CREU 2015-2016 Final Report:
Analyzing interactions between treatments, side effects, and eligibility criteria in clinical trials

Student researchers: Alexis King
Cristina Diaz

Faculty advisors: Bonnie MacKellar, Ph.D.
Christina Schweikert, Ph.D.

Institution: St. John's University

I) Goals and Purpose

Patients with serious diseases, such as cancer, may be faced with choosing from a number of clinical trials. This is particularly true for pediatric cancer patients, where participation in clinical trials, often multiple clinical trials, is very high [5]. In addition, trials usually have eligibility criteria that prevent patients from participating based on various conditions, including past treatments. This means that participation in one trial may prevent a patient from participating in another trial later on. The goal of this project is to develop and implement an algorithm that can automatically recognize potential conflicts between the treatments in one clinical trial and the eligibility criteria of another clinical trial. Since eligibility criteria are written in free text, they must first be translated into a structured knowledge representation before being processed by an algorithm. Thus, the focus of the student project was twofold: to investigate an approach to parsing eligibility criteria from the literature to determine its applicability to our problem, and to design an algorithm and knowledge representation for conflict analysis.

II) Related Work

The two main approaches to text mining are rule-based approaches and statistical or machine learning approaches [1]. Approaches to parsing eligibility criteria fall into these categories. For example, the EliXR [7] project uses machine learning methods and UMLS concept annotation to generate a semantic representation. Rule-based approaches include the ERGO project [6] which consisted of a semi-automated approach to translating criteria to a formalism developed for the project. Milian [3,4] identified a set of pattern-based rules used in eligibility criteria for breast cancer trials. These patterns were grouped into classifications allowing the ECs to be categorized.

III) Process

There were two components to the student project.

Analysis of rule-based approach to parsing

We tested the patterns against a set of trials that studied treatments for Wilm's tumor, which is a pediatric cancer, comparing our results against Milian's results when using the patterns to match breast cancer clinical trials, in order to determine if this is an appropriate approach in the pediatric cancer domain. Clinical trials for the condition Wilm's tumor were obtained by a search on ClinicalTrials.gov. These trials were retrieved in XML format. We worked with a total of 134 clinical trials. As an exploratory step, a subset of eligibility criteria were analyzed manually by looking for structural patterns in the text that fit the patterns defined by Milian. This process was important because it enabled us to see the different types of text we would encounter, how different types of text fit into patterns, and challenges

that we would need to address in the implementation phase. For the automated process, the free text was preprocessed to separate criteria with common labels such as "AGE:", "DISEASE CHARACTERISTICS:", "INCLUSION CRITERIA:", "EXCLUSION CRITERIA:". Other basic pre-processing included separating sentences, bulleted and numbered lists. A RUTA script, in an UIMA pipeline [2], was developed to recognize the patterns in the eligibility text. The regular expressions for each pattern were implemented as rules in RUTA, which matched to text that fit these expressions.

Design and implementation of the analysis algorithm

Conflicts between trials exist when a treatment administered in one trial is an exclusion criteria of another trial. To identify conflicts between trials, we developed an algorithm which compares the treatments and exclusion criteria of two trials by extracting the list of treatments and exclusion criteria from JSON files we created to represent each trial. These lists consist of UMLS concept identifiers which are associated with each unique treatment or effect. The algorithm compares each treatment identifier with each exclusion criteria identifier, and adds the identifier to a list of conflicting UMLS concepts when there is a match. This algorithm was implemented in Java.

IV) Results and Discussion

We used the full set of Milian's patterns as the basis for the RUTA rules in order to analyze their applicability to a different medical doman (pediatric cancer), comparing results against Milian's results with breast cancer, which were published in [4]. We matched a total of 134 trials to Wilm's tumor eligibility criteria, in comparison with Milian's totals for breast cancer which came to a total of 3,905 trials. We analyzed fewer trials because there are not nearly as many Wilm's Tumor trials as breast cancer trials. The number of eligibility constraint sentences in our research counted to a total of 3,412, whereas Milian's final count is 111,334. Our percentage of matched eligibility criteria is 67%, compared to 71% of matched eligibility criteria in Milian's totals, indicating that the rules work as well in this domain.

We also compared results with more specificity using semantic dimensions attached to the rules. While the numbers are similar for both datasets, there are a few interesting differences. The *time independent status* dimension refers to whether a rule specifies that a condition must be present or absent. The percentage of patterns with this dimension that successfully match ECs is much higher for the Wilm's tumor trials than breast cancer (60% vs 46%). On the other hand, more patterns with the *medical content* dimension matched ECs in breast cancer trials (22% vs 11%). This is not surprising since many of these rules are very specific to breast cancer. However, the results categorized by dimension are similar enough to indicate that these patterns serve as a valid basis for parsing eligibility constraints in the pediatric cancer domain.

With respect to the conflict algorithm, we were able to build JSON files which represented individual trials. The Java program that was developed can extract information from JSON files containing a clinical trial's data and parse it into the different objects that make up its structure. The conflict-finding algorithm can compare two trials and return a list of conflicting UMLS concepts. This code can be used to compute the number of conflicts between two trials. The conflicts between all pairs of trials in a set can be computed in order to generate a suggested ordering of trials to a patient.

V) Future Work

The work on the pattern-based parsing of eligibility criteria will be the basis for automatically generating a knowledge representation suitable for the conflict algorithm. The automated classification of eligibility content in the set of Wilm's Tumor trials can also be applied to other pediatric cancers to get a broader perspective on the type of criteria in this domain. The conflict identification strategy will be embedded in a patient-focused web system to aid in searching for trials and making an informed decision of which trial to participate in first.

VI) Web Links

- The student's blog is located on this web page: https://stjohns.digication.com/creu_research_fall_2015/

VII) Presentations and Publications

- *April 20, 2016.* Poster presentation at St. John's University's Research Month Poster Session, Queens, NY.
- *April 29-30, 2016.* "Pediatric Cancer Clinical Trial Eligibility Criteria Analysis", Cristina Diaz, Alexis King *(Faculty advisors: MacKellar, B and Schweikert, C.)*, poster, Consortium for Computing Science in Colleges – Northeastern, Hamilton College, Hamilton, NY.
- Upcoming poster presentation, "Clinical Trial Eligibility Criteria Analysis for Patient Search", at the 2016 ACM Tapia conference in Austin, TX.

VII) Bibliography

[1] K.B. Cohen, L. Hunter. Getting Started in Text Mining. PLoS Comput Biol 4(1): e20, 2008.
[2] P. Kluegl, M. Toepfer, P.D. Beck, G. Fette, F. Puppe. "UIMA Ruta: Rapid Development of Rule-based Information Extraction Applications," Natural Language Engineering, 22(1),1-40, 2016.
[3] K. Milian, A. Bucur, and F. van Harmelen. Building a library of eligibility criteria to support design of clinical trials. In EKAW, Lecture Notes in Computer Science, pages 327-336. Springer, 2012.
[4] K. Milian, R. Hoekstra, A. Ten Teije, F. van Harmelen, "Patterns of Clinical Trial Eligibility Criteria", Proceedings of the AIME'11 Workshop on Knowledge Representation for Healthcare (KR4HC11), Lecture Notes AI, 2011.
[5] R. A. Schoot, C. H. van Ommen, H. N. Caron, W. J. E. Tissing, M. D. van de Wetering, and SKION Aristocaths supportive care group the Netherlands, "Accrual in supportive care trials in pediatric oncology, a challenge!," *Support. Care Cancer*, 20(12), 3149–53, Dec. 2012.
[6] S. W. Tu et al., "A practical method for transforming free-text eligibility criteria into computable criteria.," Journal of Biomedical Informatics, vol. 44, no. 2, pp. 239–50, Apr. 2011.
[7] C. Weng et al. "EliXR: An Approach to Eligibility Criteria Extraction and Representation." Journal of the American Medical Informatics Association: JAMIA 18.Suppl 1 (2011): i116–i124. PMC. Web. 14 July 2015.