Abstract— Patient cohort is similar symptoms of group of patients over a time period. It is important to correctly identify patient cohort for observational study or interventional study. To identify patient cohort, we can easily retrieve information from large structured and unstructured data tables, but this information may not fulfill our interest. We need to extract information from the unstructured clinical notes for more accurate study. In this work, we have improved structured extraction by presenting context on a health record dataset for identifying cohort of diabetic patients. Results shows that traditional structured query-based data extraction methods accurately identified 97.14% positive patients to our question of interest where adding natural language processing supported technique have retrieved 98.37% precisely.

Keywords— patient cohort, nlp, diabetes, hemodialysis.

I. INTRODUCTION
In biomedical research, identifying patient cohort is crucial for selecting a population of patients tailored to the question of interest [1, 2]. To identify patient cohort, clinicians are leveraging the information captured by Electronic Health Records (EHRs). As EHR databases can store detailed records of huge number of patients, overall health pattern can be easily visualized for a particular cohort. Even, clinician may identify patients with higher risk of developing chronic condition and support them in early stage [3-5]. As EHRs involves patients’ private information, reliable EHRs are not available in public domain. This makes EHRs based research more challenging.

Clinicians have designed an international standard, known as ICD code [6], for easier storage and retrieval of health information which are being continually revised. ICD revision 9th and revision 10th are available in EHR databases [7]. Several studies have extracted information from EHRs using ICD codes and investigated the effectiveness of ICD codes for patient cohort identification. Sarmiento and Demoncourt [5] have also highlighted some studies [8-16] for patient cohort identification. Though Segal and Powe [8] and Eichler and Lamount [9] extracted structured data using ICD-9 with good precision, Sarmiento and Demoncourt [5] highlighted the need of NLP support for extracting structured and unstructured data. For large clinical databases like MIMIC, information extraction can be costly when we conducted extraction from several data sources [17] and applied to large cohorts [18]. Using structured queries, information can be extracted from both structured and unstructured EHR data tables easily and time efficiently. Although structured data analysis helps us but it may also contain incomplete and/or inaccurate information. For example, a clinician often assigns a patient with a diagnosis code for different condition that the patient suspected to have, but the assign diagnosis code sometimes not removed by clinicians even when further test negates the previous suspicion [19]. Which create serious problems to identify actual patient who are experiencing suspected diseases. So only used of structured data in clinical research is unreliable. But automated inspection of structured data and unstructured notes can provide reliable results using NLP and it has breakthrough success in medical research. It is a field of computing and linguistics department that aims to understand and process human (natural) language and developed more effective interaction between human and machine [20]. In clinical research, NLP has been used to extract relevant information such as medication, laboratory tests to identify patient cohorts [21]. Compared to structured query NLP yields faster results [22-27]. Identification of possible lung cancer based on patient radiology reports by using NLP techniques in [28], improving patient cohort [5] using MIMIC II database [29] and extraction of characteristics of prostate cancer patients [30] also use NLP based techniques. NLP is also used for information extraction, which helps us to improve our extraction methods [31-40].

In [28], authors used a combination of diagnostic and procedural codes and develop an NLP algorithm for information retrieval from radiology reports more accurately. NLP helps them to scan the relevant radiology reports and classify them according the presence and absence of lung nodules. In [5], authors used UMLS synonyms [41, 42], in order to identify diabetes mellitus patients who underwent hemodialysis. They searched the clinical reports and notes containing the terms “diabetes mellitus” and “hemodialysis”, and use NLP techniques with UMLS synonyms of these two terms. They also used cTakes [43] for avoid negated keywords to identify patient cohort more accurately.

In this study, we have attempted to extend the performance of the data extraction method by using NLP. We improve the identifying process of patient cohorts on the MIMIC database [44, 45]. To the best of our knowledge, use of structured data query to find patient cohorts resultant a good precision. We have used NLP with it to improve this result even more.

II. MATERIALS AND METHODS
A. Study Dataset
All data for this work were extracted from MIMIC-III database. It contains de-identified [46] data, which designed according to Accountability Act (HIPAA) privacy rules [47], and contains over 58,000 ICU patients’ data who admitted in Beth Israel Deaconess Medical Center from Jun 2001 to Oct 2012. We work on MIMIC III because it is more reliable and publicly accessible.

B. Preprocessing
We consider structured and unstructured retrieval techniques to our works, and improving unstructured retrieval by combining result of NLP with it as [19] to improve our result. We selected diabetic population as a particularly interesting feature, because the numerous cardiovascular and renal complications associated with diabetes mellitus which require hemodialysis. To identify the clinical notes similarity,
we need the use of NLP techniques to capture information correctly in clinical records.

We have extracted information from structured tables which contain discharge diagnoses and procedures codes and unstructured clinical reports and notes. We include all patients’ data in our works who were diagnosed with diabetes mellitus, patients who had diagnosed with diabetes mellitus and who had undergone hemodialysis during their admission in ICU. Patients who were under age of 18, or patients who diagnosed with diabetes insipidus only, or patients who underwent peritoneal dialysis only, or diagnosed with condition like gestational diabetes or steroid-induced diabetes without any medical history of diabetes mellitus, or patients who only had received hemodialysis prior to their hospital admission but not during admission are not part of our study.

C. Structured Data Extraction

We use ICD-9 diagnosis codes [48], and ICD-9 procedure codes to retrieve information from structured data tables. We searched illness diagnosis and procedures codes related to diabetes and hemodialysis. We used structured query language (SQL) to retrieve structured information from data tables based on specific ICD-9 codes. The ICD-9 diagnosis and procedure are shown in Table I.

TABLE I. ICD-9 CODES AND DESCRIPTIONS OF DIAGNOSED CODES FROM MIMIC III [26].

<table>
<thead>
<tr>
<th>Structured data table</th>
<th>ICD-9 code and description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes Type 1</td>
<td>Secondary diabetes mellitus discharge 249 diagnosis code includes the subsequent codes: [249, 249.0, 249.00, 249.01, 249.1, 249.10, 249.11, ..., 249.9, 249.90, 249.91]</td>
</tr>
<tr>
<td>Diabetes Type 2</td>
<td>250 diabetes mellitus includes subsequent codes: [250, 250.0, 250.00, 250.01, 250.02, 250.03, 250.1, 250.10, 250.11, 250.12, 250.13, ..., 250.9, 250.90, 250.91, 250.92, 250.93]</td>
</tr>
<tr>
<td>Hemodialysis Type 1</td>
<td>585.6 last stage renal illness (requiring chronic dialysis).</td>
</tr>
<tr>
<td>Hemodialysis Type 2</td>
<td>996.1 mechanical complication of alternative vascular device.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>E079.1 kidney dialysis for abnormal reaction of patient.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>V45.1 postsurgical nephritic dialysis detail.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>V56.0 encounter for extracorporeal dialysis.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>39.27 details of arteriovenostomy for nephritic dialysis.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>39.42 revision of arteriovenous shunt for nephritic dialysis.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>39.43 removal of arteriovenous shunt for nephritic dialysis.</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>39.95 hemodialysis.</td>
</tr>
</tbody>
</table>

D. Unstructured Data Extraction

Unstructured data include discharge summaries, ECG, radiology reports etc. We have considered most of them here. To stay focused on the NLP, we excluded notes related to any imaging results, particularly, radiology report, ECG report etc. Initially we have extracted patient identification number, admission identification number, note text, and note date/time and used SQL query for forming patient cohort to extract accurate information related to our course of interest.

E. Improving Unstructured Data Extraction

To improve the results of the target cohort identification i.e. patients with diabetes mellitus, and who underwent hemodialysis while ICU stay, we scanned the clinical notes programatically. In this process, neural word embedding model has developed for all the considered clinical notes with the help of Word2Vec [49] which support both Continuous Bag of-Words (CBOW), and Continues Skip-gram (SG) model [50]. Both models have three layers: Input, Projection and Output, shown in Fig. 1. In CBOW model, it is given the contexts \( \{w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}\} \) to predict the word \( w_t \), and in SG model, it is given the word \( w_t \) to predict the contexts \( \{w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}\} \).

![Fig. 1. CBOW (left) and SG (right) models of Word2Vec.](image)

We used SG model for finding keywords which represent diabetic and diabetic with hemodialysis to enrich our techniques of finding. For example, keywords represent the aligned context for “hemodialysis” is shown in Table II.

TABLE II. ALIGNED KEYWORDS FOR ‘HEMODIALYSIS’

<table>
<thead>
<tr>
<th>Keywords with context</th>
<th>Similarity Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haemodialysis</td>
<td>0.9678</td>
</tr>
<tr>
<td>Kidney dialysis</td>
<td>0.9613</td>
</tr>
<tr>
<td>extracorporeal</td>
<td>0.9135</td>
</tr>
<tr>
<td>dialysis</td>
<td>0.8927</td>
</tr>
</tbody>
</table>

In this case, we have given a corpus of word \( w \) for which we were interested to find similar word and the clinical notes \( c_n \) from MIMIC III. We considered conditional probabilities \( p(c_n|w) \) to find a parameter \( \theta \) of \( p(c_n|w; \theta) \) for maximizing the corpus probability using Eq. (1) to the clinical text, \( c_{text} \).

\[
\arg\max_{w, \theta} \prod_{w \in c_{text}} \prod_{c_n \in c(w)} p(c_n|w; \theta)
\]

In Eq. (1), \( c(w) \) is the set of \( w \). Alternatively –

\[
\arg\max_{w} \prod_{w \in c_{text}} p(c_n|w; \theta)
\]

Here, \( D \) is the set of all words and context pairs that we have extracted from the clinical text.

![Fig. 2. Manual process for validation of identified list of patient cohort.](image)

F. Manual Review

We manually reviewed all the notes extracted by structured data extraction method and NLP based unstructured extraction method for correctly identifying patient cohort. We have created a validation dataset that contains the positively identified patients who have diabetes mellitus and underwent hemodialysis during ICU stay (from MIMIC III). We have
used this validation database to evaluate the precision of our patient cohort identification task. Fig. 2 illustrates our manual validation process.

III. RESULTS

A. Structured Query based Records Extraction

We identified 10397 patients of diabetes according to structured analysis. There were 2589 patients who underwent hemodialysis, where 1259 patients’ data are found using ICD-9 code, and 1330 patient’s data found using unstructured clinical notes searching with the keyword “%hemodial%”. Table III and Fig. 3 depict the overall statistics of cohort identification.

TABLE III. AFFECT OF STRUCTURED EXTRACTION IN COHORT RESULT.

<table>
<thead>
<tr>
<th>Context</th>
<th>Total patients</th>
<th>Structured (using ICD – 9 code)</th>
<th>Unstructured (procedure and notes search)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>10397</td>
<td>10397</td>
<td></td>
</tr>
<tr>
<td>Underwent hemodialysis</td>
<td>1910</td>
<td>1259</td>
<td>1330</td>
</tr>
<tr>
<td>Diabetes (underwent hemodialysis)</td>
<td>1084</td>
<td>323</td>
<td>790</td>
</tr>
</tbody>
</table>

Fig. 3. Number of identified patients by structured data extraction.

B. Improvement of Structured Query with NLP Support

Total number of patients after using NLP in structured queries and unstructured records retrieval is improved; total patients who underwent hemodialysis is 2703, where 1259 patients’ data are found using ICD-9 based structured analysis, and 1444 patient’s data found using unstructured clinical notes analysis. Table IV shows the statistics of how context information has changed the overall data extraction and Fig. 4 depict the resultant cohorts.

TABLE IV. AFFECT OF NLP BASED EXTRACTION IN COHORT RESULT.

<table>
<thead>
<tr>
<th>Context</th>
<th>Total patients</th>
<th>Structured queries with NLP support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underwent hemodialysis</td>
<td>1995</td>
<td>1444</td>
</tr>
<tr>
<td>Diabetes (underwent hemodialysis)</td>
<td>1107</td>
<td>813</td>
</tr>
</tbody>
</table>

Fig. 4. Number of identified patients by structured data extraction and improved clinical procedural and notes searching using NLP.

After extracting related information, we have verified all the records manually for all patients. As total number of clinical notes is ridiculously high (>2083000) in MIMIC III, so manual verification was infeasible. In such cases, finding only precision was sufficient enough which was also mentioned in [5]. We have found better precision in case of NLP based context representation in structured queries for unstructured clinical data analysis. The Table V shows overall improvement in terms of precision metric.

TABLE V. PERFORMANCE OF PATIENT IDENTIFICATION TECHNIQUES.

<table>
<thead>
<tr>
<th>Validation data</th>
<th>Data extraction without NLP support (positive = 1084)</th>
<th>Data extraction using NLP support (positive =1107)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP)</td>
<td>1053</td>
<td>1089</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td>Precision = 100 × (TP/n)%</td>
<td>97.14%</td>
<td>98.37%</td>
</tr>
</tbody>
</table>

IV. DISCUSSION

In case of both structured data extraction and NLP supported data extraction, cohort identification achieved high precision. However, when we used NLP, performance of cohort identification improved a lot. Also, manual inspection reveals that our NLP supported technique retrieved those notes which were expressed in more natural way. As a result, without predefined medical vocabulary like UMLS which have used in some recent research, cohort have been identified with satisfactory accuracy.

However, we have to considered some limitations. We have to validate the result in the self-curated validation dataset because of the lack of gold standard EHR dataset. We could not calculate specificity because, in order to decide which patients must kept out of cohort (True Negative) and falsely kept in cohort (False Negative), we had to review the entire MIMIC III database. It requires substantial effort particularly in case of large EHR datasets.

V. CONCLUSION

In this study, we have used a large reliable health records. We have investigated how identification of patient cohort can be improved using NLP. We have focused on the better context representation when querying unstructured data and achieved satisfactory improvement without using any predefined medical vocabulary. Thus, it is anticipated that by using NLP with other medical information, cohort can be identified more reliably which in turn can be used in personalized or group-based health research.

ACKNOWLEDGMENT

Authors acknowledge Computational Physiology Lab, MIT for providing access to MIMIC III EHR dataset.

REFERENCES


