Health Informatics Meets eHealth
G. Schreier and D. Hayn (Eds.)

doi:10.3233/978-1-61499-858-7-293

Towards Phenotyping of Clinical Trial Eligibility Criteria

Matthias LÖBE [a,1], Sebastian STÄUBERT [a,b,1], Colleen GOLDBERG [a], Ivonne HAFFNER [c] and Alfred WINTER [a]
[a]*Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Universität Leipzig, Germany*
[b]*Clinical Trial Centre Leipzig (ZKSL), Universität Leipzig, Germany*
[c] *University Cancer Center Leipzig, Germany*

Abstract. Background: Medical plaintext documents contain important facts about patients, but they are rarely available for structured queries. The provision of structured information from natural language texts in addition to the existing structured data can significantly speed up the search for fulfilled inclusion criteria and thus improve the recruitment rate. Objectives: This work is aimed at supporting clinical trial recruitment with text mining techniques to identify suitable subjects in hospitals. Method: Based on the inclusion/exclusion criteria of 5 sample studies and a text corpus consisting of 212 doctor's letters and medical follow-up documentation from a university cancer center, a prototype was developed and technically evaluated using NLP procedures (UIMA) for the extraction of facts from medical free texts. Results: It was found that although the extracted entities are not always correct (precision between 23% and 96%), they provide a decisive indication as to which patient file should be read preferentially. Conclusion: The prototype presented here demonstrates the technical feasibility. In order to find available, lucrative phenotypes, an in-depth evaluation is required.

Keywords. Text Mining, Clinical Trials, Recruitment, Phenotyping, NLP, Apache UIMA, cTAKES

1. Introduction

A common problem in clinical research is slow and meager recruitment of subjects for clinical trials. There are several reasons for this. If the attending physician is also involved as an investigator in an academic trial, the risk of overestimating the number of patients can simply be a matter of enthusiasm. The researchers overestimate the number of eligible patients because they only see the underlying disease, but many subjects are unsuitable for other reasons. Furthermore, the study protocols are becoming more complex, resulting in an increasing number of inclusion and exclusion critiques and more complicated interventions. On the other hand, a large number of patients do not want to take part in clinical trials or will drop out later because the benefit is not perceived. Ultimately, patients are often not addressed at the right time, e. g. if the attending physician does not know the study during the hospital stay [1,2]. As a result, clinical studies take longer, are more expensive, compromise on meaningfulness (lower sample sizes) or are completely discontinued, resulting in the delayed availability of new drugs

[1] Corresponding Authors: Matthias Löbe and Sebastian Stäubert (equally contributed), Institute for Medical Informatics, Statistics and Epidemiology, Universität Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany; Email: matthias.loebe@imise.uni-leipzig.de, sebastian.staeubert@imise.uni-leipzig.de

and therapies. However, computers can support the investigator's work in various ways [3]. To include (or exclude) subjects for a clinical trial, a set of eligibility criteria is defined. Inclusion criteria often include factors that indicate the clinical picture studied in the study, such as demographic data, diagnoses or laboratory values. Exclusion criteria include circumstances that would adversely affect study participation, such as pregnancy, use of addictive substances, pre-existing conditions, or concomitant medication. Not all inclusion/exclusion criteria can be queried automatically on the available data in the EMR, because they are either not recorded at all or not completely (for example if they are not relevant for claims to the health insurance). In this case, cohorts of patients are often formed who meet the criteria that can be queried automatically and study nurses read the patient's file in search of further fulfilled criteria. Manual reading is time-consuming and therefore often just as incomplete. If not all the desired criteria can be taught in this way, they are collected in the context of a pre-screening visit, which is outside the focus of this paper. Our approach is to relieve the study nurse from the burden of reading large amounts of patient records by using natural language processing (NLP) techniques to identify certain characteristics in the patient file. Thus, only files of selected patients have to be investigated. The process of mining patient data to classify subjects into distinct classes by using a set of observable characteristics of an individual as a classifier is often called *phenotyping* [4-6]. Therefore, eligibility criteria can be regarded as phenotypes.

2. Objectives

The aim of the project was to create a prototype tool to support patient recruitment for clinical trials. The most time-consuming process is the search for suitable candidates from the entirety of all patients (screening). Currently, study nurses still have to read large parts of the electronic patient record of all eligible subjects in order to investigate the eligibility or reasons for not being eligible. This work can be facilitated by a computer-assisted system that identifies key words or important phrases and highlights them in the right context. To this end, inclusion and exclusion criteria are to be formalized and medical free text documents are to be analyzed by means of an NLP pipeline. Thus, the study nurses would prefer to read such files in which the relevant phenotypes (e.g. diagnoses or procedures) are at least mentioned, even if the algorithmic recognition of the circumstances does not work perfectly.

3. Methods

The following procedure was chosen for the implementation of the objectives:

1. Creation of a corpus of medical plaintext documents
2. Selection of a set of suitable studies and compilation of inclusion and exclusion criteria as target phenotypes
3. Research of suitable vocabularies to reflect the terminology used in patient records
4. Construction of an NLP pipeline und execution
5. Determination of statistical metrics und evaluation of results

Table 1. Selected studies from the university cancer center

Study name	Study type	Condition	NCT number[1]
ADAPT	phase 2, 3	Breast Cancer	NCT01779206
MATEO	phase 2	Metastatic, Esophagogastric Adenocarcinoma	NCT02128243
VARIANZ	observational	Esophageal Neoplasms, Stomach Neoplasms	NCT02305043
MONALeesa	phase 3	Advanced, Metastatic Breast Cancer	NCT02278120
OLYMPIA	phase 3	Breast Cancer	NCT02032823

The text corpus we used consists of 212 plaintext documents from 101 patients of the Leipzig University Cancer Center (UCCL). The documents were of two types: medical progress documentation and doctor's letters. The planned sample size had to be significantly reduced both in terms of volume and the number of supported document classes, because it turned out that no automatic export of all documents from a patient's record from the hospital information system is currently possible, so that the files had to be extracted manually, which of course was an additional effort. A reference corpus of German medical records for research is currently not available.

Five oncological studies were selected (cf. Table 1), all of which were actively conducted in 2015 in the UCCL and where the recruitment rate was below expectations. The free text inclusion and exclusion criteria (cf. "Eligibility Criteria" in the "Tabular View" of the 5 selected studies in the respective entry at clinicaltrials.gov) were analyzed with regard to medical domain, explicit definitions and range of values. In order to keep the prototype manageable, we focused on the two most common domains, diagnoses and laboratory values.

We were looking for suitable vocabularies that contained concepts with the textual labels corresponding to the phenotypes to be found in the medical documents. Diagnoses in Germany are coded with ICD-10-GM (German modification). ICD-10 provides only a single textual label for each diagnosis code. That's why we also used Alpha-ID. Alpha-ID is a thesaurus of colloquial German-language disease descriptions with a reference to ICD-10 codes. In addition, we wanted to distinguish between mentioning a diagnosis and excluding it. For this purpose, we have compiled a list of words for words with negating meaning based on the Duden Synonym dictionary [7]. Laboratory values usually consist of a name, a value and a measurement unit. Unfortunately, there is no uniform standard for the name of the analytes in Germany, so that a separate list of different sources has been compiled. For units, the coding system Unified Code for Units of Measure (UCUM) was used.

The next step was to plan and implement the NLP pipeline. The pipeline is based on Apache UIMA [8] and Apache cTAKES [9]. Both are open source software. The UIMA framework provides components, interfaces, data representation and patterns for the construction of such pipelines. cTAKES is an implementation of a UIMA pipeline with special components (so-called annotators), which can process clinical documents both rule-based and using machine learning methods. cTAKES supports XML, plaintext and Clinical Document Architecture (CDA) as input formats and is trained on an English clinical documentation body.

A pipeline is usually divided into several modules that are executed in succession. This approach enables the reusability of modules and flexibility in the sequence of sub-steps. Our pipeline consists of 3 sections: pre-processing, document processing and the output unit (see Figure 1). In the preprocessing section, full-text documents are read from

[1] ClinicalTrials.gov Identifier

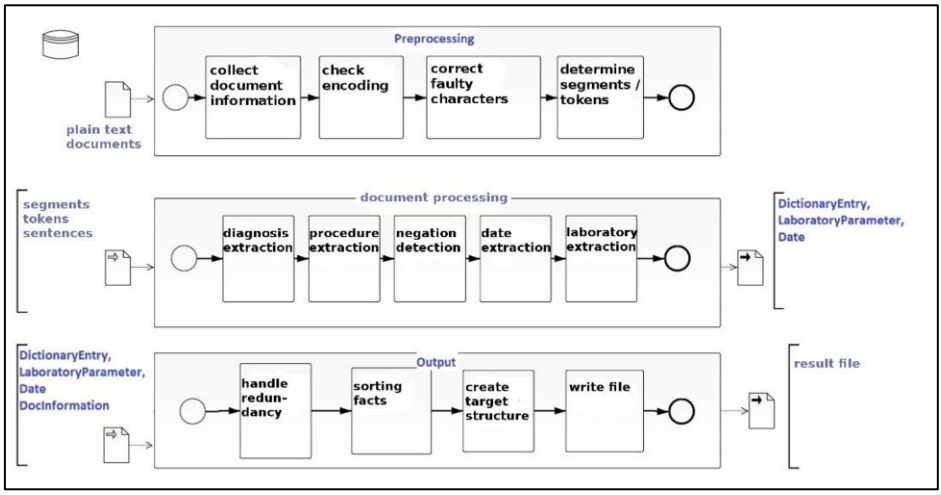

Figure 1. NLP Pipeline with its preprocessing, document processing and output units

a directory and converted from Microsoft XPS (doctor's letter) and Microsoft DOC (follow-up documentation) to plaintext. Then the encoding is checked and segments, sentences and tokens are recognized in the text. After this step, segments, records and tokens exist as objects and form the input for processing. The processing section begins with the annotation of diagnoses. Negations are searched for in the context of detected diagnoses. The negation itself does not form its own annotation type, but only a feature of the diagnosis type. Laboratory parameters are then extracted with value and unit. In the output unit, all collected information is cleaned up of redundancies and assigned to the corresponding patient. The pipeline is completed by writing the results to a CSV file. The NLP pipeline prototype is available via Leipzig Health Atlas [10].

4. Results

The available document body was divided evenly into a training set and an evaluation set. 18 documents were empty and were previously removed. Both sets of documents contained approximately the same number of patients, of doctor's letter and follow-up documentation. The total length of the documents (72,309 vs. 79,295 tokens) was similar in both sets too. Due to data protection restrictions, it was not possible to create a gold standard using annotation by external experts. The laboratory values were annotated by the authors themselves. The list of negative words has also been extended to include other expressions. The evaluation of the results of the NLP pipeline on the evaluation set was also carried out manually.

We used to following performance metrics to measure quality and completeness. All metrics have a value range from 0 to 1, the latter being the perfect value.

$$Precision = \frac{\text{TruePositives}}{\text{TruePositives+FalsePositives}} \quad (1)$$

$$Recall = \frac{\text{TruePositives}}{\text{TruePositives+FalseNegatives}} \quad (2)$$

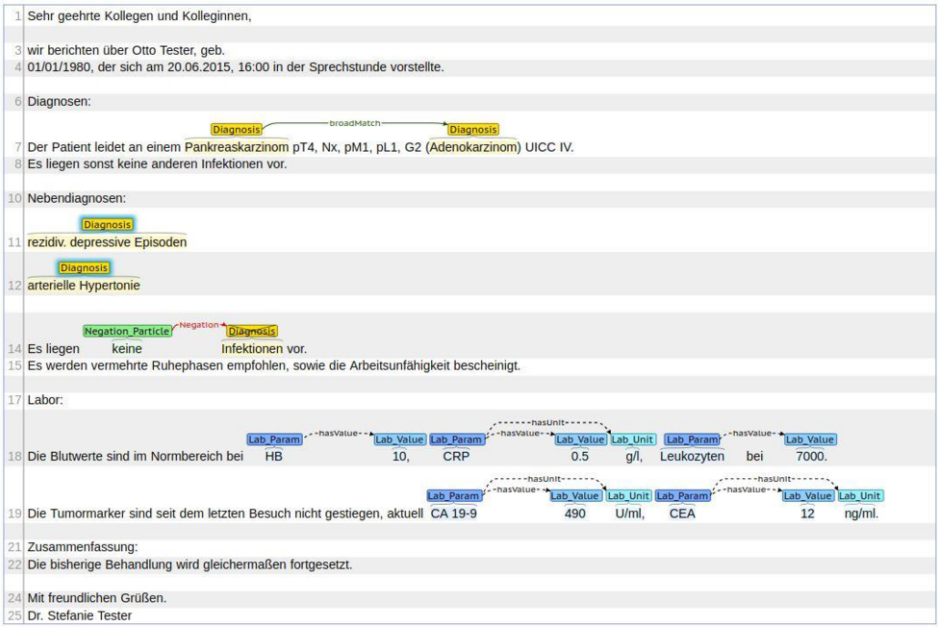

Figure 2. Example of a manual annotation with the web-based text annotation tool brat [11]

$$Accuracy = \frac{\text{TruePositives}+\text{TrueNegatives}}{\text{TruePositives}+\text{TrueNegatives}+\text{FalsePositives}+\text{FalseNegatives}} \quad (3)$$

$$F\ score = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Each of the diagnoses found by the pipeline was compared manually with the documents in the evaluation set. Each diagnosis found was divided into the group "correct annotation" or "incorrect annotation" according to the evaluation rules. A total of 1,105 annotations were found. Of these, 1.071 were correct and 34 were false. This results in a positive predictive value (precision) of 1.071/1.105 = 96.9%. The missing diagnoses were also determined. With 238 missing annotations, the sensitivity (recall) is 1.071/1.309 = 81.8%. The F score is 0,89.

Two lexicons with analyte names and units of measurement are available for the annotation of laboratory analytes. Two user defined parameters have been created in the annotator. These determine the maximum number of tokens between the analyte name and value (x) or analyte name and unit (y). The best results were obtained for x=3 and y=5. With 127 annotations, 64 were wrongly detected. This results in a positive predictive value (precision) of 63/127 = 49.6%. Of the 100 laboratory values in the evaluation set, 63 were correctly annotated, so the sensitivity (recall) is 63%. The F-score is 0.56.

For diagnoses, the system checks whether a statement is positive or negative. This is a binary qualifier. The measure that can be used here to evaluate the outcome of the negation is accuracy. For this purpose, each diagnosis of the evaluation corpus is divided into one of the categories "positive diagnosis" or "negative diagnosis". Negative statements are the TruePositives and positive statements are the TrueNegatives. The

window size to the maximum distance of the negation word from the diagnosis was 3. The accuracy was (35+1.184)/ (35+12+41+1.184) = 95.8%.

5. Discussion and Outlook

The work presented here shows a prototypical but functional pipeline for extracting diagnoses and laboratory findings from clinical documents. The study recruitment use case is particularly suitable for NLP support. Some typical metrics in statistical tests (sensitivity, negative predictive value) play a minor role here. Given the low prevalence of the "eligibility phenotype" of a typical clinical trial in the whole population of patients of a hospital department and the rapid and resource-friendly execution of computer-based analyses compared to manual reading, worthwhile patient records can be selected much more precisely.

The good performance in terms of sensitivity in the detection of diagnoses is explained by the quality of the available dictionary (Alpha-ID) and could be further enhanced by the inclusion of abbreviations [12]. Laboratory values are much more difficult to recognize, not only because of the missing dictionary, but also because of their complex expression (name + value + unit of measurement), which is complicated by typos, abbreviations or normal ranges. The good result for negation detection is mainly explained by the high number of TrueNegatives.

The current approach has a number of limitations. Firstly, only a limited number of documents could be made available, originating from only two different document classes. The 5 clinical trials all originate from the field of oncology and the inclusion and exclusion criteria are not representative. Furthermore, only diagnoses and laboratory values were examined. Diagnoses and laboratory values are typically part of the structured data available in the HIS and could be queried automatically. However, it is not uncommon for certain diagnoses not to be coded for billing reasons, for patients to move from other facilities to a cancer center, or for technical and administrative barriers to programmatic access to existing data in other departments. Other types of medical data would benefit much more from an NLP approach, such as the search for allergies, implants or certain lifestyle-related factors such as smoking or alcohol consumption.

The evaluation was not carried out by clinical personnel, but by computer scientists. To date, no tests has been carried out as to whether the approach accelerates the screening process or improves the recruitment rate by making better use of resources.

In the future, we plan to extend the NLP pipeline to include other kinds of patient data, for instance vital signs, medications, procedures and lifestyle. Another focus is the use of the resulting structured data for retrieval in the research data warehouse of the university hospital (i2b2, tranSMART). In general, more work is needed in the field of German-language text analysis. While there is a large selection of tools, document catalogs and research initiatives for English [13], there are no comparable structures for German.

Finally, electronic health records-driven phenotyping will be a decisive topic for the data integration centers of the German Medical Informatics Initiative, a new 150 million euros funded program to make data from health care available for research across sites [14]. The University of Leipzig is part of the SMITH consortium [15] of the MI Initiative. A methodical use case in SMITH is the Phenotype Pipeline. Among other things, NLP components are also planned here, which follow a similar approach to the one presented here but are much more extensive. In view of the comparability of phenotypes across site

boundaries [16], storage and availability in central open Metadata Repositories (MDR) is recommended.

Acknowledgements

The work of this project was part of the master thesis of Colleen Goldberg [17] and supported by the German Ministry of Education and Research, funding reference: 01KN1102.

References

[1] Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: A systematic review. BMJ open 2012;2(1):e000496.
[2] Thoma A, Farrokhyar F, McKnight L, Bhandari M. Practical tips for surgical research: How to optimize patient recruitment. Canadian journal of surgery. Journal canadien de chirurgie 2010;53(3):205–10.
[3] Köpcke F, Prokosch H-U. Employing Computers for the Recruitment into Clinical Trials: A Comprehensive Systematic Review. J. Med. Internet Res. 2014;16(7):e161.
[4] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association JAMIA 2013;20(1):117–21.
[5] Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association JAMIA 2013;20(e2):e206-11.
[6] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nature reviews. Genetics 2012;13(6):395–405.
[7] Duden K, Wermke M. Der Duden in zwölf Bänden: Das Standardwerk zur deutschen Sprache. Mannheim: Dudenverlag; 2002-2008.
[8] Ferrucci D, Lally A. UIMA: An architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. 2004;10(3-4):327–48.
[9] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association JAMIA 2010;17(5):507–13.
[10] Goldberg C, Löbe M, Stäubert S. NLP4CR Pipeline; Available from: https://www.health-atlas.de/en/method/nlp4cr-pipeline.
[11] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J'i. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012, p. 102–107.
[12] Long W. Extracting diagnoses from discharge summaries. AMIA Annu. Symp. Proc. 2005:470–4.
[13] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association JAMIA 2011;18(5):552–6.
[14] TMF e.V. Medical Informatics Initiative Germany. [February 11, 2018]; Available from: http://www.medizininformatik-initiative.de/en/about-initiative.
[15] Löffler M, Scherag A, Marx G. Smart Medical Information Technology for Healthcare (SMITH). [November 12, 2017]; Available from: http://www.smith.care/.
[16] Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: The eMERGE Network experience. Journal of the American Medical Informatics Association JAMIA 2011;18(4):376–86.
[17] Goldberg C. Unterstützung des Rekrutierungsprozesses für klinisches Studien durch Natural Language Processing auf klinischen Dokumenten. Master Thesis. Leipzig; 2016.