A Proof of Concept for Assessing Emergency Room Use with Primary Care Data and Natural Language Processing*

J. St-Maurice1,2; M.-H. Kuo2; P. Gooch3

1School of Health and Life Sciences and Community Services, Conestoga College Institute of Technology and Advanced Learning, Kitchener, Ontario, Canada;  
2School of Health Information Science, University of Victoria, Victoria, British Columbia, Canada;  
3Centre for Health Informatics, City University, London, UK

Keywords  
Primary care data, natural language processing, term extraction, emergency room inappropriate use, measurement methodology, service utilization

Summary  
Objective: The objective of this study was to undertake a proof of concept that demonstrated the use of primary care data and natural language processing and term extraction to assess emergency room use. The study extracted biopsychosocial concepts from primary care free text and related them to inappropriate emergency room use through the use of odds ratios.  
Methods: De-identified free text notes were extracted from a primary care clinic in Guelph, Ontario and analyzed with a software toolkit that incorporated General Architecture for Text Engineering (GATE) and MetaMap components for natural language processing and term extraction.  
Results: Over 10 million concepts were extracted from 13,836 patient records. Codes found in at least 1% percent of the sample were regressed against inappropriate emergency room use. 77 codes fell within the realm of biopsychosocial, were very statistically significant (p < 0.001) and had an OR > 2.0. Thetically, these codes involved mental health and pain related concepts.  
Conclusions: Analyzed thematically, mental health issues and pain are important themes; we have concluded that pain and mental health problems are primary drivers for inappropriate emergency room use. Age and sex were not significant. This proof of concept demonstrates the feasibility of combining natural language processing and primary care data to analyze a system use question. As a first work it supports further research and could be applied to investigate other, more complex problems.

1. Introduction  
The issue of inappropriate emergency room (ER) usage is very important. The issue places a burden on the health system and increases the workload of emergency care services [1]. Patients with non-urgent problems make up a significant proportion of ER visits. Many health administrators believe that non-urgent visits reflect “inappropriate” ER use and that non-urgent patients should be treated in other ambulatory care settings or clinics [2].

As the use of health informatics in primary care continues to become more common in Ontario, it represents a new source of health information to answer system use and patient behaviour questions. Primary care data is particularly interesting because it uses a biopsychosocial model versus the biomedical model. The biopsychosocial framework takes into account the social, psychological, and behavioural dimensions of illness versus focusing primarily on physiological dimensions of care [3]. Though harder to abstract by informaticians [4], primary care data has the capacity to offer insights regarding patient behaviour that would otherwise be impossible to ascertain or model.

It has been suggested that a straightforward method of identifying inappropriate ER caseloads could be employed to calculate population-based rates and be applied in the development of managed care programs [5]. Based on combinations of bio-

* Supplementary material published on our website www.methods-online.com
psychosocial phenomena it may be possible to accurately predict inappropriate use and achieve this goal. More broadly, predicting system-use behaviour could enable governments to pre-empt or modify patient behaviour in the interests of lowering total healthcare costs.

This study’s goal was to undertake a proof of concept that combined natural language processing (NLP) and term extraction with large amounts of primary care data to answer a health system use question. Specifically, the study’s objective was to seek out specific primary care biopsychosocial concepts and correlate them with inappropriate ER use by using existing investments in primary care computerization. This technique represents a novel approach to analyzing an international and trans-cultural research question through health informatics.

2. Background

Several studies were reviewed from 16 different countries regarding the inappropriate use of ER services by the general population. These articles were selected during a review of PubMed that included articles published in English after 1995 that measured factors associated with inappropriate ER uses. The result of this search returned 2275 articles. Articles that dealt with specific cohorts (e.g., children, homeless, veterans, etc.) or with repeated use, or with the use of the ERs for a specific service (e.g., dental) were excluded. In addition, articles were rejected if they did not present numerical results or did not specify data sources. 2190 articles were rejected based on title and abstract only. 62 articles were rejected after reviewing the entire article. In total, 23 articles met the selection criteria.

Each study [2, 4–25] performed a literature review and noted that inappropriate use of the ER or department was a costly phenomenon that has been increasing. The use of the ERs for a specific service (e.g. dental) were excluded. In addition, articles were rejected if they did not present numerical results or did not specify data sources. 2190 articles were rejected based on title and abstract only. 62 articles were rejected after reviewing the entire article. In total, 23 articles met the selection criteria.

All studies used descriptive statistics to describe their results. Fourteen studies presented their results with inferential statistics results in the form of odds ratio (crude and/or adjusted), mean differences, standard errors or beta coefficients [5, 9 –11, 13 –22]. Whenever confidence intervals were presented, 95% ranges were used. When the type of regression was stated, logistic regression was always used, except for the study by Carret et al. [6] that used Poisson Regression.

Nine studies established explicit criteria before conducting their study to categorize patients as urgent or non-urgent [7–10, 13, 15, 18–20]. Some used flow charts or established multiple criteria and categorized the patient if they met a subset of that criteria [10]. Some explicit criteria were applied by medical experts [9] while others were done systematically through a query to an existing database [8].

The sources of information for these studies have either relied on patient surveys, interviews or questionnaires [2, 4, 5, 8, 11, 13, 14, 17–23], ER registries [9, 12, 15, 16, 24, 25] or both [26]. None of the studies used primary care data, or performed large-scale concept extraction using NLP techniques in their analysis.

Several different criteria were used to define inappropriate visits to the ER. Nine studies used an existing triage system to classify urgent and non-urgent visits. The Canadian Triage Assessment System (CTAS), Australian Triage Scale (ATS), Hospital Urgencies Appropriateness Protocol (HUAP) and Emergency Severity Index (ESI) were used as pre-existing, validated tools to classify patients. When the CTAS scale was used, levels 4 and 5 were classified as non-urgent [2, 4]. A Taiwan study based its triage system on CTAS and also assessed levels 4 and 5 as non-urgent [16]. Similarly, when the ATS was used, levels 4 and 5 were considered non-urgent [26]. When using the ESI, one study categorised levels 3, 4 and 5 as non-urgent [27] while another only categorised level 4 and 5 as non-urgent [23]. Generally, when using an existing five point triage system, the studies consistently categorised levels 4 and 5 as non-urgent.

No study assessed identical factors in non-urgent use, nor did they collect data from the same source and define inappropriate use in the same way. Therefore, it is not surprising that the results and conclusions of each study are unique. Each study [2, 4–25] concluded there were reportable differences in inappropriate ER use between the groups they studied.

3. Methods

3.1 Setting and Sample

The data in primary care would be maximally useful if coded in structured formats. Unfortunately, Wilcox et al. observed that of 3,348 physicians, only 8% were storing data in structured forms [28]. However, the same study found that 75% of those physicians were entering or dictating, as a minimum, encounter notes and medication information. Given its broader use, free text in primary care medical record systems have more potential. This is reinforced by the study of Nicholson et al. that suggested that the use of free text in records might affect the results of research by providing depth to the data set. The study recommended that free text be considered as an integral part of the EMR and to be included in future research studies [29].

A clinic in Guelph, Ontario, Canada was selected for the study based on data availability and broad adoption of electronic medical records within the community. As shown in Table 1, a sample of physicians was selected to provide a variety of educational background, sex, age, and graduation year. This was done to reduce the chance that written text style and annotations might be associated with a particular style of documentation or short form in the medical record.

After an ethics review approval process and signed consent by the participating physicians, de-identified data was exported from the selected cohort. Records were exported in OntarioMD’s CDS 3.0 specification. Free text content from the data sections MyClinicalNotes, PastHealth, PersonAllHistory, ProblemList and RiskFactors were used for the analysis. As suggested by Churches and Christen [30], hashing the health card number enables the data to be correctly linked while maintaining anonymity and privacy. For linking, the de-identified primary care records included a
healthcare number hash that was generated with an unknown salt.

The primary care records were linked to a hospital report from the Guelph General Hospital. The report listed any recorded ER visits generated by the cohort during the study period. For each ER visit the report included the patient’s health card number hash (generated with the same unknown salt), the date of the visit and the visit’s CTAS score.

### 3.2 Design and Procedures

Several different NLP techniques and tools can be used to extract information. These range from simple pattern matching methods, such as regular expressions (also known as RegEx), to complete processing methods based on symbolic information and rules or based on statistical methods and machine learning. The information extracted can be linked to concepts in standard terminologies and used for coding (such as SNOMED-CT).

The MetaMap 2011 tool was used as the core NLP and term extraction tool in this study (31). MetaMap’s strength is its ability to map text to the Unified Medical Language Syntax (UMLS) database, forming a standard terminology. Both MetaMap and the UMLS database were developed and are continuously supported by the United States National Library of Medicine.

The selection of MetaMap was based on its immediate availability and accessible support, both of which are ideal for a proof of concept. MetaMap is based on a minimal commitment parser where texts are split into chunks and identified as concepts. The parser is based on the notion of a special set of so-called barrier words that indicate boundaries between phrases. These barrier words make it possible to run MetaMap without a training model [32]. Though originally designed and intended for use on Medline abstracts, it has been previously used to code problem lists [33] and to identify respiratory disease in clinical records [34]. Kang et al. reported 80.8 precision, 87.1 recall and 83.8 F-Scores for noun phrases and 74.4 precision, 83.1 recall and 78.5 F-Scores for verb phrases when MetaMap was used to analyze Medline abstracts from 1999 [32].

In addition to MetaMap, the General Architecture for Text Engineering (GATE) tool was used to pre-process data, and identify sentences. The software forwarded sentences to MetaMap 2011 through a purpose-specific plugin that was designed to enhance MetaMap’s performance for annotating UMLS concepts [35]. The GATE tool is part of a collaborative project that is broadly used for text processing in several domains and is freely available [33, 34].

A significant advantage of using GATE is its embedded edition, which supports a full-featured Application Programming Interface (API). This enables easy programmable use of GATE and, through the plugin, MetaMap. A software application was created to automate the entire workflow for the analysis, including storing results to a MySQL database. As shown in Figure 1, the application features a full graphical interface and is accessible by double clicking the JAR file.

MetaMap was configured with runtime parameters through the GATE plugin. The parameters “-Xdt –v level0and4” were used, which configured MetaMap to truncate candidate mappings, prevents the use of derivational variation on the computation word variants, specifies the ‘SPECIALIST’ parser to assist in parsing, the default settings, and also specifies the use of the level 4 SNOMEDCT vocabulary [38]. NegEx results were performed by MetaMap and exported with the annotations but were not used in the statistical analysis. The MetaMap word sense disambiguation feature was not used to simplify the analysis and improve performance during the proof of concept.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Graduation Year</th>
<th>Graduation School</th>
<th>Official Roster Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>2007</td>
<td>McMaster University</td>
<td>984</td>
</tr>
<tr>
<td>M</td>
<td>2003</td>
<td>Queen’s University</td>
<td>1157</td>
</tr>
<tr>
<td>F</td>
<td>1981</td>
<td>The University of Western Ontario</td>
<td>1390</td>
</tr>
<tr>
<td>M</td>
<td>1988</td>
<td>The University of Western Ontario</td>
<td>1747</td>
</tr>
<tr>
<td>M</td>
<td>1991</td>
<td>University of Calgary</td>
<td>1969</td>
</tr>
<tr>
<td>F</td>
<td>2000</td>
<td>University of Karachi</td>
<td>776</td>
</tr>
<tr>
<td>F</td>
<td>1987</td>
<td>University of Toronto</td>
<td>1484</td>
</tr>
<tr>
<td>F</td>
<td>1980</td>
<td>University of Toronto</td>
<td>1954</td>
</tr>
</tbody>
</table>

Figure 1
Software application screenshot
3.3 Data Analysis

For the analysis, the classification of inappropriate or appropriate ER use was computed with the same method as Field et al. [2] by using the Canadian Triage and Acuity Scale (CTAS) to assess appropriateness.

The 2008 Guidelines for CTAS categorize CTAS 1 as "Resuscitation", Level 2 as "Emergency", Level 3 as "Urgent", Level 4 as "Less Urgent" and Level 5 as "Non-Urgent" (39). As a simplistic example, a conscious patient with normal vital signs presenting with mild chronic pain would be triaged as non-urgent, whereas a patient arriving with severe respiratory or cardiac distress would be triaged as a Level 2 or Level 1. In this example, the mild chronic pain might be better suited for a community or primary care setting.

A score of 4 or 5 was considered inappropriate, and a score between 1 and 3 was appropriate. While this technique for classifying appropriateness is not perfect and has been debated, it was used as an available and reasonable proxy measure to prove the concept.

A data analysis workflow is shown in Figure 2. Through the use of the customized toolkit, de-identified data was taken from primary care records and analyzed with MetaMap. The results were then stored in MySQL database. The data was extracted in the desired format by using queries and stored procedures to export data into comma separated value (CSV) files.

Binary logistic regression was performed with R, a free, open-source statistical package [40]. The model used was \( \text{InappropriateUseYN} \sim \text{Age} + \text{Gender} + \text{BiopsychosocialCode} \), where logistic regression was used with an independent variable of biopsychosocial code, controlling for age and sex, and a dependent variable of inappropriate use. This model was selected to support the study at the proof of concept stage. The model was simple and used available data.

Data was exported from MySQL where patients formed each row, and concepts formed each column. This created a large matrix, as shown in Figure 3. In R, a for-loop was used to regress each UMLS concept as \( \text{InappropriateUseYN} \sim \text{Age} + \text{Gender} + \text{Code} \). For each column representing an UMLS code, a zero represented the absence of the code in the patient record, and a „1” presented the presence of a code in the patient record. The number of columns was limited by MySQL to a maximum of 4096 and only codes that were present in 99% (138 or more) of the patient population were used. For this dataset, 4028 UMLS codes met these criteria. Logistic regression results were imported into MySQL.

Two criteria were used to sort the resulting UMLS concepts. To minimize the effects of random error, only very statistically significant UMLS concepts were selected (\( p < 0.001 \)). In addition, to maximize clinical relevance only UMLS concepts with more than a doubling effect on the odds of using the ER inappropriately (\( \text{OR} < 0.5 \) or \( \text{OR} > 2.0 \)) were selected. Concepts that met this criteria were exported to Excel, printed and presented to the Chief of Staff of the Guelph General Hospital and sorted as either “Biological Symptoms”, “Diagnosis”, “Psychological”, “Social”, “Drugs”, “Regional Oddities”, “EMR Oddities” or “Other”.

4. Results

4.1 Data Characteristics

The patient population from the physician cohort included 13,878 records. The records from 41 patients were dropped because of null or duplicate health cards that prevented linking to ER use. The sample included 6396 men and 7437 women.

The local hospital generated a report for all ER visits over the last three years from patients who were registered to the cohort doctors. In total, 7466 visits were reported from 3976 distinct patients. There were 10 visits for 5 patients whose ER visit was recorded.
with a null health card number and considered unusable. In addition, 276 records from the ER visit data did not match any records from the primary care dataset. This was likely caused by erroneous or mismatched healthcare numbers. After removing these records, 6923 visits were used from 3690 patients.

Over a three year period (November 1, 2008 to October 31, 2011), the overall population had a 26% incidence of ER usage. During the same time, 1931 patients had used the ER with a CTAS score of 4 or 5 (13.9%). In comparison, over a one year period (November 1, 2010 to October 31, 2011), 798 patients had used the hospital with a CTAS score of 4 or 5 (5.7%). Overall, 39.13% of ER visits were CTAS 4 or 5, and considered inappropriate. A frequency distribution of the CTAS levels is shown in Figure 4.

### 4.2 Extracted Data and Concepts

There were 10,823,636 annotations extracted from the 13,836 records in the cohort. The number of codes per patient varied from 1 to 9031. A histogram of the total number of annotations per patient record is shown in Figure 5.

There were 38,263 distinct codes extracted from the 13,836 records. The number of distinct codes per record varied from 1 to 1896. A histogram of the number of distinct codes found in patient records is shown in Figure 6.

From the total of 38,263 distinct codes, 10,594 (27.68%) were only present in a single patient record. 33,446 distinct codes (87.14%) were only present in 100 different records. Only 4028 (10%) of the total distinct codes were found in at least 1% of the records. Most annotations occurred more than once in a patient record; each annotation was used between 1 and 132,904 times. On average, there were two instances of an annotation per patient.

There were 423 codes that were very significant with an odds ratio above 2. The categorization of these codes is shown in Supplementary Online File (Appendix A). Codes categorized as primary care biopsychosocial concepts are shown in Table 2. The logistic regression results for the biopsychosocial UMLS codes are shown in Supplementary Online File (Appendix B).

### 4.3 Processing Time

A total of 13,828 records were analyzed with the Data Extraction tool. The total processing time was over 7 days and was performed in an Ubuntu virtual machine on a new 12 Core, 32Gb ESXi server. As a simple measure, the time to process each record was 47 seconds. This included the time to perform language processing, to export the data to the database and to perform binary regression. This time per rec-
ord did not include the time needed for extraction and de-identification.

5. Discussion

5.1 Biopsychosocial Codes

Four hundred and seventeen concepts had a p-value below or equal to 0.001, and an odds ratio smaller than 0.5 or larger than 2.0. Of those concepts, 77 were categorized as biopsychosocial concepts.

From a clinician’s perspective, an OR of 2.0 represents doubled odds that the patient used the ER (ER) inappropriately. For example, of 139 patients that had the concept C0241331 (TALKATIVE) in their record, 46 (33.09%) used the ER inappropriately. In comparison, of 13,651 patients who did not have the talkative concept in their record, 1885 (16.02%) used the ER inappropriately. The odds of using the ER inappropriately with the code (33.09%), versus without the code (16.02%) are doubled (OR = 2.03).

Many of the concepts involved pain. The theme of pain is supported as a valid conclusion by Pereira et al. [15]. In their analysis, they noted that “emergency departments have become recognized as important source (…) for all presenting with pain or fear”. This result is not surprising, while at the same time supportive of the methodology.

The concepts categorized as social were informative. Some codes talked about specific professions (e.g. Computer Job), whereas others talked about personality traits (e.g. talkative, telling untruths, grooming self-care). Some of the codes can be considered as situational (e.g. poverty, housing, fighting, police, cultural aspects, victim of abuse). These are similar themes that were explored by other studies that considered social support [6], housing status [21] and social deprivation [18].

The psychological concepts demonstrate a dimension of mental illness in the inappropriate use of the ER. Keene and Rodriguez study reviewed the relationship between mental health problems and the use of the ER [41]. They concluded that there was an increased ER use and unmet health-related need within a total mental health population, due in part to that population vulnerability to accidents and self-harm. From this perspective, it is again not surprising to see a mental health element in inappropriate ER use, while at the same time supportive output of the validity of the methodology.

In contrast to other studies [2, 4, 6], age and sex did not have a significant effect on the inappropriate use of the ER. This was tested in the model of InappropriateUseYN ~ Age + Gender. This difference could be related to the source of the data, as previous studies relied on ER visit data or surveys [2, 4–25] and did not use the demographic information from the entire primary care population. Whereas some studies reported a small difference between male and female users, it may have been attributable to unmeasured natural imbalances in the overall population.

Given the biopsychosocial themes that surfaced in the analysis, the methodology appears to be a viable alternative to other techniques; statistically significant UMLS themes extracted from primary care records and associated to inappropriate ER use have provided similar themes found in existing literature.

5.2 Natural Language Processing Errors

There were many errors in the NLP and term extraction. These errors are a consequence of using MetaMap for term extraction with its default settings and not opting to perform pre-processing, such as noun phrase chunking. Whereas the settings selected ensured that every possible extractable code was stored and analyzed, not limiting the analysis to a specific vocabulary did introduce noise and erroneous information despite limiting for statistical significance of the results.

Word sense disambiguation was not used at the proof of concept phase and this introduced problems with word sense interpretation. The extracted term C2371406 (Kicking) is a good example. The primary definition in the dictionary for the verb “kicking” is to “strike with the foot or feet”. With this definition, the verb is violent, and potentially interesting from a variety of paradigms. However, the term “kicking” has forms including object verbs (e.g. ‘He kicked the ball’), non-object verbs (‘That horse kicks when you walk into his stall’), idioms (e.g. ‘Her refusal even to talk to me was a kick in the teeth’) and phrasal verbs (e.g. ‘kick in’, ‘kick back’, ‘kick off’). Not all of these forms should be expected in a medical record however, the verb itself is complex.

The following sentences are examples of the verb “kicking” found in the source records: ‘patient was kicked out of the mental health support group’, ‘patient was kicked in the face at the bar’, ‘Tried to leave a message for the patient but her answering machine did not kick in’ and ‘pacemaker will kick in if heart rate falls below 60’. In each case the

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Categorization of Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Codes (p &lt; 0.001, OR &gt; 2.0)</td>
</tr>
<tr>
<td>Biological Symptoms</td>
<td>Abdominal discomfort, Abdominal Pain, Bladder dysfunction, back strain, body ache, Bruising, Chronic back pain, Chronic pain, Dehydration, Diarrhea, Dizzy spells, Flank Pain, Infected, Ingurown toenail, Laceration, Muscle tenderness, Neuropathic Pain, Pain, Pain NOS, pain radiating, Paronychia, Pelvic Pain, Respiratory distress, Severe pain, Shooting pain, side pain, Sinus pain, Smell, Stomach pain, Viral illness, Vomiting, Weak</td>
</tr>
<tr>
<td>Psychological</td>
<td>ADHD, Anorexia, anxiety attacks, Bipolar, Bipolar Disorder, CBT, Chronic alcohol abuse, Cognitive, Conflict, Confusion, Hallucinations, Marihuana, psych, Suicide</td>
</tr>
<tr>
<td>Social</td>
<td>Abduction, Abuse, addictions, Assault, boyfriends, Crisis, culture, Disability, Disabled, English, Fear, fight, groom, hearing loss, Housing, Job, Looks, Lying, Mobility, Mono, Police, Poor, Pregnant, Social Work, TALKATIVE, Tearful</td>
</tr>
</tbody>
</table>
system correctly converted the verb “kicking” from the sentences, but the meaning and implication are very different in each case. The first instance is an important status indicator of a mental health treatment plan, the second would explain and qualify physical trauma, the third is trivial and meaningless from a medical standpoint and the fourth is an important component of a complex treatment plan.

The verb kicking represents the need for word sense disambiguation when analyzing clinical free text; the theme was falsely identified, statistically significant and represents an odds ratio greater than 2.0 but does not represent a meaningful relationship or result. The concept highlights the complications in interpreting results for codes that are vague or ambiguous.

There were also instances of errors where noun and verb phrases were not properly analyzed. For example, the concept C0032854 – poor had a preferred name of Poverty. In this instance, the code is intended to represent a social condition. However, the term is used in the document as an adjective (e.g. poor sleep, poor eyesight). The intended meaning of the sentence is not a poor person’s sleep or eyesight, but a lower quality of sleep or eyesight. Another example was concept C0220814 – Culture with a preferred name of Cultural Aspects. It is another example of MetaMap incorrectly associating a concept, as the meaning with the clinical texts is “blood culture”. In this case the code for culture was categorized as a social component of the biopsychosocial realm, but the analysis incorrectly attributed the code from a procedure.

Some codes appeared as a result of regional oddities. For example, the concept of C1415051 – GGH Gene is erroneous because the use of the acronym GGH in the medical record for the physicians means “Guelph General Hospital”. Though the term ‘GGH’ in a community that doesn’t use the Guelph General Hospital might genuinely mean the GGH gene, the MetaMap system was not able to understand its use correctly. This is a geographic anomaly that could be corrected through the newest version of the MetaMap system that supports a custom abbreviations file. This type of error adds noise to the results, but is easily decipherable upon inspection.

Many of these errors represent significant future work to perfect the methodology and take the proof of concept into application. A variety of opportunities exist to perfect the methodology through the use of pre-processing, to detect known acronyms before they are processed, and post processing techniques, to take negation into consideration [33]. Similarly, opportunities exist to investigate the use of other tools for the concept extraction. Tools such as IndexFinder [42] and CTAkes [43] may improve performance and reduce processing errors, but lack a full-featured Java API to integrate the analysis into the experimental workflow. It would be worthwhile to compare the processing time, results and conclusions of these alternate tools to further validate and refine the proof of concept into a robust and broadly applicable methodology. The current results could be used as a baseline for future work.

Recognizing the limitations in the code extraction analysis and the resulting errors, it is worthwhile emphasising that despite these limitations, the results still align with existing literature, are based on a substantial dataset and are limited for statistical errors associated with random error.

5.3 Processing Time

The total time for NLP and term extraction in the analysis is an important consideration for future studies. Analyzed sequentially, for a population of 100,000 patients, the total processing time would reach approximately 54 days. For 1 million records, the total time would be approximately 540 days. The majority of the processing time was consumed by MetaMap’s analysis. Though selected for the proof of concept due to its availability and documentation, MetaMap’s performance in terms of speed of processing complex sentences or those without grammatical structure is known to be problematic. Performance might be increased by passing phrase chunks to MetaMap instead of a complete sentence [31], since a ‘sentence’ can be ambiguous (e.g. large chunks of text, run-on sentences, bulleted lists, missing punctuation delimiters). Other tools may be better suited and may have a dramatic impact on performance [42, 43]. The use of pre and post processing methods should also be considered. Though the current results can be used as a baseline in future work, there are obvious opportunities and existing methods that could significantly increase the overall performance.

The data analysis architecture can also be improved. In this case all the data was analyzed sequentially, which is the least optimal approach. Efficiencies would be increased by distributing the workload to multiple computers or by taking advantage of parallel processing. While this does increase the complexity of the analysis and requires that it be managed by a central process, it is similar to cloud computing solutions used in other bioinformatics research that involved large quantities of data [44]. Architecting the data analysis strategically will be critical to achieving a scalable and highly available methodology.

It is worth noting that once a set of free text records has been analysed its data can be recycled and analyzed in a number of different ways without additional code extraction or NLP. Once the initial processing is completed once, the performance of the selected tool is less important. To this end, a centralized warehouse of results would pre-empt repetitive analysis and would enable researchers to ‘share’ their processing results. A warehouse of this type offers another opportunity to improve the methodology’s performance.

5.4 Limitations

This study is a proof of concept and first work, and is therefore limited in a number of ways. Most importantly, the categorization of ER use as appropriate or inappropriate through the use of CTAS is an area that has been debated. Though the application of CTAS for categorization of ER use is supported by previous studies by Field and Lantz [2] and Afifi et al. [4], it could be considered an oversimplification of a complex issue. Along these lines, the American College of Emergency Physicians [38] recently suggested that low scale categorizations of ER use (e.g. 4 or 5) do not necessarily mean the patient is using the ER inappropriately. They noted that...
Many 5 scale systems have lists of “non-emergency diagnoses” that include diagnoses with symptoms of serious medical conditions. In some communities, like Guelph, sutures are not normally removed in primary care. As such, the ER would treat patients seeking this service as levels 4 or 5, when in fact their use of the service was appropriate. The selection of CTAS scores for the categorization of inappropriate use in this study was based on its capacity to serve as an easy proxy measure that did not require further analysis or interpretation. Analyzing the hospital data using other methods to identify inappropriate use without CTAS would have been a substantial study in itself and would have added heavily to the already considerable volume of data to compute.

Another important limitation is the negation of codes during the NLP and concept extraction during the analysis. For example, if a record had the sentence “patient might have cancer”, but later had the sentence “not cancerous; benign growth on foot”, this study would have attributed the concept of cancer to the record. In this example the “theme” of cancer, anywhere in the chart, was used to determine the concept’s presence and was used in the analysis. This could be addressed in future work by taking negEx results, word placement and MetaMap scores into consideration. Introducing these variables into the statistical model, however, would introduce significantly complex dimensions and would require careful consideration and application.

The use of odds ratios in the univariate analysis was primitive. However, this simplification was justified given the number of variables and dimensions. The drawback is that odds ratio as a univariate measure does not take the interactions of codes into consideration, nor was it accompanied by a calibration measure.

As a result of these limitations, the results of this analysis are not necessarily ready for use in a full clinical or quality improvement process. However, as a proof of concept the methodology presented demonstrates the feasibility of this type of analysis. This study is preliminary work demonstrating a new method. The results show the capacity of using NLP and primary care data as a first step towards making use of these types of data sources to answer system use questions.

6. Applications

Though this study is a proof of concept, future uses of this technique could prove to be very practical for understanding, and even predicting inappropriate use based on biopsychosocial profiles. In this study, 417 concepts were very statistically significant (p-value below or equal to 0.001 and an odds ratio smaller than 0.5 or greater than 2.0) and 77 concepts that fell in the realm of biopsychosocial, and would normally be unavailable outside primary care contexts. There is literature to validate the biopsychosocial codes and themes discovered in the analysis. Supported by previous work in inappropriate ER use, the presented method and proof of concept support further investigation and research.

The quantity of data that was extracted during the analysis is very large: over 10 million separate annotations of over 38,000 different codes reside in a database. The analysis presented shows the tip of the iceberg; the data can be reanalyzed by separating data by CTAS level, individual physician, diagnosis, social cohort, time of visit, age or gender. Although complex, each analysis could provide further insight to inappropriate use. The question could also be reversed by assessing common concepts in patients who do not use the ER inappropriately.

This method could be applied to other areas of healthcare. Outside the context of ER use, the analysis could support improvements to the entire spectrum of hospital services by providing a richer understanding of the motivations and dimensions of service users. Even more broadly, the methodology has the potential to enable the prediction of system-use behaviour and could enable governments to reduce healthcare costs throughout regions and communities. In addition to using this technique for service-use analysis, the technique could provide very interesting data to researchers in the fields of public health and epidemiology.

Compared to previous studies regarding inappropriate use, this study offers a unique analysis method which has not been previously used. Compared to other studies that use NLP and term extraction, this study was unique where it performed an analysis on primary care records, from Ontario, on a massive scope and used the results in a statistical model. A study performing a similar analysis has not been published in this context and it is therefore difficult to compare results against previous studies. However, in future work, and as the methodology is refined, comparing the results to other studies that use term extraction will be more obvious. Whereas this study establishes a baseline for future work on the method, other studies offer opportunities for improving the performance and are complementary.

7. Conclusions and Future Work

This study was a proof of concept that extracted biopsychosocial concepts from primary care records and related them statistically to inappropriate ER use by using a software toolkit with MetaMap and GATE for NLP. Over 10 million concepts were extracted from 13,836 patient records. Analyzed thematically, the statistically significant codes involved mental health and pain related concepts. Age and sex were not significant. As part of a larger goal, the proof of concept shows the feasibility of using the extracted codes and regressing the data against other types of service usage. The use of community health services, mental health clinics, or social services are examples of other services that could be analyzed with the current data set.

This study contributes to inappropriate use literature by using data mining through NLP and primary care data to answer a question that has been previously dominated by surveys and interviews. It also uses data from the general primary care population to draw comparisons and to calculate odds ratios. Unlike other studies, this study used an undetermined set of characteristics instead of pre-determining components of inappropriate use and tailoring questionnaires to gather data.
None of the studies reviewed in the literature review used a similar technique to evaluate this research question, even though the results are similar. This study contributes to NLP literature by demonstrating the application of an analytical software tool that combines MetaMap and GATE technologies. It has demonstrated the capacity for the combination of these tools to mine large quantities of data and answer a system-use question. This study has shown that using NLP on large volumes of free text is a feasible method for data mining. Subsequent work and research is required to continue enhancing this proof of concept. In particular, instead of analyzing and considering single terms, the concepts should be grouped and analyzed to form “constellations” of terms that would depict a broader patient profile. This type of analysis introduces considerable complexity, but given the preliminary results, it would likely prove very interesting. Whereas the current use of odds ratios represented a univariate analysis of each code, to support further investigation, statistical multivariate analysis will be required. Given the quantity of data, this may prove to be a daunting and highly involved task. An appropriate calibration measure, such as the Hosmer-Lemeshow chi-squared statistic, would help characterize the way in which each code interacts with the other. There are obvious opportunities to improve the performance and reduce the noise and errors in the NLP phase of the analysis. Despite the errors, the continued investigation of this method is supported by the results achieved through this proof of concept.

References


