Analysing clinical trial outcomes in trial registries: towards creating an ontology of clinical trial outcomes

Anna Koroleva*, Corentin Masson**,
Patrick Paroubek***

*LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay; AMC, University of Amsterdam, Amsterdam, Netherlands
koroleva@limsi.fr
** LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay; AMF (French Financial Market Authority), France
corentin-masson@outlook.fr
*** LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
pap@limsi.fr

Abstract. A clinical trial is a study that evaluates the effects of one or several interventions on a certain population regarding some outcomes - variables that are monitored to assess the impact of the intervention. Trial outcomes are one of the crucial characteristics of a clinical trial. Outcomes are defined by several aspects, such as the name of the variable monitored, measurement tool used, timepoints, analysis metric, aggregation method. We propose to semi-automatically create a structured database of trial outcomes and aspects defining them, that can be used as support for outcome extraction task or to the development of Core Outcome Sets (COS). We propose to use the data from trial registries – online databases containing information about planned and conducted clinical trials, including outcomes. We apply supervised and unsupervised Natural Language Processing techniques to describe and analyse trial outcomes extracted from registries.

1. Introduction

A clinical trial is a study that evaluates the effects of one or several interventions on a certain population regarding some health-related parameters, called outcomes[1]. Outcomes in clinical trials are variables that are monitored to establish the impact of the explored intervention on the health of the population studied. Trial outcomes are one of the crucial characteristics of a clinical trial as they reflect the research question and the explored hypothesis of a trial.

Outcomes are defined by several dimensions. The description of an outcome always comprises a definition of the variable monitored. It can be numerical (temperature), binary (occurrence of an event), or qualitative (quality of life). Some outcomes can be difficult to measure directly, so various measurement tools can be used, such as questionnaires or scales. Outcomes can be measured objectively or subjectively, recorded by a clinician or patient-reported. An outcome is measured several times during a given trial, and these timepoints should be specified for each outcome.

Various analysis metrics can be used for analysing an outcome at the participant level: change from baseline, final value, time to event. At the group level, outcomes are analyzed using some method of aggregation (mean, median, proportion). For the final analysis of the studied population, two main types of analysis can be used: intention-to-treat analysis[2] (all the patients enrolled are analyzed, including those who dropped out of the trial) and per-protocol analysis[3] (only patients who followed the clinical trial instructions are included into the analysis).

In this paper, we propose to create semi-automatically a database containing information about outcomes used in randomized controlled trials (RCTs), related measurement tools, timepoints, analysis metrics and aggregation methods used. Such a database could be used as support for outcome extraction task (Blake and Lucic, 2015; Demner-Fushman et al., 2006; Blake and Lucic, 2016; Summerscales et al., 2009) or could contribute to the development of Core Outcome Sets (COS) – agreed standardised sets of outcomes (and related measurement tools) that should be reported for each specific medical domain[4] to facilitate summarising and practical use of research results (Clarke and Williamson, 2016).

The structure of this paper is as follows: first, we describe the data source that we propose to use to build a database of outcomes. After that, we describe the textual features of outcomes in registries, and we report on our first experiments on building

[1]https://www.who.int/topics/clinical_trials/en/

[2]https://www.nice.org.uk/glossary?letter=i

[3]https://www.nice.org.uk/Glossary?letter=P

[4]http://www.comet-initiative.org/glossary/cos/

a structured database of trial outcomes, using unsupervised or semi-supervised clustering to normalize the outcome descriptions.

2. Data

We propose to use the data from trial registries – online databases containing information about planned and conducted clinical trials, such as studied medical condition, treatment(s), population, outcomes, etc. Information in registries is presented in a structured form; all the registries have data fields for outcomes, usually with division into primary (the most important) and secondary outcomes.

Our starting point is a corpus of 3,938 articles from PubMed Central[5] with the publication type "Randomized controlled trial". For 2,701 articles from this corpus, we were able to find the trial registration number in the text using regular expressions. In some texts, there were several registration numbers mentioned (reporting several trials in one paper, referring to previous trials etc.); for some registration numbers, entries were found in several registries. We searched 13 trial registries and the WHO portal. We downloaded and parsed the data from corresponding trial registries for the obtained registration numbers, and we extracted the fields describing primary and secondary outcomes. If the data for the same trial registration number was available in several registries, we downloaded all the versions, since for a given outcome the wording or the structuring of the description may differ. This work resulted in a corpus of 17,515 outcome descriptions (11,182 unique outcome entries).

3. Textual features of outcomes

The level of detail in the outcome field varies. The field can contain only a noun phrase naming the measured variable (e.g. "*body weight*"), or a free-text description of the outcome and related information elements (e.g. "*The outcome of interest was self-reported medication side effects ever up until the time of interview in 1994, and was recorded as Yes or No.*"). The length of outcome descriptions in our dataset

Primary outcome [1]	Part 1: The primary endpoint of the study is safety assessed by incidence and/or clinically significant changes of a combination of ocular and non-ocular adverse events of single ascending PO doses of STG-001. Systemic adverse reactions will be assessed by physical exam, vital signs, EKG and blood testing (CBC, chemistry, and urinalysis). Ocular adverse reactions, including delayed dark adaptation, will be assessed by ocular exam, visual acuity and color vision testing, intraocular pressure testing, retina exam and a night vision questionnaire. Additional diagnostic testing to monitor ocular adverse events will include ocular coherence tomography, fundus autofluorescence, dark adaptometry and electroretinogram.
Timepoint [1]	Physical Examination: Screening and Day 3 Vitals: Screening, Day1,2,3,4 and Day 8 ECG: Screening, Day, 2, 3, Day 8 Telemetry: Day 1, 2 and 3 Lab assessments: Screening, Day 2, Day 4 and Day 8 Visual Acuity , color vision and intra ocular pressure: Screening, Day 3 and Day 8 FAF will be performed at Screening, Day 3 and Day 8 DA will be performed at Screening, Day -1, Day 3 and Day 8 ERG will be performed at Screening, Day 3 and Day 8 OCT will be performed at Screening, Day 3 and Day 8 AE: Screening, Day 1, 2, 3, 4 and Day 8 Night Vision Questionnaire: Screening, Day 2, Day 3 and Day 8

Figure 1: An outcome entry

[5]https://www.ncbi.nlm.nih.gov/pmc/

ranges from 2 to 6,606 characters (median = 81, mean = 197.4). Shorter outcomes (up to 6 symbols) are often represented by an abbreviation.

Structure of registries differs. Some registries have a separate field for each of trial outcomes, others have only one field where a list of outcomes is recorded. Each item of the list can contain several sentences, describing all the outcome-related information. Some registries have separate fields for outcome timepoints or for outcome measurement tools, while in others all the outcome-related information is recorded in one field.

Figure 1 shown an example of an outcome entry from the Australian New Zealand Clinical Trials registry.

4. Methods

We propose to use Natural Language Processing (NLP) techniques (rules, deep learning and clustering methods) to create a database with structured information on outcomes, based on data extracted from trial registries. We address normalisation of primary outcomes extracted from trial registries and extracting related information.

4.1 Clustering

To assess the variability of outcome descriptions, we used unsupervised clustering. There are several methods of clustering:

1. *Content Mapping methods*: transformation of words to concepts extracted from ontologies (WordNet) to obtain a vector containing each concept representing every document (outcome entry). The vectors can be analysed using Bag-of-words and TF-IDF approaches. Singular value decomposition (SVD) can be applied to reduce the dimensionality to improve clustering with K-means or hierarchical agglomerative clustering (HAC) (cf. Termier et al., 2001). The clustering algorithm can be modified to change the used disance (cosinus, euclidian) to graph distances like Wu-Palmer so that the algorithm can exploit semantic distance to identify clusters.
2. *Word embeddings methods*: language models trained with neural networks, such as word2vec, can be used to obtain word embeddings without using ontologies.
3. *Hybrid methods*: combining classic and word embeddings methods

4.2 Rules

To normalise an outcome entry, we first need to determine if a description contains one outcome or a list of outcomes. Normalising single short outcome descrip-

tions is a rather simple task for which we perform abbreviation expansion by simple regular-expression-based approach using the text of the article related to a registry entry to search for possible expansions of abbreviations.

Lists of outcomes should be divided into single outcomes. It should be taken into account that a list may be present within a description of a single outcome, e.g. a list of measurement tools used, which should not be separated at this stage.

Items describing an outcome may be defined in several sentences, e.g.:

The primary outcome is the change in child problem behavior after intervention. The following instruments will be applied: 1. Strengths and Difficulties Questionnaire (SDQ); 2. Eyberg Child Behaviour Inventory (ECBI).

Although such cases are more difficult for analysis than one-sentence entries, the number of constructions used to describe an outcome and related information elements is limited, allowing to create a set of rules to extract the information.

4.3 Supervised machine learning

Supervised machine learning can be used to extract information from long free-text outcome descriptions, using an annotated corpus to train. For this goal, we annotated 2000 sentences for mentions of primary outcomes, e.g. (outcome is in bold):

The primary outcome was ***the change from baseline in airway resistance (sRaw) at 12 hrs post dose measured by whole body plethysmography***.

We annotated text spans containing all the outcome-related information (outcome and measurement tool name, timepoints, etc.).

We focused on primary outcomes in our annotation efforts as they are the most important information element for our main goal of outcome switching / spin detection (Koroleva and Paroubek, 2018).

Our experiments on applying supervised machine learning to the outcome extraction task are described in detail elsewhere. We compared several approaches and models to choose the best performing method. In brief, the chosen method, proposed by Devlin et al. (2018), consists in pre-training deep bi-directional language representations on a large unannotated corpus and consequently fine-tuning them on a rather small annotated corpus for a supervised task. We compared a number of language models, including BERT (Devlin et al., 2018), BioBERT (Lee et al., 2019) and SciBERT (Beltagy et al., 2019).

5. Results

5.1 Clustering

The first experiment was based on Content Mapping Method. Using POS-Tag techniques, disambiguation and WordNet, we transform each outcome into a list of synsets. In the first approach, we mapped those synsets into vectors using TF-IDF; in the second approach, we mapped this TF-IDF into a smaller matrix using SVD. Results are not satisfying. As expected, intra-cluster variance is decreasing with the number of clusters (cf. Table 1 and 2), but there is no significant drop that would allow us to select an optimal number of clusters.

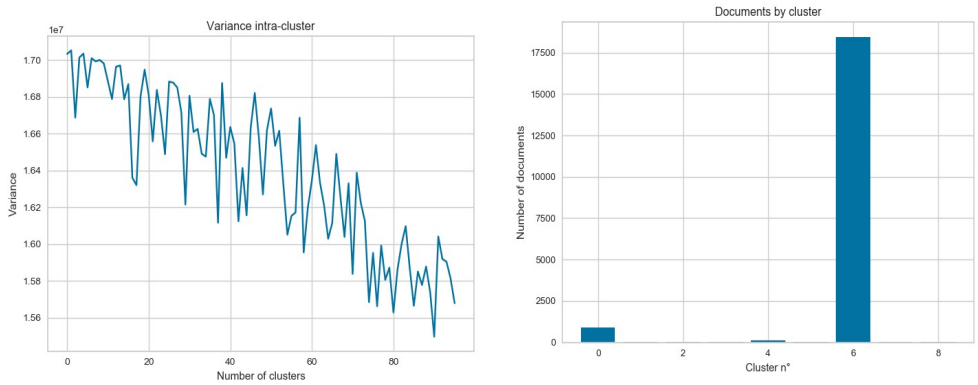

Table 1: TFIDF: Variance depending on number of clusters; cluster sizes

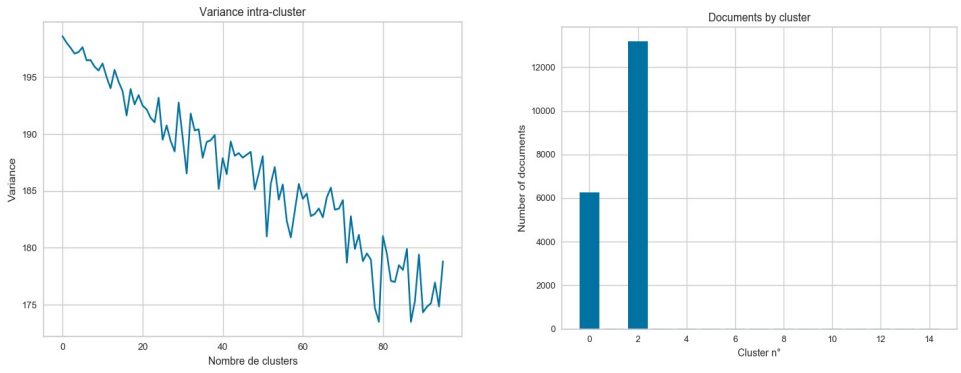

Table 2: SVD: Variance depending on number of clusters; cluster sizes

In our second experiment, we tried to add hypernyms into token vectors to improve the results. Adding those tokens, we hoped to add generality and see clusters merging as we go from hypernym to hypernym. We tried to keep only the first hypernyms along original synsets, or to keep everything, getting a large token vector. Figure 2 shows intra-cluster variance as we add more clusters in K-means. Each curve represent one more hypernym taken into account. The variance decreases as we take more hypernyms. We still can not find an optimal number of clusters.

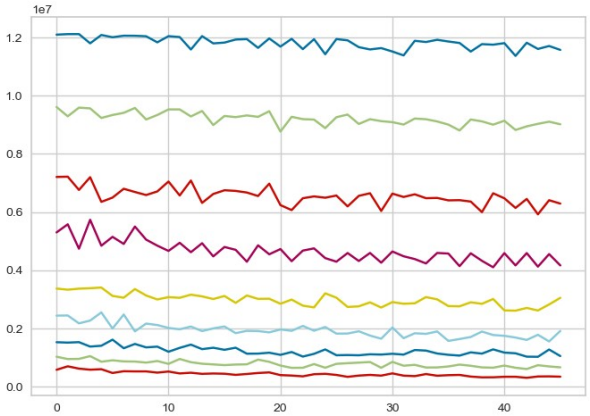

Figure 2: Variance when including hypernyms in vectors

5.2 Supervised machine learning and rules

The best performing deep-learning model (BioBERT fine-tuned for primary outcome extraction) showed the precision of 86.99%, recall of 90.07% and F-measure of 88.42%.

We suggest to use the deep learning algorithm to extract outcome mentions (such as "*the change from baseline in airway resistance (sRaw) at 12 hrs post dose measured by whole body plethysmography*") and to consequently use simple pattern-based rules to extract outcome-related information. For example, measurement tool name can be extracted using a regular expression pattern "*(.*)\s*,?\s*(?:as|which \ w+|that \w+)? (?:measured|assessed|defined|rated|quantified|marked|tested| recorded) (?:with|by|as|using|on|through) (.*)*". Timepoints can be extracted as prepositional phrases containing words with semantics of time (*day*, *month*, *baseline*, etc.). Aggregation method, analysis metrics and type of analysis can be extracted using a dictionary of relevant words (*mean*, *change*, *per-protocol*, etc.).

6. Discussion

Issues encountered

Ones of the encountered difficulties consists in separating coordinated outcomes (e.g. “BMI and aerobic fitness”). Syntactic analysis of such phrases to identify coordinated entities is not likely to be useful because of common errors in parsing incomplete sentences. The task is further complicated by possible presence of coordination within one outcome, which does not need to be divided, and by the need for ellipsis analysis to obtain correct outcome names for some cases (e.g. “local and regional control” which would need to be divided into “local control” and “regional control”). At the current stage, we have not resolved this issue.

Another important issue raised by this work is the absence of uniformity of describing outcomes in registries regarding the length, the details included, and the structure of descriptions (free text vs. noun phrases).

We faced some problems during our clustering experiments. First, outcomes are often represented by short texts containing a high proportion of specific words. Our approach based on WordNet is not efficient for domain-specific documents because a high percentage of words are not present in WordNet. A way to solve this issue would be to use a biological ontology or switch to word embedding methods, potentially using BioBERT representations.

For the experiments using hypernyms to generalize a document, the problem we faced is whether to keep all levels of hypernyms or not. Each word in a document being at a different depth of the WordNet, iteratively taking hypernyms for all words does not result in the same level of generality for each original synset. A word being level deeper than another one will not merge using this technique, thus we have to select the optimal hypernym for each word. TF-IDF & cosine similarity do not give better results. We should try to use Wu-Palmer as distance for the clustering algorithm.

Future work

The current experiments get its inspiration from the Lesk Algorithm and the paper of Scheepers et al. (2018). The idea behind it is to be able to extract the good level of hypernyms without adding noise to the document. For this end, we take the definition of each extracted synset, which will represent sense of the word. For each level of hypernym until reaching the root, we measure the distance between the definition of the current hypernym with the one of the original synset. We take into account each hypernym until the distance goes beyond a certain threshold. We can choose a distance measure, based on Wordnet (Wu-Palmer similarity) or based on word-embeddings (e.g. Glove, Paragram, Bert, Elmo). When the procedure is ac-

complished, we should have a generalized document that might be more adequate to clustering, using TF-IDF, SVD or even on word-embeddings.

7. Conclusion

In this paper, we described the task of creating a structured database of trial outcomes on the basis of data recorded in trial registries. Outcomes extracted from registries vary significantly in terms of their length, level of detail included in the definition of an outcome, and syntactic structure. The absence of uniformity in defining outcomes in registries makes the creation of a structured database a difficult task.

We described our first experiments on clustering of the extracted outcomes and the difficulties encountered. Due to the mentioned absence of uniformity in defining outcomes, finding an optimal number of clusters proved to be difficult in our current experiments.

We outlined some machine learning and rule-based methods that we consider useful for creating a database of outcomes. We propose to extract a complete definition of an outcome from the free-text descriptions in registries using a deep learning method, and to consequently extract information on time points, measurement methods etc. using simple rule-based techniques.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

Beltagy I., Cohan A., Lo K. Scibert: Pretrained contextualized embeddings for scientific text. arXiv:arXiv:1903.10676 2019.

Blake, C., Lucic, A. Automatic endpoint detection to support the systematic review process. J. Biomed. Inform. 2015; 56, 42–56.

Clarke M., Williamson P. R. Core outcome sets and systematic reviews. Systematic Reviews 2016; 5:11. doi:10.1186/s13643-016-0188-6.

Demner-Fushman D., Few B., Hauser S. E., Thoma G. Automatically identifying health outcome information in MEDLINE records. Journal of the American Medical Informatics Association 2006; 13(1):52–60.

Devlin J., Chang M., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 2018; arXiv:1810.04805. URL http://arxiv.org/abs/1810.04805

Ferreira J. C., Patino C.M. Types of outcomes in clinical research, Jornal Brasileiro de Pneumologia 2017; 42-6:5.

Koroleva A., Paroubek P. Automatic detection of inadequate claims in biomedical articles: first steps. Workshop on Curative power of Medical Data (MEDA) 2018. http://doi.org/10.5281/zenodo.1164680

Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746 2019.

Lucic A., Blake C.L. Improving Endpoint Detection to Support Automated Systematic Reviews. AMIA Annu Symp Proc. 2016; 1900–1909.

Scheepers T., Kanoulas E., Gavves E. Improving Word Embedding Compositionality using Lexicographic Definitions. WWW 2018.

Summerscales R, Argamon S, Hupert J, Schwartz A. Identifying treatments, groups, and outcomes in medical abstracts. The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009) 2009, Bloomington, IN, USA: Indiana University

Termier A., Rousset M.-C., Sebag M. Combining Statistics and Semantics for Word and Document Clustering. IJCAI'2001 Workshop on Ontology Learning 2001, Seattle, USA.

Résumé

Un essai clinique est une étude qui évalue les effets d'une ou de plusieurs interventions sur une population donnée en ce qui concerne certains «outcomes» - des variables contrôlées pour évaluer l'impact de l'intervention. Les outcomes sont l'une des caractéristiques essentielles d'un essai clinique. Les résultats sont définis par plusieurs aspects, tels que le nom de la variable contrôlée, l'outil de mesure utilisé, les points horaires, la métrique d'analyse, la méthode d'agrégation. Nous proposons de créer de manière semi-automatique une base de données structurée des outcomes des essais et des aspects les définissant, qui peut être utilisée comme support pour la tâche d'extraction automatique des outcomes ou pour le développement de «Core Outcome Sets» - ensembles de outcomes de base. Nous proposons d'utiliser les données des registres d'essais - des bases de données en ligne contenant des informations sur les essais cliniques, y compris les outcomes. Nous appliquons des techniques de traitement des langues supervisées et non supervisées pour décrire et analyser les outcomes extraits des registres.