured data extraction technique.

We identified several variables that may have resulted in a lower precision when using SQL only in identifying patient cohorts such as the kind of illness and the kind of intervention, the presence of other conditions similar to diabetes (i.e., diabetes insipidus, gestational diabetes), and the presence of other interventions similar to hemodialysis (i.e., peritoneal dialysis, continuous renal replacement therapy). The temporal feature of the intervention also added to the complexity of the cohort identification process.

Extracting and using the UMLS synonyms for "diabetes mellitus" and "hemodialysis" in performing NLP on the clinical notes helped increase the number of patients included in the final cohort. Knowing that clinicians often use acronyms, such as "DM" to refer to diabetes mellitus and "HD" for hemodialysis, and abbreviations, such as "cont" for the word 'continue' when taking down notes helped us refine our final query parameters.

There are several limitations to this case study. Specificity could not be calculated because in order to determine the TN and FN, the entire MIMIC-III database would need to be manually validated. Though it can be argued that the ones in the validation database that were missed by either method could be considered as FN, this may not be the true FN count in MIMIC-III because those that could be found outside of the validation database have not been included. Moreover, since the validation database used was not independent of the two methods, the TP and FP counts as well as the precision and recall may have been overestimated.

Another limitation is the lack of a gold standard database for the specific patient cohort we investigated. Without it, we were not able to fully evaluate the cohort identification methods we implemented. The creation of a gold standard database, one that is validated by clinicians and includes patients in the MIMIC-III database that have been correctly identified as TN and FN, for this particular patient cohort will help to better evaluate the performance of the methods used in this case study. Having a gold standard database will also help calculate the specificity for both methods.

Another limitation is that we focused on discharge diagnosis and procedure events especially in the structured data extraction method. Other data sources in MIMIC-III such as laboratory results and medications may help support the findings or even increase the number of patients identified when using SQL.

Furthermore, although we used a large database, our data originated from a single data source. Comparing our results found using MIMIC-III to other publicly available databases containing EHR data may help to assess the generalizability of our results.

28.5 Conclusions

NLP is an efficient method for identifying patient cohorts in large clinical databases and produces better results when compared to structured data extraction. Combining the use of UMLS synonyms and a negation detection annotator in a clinical NLP tool can help clinical researchers to better perform cohort identification tasks using data from multiple sources within a large clinical database.

Future Work

Investigating how clinical researchers could take advantage of NLP when mining clinical notes would be beneficial for the scientific research community. In this case study, we found that using NLP yields better results for patient cohort identification tasks compared to structured data extraction.

Using NLP may potentially be useful for other time-consuming clinical research tasks involving EHR data collected in the outpatient departments, inpatient wards, emergency departments, laboratories, and various sources of medical data. The automatic detection of abnormal findings mentioned in the results of diagnostic tests such as X-rays or electrocardiograms could be systematically used to enhance the quality of large clinical databases. Time-series analyses could also be improved if NLP is used to extract more information from the free-text clinical notes.

Notes

1. cTAKES is available from the cTAKES Apache website: http://ctakes.apache.org/downloads.cgi. A description of the components of cTAKES 3.2 can be found on the cTAKES wiki page: https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+Component+Use+Guide [28].

Code Appendix

All the SQL queries to count the number of patients per cohorts as well as the cTAKES XML configuration file used to analyze the notes are available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-

data-book. Further information on the code is available from this website. The following key scripts were used:

- *cohort_diabetic_hemodialysis_icd9_based_count.sql*: Total number of diabetic patients who underwent hemodialysis based on diagnosis codes.
- *cohort_diabetic_hemodialysis_notes_based_count.sql*: List of diabetic patients who underwent hemodialysis based on unstructured clinical notes.
- *cohort_diabetic_hemodialysis_proc_and_notes_based_-count.sql*: Total number of diabetic patients who underwent hemodialysis based on unstructured clinical notes and procedure codes.
- *cohort_diabetic_hemodialysis_proc_based_count.sql*: Total number of diabetic patients who underwent hemodialysis based on procedure codes.
- *cohort_diabetic_icd9_based_count_a.sql*: List of diabetic patients based on the ICD-9 codes.
- *cohort_hemodialysis_icd9_based_count_b.sql*: List of patients who underwent hemodialysis based on the ICD-9 codes.
- *cohort_hemodialysis_proc_based_count_c.sql*: Lists number of patients who underwent hemodialysis based on the procedure label.
- *CPE_physician_notes.xml*: cTAKES XML configuration file to process patients' notes. Some paths need to be adapted to the developer's configuration.

References

1. Kury FSP, Huser V, Cimino JJ (2015) Reproducing a prospective clinical study as a computational retrospective study in MIMIC-II. In: AMIA Annual Symposium Proceedings, pp 804–813
2. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 33(7):1123–1131
3. Segal JB, Powe NR (2004) Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol 75 (1):12–17
4. Eichler AF, Lamont EB (2009) Utility of administrative claims data for the study of brain metastases: a validation study. J Neuro-Oncol 95(3):427–431
5. Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, Aron D, Pogach L (2006) Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 41(2):564–580
6. Zhan C, Eixhauser A, Richards CL Jr, Wang Y, Baine WB, Pineau M, Verzier N, Kilman R, Hunt D (2009) Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. Med Care 47(3):364–369
7. Floyd JS, Heckbert SR, Weiss NS, Carell DS, Psaty BM (2012) Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. J Am Med Assoc 307(15):1580–1582

8. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A, Forster AJ (2010) The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. J Clin Epidemiol 63(12):1332–1341
9. Tieder JS, Hall M, Auger KA, Hain PD, Jerardi KE, Myers AL, Rahman SS, Williams DJ, Shah SS (2011) Accuracy of administrative billing codes to detect urinary tract infection hospitalizations. Pediatrics 128:323–330
10. Rosen LM, Liu T, Merchant RC (2012) Efficiency of International Classification of Diseases, Ninth Revision, billing code searches to identify emergency department visits for blood and body fluid exposures through a statewide multicenter database. Infect Control Hosp Epidemiol 33:581–588
11. Lamont EB, Lan L (2014) Sensitivity of Medicare claims data for measuring use of standard multiagent chemotherapy regimens. Med Care 52(3):e15–e20
12. Bache R, Miles S, Taweel A (2013) An adaptable architecture for patient cohort identification from diverse data sources. J Am Med Inform Assoc 20(e2):e327–e333
13. Sada Y, Hou J, Richardson P, El-Serag H, Davila J (2013) Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care
14. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ (2014) Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 21(5):801–807
15. Jurafsky D, Martin H (2008) Speech and language processing, 2nd edn. Prentice Hall, Englewood Cliffs, NJ
16. Voorhees EM, Tong RM (2011) Overview of the TREC 2011 medical records track. In: The twentieth text retrieval conference proceedings (TREC 2011). National Institute for Standards and Technology, Gaithersburg, MD
17. Wilbur WJ, Rzhetsky A, Shatkay H (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinform 7:356
18. Buchan NS, Rajpal DK, Webster Y, Alatorre C, Gudivada RC, Zheng C, Sanseau P, Koehler J (2011) The role of translational bioinformatics in drug discovery. Drug Discov Today 16:426–434
19. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural language processing: an introduction. J Am Med Inform Assoc 18:544–551
20. Uzuner Ö, South BR, Shen S, Duvall SL (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 18(5):552–556
21. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK (2012) Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 7:1257–1262
22. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, Slezak J, Porter K, Jacobsen SJ, Chien GW (2014) Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol 32(1):99–103
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
24. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD (2008) Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8:32
25. Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164 (2002) http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privruletxt.txt. Last accessed 6 Oct 2015
26. MIMIC. https://mimic.physionet.org/gettingstarted/access. Last accessed 19 Feb 2016
27. The Web's Free 2015 Medical Coding Reference. http://www.icd9data.com. Last accessed 7 Oct 2015

28. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
29. Apache cTAKES™. http://cTAKES.apache.org/index.html. Last accessed 3 Oct 2015
30. Lindberg DA, Humphreys BL, McCray AT (1993) The unified medical language system. Meth Inf Med 32(4):281–291
31. Unified Medical Language System® (UMLS®) The Metathesaurus. https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_001.html. Last accessed 7 Oct 2015
32. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno JF, Darmoni SJ (2012) Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed. BMC Med Inform Decis Mak 12:12
33. cTAKES 3.2 Component Use Guide. https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+Component+Use+Guide. Last accessed 7 Oct 2015