Comparison of Natural Language Processing Techniques in Analysis of Sparse Clinical Data: Insulin Decline by Patients

Shervin Malmasi, PhD1,2; Wendong Ge, PhD1,2; Naoshi Hosomura, DDS, DMSc1,2; Alexander Turchin, MD, MS1,2
1Brigham and Women’s Hospital, 2Harvard Medical School; Boston, MA

Abstract

We present a comparative evaluation of a range of popular Natural Language Processing (NLP) approaches for Information Extraction (IE) in clinical documents to detect cases of patients declining medication that has been recommended by their providers. More specifically, we tackle the task of identifying diabetics who decline insulin, using a training set of 51k randomly selected provider notes. Analysis shows that decline of insulin by patients is a rare phenomenon, with a document-level prevalence of approx. 0.1%. We examine the effectiveness of some of the most popular IE approaches, including sentence-level support vector machines (SVM)-based classification, token-level sequence labelling using conditional random fields (CRFs), and rule-based detection based on encoding human knowledge. Our results on a held-out test set show that the generalization of rule-based approach (F1=0.97) outperforms the SVM (F1=0.61) and CRF models (F1=0.40).

1 Introduction

The rise of Electronic Health Records (EHRs) has led to an explosion in medical informatics research and applications. While the majority of these approaches use the structured parts of these records, an omnipresent obstacle in the application of these quantitative analytics techniques has been that many crucial observations are often only available within EHR narrative documents. Consequently, investigators have realized the potential value of the information present in the free-text narratives written by providers.

This use of narratives over codes is borne out of the flexibility required by providers in describing their nuanced observations, diagnoses, the patient’s perspective, and discussed treatment strategies. However, the unstructured natural language that they create is not directly usable in quantitative analyses. To use this narrative data, it would need to be manually abstracted via labor-intensive chart review. This manual review can be time consuming, error prone, and require extensive resources. This issue is particularly pronounced for identifying low-prevalence information where hundreds or thousands of notes would have to be reviewed to find a single instance of the target information.

In many scenarios the amount of EHR data available is so large that chart review becomes unfeasible. Such limitations have led to the development of computational methods to automatically process and mine EHR documents for information of interest. This is particularly pertinent in cases where the information being sought is not highly prevalent. One such example, and the focus of this study, is the decline or rejection of medications by patients. While it is anecdotally known that patients frequently decline medications that are recommended by their providers, little information is available on this phenomenon. It is not completely clear how frequently patients decline medications and how commonly they ultimately receive medications they initially refused.

While numerous approaches to information extraction have been proposed, there is no clear method that is superior across all tasks. The primary objective of this study is to comparatively evaluate the effectiveness of the most popular approaches that have been identified in the literature. Such an analysis will allow us to compare and contrast these methods within the context of our specific NLP task.

In summary, the key aims of the present study are to:

1. Determine how the identification of decline of insulin therapy by patients can be framed as an IE task, allowing for the quantification of its prevalence;
2. Identify a range of computational methods that can be used to achieve this task;
3. Evaluate their individual performance on identifying such rare phenomena.
The rest of this manuscript is organized as follows: we continue with a review of related work in section 2, followed by a detailed description of our data and methodology in section 3. Our results are presented in section 4 and we conclude with a discussion in section 5.

2 Background Related Work

In this section we briefly discuss some of the related topics and relevant previous work that have motivated the present study.

2.1 Information Extraction

A key benefit of the rise of NLP methods has been their direct application in data mining and information extraction from clinical documents. A wide range of methods have been applied by researchers, varying by their requirements and the specific task being addressed.

A recent systematic review of the medical information extraction literature related to case detection found that the most common approaches include simple keyword searches, rule-based text mining, and machine learning based methods. It was also reported that while 67% of the studies incorporated rule-based components, only 9% made use of machine learning methods. Regardless of the methods used, information extraction from narratives substantially improved cases identification over using structured fields alone. The authors also state:

“Different methods of information extraction were reported, ranging from manual review of records to both rule-based algorithms and probabilistic or statistically driven models using machine learning methods. No particular type of algorithm stood out as particularly better than any other. Accuracy also varied by condition, but no clear pattern was evident.”

The prevalence of the target condition is an important factor that affects the performance of the various IE methods. In this context, the overarching aim of the present investigation is to assess the performance of several widely-used IE approaches for the identification of rare target information.

In the following section we will briefly look at the different machine learning and rule-based approaches.

2.1.1 Classification Approaches

Within machine learning approaches, classification methods have been the most popular approach. More specifically, Support Vector Machines (SVMs) are considered to be a state-of-the-art approach for text classification due to their ability to deal with extremely high-dimensional feature spaces across large training data. They are widely used for dealing with unstructured notes within the medical domain.

Pakhomov et al. considered the task of predicting quality of life from free-text provider notes at the document level. Using an SVM model trained on bag-of-words and bag-of-concepts features, they compared the machine learning results against standardized assessment instruments (e.g. questionnaires) completed by the patients. The results showed that the SVM model results from the provider notes had good concordance with the instruments completed by the patients.

In other tasks the concepts or information of interest is expected to appear in a limited, local context and considering information at the document level may not be helpful and introduce noise that results in false positives. In such cases the classification can be performed at the sentence level. In one study of identifying clinical phenotypes in narrative data, researchers applied various SVM models to identify discussion of breast cancer. Their results showed that a sentence-level SVM model yielded the best results on the test set, with performance far surpassing a simple string matching baseline.

Sentence-level SVM models have also been used in various other medical IE tasks, such as the categorization of sentence types in the abstracts of randomized controlled trials, sentence extraction for question answering, and smoking status detection.

2.1.2 Sequence Labelling Approaches

Another popular machine learning approach to information extraction is based on sequence labelling. While the classification approach described in the above section assigns a single discrete class to each sentence, the sequence labelling approach treats each sentence as a sequence of observations, with each word being an individual observation. These models then assign a discrete label to each observation, taking into account the labels assigned to the previous observations as well as additional contextual features related to the preceding and following words. It is this context-
aware nature of these models that makes them highly suited for tasks such as Named Entity Recognition (NER) which require identifying relevant fragments of text.

Hidden Markov models (HMMs) are one type of probabilistic generative model that perform this task. However, Conditional Random Field (CRF) models, a discriminative variation of HMMs, have proven to be most popular in practice. One important factor is their ability to take into account longer range dependencies between contextual features. Another key consideration is their capacity to model conditional relations between labels and features, as opposed to the independence assumption required to make generative models such as HMMs practical.

In a broad study of medical concept extraction spanning multiple datasets and concepts such as “Problem”, “Treatment” and “Test”, Kim et al. employed a range of machine learning methods, including CRFs and SVMs. While the results were close, the CRF models outperformed the SVMs in most cases. It is also worth noting that the concepts they were extracting had a very high prevalence in their documents, with over 75,000 annotated concepts.

Another study examined the extraction of temporal constraints from free-text notes in examining eligibility criteria. CRFs were used to automatically identify temporal expressions, achieving high performance on a held-out test set.

Recently, another study looked at the application of CRF models to extract prescription information from clinical documents. This task involves the identification of medication names, dosage details, mode of administration, frequency, and reason for prescription. Comparing CRFs, SVMs and Decision Trees the authors found that the CRF models were better on the whole, while the performance of the SVM models was close.

2.1.3 Rule-based Approaches

While machine-learning approaches to NLP have been evaluated by many investigators, rule-based approaches have remained popular for a number of reasons.

A prevalent issue is that statistical NLP methods cannot be easily adapted for the medical domain, due to the idiosyncrasies of medical data as well as the concise nature of the texts which often have limited use of syntax as they are written under time pressure. Adapting standard NLP models for these tasks would require substantial work to retrain the statistical models and researchers have reported that this process is considerably time consuming. A secondary issue is that the technical skills required for the development or use of machine learning and NLP models present a major barrier for researchers and clinicians wishing to employ these methods.

It has also been demonstrated that the prevalence of the target information affects the learning rate of machine learning algorithms. In particular, it has been shown that this effect is particularly significant for “rare outcome states” with a prevalence rate of < 10% in the training data. While the use of larger datasets can help here, this effect is mostly independent of the total sample size, relating more to the distribution of the classes within the training data.

These shortcomings are often mitigated by the use of rule-based processing that explicitly encodes expert knowledge. For example, the second i2b2 challenge focused on the extraction of obesity and related co-morbidities from “sparse” (i.e. low prevalence) data. An analysis of the results by the organizers showed that “rule-based approaches played a significant role” in the top performing systems, highlighting the challenges faced by machine learning systems on this type of data. One reason for the success of these approaches is the relatively low prevalence of the target information; machine learning approaches would require large amounts of data in order to induce models that are able to generalize sufficiently well to achieve good performance. On the other hand, rule-based approaches take advantage of the background knowledge of language and specifically medical discourse by the human rule designer to generalize the language model based on the limited available data.

Another study aiming to identify patients with colorectal cancer used NLP to develop a case detection system using both clinical narratives and structured data. At each step in the development of their system they compared both rule-based and machine learning approaches to the tasks. The authors report that “[t]he rule-based method achieved high performance on both training and test data sets (F-measure of 0.996), indicating such rules were generalizable”. Similar approaches have also been found to be useful for other NLP tasks such as extracting semantic relationships between medical entities, and negation detection.

In other areas, such as the Third i2b2 Challenge on NLP for Clinical Records, hybrid systems augmenting rule-based approaches with statistical methods have proven to achieve the best results. Finally, it is also worth noting that two of the most widely used tools in the field, MetaMap and MedLEE, are both based on rule-based approaches.
2.1.4 Summary of Information Extraction Methods

In summary, this brief review of the literature demonstrates that approaches to information extraction broadly fall into two categories: those relying on human-developed rules and others using statistical machine learning approaches to automatically induce decision rules. Among the machine learning approaches two of the most common ones have been classification methods and sequence labelling models.

Based on the existing literature, there is no definitive finding or hard-and-fast rule about which methods work best. On the whole, no method is superior to others in all scenarios; this is largely dependent upon task-specific variables, such as prevalence of the outcome classes, the amount of available data, the complexity of the task, and the level of variance in expressing the target information.

The identification of the best approach in such scenarios should generally be based on empirical results from the application of the methods. Accordingly, the present study focuses on the evaluation and comparison of these approaches for our specific task, the identification of patients who have rejected their provider’s recommendation of using insulin.

2.2 Insulin Decline by Patients

Type 2 diabetes is a naturally progressive disease. As beta cell function declines and insulin resistance grows with time, patients’ needs for pharmacotherapy increase. Though multiple classes of medications are available to treat patients with type 2 diabetes, their impact on glucose control is limited, while adverse reactions and contraindications constrain the number of patients who can take them. Insulin therapy, on the other hand, has no contraindications and few side effects. Consequently, for many patients time eventually comes when adding insulin to their treatment regimen becomes the best option for achieving glucose control.

However, studies show that transition to insulin therapy is frequently delayed. Clinical experience and initial research suggest that many patients with type 2 diabetes initially decline their healthcare providers’ recommendation to start insulin – possibly an important contributor to delays in insulin initiation. Studying decline of insulin therapy by patients has been challenging because it is only documented in narrative documents – there are no prescriptions written or insurance claims made that would leave a trace in structured electronic data. On the other hand, identification of patients who had declined insulin therapy is important to be able to conduct studies to understand the risk factors and long-term outcomes of this phenomenon, and to design population-level interventions that could help patients make an informed decision about treatment of their diabetes. In order to identify insulin decline by patients, it was therefore necessary to develop NLP tools to abstract this information from narrative electronic provider notes. Using NLP for this purpose was particularly challenging because, even though decline of insulin therapy may be common at the patient level, it is recorded in only a few of out of dozens or hundreds of provider notes a given patient might have, and is consequently rare at the document level. Therefore, in order to establish the optimal technique for identification of patients who had declined insulin therapy, we developed several methods, including classification- and sequence labelling-based machine learning techniques as well as a rule-based approach, for this purpose, and evaluated their accuracy.

3 Methods and Materials

3.1 Experimental Design

We frame our experiments as sentence-level information extraction tasks, similar to the previous studies outlined in section 2. Here each sentence in the data must be examined by a computational model which must decide whether it contains a mention of a patient refusing insulin or not. This is in essence a binary classification task, where the insulin rejection sentences are assigned to the positive class and everything else to the negative class. This classification is operationalized in different ways in each computational approach, and we describe these later in this section.

3.2 Data

3.2.1 Study Setting

Data for this study comes from the clinical records of adult patients with diabetes treated in primary care practices affiliated with Massachusetts General Hospital and Brigham & Women’s Hospital between 2000 and 2014.

3.2.2 Data Collection Criteria
For training and testing the NLP methodologies we used provider notes of patients who were at risk for declining insulin therapy. We identified them by selecting patients who had blood glucose levels above the recommended targets (HbA1c ≥ 7.0%) and had no previous record of treatment with insulin. These criteria selected 2,358,453 notes of 75,412 patients, from which the training and test data sets were subsequently drawn.

We randomly selected 50,046 notes into the training set and 1,501 notes into a non-overlapping test set. Both sets of notes were independently annotated by trained senior pharmacy and medical students. The reviewers marked all sentences describing patients who declined insulin therapy. The “gold standard” test set was reviewed by two independent reviewers, whose ratings were subsequently reconciled. The distribution of the data across both sets is listed in Table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (Insulin decline present)</td>
<td>535</td>
<td>19</td>
</tr>
<tr>
<td>Negative (No insulin decline)</td>
<td>2,660,475</td>
<td>86,487</td>
</tr>
</tbody>
</table>

These values demonstrate the rarity of the phenomena we seek to identify: in both set the negative class instances make up 99.98% of the data. This also highlights the difficulty associated with identifying this information through manual chart review and reinforces the need for computational methods to extract this information.

### 3.3 Evaluation

The primary aim of our evaluation is to assess the accuracy with which instances of insulin decline by patients can be detected. Given the extremely low prevalence of this information (as given in section 3.2), we utilize the balanced F-score (F1) of the positive class as our evaluation metric. To this end, for all models being evaluated we will report precision (positive predictive value; PPV), recall (sensitivity), and the F1 value which is the harmonic mean of the precision and recall values.

This evaluation was conducted against the aforementioned held-out gold standard test set. This data is annotated at the sentence level, so our measures quantify how well the methods are able to detect sentences that mention insulin decline, across all notes.

### 3.4 Classification Methods

We model the classification task at the sentence level, similar to the approaches described earlier in section 2. More specifically, a linear Support Vector Machine model is used, as they have been shown to achieve state of the art results for various text classification tasks. L2 regularization, with the default regularization parameter (C=1) is used to avoid overfitting. This was implemented by using the LIBLINEAR library.

Tokenization was first performed on the documents. For features we employed a bag-of-words approach by extracting word unigrams from our data. We also experimented with extracting the unigrams from the document lemmas instead of the original tokens (i.e. a bag-of-lemmas approach) which reduces the sparsity of the feature vector and may improve model generalization. The lemmatization process was performed using the Natural Language Toolkit (NLTK) library.

In some cases task-relevant information may be contained within the previous sentence, and not just the current sentence. To address this we also experimented with models where a feature vector representing the words/lemmas from the previous sentence is concatenated with that of the current sentence.

### 3.5 Sequence Labeling Methods

In order to frame our information extraction task as a sequence labelling problem we focused on assigning labels to all mentions of the word “insulin” and its synonyms, such as “Lantus”. More specifically, we labeled these words as belonging to the positive class if they appeared in a context that indicated the medication is being rejected or refused by the patient. Other mentions of these keywords were assigned to the negative class.
Tokenization was first performed and the sentence labels were then transferred to the keywords in each sentence. This is possible since our data annotation was performed at the sentence level, as described in section 3. This process allows us to assign a label to each word in a sentence: “P” for positive mentions of insulin being rejected, “N” for negative mentions of insulin in all other contexts, and “O” for all other words. For example, the sentence “He declines insulin therapy.” Will be assigned the following label sequence [O, O, P, O].

We then trained a Conditional Random Field (CRF) model to solve this task. We extracted features with a window size of [-5, +5] to capture the preceding and following contexts for each word. This model was implemented using the CRF++ library. L2 regularization was applied to reduce overfitting.

3.6 Canary Model

The final approach to this problem was based on the Canary information extraction software. The Canary software provides a platform that is designed for clinicians and researchers with limited technical expertise, allowing them to design natural language processing tools through the definition of a vocabulary and set of rules to match their target information. A screenshot of the software’s graphical user interface is shown in Figure 1.

This approach can be useful in scenarios where machine learning approaches are not feasible due to a lack of sufficient training data. An experienced clinician with a clear understanding of how the target information can manifest itself across documents can propose a set of criteria for identifying this information, in essence crystallizing the human intelligence and knowledge into a set of Canary criteria.

In the present study, this information modelling was performed by a clinician without prior NLP or information extraction experience. The researcher was provided the same training data described above in section 3.2 and asked to develop a Canary language model for identification of documentation of insulin decline by patients based on this information.

4 Results

We first report the performance of our machine-learning models on the training data. This was performed under stratified 10-fold cross-validation. This was then followed by an evaluation on the held-out test set to assess their generalization.

4.1 Cross-Validation Experiments

In this section we report our cross-validation results using the training data.

4.1.1 SVM Classification Model

As described earlier, we experimented with training linear SVM models using words, lemmas, and features from the previous sentence to capture any relevant contextual information. The results for these models are presented below in Table 2.
We observe that the performance between words and lemmas is very similar. However, including information from the previous sentence decreases performance for both feature categories. This indicates that these sentences do not usually contain relevant information. To the contrary, it appears that the model is in fact negatively affected by potentially irrelevant information from these sentences.

We also repeated these experiments with sentence filtering to remove instances that did not contain relevant keywords (e.g. ‘insulin’ and ‘Lantus’). This filtering greatly reduces the size of the negative class to approximately 10k sentences. However, this downsampling did not improve model performance relative to using the full dataset.

As noted earlier, our data is highly imbalanced and SVM models are sensitive to such imbalances. In particular, it is known that in such cases they produce models that favor the majority class. Data resampling methods are one way to mitigate such issues by oversampling the minority class or undersampling the majority class.

Since our training set only included 535 samples from the positive class, we tested increasing this number via oversampling. To achieve this we took our best model from the previous experiments (SVM with word features) and applied the SMOTE oversampling method, resampling the minority (positive) class to increase it to 2k, 4k, and 10k samples. The results for these oversampled models are listed below in Table 3.

We observe that oversampling with 4k provides a significant boost over the same model with just the original data. The increase in the F1 score is due to a substantial increase in the model’s sensitivity (recall) and a decrease in precision. Accordingly, we evaluated the word-based SVM with and without data resampling against the test set.

4.1.2 CRF Sequence Labeling Model

The results for our CRF models, using both word and lemma features, are presented in Table 4. Compared to the SVM model from the previous section, the CRFs achieve higher precision but lower recall. This may be preferable in scenarios where we wish to minimize Type I errors (false positives) at the risk of having a higher incidence of Type II errors (i.e. false negatives).
4.1.3 Canary Language Model

Our final method is a clinician-designed language model created using the Canary platform described above. The clinician received the annotated training set that was used to train the machine learning models and analyzed the data by manually inspecting instances of insulin decline by patients. Next, Canary vocabulary and information extraction criteria were created. These criteria were designed to detect the specific language used to document insulin decline in the manually identified instances with maximum accuracy and generalizability.

This process was performed iteratively until the clinician determined that no significant further improvements could be made. At the end of the process it resulted in a set of 148 word classes and 284 rules. When evaluated on the training data this model achieved a precision rate of 0.93 and a recall of 0.99, resulting in an F1-score of 0.96.

4.2 Test Set Evaluation

We next evaluated all of our models on the test set to see how well they generalize beyond the training data. The results for all models are shown below in Table 5.

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (words)</td>
<td>0.75</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>SVM (words + SMOTE oversampling)</td>
<td>0.71</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>CRF (words)</td>
<td>0.55</td>
<td>0.32</td>
<td>0.40</td>
</tr>
<tr>
<td>Canary</td>
<td>0.95</td>
<td>1.00</td>
<td>0.97</td>
</tr>
</tbody>
</table>

The evaluation shows that the cross-validation patterns hold, with the Canary model achieving the best results. This high performance indicates that the rules developed during training generalize very well to additional data. On the other hand, we note that the performance of the machine learning models is lower on the test set compared to cross-validation, suggesting that they do not generalize as well.

The SVM model performs better with oversampling, highlighting the importance of class prevalence and imbalance for such information extraction tasks. The CRF model achieved the worst performance on this task.

5 Discussion and Conclusion

In this study we presented a comparative exploration of different information extraction approaches for identifying rare clinical data in narrative notes. Our problem was novel in terms of the extremely low prevalence of the target information. A comparison of two classes of machine learning approaches and a rule-based approach based on human expertise found that the latter approach significantly outperformed the statistical methods on this task.

The reasons for the poor performance of the statistical methods are likely due to the rarity of the targeted outcome class, as pointed out in section 2. This low prevalence means that a very large amount of training data would be required to train a machine learning technique that can effectively capture this information.

This is further supported by a qualitative ad-hoc analysis of the positive examples from the training data, showing that there is a large amount of variance in how different providers express information about a patient’s refusal to accept a medication. This is sometimes stated in straightforward, indicative manner (e.g. The patient has refused to take insulin.) while in other cases it is expressed in terms of the patient’s concern or narrative (e.g. He is concerned about gaining weight if he goes on insulin.) These variations make it harder to train effective machine learning models without having large amounts of data.
The CRF model may have been additionally hampered by the size of the window it considered. Identification of the optimal window size should be a subject of future research. However, it would be unlikely to close the performance gap of this magnitude.

In many cases the manual annotation of the required amount of data would be prohibitive in terms of time and cost. This is reflected in the machine learning performance achieved in this study, given the relatively large amount of data that we had to process to find several hundred positive instances.

As we argued earlier, computational models are needed to extract this information since manual chart review would be too time consuming due to the low prevalence. Consequently, the rarity of the information being sought means that failing to identify the relevant cases is costly. This issue is particularly important for scenarios where it is critical to maximize the number of identified cases (i.e. sensitivity) such as when studying rare phenomena. Errors in the initial identification phase can have a significant impact on epidemiological studies, leading to biased findings and incorrect conclusions.

In this sense it could be argued that sensitivity is very important for this task as using a model with low recall exacerbates the difficulty of finding rare information. In practice this would mean that we would prefer models with a higher Type I error (i.e. false positive) rate than one with a higher Type II error (i.e. false negative) rate. In many cases it would be much more effective for us to weed out the false positive from a small set of results with high sensitivity than to perform the task manually, or to miss out on many true positives at the cost of not reviewing the results. This has implications for the low sensitivity achieved by the machine learning models in this study, while the rule-based approach achieved much higher results.

There are several avenues for future work. One possibility is to explore additional CRF models with longer distance dependencies. The machine learning models could also be augmented with additional syntactic features, such as part-of-speech tags. Another promising avenue for future work is the application of deep learning models for this task. More specifically, Recurrent Neural Network methods have shown great promise for improving information extraction performance and we plan to conduct additional experiments to assess their performance.

References