ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-226

Extraction and Prevalence of Structured Data Elements in Free-Text Clinical Trial Eligibility Criteria

Christian GULDEN [a,1], Inge LANDERER [b], Azadeh NASSIRIAN [c], Fatma Betül ALTUN [d], and Johanna ANDRAE [b]
[a] *Medical Informatics, University of Erlangen-Nürnberg, Erlangen, Germany*
[b] *Institute of Medical Biometry and Statistics, Medical Center and Medical Faculty, Albert-Ludwigs-University Freiburg, Freiburg, Germany*
[c] *Dresden University of Technology, Dresden, Germany*
[d] *Medical Informatics Group, University Hospital Frankfurt, Frankfurt, Germany*

Abstract. Understanding the prevalence of structured data elements within clinical trial eligibility criteria is a crucial step for prioritizing integration efforts to supported automated patient recruitment into clinical trials based on electronic health record data. In this work, we extract data elements from 50 clinical trials using a collaborative, crowd-sourced, and iterative method. A total of 1.120 criteria were analyzed, and 204 unique data elements were extracted. The most prevalent elements were diagnosis code, procedure code, and medication code, occurring in 414 (37 %), 112 (10 %), and 91 (8 %) of eligibility criteria respectively. The results of this study may aid in optimizing data integration and documentation efforts in the EHR to support clinical trial eligibility determination.

Keywords. eligibility criteria, clinical trials, data elements, MIRACUM, MI-I

1. Introduction

Clinical trials are the foundation of evidence-based medicine and the gold standard for the advancement of medical knowledge [1]. Thus, it is a crucial task of medical informatics to support and optimize their execution. One aspect is the use of routine data stored within the electronic health record (EHR) for patient recruitment. Eligibility criteria specify the characteristics relevant for a patient to be considered suitable for participation in a clinical trial and consequently also describe the data elements an EHR should contain to support automated recruitment. These criteria are usually expressed as unstructured text in trial protocols, mixing discrete data elements with temporal and conditional modifiers in a semantically complex way [2]. An understanding of the occurrences of common data elements is useful to prioritize data harmonization and integration efforts to support automated patient recruitment. This work identifies these common data elements and describes their distribution, answering the research question of what the most common data elements in clinical trial eligibility criteria are.

[1] Corresponding Author: Christian Gulden, Medical Informatics, Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), Wetterkreuz 13, 91058 Erlangen, Germany; E-Mail: Christian.Gulden@fau.de

2. Methods

We extracted structured data elements from clinical trial eligibility criteria using a collaborative, crowd-sourced approach over 10 participating sites and present the distribution of data elements at criteria granularity. The crowd-sourced extraction process was conducted on an installation of the Atlassian Confluence platform [3], a web-based tool allowing for collaborative editing of documents and pages.

2.1. Selection of trials

The initial step for analysis was the selection of 50 clinical trials whose criteria we set out to analyze. The clinicaltrials.gov registry was used as a source of trials. It provides a search functionality and structured metadata for registered trials, which simplified the trial selection process.

The first filtering step included trials on the condition that the study was ongoing and recruiting at the time and at least one of the 10 analyzing sites was participating in the trial. Next, the list of trials was manually reviewed to generate a final list of 50 trials that span a broad spectrum of medical specialties, allowing us to extract a potentially wide and varied range of data elements.

2.2. Extraction of data elements

Around 5 trials were assigned for further processing to each of the sites and the eligibility criteria were extracted from the selected trials and imported into the collaboration platform. Tables of the same structure were created for both the inclusion and exclusion criteria with columns for the raw free-text criteria, the simplified criteria, an indication of the formalizability of the criteria, the data group, and the extracted data elements. Each row thus represented one (simplified) criteria and its associated data elements. The structure of this table for a select clinical study is shown in Table 1.

Table 1. Excerpt from a table used to extract data elements from the eligibility criteria.

Criterion	Simplified	Formalizable	Data Group	Data Element
Hepatorenal syndrome (type I or II) or screening serum creatinine >2 mg/dL (178 μmol/L)	Hepatorenal syndrome (type I or II)	Yes	Diagnosis	Diagnosis Code
	screening serum creatinine >2 mg/dL (178 μmol/L)	Yes	Laboratory Findings	Creatinine in serum

The criteria were simplified such that one criterion ideally corresponds to one logical data element, for example, the criterion “A mean total bilirubin > ULN and ≤ 3x ULN or an ALP > 5x ULN” was simplified to two distinct statements: “A mean total bilirubin > ULN and ≤ 3x ULN” and “ALP > 5x ULN”. These statements correspond to the data elements “Total Bilirubin in Serum” and “Alkaline Phosphatase” respectively. The

simplification discards the logical relationship between the simplified criteria as they are not relevant for extracting data elements.

Some eligibility criteria encode concepts that are not suitable for automatic eligibility assessments, for example, those relying on a physician's judgment or ones referring to plans of the participants to become pregnant or father children. We considered these not formalizable and no data elements were extracted. Similarly, redundant criteria were excluded from analysis.

The data elements were categorized according to the data inventories for patient identification and recruitment generated by the EHR4CR project [4,5]. If no suitable elements were found the categorization by Luo et al. [6] was used. If neither contained a suitable data element, new ones were introduced.

2.3. Review and harmonization

To ensure quality and consensus between participants, the extraction results of each site were reviewed by one additional site. The comment functionality of Confluence was used to discuss the subjective assessment of formalizability and the chosen data elements. This allowed for an iterative improvement of the extraction process until agreement was reached.

When the reviews were completed, all tables were exported to CSV files and processed by custom scripts. The resulting set was grouped by data elements and their total occurrences were aggregated. This summarized data was imported into Confluence and used to finally harmonize the data elements.

Harmonization entailed the manual identification of spelling errors and semantically identical but differently expressed data elements. For example, no data element for describing the presence of the hepatitis B virus exists in the original data inventory, this caused some sites to introduce this item as "hepatitis b" and others as "hepatitis b virus (hbv)". The harmonization step ensured that these identical elements were mapped to the same identifier.

3. Results

3.1. Selected trials

The initial filtering step reduced the total number of trials registered on clinicaltrials.gov from 277.228 to 416. Manual filtering reduced this to the final 50 trials. A total of 1.120 free-text eligibility criteria were imported from them. The mean number of criteria per study was 22 (Inter-Quartile Range=18, Range=6-49). 130 (12 %) criteria were identified as non-formalizable.

3.2. Prevalence of data elements

A total of 1.625 data elements were extracted, after aggregation and harmonization, 204 of those remained as unique items. Figure 1 shows the relative occurrence of common elements in the analyzed eligibility criteria and their associated data group.

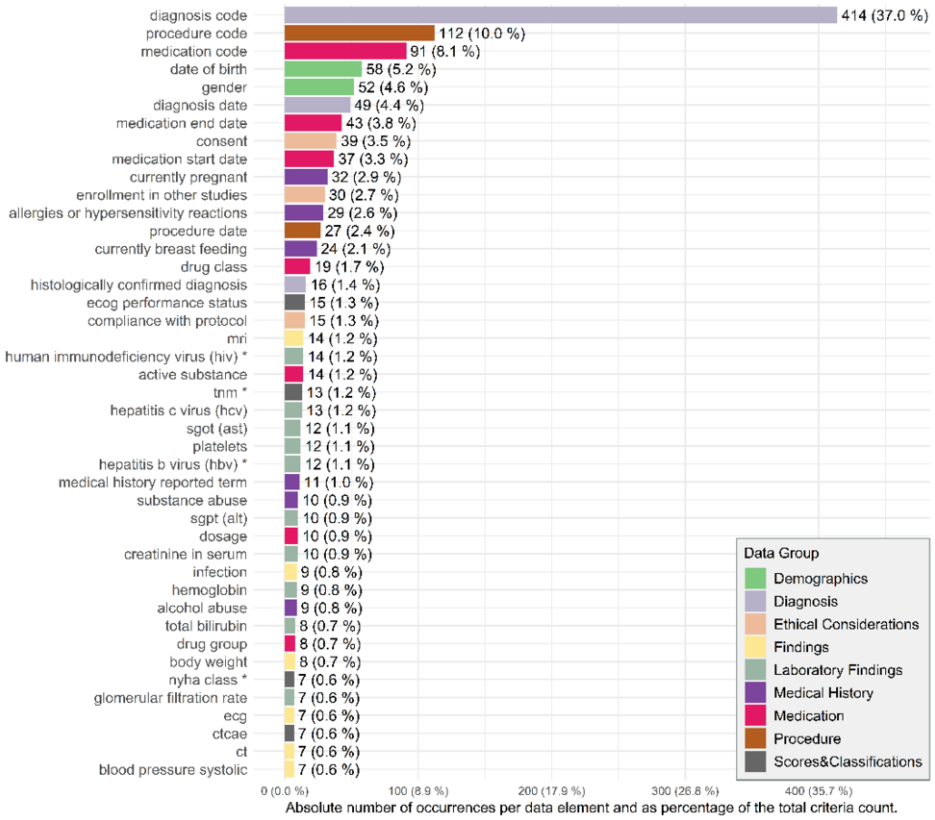

Figure 1. Prevalence of the top 40 structured data elements within the analyzed eligibility criteria. Novel data elements that were not part of the data inventories used are indicated by an asterisk.

4. Discussion

The frequencies of common elements extracted in this work reveals insights into the kind of data required to cover eligibility criteria and allows for the generation of a prioritized list containing the most commonly occurring parameters relevant for patient recruitment.

While some of the common data elements identified in this study may be readily available in the EHR it is often insufficient to determine the eligibility of an individual patient as additional screening may be necessary, given that 130 (12 %) of criteria were identified as non-formalizable. Further, while structured data elements such as diagnosis code and procedure codes may be documented for reimbursement purposes in German hospitals, their quality beyond data completeness [7] remains a subject for future work.

The collaborative method presented in this work is ideal for the given task, as it allows frequent exchange between medical experts and computer scientists. However, the method still requires significant manual labor, which, while benefitting the quality of the results, makes it less scalable than fully automated approaches.

5. Conclusion

Extracting structured data elements from free-text clinical trial eligibility criteria is a challenging task requiring an understanding of both medical domain concepts and the capabilities of formal, computable representations. Tackling this task with an online collaboration tool allowed us to incrementally extract and harmonize data elements using the expertise of participants from 10 university hospitals across Germany. The generated list of data elements and their prevalence is the foundation of future work optimizing the use of EHR data for patient recruitment into clinical trials.

Acknowledgments

This research has been conducted within the MIRACUM consortium. MIRACUM is funded by the German Federal Ministry of Education and Research (BMBF) under the Funding Number FKZ 01ZZ1801A. The present work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (CG).

References

[1] L.M. Friedman, C.D. Furberg, D.L. DeMets, D.M. Reboussin, and C.B. Granger, Fundamentals of Clinical Trials, Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-18539-2.

[2] J. Ross, S. Tu, S. Carini, and I. Sim, Analysis of eligibility criteria complexity in clinical trials., *AMIA Jt. Summits Transl. Sci. Proc.* **2010** (2010) 46–50. http://www.ncbi.nlm.nih.gov/pubmed/21347148.

[3] Confluence - Team Collaboration Software | Atlassian, (n.d.). https://www.atlassian.com/software/confluence (accessed January 14, 2019).

[4] J. Doods, C. Lafitte, N. Ulliac-Sagnes, J. Proeve, F. Botteri, R. Walls, A. Sykes, M. Dugas, and F. Fritz, A European inventory of data elements for patient recruitment, *Stud. Health Technol. Inform.* **210** (2015) 506–510. doi:10.3233/978-1-61499-512-8-506.

[5] P. Bruland, M. McGilchrist, E. Zapletal, D. Acosta, J. Proeve, S. Askin, T. Ganslandt, J. Doods, and M. Dugas, Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting, *BMC Med. Res. Methodol.* **16** (2016) 159. doi:10.1186/s12874-016-0259-3.

[6] Z. Luo, M. Yetisgen-Yildiz, and C. Weng, Dynamic categorization of clinical research eligibility criteria by hierarchical clustering, *J. Biomed. Inform.* **44** (2011) 927–935. doi:10.1016/j.jbi.2011.06.001.

[7] F. Köpcke, B. Trinczek, R.W. Majeed, B. Schreiweis, J. Wenk, T. Leusch, T. Ganslandt, C. Ohmann, B. Bergh, R. Röhrig, M. Dugas, and H.-U. Prokosch, Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence., *BMC Med. Inform. Decis. Mak.* **13** (2013) 37. doi:10.1186/1472-6947-13-37.