electronic Journal of Health Informatics
www.eJHI.net
2014; Vol 8(1):e6

A Method for Matching Patients to Advanced Prostate Cancer Clinical Trials

Amol S Wagholikar[1], Anthony Nguyen[1], and Maggie Fung[2]

[1] The Australian e-Health Research Centre, CSIRO Computational Informatics, Queensland, Australia
[2] The Australian Prostate Cancer Research Centre, Queensland, Queensland University of Technology, Queensland, Australia

Abstract

Objective: To illustrate a new method for simplifying patient recruitment for advanced prostate cancer clinical trials using natural language processing techniques. Background: The identification of eligible participants for clinical trials is a critical factor to increase patient recruitment rates and an important issue for discovery of new treatment interventions. The current practice of identifying eligible participants is highly constrained due to manual processing of disparate sources of unstructured patient data. Informatics-based approaches can simplify the complex task of evaluating patient's eligibility for clinical trials. We show that an ontology-based approach can address the challenge of matching patients to suitable clinical trials. Methods: The free-text descriptions of clinical trial criteria as well as patient data were analysed. A set of common inclusion and exclusion criteria was identified through consultations with expert clinical trial coordinators. A research prototype was developed using Unstructured Information Management Architecture (UIMA) that identified SNOMED CT concepts in the patient data and clinical trial description. The SNOMED CT concepts model the standard clinical terminology that can be used to represent and evaluate patient's inclusion/exclusion criteria for the clinical trial. Results: Our experimental research prototype describes a semi-automated method for filtering patient records using common clinical trial criteria. Our method simplified the patient recruitment process. The discussion with clinical trial coordinators showed that the efficiency in patient recruitment process measured in terms of information processing time could be improved by 25%. Conclusion: An UIMA-based approach can resolve complexities in patient recruitment for advanced prostate cancer clinical trials.

Keywords: Clinical Trials; MetaMap; Natural Language Processing; UIMA

1 Introduction

There is always a strong need for better cancer treatment options to find a cure, minimize the risks or prolong the life of cancer suffers. Clinical trials offer an opportunity for participation by cancer patients. They may result in discovery of newer medications or treatment options and thus it can improve the healthcare of cancer patients [1]. However, it is widely reported that recruitment of participants for clinical trials is a major factor in delaying new drug discovery and new treatments [2-5]. Operational delays in participant's recruitment have a negative impact on the medication patent exclusivity time, commercial benefits and most importantly availability of treatment options. The healthcare benefits strongly justify the need of any information technology based efforts to minimize the delays in the clinical trial process.

Our research shows that unstructured patient data puts a computational challenge in automating the patient identification process fully or partially. The other challenge in automating the process is that the data usually resides in disparate electronic or paper-based systems at various clinicians' practices. This adds another

layer of complexity in automating the patient identification process. We propose an approach that can address these challenges and provide a pragmatic solution for improving efficiency in patient identification tasks. We describe the application of our approach for advanced prostate cancer clinical trials.

1.1 Background

1.1.1 Issues in patient recruitment

Identification of patients for advanced prostate cancer clinical trials is a challenging task for clinicians as well as clinical data custodians. The process of assessing suitability of the patient against a clinical trial's inclusion and exclusion criteria is highly labor-intensive. The data usually resides in disparate electronic or paper-based systems both at Urologists and Oncologists. As a consequence, the clinicians find it difficult to refer a patient for a clinical trial and in some cases patients may not have the opportunity to participate in a relevant clinical trial. This may put certain limitations on the treatment options of patients with advanced prostate cancer. The inclusion and exclusion criteria for a clinical trial are usually described in a free-text format. It is mostly published in an un-structured format. There is also variability in describing inclusion and exclusion criteria for each clinical trial. The investigators may describe the same criteria differently. The complete automation of the process of evaluating inclusion/exclusion criteria is challenging due to variable and un-structured format of the data. Some advances have recently been made with web-based clinical trial finders [6]; however, this requires manual review through a checklist of common inclusion and exclusion criteria for each trial candidate to assess their eligibility for a clinical trial.

The variability in clinical trial eligibility criteria can be managed by mapping various semantically related clinical terms to a standard set of clinical terminology. We propose that Systematized Nomenclature of Medicine – Clinical Terminology (SNOMED CT) can be used to evaluate inclusion/exclusion criteria [7]. It allows a consistent way to index, store, retrieve, and aggregate clinical trial data. SNOMED CT also helps organizing the content of medical records, reducing the variability in the way data is captured, encoded and used for clinical care of patients and research.

1.1.2 Existing State-of-the-art

There have been efforts made in developing computational methods for automating the clinical trial recruitment process either fully or partially [8]. The level of automation can be measured by the amount of manual efforts required for a clinical trial coordinator to determine a patient suitable for the clinical trial. A fully automated computational method means that all the information processing tasks are automatically by the system giving the final list of matching clinical trials to the clinical trial coordinator. The approaches include rule-based methods [9], query-based methods [9] and Bayesian methods [10]. The final outcome of these methods is to provide decision support to the clinicians engaged in the patient recruitment process. The rule-based methods use inference rules in "if-then" format to identify eligible participants. These methods also integrate the inference rules within the clinical workflow in order to identify any potential candidate for a clinical trial [11, 12]. Some approaches suggest the application automatic reasoner and description logic [13, 14]. The semantic approach proposed in these methods certainly suggests automatic processing of patient and clinical trial data [15]. The query-based methods search a patient's electronic health record (EHR) for identifying specific conditions. These methods mainly propose SQL-based queries to identify potential participants using their EHR stored in the hospital information system. Some of these methods also propose a high level of semantic abstraction and use of RDF and SPARQL languages.

However, our research shows that these methods are not fully evaluated in a real-world setting and these methods are used mainly only in research communities. Our contribution to the research is a UIMA-MetaMap based mechanism to annotate key clinical terminology concepts in the text and a preliminary evaluation for a real-world clinical trial setting. Our mechanism identifies the value of the key inclusion criteria to filter out patient records and thus minimize the delays in identifying participants eligible for clinical trials. The novelty of our approach lies in the software architectural framework which proposes to annotate clinical trial descriptions and patient data for common inclusion criteria using Unstructured Information Management Architecture (UIMA) [15] and Metamap [16, 17]. Metamap is a widely used program for mapping clinical text to concepts in Unified Medical Language System (UMLS). UMLS is an overarching medical language system comprising of file and software that enables many health and biomedical vocabularies and standards to interoperate between various computerized clinical information systems. SNOMED CT is one of the controlled vocabularies of clinical terms embedded in UMLS. This work is the extension of our work reported earlier [18]. The earlier work reported about computational challenges in developing patient matching algorithms for advanced prostate cancer clinical trials.

Common Inclusion Criteria	Rules for extraction
ECOG Performance Status	0-2
PSA	Greater than 26
Diagnosis	Progressive Prostate Cancer
Gleason Score	3+3 or 8
Age	Greater than 50
Sex	Male

Table 1: Common Inclusion Criteria

2 Methods

This research describes an application of natural language processing (NLP) techniques to determine patient eligibility for advanced prostate cancer clinical trials. The clinical data of advanced prostate cancer patients is mostly in free-text form. It resides in radiology report, pathology reports and clinician correspondence. A clinical trial consists of inclusion and exclusion criteria to determine eligibility of participants. We used Java technologies [19] and a text-mining development framework, namely, UIMA to demonstrate our approach. UIMA is an architecture platform used to develop NLP-based pipeline applications. We applied MetaMap to map the clinical text to a SNOMED CT concept. We are proposing an UIMA pipeline application for extracting values of common inclusion and exclusion criteria from the patient free-text.

3 Data Collection

The common criteria for advanced prostate cancer clinical trials were identified through consultations with clinical researchers and clinical trial coordinators. Table 1 shows the common clinical trial criteria for the advanced prostate cancer clinical trials.

The common exclusion criteria are highly specific to a clinical trial. It is difficult and challenging task to determine the common exclusion criteria. Therefore, we show our approach using highly common inclusion criteria. We identified an advanced prostate cancer clinical trial through consultation with the clinical trial coordinator at the Australian Prostate Cancer Research Centre-Queensland (APCRC-Q). Sample patient data were collected from a private urology clinic. The source of the patient data was also determined through consultations with the clinical researchers. The data sources for evaluating the patient eligibility were mostly radiology and pathology investigations.

4 Data Processing

The common set of inclusion criteria shown in Table 1 was used to identify semantically equivalent clinical information from the patient's free text record. That is, clinical concepts used to describe the inclusion/exclusion criteria were used to test for occurrence of clinical concepts identified in the free text. The following steps were performed to process the clinical trials data as well as the patient data:

1. Process the clinical trials criteria text using UIMA MetaMap framework.
2. Identify and extract the best matching SNOMED CT concepts semantically related to the common inclusion criteria as specified in Table 1.
3. Process the patient data using the UIMA-MetaMap wrapper.
4. Identify and extract the common criteria and their values.
5. Match the values of the concepts extracted in step 2 and step 4.

An aggregate score (Pat-Match score) resulting from the inclusion/exclusion criteria matching process can be used to identify or rank participants for possible enrolment in clinical trials. This effectively automates the short-listing of participants eligible for clinical trials, whereby clinicians will be then able to evaluate the participants for a given clinical trial. The UIMA-Metamap wrapper was implemented to analyse the sample data. Figure 1 shows the proposed application pipeline.

5 Results

The system outcomes are shown in Figure 2.

The common clinical inclusion criteria such as ECOG Performance status is annotated by our system. The other criteria such as PSA can be extracted from the pathology investigation reports. The criteria such as age and sex are not usually presented in the patient reports but are easily extracted from patient metadata. Figure 2 shows that our proposed approach is able to annotate key inclusion criteria. The patient records extracted from a private urology clinic were manually analysed to confirm the values of the annotations. For example, the above annotation for ECOG in the clinical trial description was manually checked against the patient data. The annotations helped in navigating large quantity of clinical textual data in much lesser time than a complete manual process.

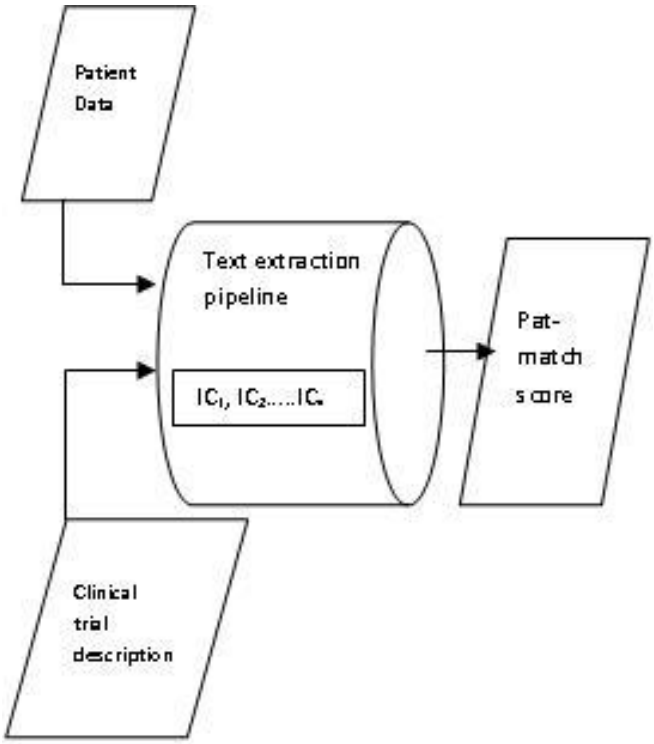

Figure 1: Proposed application pipeline

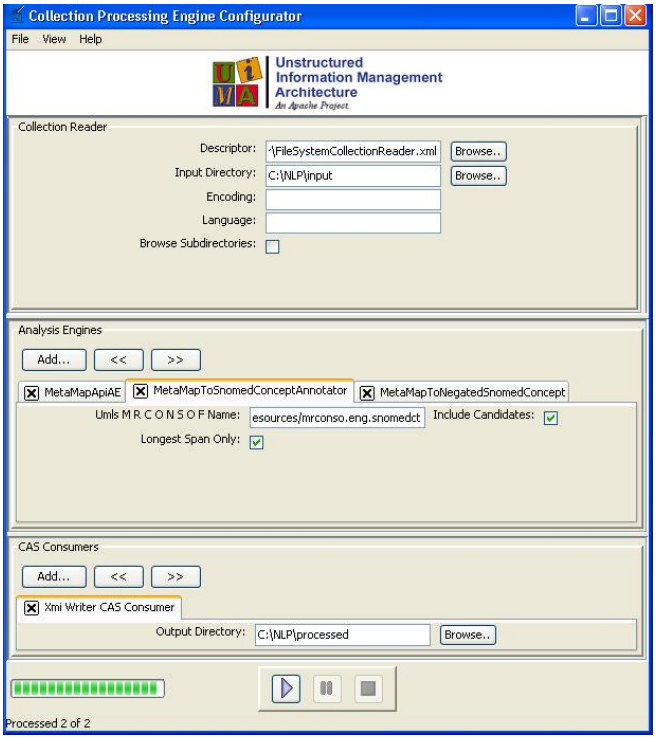

Figure 3: UIMA Collection Processing Engine

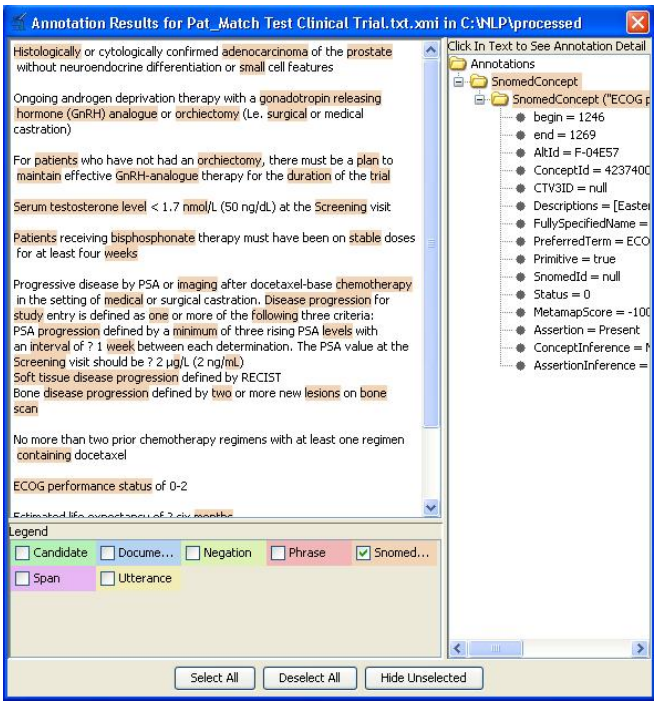

Figure 2: SNOMED CT annotations for sample clinical trial inclusion criteria

6 Discussion

Our early experimental research shows that patient data as well as clinical trial description data can be annotated. The annotations using highly common inclusion criteria for prostate cancer patients are a filtering mechanism to identify potential participants eligible for clinical trials. The evaluation of each and every mandatory criterion is necessary for determining a participant's eligibility. The complete automation of each and every criterion is challenging as each criteria must have an acceptance value for the clinician as well as the pharmaceutical company conducing the trial. Our research shows that there is a need of an individual annotator for every clinical trial inclusion criterion. The annotators combined with a rule or condition in the clinical trial criteria can improve the efficiency in the overall patient evaluation process. The implementation of our application pipeline in UIMA is shown in Figure 3.

The proposed architecture was used to evaluate 84 advanced prostate cancer patients from a multidisciplinary team clinic (MDT) at a tertiary hospital. The data for preliminary evaluation of our technology were collected from the advanced prostate cancer patients. The data collected were in a semi-structured format. The patient data showed the values of common clinical criteria specified in Table 1 in semi-structured format. The patient data were used to identify eligible participants for potential clinical trials being planned at an

adult tertiary hospital. The execution of our proposed mechanism and review by the clinical trial coordinator showed that the sample size of 84 participants can be reduced to potential 5 to 10 participants. These short-listed participants can be further reviewed to evaluate their eligibility. The main contribution of our proposed mechanism is the filtering out ineligible patient records. The filtering mechanism improves the efficiency in the overall patient recruitment process. Specifically, our mechanism is particularly very useful to the clinical trial coordinators.

Our discussion with the clinical trial coordinators at a tertiary hospital showed that the proposed mechanism can improve the efficiency in patient identification. The consultation with the clinical trial coordinators suggests that at least 25 percentage of effort measured in time can be saved if proposed mechanism in integrated within the patient identification and recruitment process.

7 Limitations

The authors report the limitations encountered during this research, which constrained the development of the proposed computational approach. The authors could not get sufficient sample size of advanced prostate cancer patient records to validate the proposed approach. The advanced prostate cancer patient records were not accessible in the required format to build NLP pipeline. The authors also acknowledge lack of quantitative results and quantitative comparison for the proposed approach. The lack of quantitative comparison of the results was mainly due to unavailability of baseline representative sample of advanced prostate cancer patients. The data set for advanced prostate cancer is mainly not available for secondary use due to lack of data registries for advanced prostate cancer in Australia and worldwide. This is a well-known constraint among advanced prostate cancer research community. This work presents a baseline research that can build a foundation for building clinical text processing application pipelines.

8 Conclusion

The initial analysis of the patient data and active clinical trials shows that the evaluation of participants using a fully automated solution is a challenging task. A semi-automated solution can be developed using inputs from the clinicians. The computational approach will aim at identifying a value of the inclusion/exclusion criteria using the SNOMED CT terminology. Our proposed text-mining and clinical terminology-based methods can be used to further evaluate the patient data [20, 21]. Our proposed method can be useful when patient data is coded to SNOMED CT in the source electronic health record. It is anticipated that the proposed baseline solution will simplify the process of participant recruitment for clinical trials.

Acknowledgements

This research is a collaborative project between the Australian Prostate Cancer Research Centre-Queensland(APCRC-Q) at Queensland University Technology's Institute of Health and Biomedical Innovation and the Australian e-Health Research Centre(AEHRC) at CSIRO. The researchers sincerely thank the collaborators of this project for their valuable support.

References

1. Cutting Edge Information, Rep. Streamlining Clinical Trials. 2008: 225 pages.

2. Dugas M, Amler S, Lange, Gerss J, Breil B, Köpcke W. Estimation of patient accrual rates in clinical trials based on routine data from hospital information systems. Methods of Information in Medicine. 2009, 48(3):263–266.

3. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomization. British medical journal Clinical research ed. 1984, 289(6454):1281–1284.

4. Bjornson-Benson WM, Stibolt TB, Manske KA, Zavela KJ, Youtsey DJ, Buist AS. Monitoring recruitment effectiveness and cost in a clinical trial. Controlled Clinical Trials. 1993, 14(2) Suppl.

5. Campbell MK, Snowdon C, Francis D, Elbourne D, Mcdonald A, Knight R. Recruitment to randomised trials: strategies for trial enrolment and participation study. Business. 2007, 11: 48.

6. Patel C, Gomadam K, Khan S, Garg, V. TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records. Web Semantics: Science, Services and Agents on the World Wide Web. 2010, 8(4):342-347.

7. Cornet R, De Keizer N. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making. 2008, 8(1):S2.

8. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, Kershenbaum A, Ma L, Schonberg E, Srinivas K. Matching Patient Records to Clinical Trials Using Ontologies. Lecture Notes in Computer Science. 2007, 4825:816–829.

9. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. International Journal of Medical Informatics. 2011, 80(6):371–388.

10. Afrin LB, Oates JC, Boyd CK, Daniels MS. Leveraging of Open EMR Architecture for Clinical Trial Accrual. AMIA Annual Symposium Proc. 2003, 2003:16–20.

11. Hernandez ME, Carini S, Storey Margaret-Anne, Sim I. An interactive tool for visualizing design heterogeneity in clinical trials. AMIA Annual Symposium Proc, 2008:298–302.

12. Gibbons J, Calinescu R, Harris S, Davies J. Cross trial query system for cancer clinical trials. Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Springer Netherlands. 2007:385-390.

13. Ash N, Ogunyemi O, Zeng Q, Ohno-Machado L. Finding appropriate clinical trials: evaluating encoded eligibility criteria with incomplete data. AMIA Annual Symposium Proc. 2001, 1067-5027:27–31.

14. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Martin Harris, C. Effect of a clinical trial alert system on physician participation in trial recruitment. Archives of Internal Medicine. 2005, 165(19):2272–2277.

15. Patel CO, Weng C. ECRL: an eligibility criteria representation language based on the UMLS Semantic Network. AMIA Annual Symposium Proc, 2008:1084.

16. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004, 10(3-4):327–348.

17. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010, 17(3):229–236.

18. Wagholikar A, Nguyen A, Fung M. Patient Identification for Advanced Prostate Cancer Clinical Trials. The Australian-Canadian Prostate Cancer Research Alliance Symposium Proc. Day Dream Island, Queensland. 2012:46.

19. Oracle, Java Technologies (online), http://www.oracle.com/us/technologies/java/overview/index.html. Accessed February 2012.

20. Nguyen A, Lawley M, Hansen D, Colquist S. A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text. Health Informatics Conference Proc. Canberra, Australia; 2009: 188-193.

21. Lawley MJ. Exploiting Fast Classification of SNOMED CT for Query And Integration Of Health Data. KR-MED, Phoenix, AZ: CEUR-WS. 2008:8–14.

Conflict of Interests

None declared.

Correspondence

Dr Amol Wagholikar
The Australian e-Health Research Centre, Computational Informatics, CSIRO
+61 7 3253 3604
amol.wagholikar@csiro.au