Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210183

Automatic Detection of Metadata Errors in a Registry of Clinical Studies Using Shapes Constraint Language (SHACL) Graphs

Daniel KEUCHEL[a] and Nicolai SPICHER[b,1]
[a] *University of Applied Sciences and Arts Dortmund, Dept. of Computer Science, Dortmund, Germany*
[b] *Technical University of Braunschweig and Hannover Medical School, Peter L. Reichertz Institute for Medical Informatics, Braunschweig, Germany*

Abstract. Registries of clinical studies such as ClinicalTrials.gov are an important source of information. However, the process of manually entering metadata is prone to errors which impedes their use and thereby the overall usefulness of the registry. In this work, we propose a generic approach towards detection of errors in the metadata by using the Shapes Constraint Language for defining rule templates covering constraints regarding value type and cardinality. We developed a Python 3 algorithm for the automatic validation of 15 rule instances applied to the whole ClinicalTrials.gov database (355,862 studies; 27th October 2020) resulting in more than 5 million metadata verifications. Our results show a large number of errors in different metadata fields, such as i) missing values, ii) values not coming from a predefined set or iii) wrong cardinalities, can be detected using this approach. Since 2015 approximately 5% of all studies contain one or more errors. In the future, we will apply this technique to other registries and develop more complex rules by focusing on the semantics of the metadata. This could render the possibility of automatically correcting entries, increasing the value of registries of clinical studies.

Keywords. ClinicalTrials.gov, Clinical studies, Metadata, Data quality, Big data

1. Introduction

Clinical trials are the gold standard for evaluating the safety and efficacy of a therapy. Therefore, registries of clinical trials, such as ClinicalTrials.gov, are an important source of information for scientists as well as non-experts. Especially in the current COVID-19 epidemic there is a high interest of the latter group in clinical studies. If they are not reported adequately, this can influence public opinion or health politics negatively.

ClinicalTrials.gov is provided by the U.S. National Library of Medicine and one of the largest registries containing 362,413 (31th December 2020) studies conducted in 291 countries. Each study record contains metadata describing properties of the study, such as the disease of interest, the type of intervention, eligibility criteria of the participants,

[1] Corresponding Author: Nicolai Spicher, Technical University of Braunschweig and Hannover Medical School, Peter L. Reichertz Institute for Medical Informatics, Braunschweig, Germany; E-mail: nicolai.spicher@plri.de

and contact information. For an optimal use of the metadata for manual analysis of studies or automatic integration into systems with patient data (e.g. electronic health records), the metadata needs to be syntactically and semantically correct. Hence, there is a "Guided tutorial" as well as technical support during the entering of the study metadata in the ClinicalTrials.gov database. For example, there are automated validation messages ("Note, "Warning", "Error"), required fields are highlighted, radio buttons are used for Boolean values and drop-down menus for fields with a limited value set.

However, in the past multiple issues regarding the data quality within the studies reported in ClinicalTrials.gov have been shown: Often data is missing such as information on the principal investigators [1], many clinical trials are reported too late [2], or even bogus entries are added, e.g. for promoting unapproved interventions [3]. Additionally, as the metadata that needs to be submitted is not strongly enforced, e.g. by the use of ontologies or MeSH terms, it can only be re-used to a limited extend [4]. Thereby, several works proposed methods for automatically analyzing the data of ClinicalTrials.gov for detecting unusual patterns due to policy changes [5] or errors in metadata [1][4]. However, these works are based on manually crafted rules or data analysis which cannot be re-used or adjusted without effort.

Recently, the Shapes Constraint Language (SHACL) specification has been proposed by the World Wide Web Consortium (W3C) for the validation of graph-based data. SHACL graphs consist of two parts: a "data graph" which contains the data to validate, and a "shapes graph" that contains the properties the data has to fulfill to be a valid graph. It follows the Closed World Assumption and thereby allows to express complex conditions enforced on the data graph, e.g. checking types of values, numeric ranges, string matches, or if a value stems from a subset of valid values. It has already been used successfully for detecting errors in electronic health record clinical models [6] and real-time error detection during the generation of Business Process Model Notation models [7] of healthcare processes [8].

In this work, we aim for using SHACL for detecting errors in the metadata reported in ClinicalTrials.gov. The requirements for submitting a trial to ClinicalTrials.gov are well-documented and publicly available. Based upon that, we propose a generic algorithm for checking requirements fully automatically by defining generic rule templates via SHACL which are instantiated as 15 concrete rules that are checked against the complete ClinicalTrials.gov database.

2. Material and Methods

In the following, we first present the processed data, followed by the rules defined in SHACL and lastly the developed algorithm.

ClinicalTrials.gov data was downloaded as Extensible Markup Language (XML) files with each XML file representing one study entry (355,862 entries; 27th October 2020) with a total file size of 14.1 gigabytes (compressed 1.8 gigabytes).

The data graph was constructed by using results of the BIO2RDF initiative as starting point [9]. They released a dataset describing ClinicalTrials.gov in Resource Description Format (RDF) format which is freely available (https://download.bio2rdf.org/files/release/3/clinicaltrials/clinicaltrials.html). The dataset was acquired and converted to Terse RDF Triple Language (Turtle, (TTL)) Syntax, constituting the data graphs. The shape graphs containing the rules the data has

to follow were manually inserted into the corresponding files. We constructed three types of generic rule templates based on the SHACL core constraint components *sh:hasValue*, *sh:in*, and *sh:maxCount*.

```
#data graph: gender
<http://bio2rdf.org/clinicaltrials_vocabulary:gender>
  rdf:type <http://bio2rdf.org/clinicaltrials_vocabulary:Resource>;
  rdf:type owl:DatatypeProperty;
  rdf:type sh:PropertyShape;
  <http://bio2rdf.org/bio2rdf_vocabulary:identifier> "gender";
  <http://bio2rdf.org/bio2rdf_vocabulary:namespace> "clinicaltrials_vocabulary";
  <http://bio2rdf.org/bio2rdf_vocabulary:uri>
            "http://bio2rdf.org/clinicaltrials_vocabulary:gender";
  dc:identifier "clinicaltrials_vocabulary:gender";
  dc:title "gender"@en;
  void:inDataset <http://bio2rdf.org/clinicaltrials_resource:
            bio2rdf.dataset.clinicaltrials.R3>;
  rdfs:label "gender[clinicaltrials_vocabulary:gender]";
  rdfs:label "gender[clinicaltrials_vocabulary:gender]"@en;
```

Figure 1a. Data graph based on results of BIO2RDF. Line 5 was manually added to include the shape graph shown in Figure 1b).

```
#shacl properties
sh:property [
  sh:path clinicaltrials_shape:Gender;
  sh:hasValue clinicaltrials_shape:Value;
  sh:in (clinicaltrials_shape:Male clinicaltrials_shape:Female clinicaltrials_shape:All)
  ]
.

#shape graph: gender
clinicaltrials_shape:Gender
  rdf:type rdfs:Class;
  rdf:type sh:NodeShape;
  rdfs:label "Gender";
  rdfs:subClassOf rdfs:Resource;
.
#shape graph: female
clinicaltrials_shape:Female
  rdf:type rdfs:Class;
  rdf:type sh:NodeShape;
  rdfs:label "Female";
  rdfs:subClassOf MinimalBeispiel:Gender;
.
```

Figure 1b. Shape graph. Shapes for *"Male"* and *"All"* are not shown but are similar to the shown *"Female"* shape.

The first two constitute value type constraints while the latter constitutes a cardinality constraint. Hence, the rule templates are:

1) A field must contain a value (*sh:hasValue*)
2) A field must contain a value from a set of possible options (*sh:in*)
3) A field occurring multiple times must be filled with a value only once (*sh:maxCount 1*).

We derived 15 rule instances from these three rule templates, respectively:

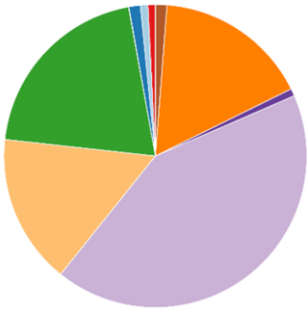
Rules 1-13) The fields *"OrgStudyId", "StudyType", "BriefTitle", "StatusVerifiedDate", "OverallStatus", "PrimaryCompletionDate", "ResponsiblePartyType", "LeadSponsorName", "BriefSummary", "Condition", "Gender", "OverallOfficialRole", "PrimaryOutcomeMeasure"* must contain a value.

Rule 14) The field *"Gender"* must contain a value from the set *{"Male" | "Female" | "All"}*.

Rule 15) The field *"OverallOfficialRole*" - which occurs multiple times depending on the number of study officials - must be filled only once with the value "Principal Investigator".

These rule instances are examples of typical problems with metadata in the ClinicalTrials.gov database that have already been reported by others [1][4]. As an example, Figure 1 shows the TTL file corresponding to rule 14 with the data graph being shown in a) and the shape graph in b). Both are stored in the same TTL file.

A Python 3 program was developed to fully automatically check each of the 15 rules on each study, i.e. XML file. The program reads the XML files using the *ElementTree* API which is part of the Python Standard Library and the TTL files were read using *RDFLib* (https://github.com/RDFLib/rdflib). Subsequently, the rules are extracted from the TTL files using SPARQL. Each rule is then checked on each XML file and the corresponding result (rule [is / is not] violated + details) is stored in a log file.



	Rule 1: *OrgStudyId* – missing value
	Rule 2: *StudyType* – missing value
	Rule 6: *PrimaryCompletionDate* – missing value
	Rule 7: *ResponsiblePartyType* – missing value
	Rule 9: *BriefSummary* – missing value
	Rule 10: *Condition* – missing value
	Rule 11: *Gender* – missing value
	Rule 13: *PrimaryOutcomeMeasure*
	Rule 14: *Gender* – value not part of set {"Male" | "Female" | "All"}
	Rule 15: *OverallOfficialRole* – filled multiple times with "Principal Investigator"

Figure 2a. Distribution of rule violations. Some occur too rarely to be seen.

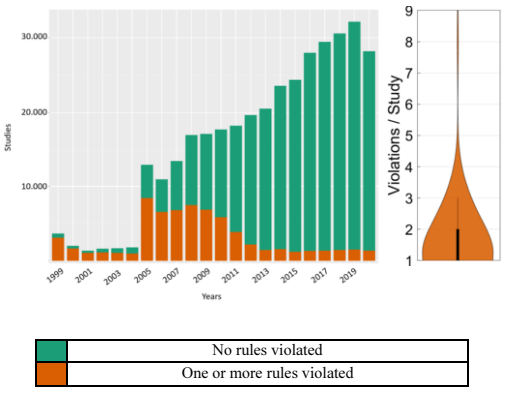


Figure 2b. *Left:* Bar plot of results. Each bar represents all studies registered in ClinicalTrials.gov database for the given year. The green portion shows studies without violated rules and the orange portion reflects studies with one or more violations. It should be noted that data was acquired on 27th October 2020, therefore results for 2020 are incomplete. *Right:* Violin plot showing the distribution of violations per study.

3. Results

The developed algorithm validated the 15 rules on each of the 355,862 XML files on an off-the-shelf computer (Windows 10 64-bit; AMD Ryzen 7 3700X; 16GB RAM) resulting in 5,337,930 verifications. The whole process was performed ten times, resulting in an average run-time of 2,796±191s (mean ± standard deviation) which is approximately 46±3min. In total, 68,438 studies (19.23%) contained one or more violated rules. Log files were analyzed using R and the ggplot2 package for visualization.

Figure 2a shows the distribution of the 108.085 detected violations. Missing values for the *ResponsiblePartyType* (42.27%), *PrimaryCompletionDate* (20.41%), *PrimaryOutcomeMeasure* (15.99%), and cardinality violation for the *OverallOfficialRole* (16.48%) data fields represent most of the errors. Figure 2b shows the number of annual studies with and without rule violations. The large increase in 2005 is due to policy changes and was reported by others as well [5]. The proportion of studies with violations is high in early years and is reduced over time resulting in 4.67% in 2019. Furthermore, Figure 2b displays all studies containing one or more rule violations by means of a rotated kernel density plot showing 1.58±1.02 and a maximum of 9 violations.

4. Discussion

The implemented rules refer to the current state of the ClinicalTrials.gov guidelines, so it should be noted that earlier registered studies may have been subject to different guidelines. Therefore, the high amount of studies containing violations in early years of ClinicalTrials.gov (Figure 2b: orange bars) should be taken with a grain of salt.

We further analyzed the distributions of violations for each year (results not shown). In recent years, they show that the violated cardinality of the *OverallOfficialRole* data field occurred most frequently (2018: 56.44%; 2019: 51.42%; 2020: 50.57%). This metadata field has been analyzed by others as well; Miron et al. reported that 12% of studies did not contain a principal investigator [4] and Chaturvedi et al. reported 17% of "junk" information in the fields [1].

Our work has a clear limitation regarding the analysis of results. Due to the sheer size of detected rule violations, we cannot verify that each detected violation is an actual "True positive" (violation was detected and error is at hand). "False positive" (violation was detected but error is not at hand) and "False negative" (violation was not detected but error is at hand) results could have occurred. However, we visually analyzed the log files and did not observe any problems. Moreover, our results agree to a certain extend with results from literature reporting similar issues [1][4].

5. Conclusion

Registries of clinical trials are an important source of information for both experts and interested non-experts. Although there are validation rules during the submission process, metadata errors are which complicate their re-use [4]. The automatic detection of errors has been proposed by various groups [1][4]. However, these works were mostly focused on a single registry of individual, manually crafted rules. In this work, we provided a more generic approach which allows to manually define rule templates which can be instantiated effortlessly. We fully automatically analyzed more than 5 million metadata entries and detected a high number of rule violations. This detection could be a first step towards giving users more sophisticated feedback during submission or – eventually – automatic correction of violations. In future work, we will i) target more complex rules focusing on the semantics of the metadata and ii) apply the algorithm to other registries.

References

[1] Chaturvedi N, Mehrotra B, Kumari S, et al. Some data quality issues at ClinicalTrials.gov. Trials, 2019;20: 378.
[2] Wise J. Half of all European clinical trials break rules on reporting results within a year. BMJ. 2018 Sep 12;362:k3863.
[3] Turner L. ClinicalTrials.gov, stem cells and 'pay-to-participate' clinical studies. Regen Med. 2017 Sep;12(6):705-719.
[4] Miron L, Gonçalves RS, Musen MA. Obstacles to the reuse of study metadata in ClinicalTrials.gov. Sci Data. 2020 Dec 18;7(1):443.
[5] Zarin DA, Tse T, Ide NC. Trial Registration at ClinicalTrials.gov between May and October 2005. N Engl J Med. 2005 Dec 29;353(26):2779-87.
[6] Martínez-Costa C, Schulz S. Validating EHR clinical models using ontology patterns. J Biomed Inform. 2017 Dec;76:124-137.
[7] Keuchel D, Böckmann B, Spicher N.Semantic verification during BPMN modeling of healthcare processes by integrating Shapes Constraint Language (SHACL) graphs.In: Proceedings of the 64th Conference of the German Association for Medical Informatics,Biometry and Epidemiology;2019.p. 227.
[8] Cornelissen CG, Spicher N, Vollmer T, et al. DIGIVENT - Digitalisiertes, individualisiertes Therapie-unterstützungssystem zur nicht-stationären Indikationsstellung, Einleitung und Kontrolle einer außerklinischen Beatmung für COPD Patienten. In: Proceedings of the 61st Annual Conference of the German Respiratory Society, Leipzig, Germany; 2020.
[9] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Oct;41(5):706-16.