Combining Unsupervised, Supervised, and Rule-based Algorithms for Text Mining of Electronic Health Records: A Clinical Decision Support System for Identifying and Classifying Allergies of Concern for Anesthesia During Surgery

Geir Thore Berge *geir.thore.berge@sshf.no*
Department of Information Systems, University of Agder
Department of ICT, Sørlandet Hospital Trust
Kristiansand, Norway

Ole-Christoffer Granmo *ole.granmo@uia.no*
Department of ICT, University of Agder
Grimstad, Norway

Tor Oddbjørn Tveit *tor.tveit@sshf.no*
Department of ICT & Department of Anaesthesia and Intensive Care, Sørlandet Hospital Trust, Kristiansand, Norway

Abstract

Undisclosed allergic reactions of patients are a major risk when undertaking surgeries in hospitals. We present our early experience and preliminary findings for a Clinical Decision Support System (CDSS) being developed in a Norwegian Hospital Trust. The system incorporates unsupervised and supervised machine learning algorithms in combination with rule-based algorithms to identify and classify allergies of concern for anesthesia during surgery. Our approach is novel in that it utilizes unsupervised machine learning to analyze large corpora of narratives to automatically build a clinical language model containing words and phrases of which meanings and relative meanings are also learnt. It further implements a semi-automatic annotation scheme for efficient and interactive machine-learning, which to a large extent eliminates the substantial manual annotation (of clinical narratives) effort necessary for the training of supervised algorithms. Validation of system performance was performed through comparing allergies identified by the CDSS with a manual reference standard.

Keywords: Electronic Health Record, clinical decision support systems, structured data, unstructured information, narrative, machine learning, unsupervised machine learning, supervised machine learning, semi-supervised machine learning, rule-based algorithms.

1 Introduction

Undisclosed allergic reactions of patients are a major risk when undertaking surgeries in hospitals [13]. Adverse drug reactions (ADEs) perceived as a type of allergic reaction occur in 10% to 15% of hospitalized patients worldwide [26], and significant risks, costs and increased hospital stays are associated with unknown ADEs [25].

Although critical patient allergy information has been captured and recorded in the patient's Electronic Health Record (EHR), it may still be overlooked by physicians [24]. Critical information in the form of structured data in EHRs containing information about patient allergies is often the first physicians automatically encounter or manually look-up due to e.g. system alarms being triggered. However, such information may not be updated or complete, and may also be prone to inaccuracies increasing clinical risk [19]. Although the patient narrative is the primary, preferred, and richest source of patient information [15] and may contain detailed information about patient allergies, the clinical language it contains is

voluminous, unstructured, complex, and varied. Performing manual search for and identification of clinical information in the patient narrative demand much attention from busy physicians, potentially disrupting clinician workflow or patient-clinician communication [19]. Competing work tasks may thus compromise a thorough examination of patient narratives, leaving the exercise inconsistent and incomplete. There is also a lack of robust search engines in todays' EHRs, which typically only allow simple searching for explicitly stated words or phrases one at a time, while also being restricted to certain document types or EHR modules [16]. Although there is a trend towards using more structured data in the EHR, the unstructured narrative still excels when it comes to e.g. contrasting details, which makes its elimination or even decimation unlikely in the short or medium term [3]. Thus, there is a need for developing more robust methods for the retrieval of valuable clinical data from the narrative part of the EHR.

Clinical decision support systems (CDSSs) driven by natural language processing (NLP) have shown promise in leveraging information from the clinical narrative [7]. Traditional rule-based expert systems have been used extensively in healthcare [18], [20], [27]. As explored further in Section 2, such systems have been shown to be very accurate, while depending on controlled medical vocabularies which can be both demanding to develop and maintain [18], [10]. Machine learning-based systems are generally a more recent phenomena [20], but demand expert labeling of relatively huge amounts of data associated with high costs [21]. This paper presents our early experience and preliminary findings in developing a Clinical decision support system (CDSS) in a Norwegian Hospital Trust. The system incorporates a novel, but potentially high-performing, combined algorithm-based approach for text mining of the patient narrative for identifying and classifying allergies of concern for anesthesia during surgery. Our approach is novel in that it utilizes unsupervised machine learning algorithms to analyze large corpora of clinical narratives to automatically build a clinical language model containing words and phrases of which meanings and relative meanings are also learnt. The CDSS also implements rule-based algorithms, and a semi-automatic annotation scheme for efficient and interactive machine learning, which to a large extent eliminates the substantial manual annotation (of clinical narratives) effort necessary for the training of supervised algorithms.

2 Background

While several definitions exist for CDSSs, the pragmatic definition adopted here is that it is any computer program designed to help healthcare professionals to make clinical decisions [17]. Even though NLP-based techniques have been successful in retrieving clinical data from patient narratives [10] , [15] , [20], yet few have utilized its methods to detect allergies recorded in patient narratives [9], [12]. Our approach, however, differs from the methods used by Epstein et al. [9] and Goss et al. [12], whom both used primarily rule-based techniques and tagging of medical concepts by the use of dictionaries in their studies. Fundamental for our research is also that the bulk of relevant research in the field has been carried out on English text only [10].

Until quite recently, the majority of the NLP-based efforts in healthcare have revolved around using rule-based methods to automatically annotate medical concepts in the narrative [18], [20], [27]. Although such expert systems have been shown to be very accurate [27], they are also known to depend on specialized clinical vocabularies or dictionaries which may not be available to all countries and which are also demanding to develop and maintain [10],[18]. English has several readily available resources which can be used to support NLP keyword-driven text mining efforts, such as the Unified Medical Language System (UMLS), Medical Subject Headings (MeSH), and SNOMED CT. For instance, Goss et al. [12] compared five English vocabularies in their ability to represent drug allergies in medical records, and found RxNorm to provide the greatest concept coverage for allergens. Except for MeSH which is currently being translated into Norwegian (18 100 out of 27 500 terms translated at the time of writing), none of these vocabularies (or any other comparable) are available in Norwegian. While we could have developed a custom dictionary containing allergy related terms, such undertakings have been shown to be very resource-demanding and time-consuming. Eriksson

et al. [10], for instance, reported that one person spent half a year to develop the custom Danish ADEs dictionary which they used in their study.

Accurate mapping of allergy related information in clinical records to concepts in controlled vocabularies can be useful for clinical tasks (e.g. search retrieval and decision support), but is very challenging because clinical records exhibit a range of different styles and grammatical structures [10]. Expert systems are likely to have particular challenges with misspellings, compound words and lexical variants [12], and may suffer performance issues if words or phrases that appear in the narrative text are not accounted for in dictionary sources [11]. Achieving high performance using a dictionary-based approach further requires expert domain knowledge [3], i.e. demanding significant involvement of physicians from the target clinical domain for quality assurance of dictionary contents and extracted data [14].

Machine learning-based methods for NLP to automatically annotate medical concepts in the narrative may provide a partial solution to the outlined challenges associated with using controlled clinical vocabularies [14]. By using certain features of the text related to distributional semantics such as e.g. counts and co-occurrence of words, a clinical language model containing concept relevant words and phrases can automatically be built. Provided that a large enough text corpus is available for building and training the model, commonly misspelled relevant words and phrases can also be covered. A major challenge, however, with most machine learning-based methods which have been explored in healthcare is that they typically rely on huge amounts of manually annotated patient narratives for training [27]. As well as causing privacy concerns [14], such labeling of data requires expert knowledge and is both expensive and time-consuming [21]. The cost associated with the labeling process may thus render a fully labeled training set infeasible [29]. While not being without precedent, the idea of combining supervised and unsupervised learning for extracting clinical concepts from the narrative has not yet been widely adopted [14]. Semi-supervised learning utilizing both unsupervised and supervised learning techniques typically uses only a small amount of labeled data in conjunction with a large amount of unlabeled data, and has the potential for learning accuracy while avoiding the cost problems associated with annotation of the narrative [14].

In-part guided by these possibilities and limitations, our approach is different and novel in that it utilizes unsupervised machine learning algorithms to analyze large corpora of clinical narratives to automatically build a clinical language model containing words and phrases of which meanings and relative meanings are also learnt.

3 Method

3.1 System Architecture

The data in this study were obtained from Sørlandet Hospital Trust's enterprise-wide integrated EHR system, which stores data electronically as either structured data or narrative (free text) data. The narrative part of the system contains a copy of all the clinical documents for hospitalized patients admitted to either somatic, psychiatric or radiology departments. Since the system's inception in 1992, 39 570 425 clinical documents (at the time of writing), have been stored in the system across 2298 different document types. Common document types include (but are not restricted to) hospital admission and discharge summaries, progress notes, outpatient clinical notes, medication prescription records, radiology reports, laboratory data reports, surgery notes, anesthesia and intensive care journals, physician referrals, and a range of different specialized forms such as the pre-operative assessment and planning form (POAPF) containing structured data and/or unstructured information.

Figure 1 shows the overall CDSS architecture, which combines EHR data extraction techniques, pre-processing techniques, unsupervised and supervised machine learning techniques with rule-based techniques.

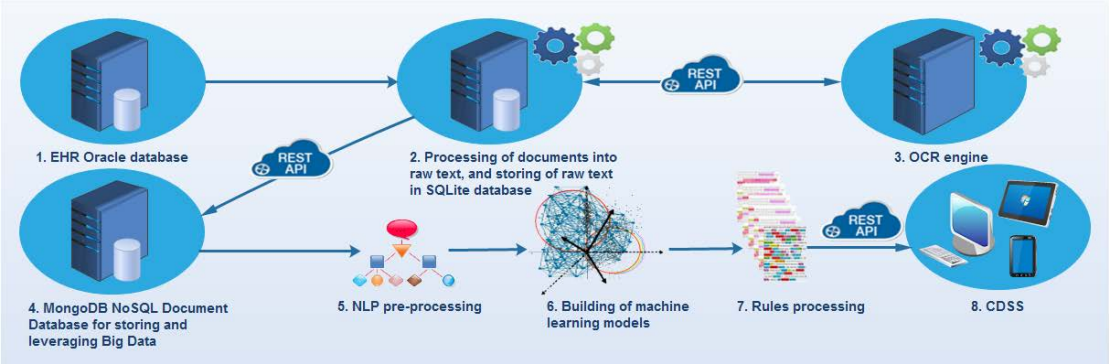

Fig. 1. The overall system architecture covering the whole processing pipeline.

3.2 Study Population and Extraction of EHR Data

9267 incidents of planned patient surgery performed between January 1st 2014 and December 31st 2015 with physician-assigned orthopedic surgical procedure codes (NOMESCO Classification of Surgical Procedures, NCSP codes in chapter A, N, Q and T) were identified through the Sørlandet Hospital Trust's integrated EHR system. The incidents were distributed across 4101 distinctive patients, and constitute the study population of the present study. All incidents had data recorded in corresponding POAPFs detailing risk factors such as e.g. types of allergy, and each of the corresponding patients also had narrative text recorded in the EHR system. We queried the patient POAPFs via NCSP codes, and a total of 863 937 corresponding EHR documents were extracted and processed via NLP techniques. As the NLP techniques and machine learning algorithms only supported processing of plain text, text had to be extracted from different document formats stored in the EHR system's Oracle database for it to be computable. A range of different document formats such as (not exhaustive list) the Portable Document Format (PDF), Jetform, Rich Text Format (RTF), Extensible Markup Language (XML) and also some proprietary EHR vendor document formats were processed. Printed text was also extracted from 146 400 scanned patient Tagged Image File Format (TIFF) documents by performing optical character recognition (OCR) using a commercially available OCR software solution. Further, as no off-the-shelf solutions readily exist to perform the necessary extraction and transformation of data from the EHR system, several customized C# and Python based software solutions were developed as part of the research project to integrate and automate these steps. Finally, the full dataset containing the patients' (in the study population) EHR documents in raw text format was imported into a MongoDB document database for further text mining pre-processing and building of machine learning models.

3.3 Pre-processing of Words and Sentences

The text mining pre-processing pipeline which we implement uses several NLP-methods including lowering case, removal of non-informative terms and punctuations, sentence boundary detection and splitting, and tokenization. Additionally, N-gram (unigram, bigram, and trigram) models necessary for machine learning feature generation are built by performing chunking of tokens [15]. This is partly an iterative process, where domain knowledge of the language and vocabulary being normalized is also helpful in filtering out noise.

3.4 Combining Unsupervised, Supervised and Rule-based Algorithms to Identify and Classify Allergy Concept Related Information

Algorithmic processing starts by first feeding the normalized document collection to the unsupervised learning algorithm for training to facilitate automated modeling of allergy related concepts. By using certain features, all the words and their relations in the patient narratives are mapped, and a clinical language model is built. Several features such as frequency of words, word location, and word co-occurrence related to the usage of terms and phrases are extracted,

and are used to build a vector space model of semantics [8], [28]. Specifically, co-occurrences of terms in the text are used as a metric of similarity, and inferring the word distribution in the set of words the text contains their frequencies are used for document clustering [1]. While some of the features can be automatically extracted, others need human labeling. In all brevity, a small number of allergy concept related example terms and phrases (e.g. different allergen terms or phrases reflecting the occurrence of allergic reactions), typically occurring in the patient narratives are provided as input to the supervised learning algorithm as labels to learn a predictive function (i.e. representing the relation between the features and particular allergy concept related terms and phrases) so that it can cluster or classify any word or phrase in the narratives as belonging to the clinical concept of allergy. Next, the clustering of data is constrained by interactive and iterative user feedback, which allows guiding of the clustering process towards a more precise modeling of the clinical concept of allergy [6]. Manually labeling text is costly and time-consuming to generate. When there is a corpus of manually labeled text available, usually there will also be a much larger amount of unlabeled data available, a resource not utilized by purely supervised training algorithms [4]. The combined unsupervised and supervised learning approach ("guided" semi-supervised clustering), as opposed to only using supervised learning, allows us to utilize both labeled and unlabeled text in the training to create a highly accurate tagger. The result is the grouping of documents into clusters of documents, where allergy concept related words or phrases occur, after a "must-link" and a "cannot-link" based keyword filter [22] using only a small amount of manually tagged text. Finally, 35 rule-based algorithms (see Table 1) are employed to detect allergy relevant information by e.g. combining multiple tagged allergy concept related words and phrases in close proximity to filter documents, identifying relevant windows of context, and paragraph/sentence starts/stops to remove obvious false positives. Since it can be used to group and filter allergy concept relevant documents together, the combined method is used by the CDSS to present physicians with relevant allergy related information either as highlighted text in a reduced narrative document collection or as filtered, classified data (relevant terms and phrases as keywords classified according to allergen type) in a graphical view.

Table 1. A subset of the 35 rules used by the CDSS to detect relevant allergy information.

Rule	Description	Comments
1. Document filtering	Documents must contain allergy concept related words or phrases associated with e.g. "allergy", "allergen" "allergic reaction" or "symptom". While words or phrases of type 1 (strong indicators like e.g. "Anaphylaxis") are allowed to "exist alone" in a document, other types must conform to rules 2 and 3.	Words and phrases can be of type 1) Exist alone, 2) Primary (exist when supported by 1 or 3), 3) Secondary (depend on 2 for existence), and 4) Negation.
2. Window of context	Allergy concept related words or phrases must be located within the same sentence, or if located in adjacent sentences must be in proximity (within a +/- 6 word distance), of other identified allergy concept related words or phrases.	Distance tolerance can easily be adjusted in the system. We experimented with different scopes. We found a six to ten word distance to be optimal.
3. Negation	Detection of positive/negative contexts is handled by checking for the existence of negations in the text.	E.g. "reacts to Penicillin" vs. "does not react to Penicillin".

3.5 Identification of Patient Cases for Training

Patient allergies are continuously documented in the patient narrative as they are identified by physicians, and all patient allergies discovered by reading a patient's narrative are manually registered by physicians as part of the pre-operative assessment and planning routine in the POAPF. At the time of conducting patient surgery, the POAPF is thus considered to contain a correct, updated, and complete picture of known patient allergies and becomes the reference standard for known patient allergies.

In accordance with the method described in Section 3.4, training of the system consists of providing the system with multiple keywords (unigrams, bigrams and trigrams) representing

allergy related terms and phrases occurring in the patient narratives. In order to maximize sensitivity when training the system, we focused on incidents or POAPFs which had one or several allergies registered, and only the last registered POAPF for each of the patients in the subsample was included. The goal when training the algorithms was not to determine whether or not an allergy existed, but to identify keywords used to classify allergy information into categories, given that one or several were there. Thus, there was no need to include allergy-free cases in the training set [30]. To identify random patient cases with recorded allergy reactions, we performed a query on the patient cases' POAPFs, identifying all patients with data registered in the structured data field "Allergies of concern". The result of this query was the identification of 1412 POAPFs (15.2% of all incidents), distributed across 735 (17.9% of all patients) unique patients. Of these unique patient cases, 100 were randomly sampled to be included in the training set.

3.6 Simplified Annotation Scheme for Training

Annotations were done by querying the patients' narratives (the training set corpus) for allergies (types or specific allergens) described in the POAPFs. The annotations were registered, analyzed and systematized by two health professionals (an anesthetist and a nurse with special training) into categories of keywords reflecting that there had either been an "allergy" or an "allergic reaction" of some kind (allergy related terms and phrases, regular expressions consistent with allergic reactions, and symptoms confluent with allergy), together with ten categories constituting different types of reactive allergens. To validate our findings, relevant literature covering the topic was also consulted during this process [13]. In-part inspired by Goss et al. [12], we further abstracted the allergen categories into a smaller set of allergen types which system performance measurements are based on (see Table 2).

Table 2. A subset of the allergen types (as categories of keywords) used to train the system.

Allergen Type	Category	Keywords
	"Allergy" related terms	Allergy(ies), anaphylaxis, side effects, allergy compound words, etc.
	"Symptoms" typic. associated with allergic reactions	eczema, rash, hives, urticaria, allergic rhinitis, asthma, hayfever, etc.
Drug/Contrast Media	Drug	Penicillin, Diural, Sulfa, Morphine, Zocor, etc.
	Contrast Media	contrastfluid, contrast fluid, barium contrast, etc.
Food	Food	milk, lactose, casein, gluten, nuts, egg, shrimps, soy, shellfish, corn, etc.
Environmental	Animal and pet	cat, dog, worms, animal hair, etc.
	Tree, flower and mold	pollen, birch, red alder, molds, etc.

Taken together, the categories and corresponding keywords constitute the concept of "allergy". The categories with corresponding keywords were used by the CDSS as labels to train the supervised learning algorithm on the clinical language model. We developed a graphical user interface module in the CDSS specifically used for training clinical concepts, where the size of predicted words and phrases similar to the abstraction of a word cloud depends on strong or weak concept weighted association. During the training, the scope of clinical concepts can also be constrained by providing the supervised algorithm with discriminating words or phrases [6]. The training of the supervised algorithm is very fast, which allows training to be done iteratively and interactively until the desired level of recall and precision is achieved. The described process represents a semi-automatic and rule-based annotation scheme (see Figure 2) for efficient training, which to a large extent eliminates the traditional expensive and time-consuming annotation of narratives necessary when training and testing supervised algorithms [21].

Categories and corresponding keywords were furthermore adjusted and refined by iteratively assessing the performance of the method on the training set corpus containing a total of 22 821 narrative documents until a sufficient level of recall and precision was achieved.

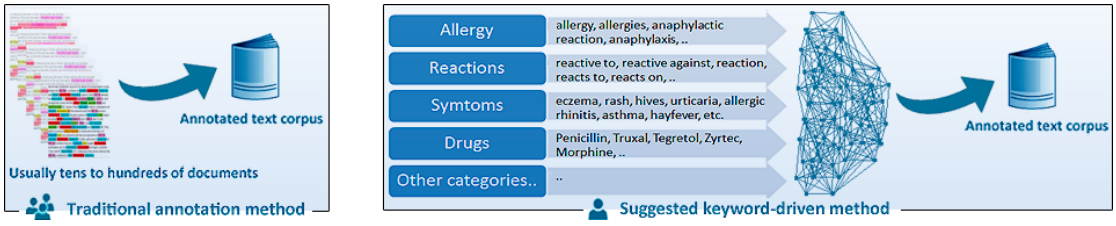

Fig. 2. The system architecture for building medical concepts and annotating text.

With sufficient level of recall and precision is reflected a state where the CDSS after error correction (see the error analysis in Section 4.1 for details) was able to identify as true positives 94.7% (recall), with a precision level of 93.8% and a F-measure of 94.2%, of all allergies recorded in the POAPF for the patients in the training set. Although aimed for initially, 100% recall could not be achieved during the training phase because some of the allergies in the reference standard could not be identified as true positives due to e.g. certain allergy concept related terms and phrases missing from the clinical language model (see Section 5 for a further discussion of this point).

3.7 Evaluation Metrics and Error Analysis

The performance of the CDSS was evaluated according to the common text mining metrics recall, precision, and F-measure (the harmonic mean of recall and precision). Validated cases (true positives) were those patient allergies identified by the CDSS that met the reference standard's definition for allergy. That is, the specific reaction with the patient had been classified as an "allergic" one by trained physicians and had been recorded as an allergy (type or specific allergen) in the POAPF. False positives were those allergies identified by the CDSS but which were not reported in the POAPF, while false negatives were those registered in the POAPF but not identified by the CDSS.

An error analysis was also conducted to better understand the limitations of the trained algorithmic models. Incorrect classifications were categorized based on error cause. Based on the results of the error analysis, evaluation results were recalculated. The rectified results reclassify false positives reported by the CDSS as true positives where the error analysis suggested that the allergy actually is a true positive missing from the reference standard. Because the focus here is on measuring the performance of the combined algorithmic method for text mining of EHRs (and not OCR performance per se), false negatives caused by poor OCR performance were also discarded from the final results.

Definite detection and/or extraction of patient allergies from the narrative is challenging because allergy as a clinical concept has no universally accepted definition or accepted criteria [23]. Allergies are heterogeneous in terms of both their underlying pathophysiology and their clinical manifestations (ranging from mild rashes to life-threatening anaphylaxis) [23]. Physicians may thus have different opinions as to what constitutes an allergic reaction. Important to note here is that there were a number of false positive allergies (28) reported by the CDSS which were discarded due to their uncertain nature. Typically, these reflected reasoning processes focused around symptoms as perhaps being caused by patient allergies, or the reporting of possible adverse effects of drugs, but where there were no clear conclusions of allergies being present. As it would cause unfair bias to count these findings both as false positives (before error correction) or true positives (after error correction), they were discarded. Also other strictly speaking false positive allergens (e.g. drugs) being highlighted because of their proximity with a true positive allergy finding were discarded when there was clear indication of them serving an auxiliary function such as providing supplementary or additional help and support (e.g. other drug mentioned as a substitute because of allergy for a specific drug).

4 Results

Evaluation of system performance was conducted by comparing patient allergies detected by the CDSS against the allergies (types or specific allergens) registered in the corresponding

POAPFs (the reference standard) for 329 randomly subsampled patients from the study population. The patient validation test set had a text corpus consisting of 58 531 documents. We made sure that the patients used in the test set did not overlap with the patients in the training set. No adjustments or refinements to categories and corresponding keywords were allowed in the testing phase.

Table 3. CDSS system performance on processing EHR notes for allergen names and no known allergies after error correction (results before error correction are reported in parenthesis).

Allergen type	Total #	Recall (%)	Precision (%)	F-Measure (%)
Drug/Contrast Media	206 (293)	94.4 (92.1)	81.6 (23.9)	87.5 (37.9)
Food	41 (43)	87 (88)	97.6 (51.2)	92 (64.7)
Environmental	163 (170)	92.3 (92)	95.7 (27.1)	94 (41.8)
Total	410 (506)	92.6 (91.4)	88.8 (27.3)	90.7 (42)

The results achieved for the patient validation test set before and after error correction are reported in Table 3. Overall, recall results for the test set after error correction (recall 92.6 ± 2.7 and precision 88.8 ± 3.1, with 95% confidence interval) corresponded well with the results achieved for the training set. While precision score was somewhat lower ($\Delta p = 5\%$), it was within tolerable limits, indicating that the method for identifying allergies we have used can be generalized at least to the clinical narratives in our study population.

Evaluation of system performance differed somewhat by allergen types, indicating that different types have varying system performance. Both before and after error correction, recall was highest for drug/contrast media (92.1-94.4%) and lowest for food (88-87%), while precision was highest for food (51.2-97.6%) and lowest for drug/contrast media. After error correction we achieved the highest overall F-measure for environmental allergens, and the lowest for drug/contrast media.

4.1 Error Analysis

An analysis of the named entities predicted wrongly revealed that there were multiple reasons for false positive and false negative identification [18] of allergies by the CDSS. These include issues (Table 4 examplifies a subset of these) such as: terms and phrases missing from the clinical language model; omitted or erroneously rendered characters/words in the OCR-scanned documents; problems caused by the look-up window scope being either too strict or loose; missing punctuations, line feeds and/or carriage returns; and allergies missing from the reference standard (the patients' POAPFs).

True allergies missing from the reference standard were of a greater number than originally anticipated, and were the major contributor (71.5%) to the false positives occurring in the results before error correction. Systems based on unsupervised machine learning algorithms for building language models may suffer from the issue of data sparseness [20] because frequency of terms is often used as a feature when building language models. While we had 863 937 EHR documents available in our study, the error analysis revealed data sparseness to be the primary cause (61.3%) for missing words or phrases in the clinical language model causing false negatives. Another important finding was also that current state-of-the-art OCR technology is still far from being perfect (i.e. not comparable to human levels of recognition), and recall suffered greatly (32.3%) from missing or wrongly rendered allergy information in OCR-scanned documents due to e.g. handwritten text and subnormal quality of scanned text. Imprecise boundary detection, e.g. faulty detection of paragraph or sentence start/stop, and too loose look-up window scope, further contributed considerably to the number of false positives (23%) reported by the CDSS. For example, several cases occurred where drugs were falsely tagged as allergens near unrelated clinical descriptions. On the other side, a too strict look-up window scope sometimes also caused true positive allergens (e.g. drugs) in adjacent sentences to miss highlighting.

Table 4. Issues with the CDSS: false positive and negative allergens.

False negative /positive	Cause	Examples/Comments
False negative	N-grams (incl. compound words) not in models (i.e. occurs < 10 times in the study pop. narratives, or because of assoc. problems between words serving the same semantic role)	E.g. "waspallergy", "tree- and grasspollenallergy," "multiallergic," "Xiapex," "Glutenissues," "allergyissues,", "grasspollenallergy."
False negative	Missing or wrong allergy information in OCR-scanned documents	E.g. handwritten text and subnormal quality of scanned EHR documents; "Apicilln" instead of "Apocillin."
False negative	Spelling mistakes and uncommon abbreviations	Spelling mistakes such as e.g. "berch" instead of "birch" and "Pencillin" instead of "Penicillin". Uncommon abbreviations such as e.g. "P. forte" for Paralgin forte.
False negative	Too strict look-up window scope	E.g. two adjacent complementary allergy containing sentences, but where only allergens in one of the sentences are highlighted.
False positive	Allergies missing from the reference standard	Allergies recorded in the patient narratives, while missing in the POAPF forms.
False positive	Negations	"skin prick test shows negative for birch allergy."

5 Discussion

The system implements a novel method combining unsupervised, supervised and rule-based algorithms to identify allergy related concepts in patient narratives. The method demonstrated an overall F-Measure score of 90.7% (after error correction) when tested on patient narratives extracted from the integrated EHR system of a Norwegian hospital trust. Although results are not directly comparable due to e.g. differences in research design and data, the overall recall score of 92.6% achieved after error correction was higher than what has earlier been reported by Epstein et al. [9] (88.61%) and Goss et al. [12] (91%). The precision score of 88.8% is also somewhat higher than what Goss et al. [12] achieved (84.4%), whereas lower than the 99.94% score reported by Epstein et al. [9]. However, while our corpora consists of all the different document types contained in an enterprise-wide hospital integrated EHR system (excluding OCR-scanned documents after error correction), Epstein et al. [9] focused only on data in a perioperative management system (containing data collected from other systems). Based on our findings, we speculate that the comparatively lower precision scores we achieved to some degree can be explained by the much greater variety and complexity of our text corpora compared with what was used in the other study.

Clinical language contains short entries with diverse structures and styles. It is filled with abbreviations, shorthand and acronyms, and meaning is often ambiguous depending on the context. There are often issues with non-conformity with standard grammar, narratives are likely to contain more spelling and type errors than published text, and Norwegian like other Germanic languages is a compound-rich language [10]. While expert systems are known to suffer performance issues if words or phrases that appear in the narrative text are not accounted for in dictionary sources [11, 12], machine learning-based systems have the capacity to automatically create customized conceptual dictionaries for words and phrases in the narratives provided that a large enough text corpus for building and training of models is available [14]. However, parallel to Ramesh et al. [20] who found data sparseness to be the leading type of error (35%) when recognizing medication and ADEs, data sparseness was also found to be the leading cause affecting recall results negatively in our analysis. The error analysis showed that 61.3% percent of the false negatives were caused by N-grams not being included in the models because they did not occur above the threshold frequency of ≥ 10 in the text corpus, or because of association problems between words that serve the same semantic role (paradigmatic associations). As our study includes only about 2.3% of the total

available documents in the hospital EHR, we hypothesize that the problem of data sparseness will resolve with more data in a production setting [20]. During the error analysis, we also discovered that recall suffered substantially (32.3% of the false negatives) because important patient allergy information was lost or rendered erroneously in the data pre-processing step due to 1) subnormal quality of scanned documents and 2) handwritten text in scanned documents. Although great strides have been made recent years to improve handwriting recognition using deep learning techniques [5], the processing power needed to achieve high-speed processing of documents is still not largely available, and made it unsuitable to include as part of this empirically oriented research project.

To some extent, achieving a lower precision score than recall score was an anticipated finding because we were more concerned with achieving high levels of recall than precision when designing the system; in clinical practice losing out on information is considered worse than having a little bit too much. Allergy related concept terms and phrases missing from the reference standard served as the main contributor to lowering the precision score (71.5% of the false positives) before error correction. Although a clinically derived reference or gold standard varies in quality and practicality [2], the reference standard used here is supposed to contain a valid and complete picture of patient allergies at the time of surgery. However, during training and validation of the system, we found that this was not always the case; i.e. some of the pre-assessment and planning forms missed out on one or several true positive patient allergies detected by the CDSS. While it may be tempting to attribute this effect to superior system performance, we believe any such conclusion would be premature and also presuppose further investigations which are beyond the scope of the current paper. As for now, we refer to the challenges associated with defining the clinical concept of allergy in Section 3.7 for a possible explanation. Another factor contributing to lowering precision both before (23% of the false positives) and after error correction was imprecise boundary detection. Missing punctuations, line feeds and/or carriage returns typically occurred in many of the OCR-scanned documents, but also to varying degrees in other documents. Several of the rules which we implement for boundary detection depend on such text markings normally occurring in the narrative text to e.g. navigate the documents, and to identify relevant headings and paragraph/sentence starts/stops. Whenever they lack in documents, highlighting of relevant allergy related terms and phrases has a tendency to become more pronounced, and also less precise. There are other rules implemented to counter this effect to some extent, but these are general in nature as the plurality of structures and styles found in clinical documents makes it unfeasible to implement rules for every situation occurring.

5.1 Limitations and Strengths

There are several limitations to our study. First, the study is conducted on EHR data from only one single hospital in Norway, and the EHR data in our study population only includes 863 937 or about 2.3% of the total available documents in the hospital's integrated EHR. Although a literature review conducted as part of the research confirmed the adequacy of the corpus size compared with other relevant data mining research done in the past [9, 10], [12], we still found it insufficient to build a clinical language model containing all the allergy concept relevant words and phrases used. Second, during the study we discovered that the reference standard used to validate system performance had errors. This is however not an uncommon finding in clinical related research where reference or gold standards may be derived from data collected from clinical practice, and "is only as good as it gets" [2]. We also note that in some cases relevant allergy information was missing from the OCR-scanned documents due to e.g. handwritten text and below average quality of the particular scanned EHR documents. Based on the findings in the error analysis, we performed additional manual annotation of the patient narratives and did further data analysis, rectified the reference standard, and performed recalculations. Nevertheless, our measurement results may still be biased to a certain extent, and should be judged accordingly.

As far as generalizability concerns, our method also has some strengths. First, unlike other similar research done in the clinical domain in the past [10], [12], [15], the study

comprises all the narrative documents (associated with the study population) available in a Norwegian medium sized hospital trust's integrated EHR. Thus, the clinical language model should be more robust to tackle the range of different styles and grammatical structures used to record information in the patient narrative [10]. Second, by using an unsupervised learning algorithm to build a clinical language model and a supervised algorithm to guide the model, we are able to create both a customized while dynamic "dictionary" of allergy concept related words and phrases. Using frequency and co-occurrence of words as features, associations/links are created between words independent of language, which are also automatically updated (e.g. a new drug is quickly assimilated into the model as it starts to occur a small number of times in the patient narratives). The method with some rule-based adjustments should therefore be flexible enough to be transferable to other hospitals, countries and languages. As opposed to traditional rule-based expert systems depending on dictionaries which have to be manually updated continuously, the method suggested here furthermore requires less maintenance, and should thus also be available to smaller healthcare facilities with less available resources to spend. Finally, the method is independent of the focus on allergies specific to this study, purposely used for initial research and testing. The CDSS and the method it utilizes can easily be expanded and adapted to search for other clinical concepts in the patient narrative, something which we plan to explore in upcoming research projects.

6 Conclusion

We have presented our early experience and preliminary findings for a CDSS in development, incorporating a novel algorithm-based approach for text mining of the patient narrative for identifying and classifying allergies of concern for anesthesia during surgery in a Norwegian Hospital Trust. Performance of the system was evaluated using standard text mining metrics. The system is capable of detecting and presenting potentially crucial patient allergy information, with a high degree of recall at an acceptable level of precision, and with a much faster speed than what the physicians in the hospital otherwise routinely would achieve by manually reading through the patient narrative. Thus, the system is able to support physicians with improved clinical decision making and increase safety for those patients undergoing surgery. Based on the promising results for the CDSS so far, plans for implementing the system in the hospital trust are currently being discussed.

References

1. Angelov, P.: Autonomous Learning Systems. John Wiley & Sons (2013).
2. Bellazzi, R., Ferrazzi, F., Sacchi, L.: Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(5), 416-430 (2011)
3. Berge, G. T.: Drivers and Barriers to Structuring Information in Electronic Health Records. In: PACIS 2016 Proceedings (Vol. 2016, paper 18). AIS (2016).
4. Brill, E.: Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In: Proceedings of the third workshop on very large corpora, Vol. 30, pp. 1-13. Association for Computational Linguistics, Somerset, New Jersey (1995)
5. Ciresan D.C., Schmidhuber J. Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification. Preprint arXiv:1309.0261, 1 Sep 2013.
6. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Constrained Clustering: Advances in Algorithms, Theory, and Applications, 4(1), 17-32 (2003)
7. Demner-Fushman, D., Chapman, W. W., McDonald, C. J.: What can natural language processing do for clinical decision support? J. biomed. Inform., 42(5), 760-772 (2009)
8. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19(1), 61-74 (1993)
9. Epstein, R. H., St Jacques, P., Stockin, M., Rothman, B., Ehrenfeld, J. M., Denny, J. C.: Automated identification of drug and food allergies entered using non-standard terminology. J. of the American Med. Inform. Assoc., 20(5), 962-968 (2013)

10. Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J., Brunak, S.: Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J. of the American Med. Inform. Assoc., 20(5), 947-953 (2013)
11. Farkas, R., & Szarvas, G. Automatic construction of rule-based ICD-9-CM coding systems. BMC bioinformatics, 9(Suppl 3), S10 (2008)
12. Goss, F. R., Plasek, J. M., Lau, J. J., Seger, D. L., Chang, F. Y., Zhou, L.: An evaluation of a natural language processing tool for identifying and encoding allergy information in emergency department clinical notes. In: AMIA Annual Symposium Proceedings, pp. 580. American Medical Informatics Association (2014)
13. Hegvik, J. A., Rygnestad, T.: Treatment of serious allergic reactions. Tidsskrift for den Norske lægeforening: tidsskrift for praktisk medicin, ny række 122(10), 1018 (2002)
14. Jonnalagadda, S., Cohen, T., Wu, S., Gonzalez, G.: Enhancing clinical concept extraction with distributional semantics. Journal of biomedical informatics 45(1), 129-140 (2012)
15. Kovačević, A., Dehghan, A., Filannino, M., Keane, J. A., Nenadic, G.: Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J. of the American Med. Inform. Assoc., 20(5), 859-866 (2013)
16. Mandl, K. D., Kohane, I. S.: Escaping the EHR trap—the future of health IT. New England Journal of Medicine, 366(24), 2240-2242 (2012)
17. Musen, M. A., Middleton, B., Greenes, R. A.: Clinical decision-support systems. Biomedical informatics, 643-674. Springer London (2014)
18. Nadkarni, P. M., Ohno-Machado, L., Chapman, W. W.: Natural language processing: an introduction. J. of the American Med. Inform. Assoc. 18(5), 544-551 (2011)
19. Pradhan, H., Stokes, J.: Does Your Electronic Health Record System Introduce Patient Safety Risks? Washington Patient Safety Coalition. (2015)
20. Ramesh, B. P., Belknap, S. M., Li, Z., Frid, N., West, D. P., Yu, H.: Automatically recognizing medication and adverse event information from FDA's adverse event reporting system narratives. JMIR medical informatics 2(1), e10 (2014)
21. Rebholz-Schuhmann, D., Kirsch, H., Couto, F.: Facts from text—is text mining ready to deliver? PLoS Biol, 3(2), e65 (2005)
22. Rowley, J. E.: The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180 (2007)
23. Chafen, J.J.S., Newberry, S.J., Riedl, M.A., Bravata, D.M., Maglione, M., Suttorp, M.J., Sundaram, V., Paige, N.M., Towfigh, A., Hulley, B.J., Shekelle, P.G.: Diagnosing and Managing Common Food Allergies: A Systematic Review. JAMA 303(18), 1848-1856 (2010)
24. Slight, S. P., Seger, D. L., Nanji, K. C., Cho, I., Maniam, N., Dykes, P. C., & Bates, D. W.: Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PloS one, 8(12), e85071 (2013)
25. Sultana, J., Cutroneo, P., Trifirò, G.: Clinical and economic burden of adverse drug reactions. Journal of Pharmacology and Pharmacotherapeutics, 4(5), 73 (2013)
26. The World Allergy Organization (WAO). Drug Allergies (2016). http://www.worldallergy.org/professional/allergic_diseases_center/drugallergy/. Accessed January 17, 2017.
27. Torii, M., Wagholikar, K., Liu, H.: Using machine learning for concept extraction on clinical documents from multiple data sources. J. of the American Med. Inform. Assoc., 18(5), 580-587 (2011)
28. Turney, P. D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1), 141-188 (2010)
29. Wang, H., Qi, J., Zheng, W., Wang, M. Semi-supervised cluster ensemble based on binary similarity matrix. In: Information Management and Engineering (ICIME), 2010. The 2nd IEEE International Conference on, pp. 251-254, IEEE (2010)
30. Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)