Improving Patient Cohort Identification Using Natural Language Processing

RAYMOND FRANCIS R. SARMIENTO, MD
Director (OIC), UP National Telehealth Center
University of the Philippines Manila
April 25, 2017

Data Analytics in Clinical Settings

Davenport & Harris:

“The extensive use of data, statistical, and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions”

IBM:

“The systematic use of data and related business insights developed through applied analytical disciplines (e.g., statistical, contextual, quantitative, predictive, cognitive, other models) to drive fact-based decision making for planning, management, measurement, and learning.”

Top Uses of Analytics in Health Care

1. Identify patients for care management – 66%
2. Clinical outcomes – 64%
3. Performance measurement – 64%
4. Clinical decision making at the point of care – 57%

Source: http://www.healthcareitnews.com/news/pop-health-analytics-top-aco-priority

Clinical NLP 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.

- To identify a specific patient cohort from the MIMIC-III database by searching the structured data tables using ICD-9 diagnosis and procedure codes.
- 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.
- To evaluate the performance of the structured data extraction method and the NLP method when used for patient cohort identification.

Introduction

The widening scale of electronic health records (EHR) databases, that contain both structured and unstructured information, has been an active area of research in the biomedical informatics community and become beneficial to clinical researchers.

- ✓ Identify eligible participants for clinical trials and retrospective studies to validate results at a fraction of the cost and time.
- ✓ Helps clinicians identify patients at a higher risk of developing chronic disease, especially those who could benefit from early treatment.

Natural Language Processing (NLP)

A field of computer science and linguistics that aims to understand human (natural) languages and facilitate more effective interactions between humans and machines.

Advantage
✓ Examines large volume of clinical notes (i.e., laboratory results, medications, and diagnoses) from de-identified medical patient record to identify eligible patient cohorts in clinical research studies ✓ Yields faster results when compared to human chart review of medical records ✓ Facilitates disease and intervention diagnosis of chronic conditions (i.e., DM, lung and prostate cancer) ✓ Can capture and automatically analyze unstructured data correctly (i.e., medical abbreviations and acronyms)

NLP: Limitations of handwritten rules

1. **NLP must ultimately extract meaning ('semantics') from text**: formal grammars that specify relationship between text units and parts of speech such as nouns, verbs, and adjectives

2. **Handwritten rules handle 'ungrammatical' spoken prose** and (in medical contexts) the highly telegraphic prose of in-hospital progress notes very poorly, although such prose is human-comprehensible.

- THE RISE OF STATISTICAL NLP -

Source: J Am Med Inform Assoc 2011;18:544e551. doi:10.1136/amiajnl-2011-000464

NLP: High level tasks

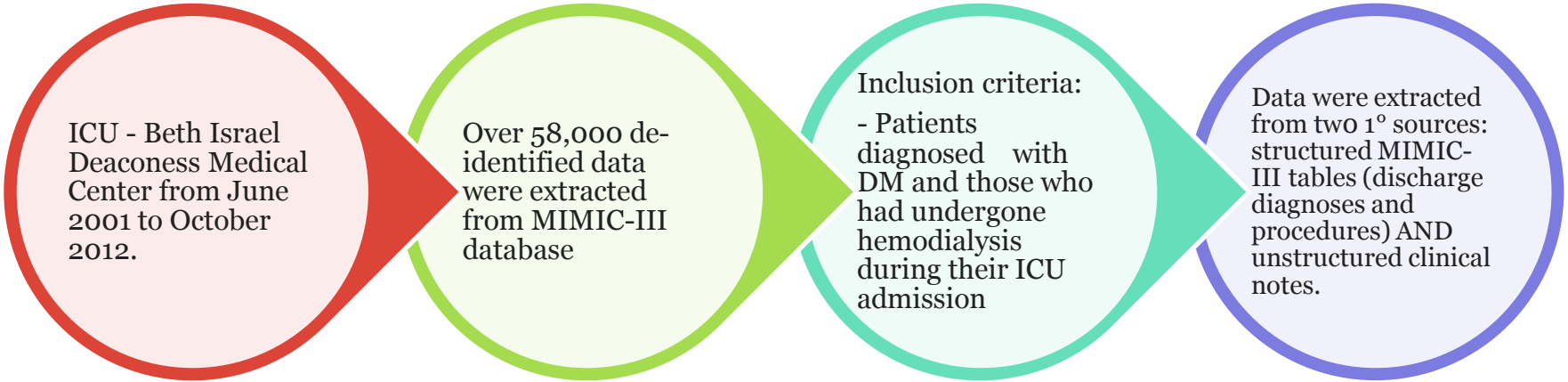
1. Spelling/grammatical error identification and recovery
2. Named entity recognition (NER): need to map entities to a vocabulary (issues: word/phrase order variation; derivation; inflection; synonymy; polysemy)
3. Word sense disambiguation
4. Negation and uncertainty identification
5. Relationship extraction
6. Temporal inferences

Source: J Am Med Inform Assoc 2011;18:544e551. doi:10.1136/amiajnl-2011-000464

WHO International Classification of Diseases (ICD-9)

Advantage	Disadvantage
✓ Provides accurate and structured data – good recall, precision, and specificity – when identifying distinct patient populations	✓ For large clinical databases – time consuming, costly, and impractical – when conducted across several data sources and applied to large cohorts ✓ Isolated data element may not be able to provide information on the overall clinical context ✓ When diagnosis code is viewed without context, it prohibits the ability of investigators to accurately identify patient cohorts and to utilize the full statistical potential of the available populations

Methodology



Exclusion Criteria

- Below 18 years old
- Patients with diabetes insipidus only and not diabetes mellitus
- Underwent peritoneal dialysis only and not hemodialysis
- Diagnosed with transient conditions (i.e., gestational diabetes or steroid-induced diabetes) without any medical history of diabetes mellitus
- Patients who had received hemodialysis prior to their hospital admission but did not receive it during admission

Table 1. ICD-9 codes and descriptions indicating a patient was diagnosed with DM and who potentially underwent hemodialysis from structured data tables in MIMIC-III

Structured data table	ICD-9 codes 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

Table 2. Unstructured Data Extraction from Clinical Notes from MIMIC-III

Criteria	Number of clinical notes
Discharge summaries	52,746
Nursing progress notes	812,128
Physician notes	430,629
Electrocardiogram (ECG) reports	209,058
Echocardiogram reports	45,794
Radiology reports	896,478

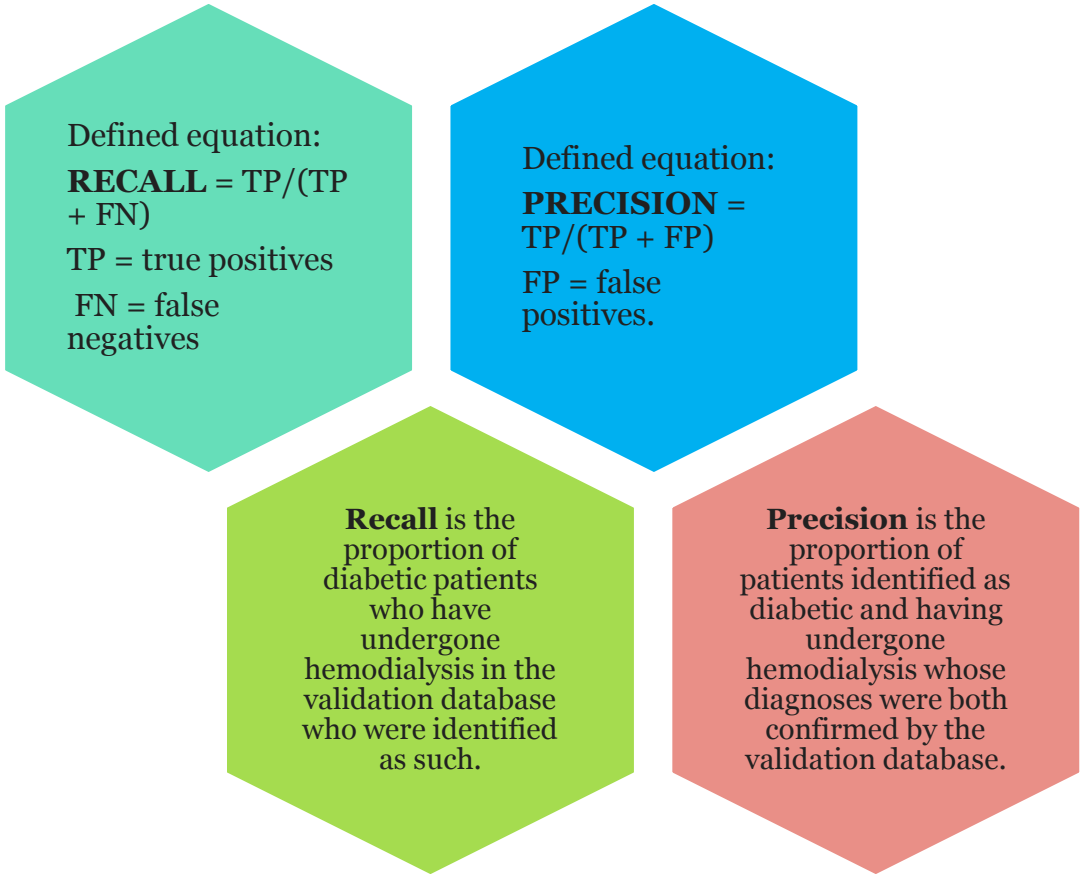
*Excluded imaging results (i.e., ECG_Report, Echo_Report, and Radiology_Report)
*Data elements extracted using SQL (i.e., SUBJECT_ID, HADM_IDs, ICUSTAY_ID, note type, note date/time, and note text)

cTAKES

- An open-source natural language processing system that extracts information from clinical free-text stored in electronic medical records with access to Unified Medical Language System (UMLS) concepts to use the negation detection annotator when searching the note text.
- It accepts either plain text or clinical document architecture (CDA)-compliant extensible markup language (XML) documents and consists of several annotators (i.e., 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)

cTAKES

Analysis


Defined equation:
RECALL = TP/(TP + FN)
TP = true positives
FN = false negatives
Defined equation:
PRECISION = TP/(TP + FP)
FP = false positives.
Recall is the proportion of diabetic patients who have undergone hemodialysis in the validation database who were identified as such.
Precision is the proportion of patients identified as diabetic and having undergone hemodialysis whose diagnoses were both confirmed by the validation database.

Results

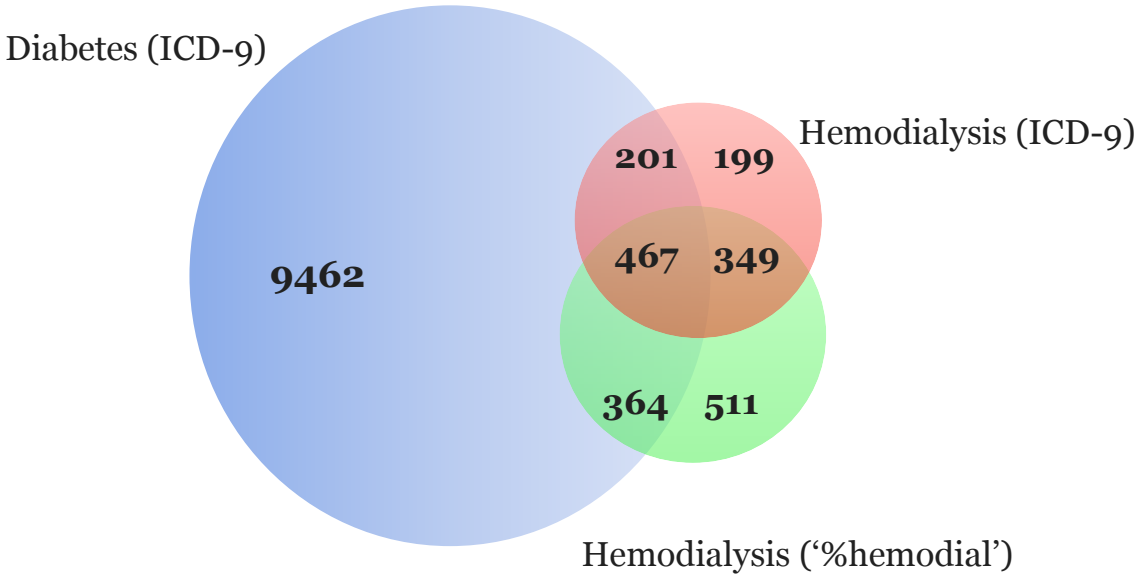


Figure 1. Patients identified by structured data extraction, clockwise from left diagnosed with diabetes mellitus using ICD-9 diagnosis codes, underwent hemodialysis using ICD-9 discharge diagnosis and procedure codes, and underwent hemodialysis using the string '%hemodial%'.

Results

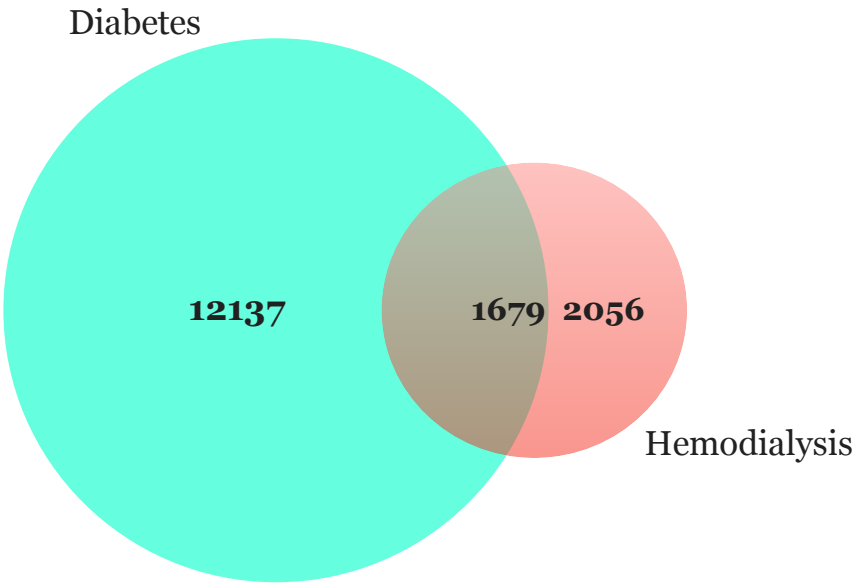


Figure 2. Patients identified by clinical NLP method, from left diagnosed with diabetes, diagnosed with diabetes and who underwent hemodialysis, and who underwent hemodialysis.

Results

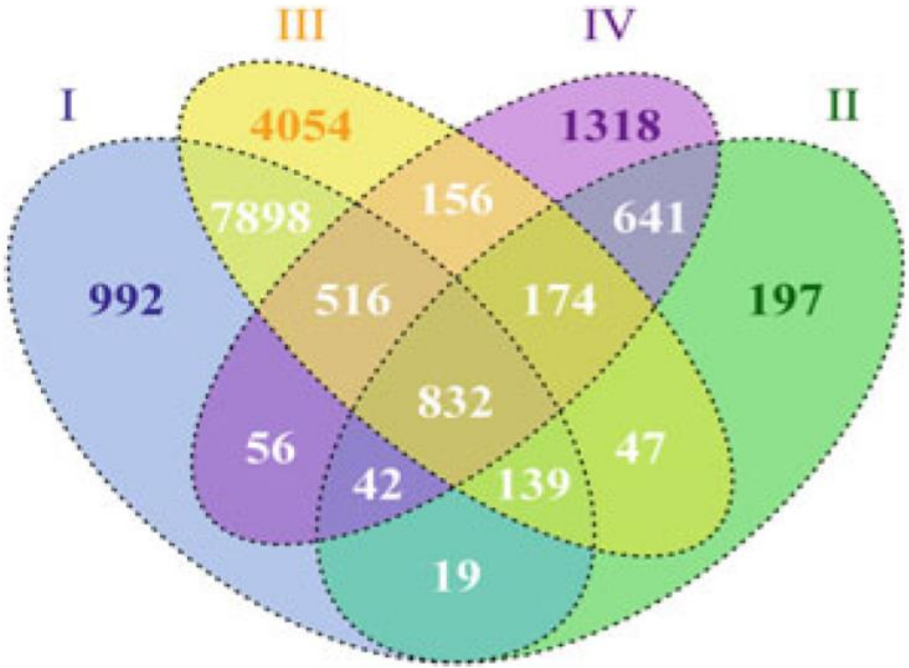


Figure 3. Patients identified by structured data extraction and clinical NLP methods: I—diabetes patients found using SQL; II—patients who underwent hemodialysis found using SQL; III— diabetic patients found using cTAKES and; IV—patients who underwent hemodialysis found using cTAKES.

Results

Table 3. 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%

Discussions

- Clinical NLP method exhibited better precision and higher recall in a more time-saving and efficient way compared to the structured data extraction technique

- Helps increase the number of eligible patients in the cohort utilizing the UMLS synonyms in performing NLP on the clinical notes

- Analyzes and refine medical abbreviations and acronyms (i.e., “DM” diabetes mellitus, “HD” hemodialysis, and “cont” for continue)

- There were several limitations identified in this case study: 1) specificity could not be calculated because the entire MIMIC-III database would need to be manually validated to determine the TN and FN, 2) TP and FP counts as well as the precision and recall may have been overestimated because the validation database used was not independent of the two methods, 3) lack of a gold standard database for the specific patient cohort, 4) focused only on the discharge diagnosis and procedure events especially in the structured data extraction method, and 5) comparing the results to other publicly available databases containing EHR data may help assess the generalizability of the results.

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

- ✓ The use of NLP is highly beneficial to various scientific and clinical researches, especially for patient cohort identification tasks.
- ✓ 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

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

THANK YOU !!!

RAYMOND FRANCIS R. SARMIENTO, MD
University of the Philippines Manila
rrsarmiento@up.edu.ph
April 25, 2017