P. Otero et al. (Eds.)

doi:10.3233/SHTI220038

Using an Ontological Representation of Chemotherapy Toxicities for Guiding Information Extraction and Integration from EHRs

Alice Rogier[a,b], Adrien Coulet[a,b], Bastien Rance[a,b,c]

[a] Centre de Recherche des Cordeliers, Inserm, UMRS 1138, Université Sorbonne Paris Cité, Université de Paris, Paris, France
[b] INRIA Paris, Paris, France
[c] Department of Medical Informatics, Hôpital Européen Georges Pompidou, AP-HP, Paris, France

Abstract

Introduction. Chemotherapies against cancers are often interrupted due to severe drug toxicities, reducing treatment opportunities. For this reason, the detection of toxicities and their severity from EHRs is of importance for many downstream applications. However toxicity information is dispersed in various sources in the EHRs, making its extraction challenging.

Methods. We introduce OntoTox, an ontology designed to represent chemotherapy toxicities, its attributes and provenance. We illustrated the interest of OntoTox by integrating toxicities and grading information extracted from three heterogeneous sources: EHR questionnaires, semi-structured tables, and free-text.

Results. We instantiated 53,510, 2,366 and 54,420 toxicities from questionnaires, tables and free-text respectively, and compared the complementarity and redundancy of the three sources.

Discussion. We illustrated with this preliminary study the potential of OntoTox to guide the integration of multiple sources, and identified that the three sources are only moderately overlapping, stressing the need for a common representation.

Keywords:

Knowledge Discovery, Drug-Related Side Effects and Adverse Reactions, Electronic Health Records

Introduction

Chemotherapies against cancers are often interrupted or reduced due to the onset of severe drug toxicities, reducing treatment opportunities for patients, and limiting the therapeutic choices for the clinicians. Chemotherapy regimens involve complex combinations of drugs. Iatrogenic toxicities (adverse drug reactions) are reported and categorized using the Common Terminology Criteria for Adverse Events (CTCAE) [1] according to their severity into five grades, from one being benign side effects, to five the death of the patient.

Toxicities are collected in a very controlled and structured manner in clinical studies, but not in every day care. This makes challenging the secondary use of this information for retrospective studies or for the development of clinical decision support systems.

Today Electronics Health Records (EHRs) offer unprecedented opportunities for using patient data to study variable patient outcomes, including drug response. Information about chemotherapy response and adverse drug reaction occurrences is present, but not directly available in CDWs (Clinical Data Warehouses). This information is present in a variety of sources and formats from structured to unstructured, including structured clinical questionnaires, tables and text of narrative reports.

Several terminologies and ontologies have been developed to facilitate the reporting of adverse drug reactions. The Medical Dictionary for Regulatory Activities (MedDRA) [2] is used to describe adverse events in all types of pathologies. Prior to MedDRA, the World Health Organisation Terminology (WHOART) was used as the reference to code adverse reactions. In oncology, physicians can use the CTCAE to categorize the severity of toxicities, using precise description and laboratory values. Note that the CTCAE shares links to MedDRA. The representation of chemotherapy treatments and adverse reactions has received a great deal of attention over the years. For example, HemOnc [3] was introduced as an ontology on general information relevant to oncology. HemOnc recently adapted its chemotherapy content to the OMOP data model. Meanwhile, there has also been a growing interest in results provided by narrative reports extraction associated with NLP tools [4] and not only in the oncology field [5].

NLP tools have also been used to populate medical ontologies. In particular, Monnin *et al.* [6] created a formal ontology that supports the comparison of pharmacogenomics adverse events heterogeneously reported in the literature.

The correct detection of toxicity events and of their severity is of utmost importance to improve cancer treatment. However, to the best of our knowledge, there is no ontology that aims at describing chemotherapy toxicities and their attributes. Although, such a representation is needed to allow the integration of information from heterogeneous sources.

In the work reported here, our objectives are twofold: (1) we introduce OntoTox, a simple ontology designed to represent chemotherapy toxicities and their attributes, and (2) we demonstrate the interest of OntoTox in a clinical use-case to gather and compare information regarding chemotherapy toxicities found in three types of heterogeneous sources from EHRs: structured clinical questionnaires, tables and free-text from narrative reports (semi-structured and unstructured respectively).

Materials

The Clinical Data Warehouse at the European Hospital Georges Pompidou

The European Hospital Georges Pompidou (or HEGP) is a 700 beds teaching hospital located in Paris. It is specialized in oncology, cardiovascular diseases and emergency medicine. The hospital has deployed since 2008 a clinical data warehouse based on i2b2 [7] integrating virtually all the data generated during everyday care. Among all the data sources, patients in oncology are associated with narrative reports, such as discharge summaries or histology reports, which include free text sections and semi structured tables and clinical questionnaires. Such questionnaires are well structured and some of them collect chemotherapy toxicities associated with their grades. Answers for such questionnaires are collected by caregivers the day before chemotherapy administrations.

Cohort definition

We queried the data warehouse using ICD code C34 ("Malignant neoplasm of bronchus and lung"), and its descendants to identify 3,239 patients treated for lung cancers. Out of the 3,239, we identified 470 patients with at least one narrative report and one questionnaire about chemotherapy toxicity. Out of 470, we randomly selected 330 patients to constitute our studied cohort. We left out 140 for a future evaluation (out of the scope of this work). For each patient, all narrative reports and toxicity questionnaires were extracted, representing 11,819 narrative reports and 71,140 questionnaire items.

Toxicities vocabulary sources

We relied on two reference terminologies to identify toxicity terms in French: the 5th version of CTCAE and WHOART. These two terminologies respectively brought 366 and 1,827 terms.

Methods

In this section, we first introduce OntoTox, our simple ontology to represent chemotherapy toxicities and their attributes. We then present a use case with the extraction of toxicities and their grading from three heterogeneous sources: clinical structured questionnaires, semi-structured tables extracted from narrative reports, free-text from same reports; and their integration following the schema defined by OntoTox.

The OntoTox Ontology a shared conceptualisation

Specification

The aim of the OntoTox ontology is to unify the information extracted toward toxicities and their grade from distinct sources. To this aim, we need OntoTox (1) to guide the normalization of extracted toxicities (2) to encode provenance information.

Conception, implementation and import of external ontologies

The ontology was implemented in OWL using Owlready [8], a module for ontology management in Python and Protégé, the ontology editor [9]**.** OntoTox is composed of 11 classes which are organized around the central class ChemotherapyToxicity. OntoTox also includes 8 object and data properties that serve to qualify toxicities. To this aim, the ChemotherapyToxicity class can be linked to Grade, StartDate and Patient classes through object properties. Grade class has 7 subclasses that are Grade0, Grade1, Grade2, Grade3, Grade4, Grade5 and GradeNull. GradeNull corresponds to the absence of a detected grade, whereas Grade0 denotes the explicit report of a grade 0, *i.e.* the absence of this toxicity, which is commonly found in questionnaires and tables. ChemotherapyToxicity instances can be associated with different data properties to characterize the context of the extraction of the toxicity (*e.g.*, isNeg, isHyp qualify the fact that the toxicity may be extracted as a negated or a hypothetical fact). StartDate and Patient classes are instantiated using documents metadata. To identify the set of classes and properties necessary to represent toxicity information of various provenance, we selected randomly a small set of EHR (*n=20*) that we reviewed to instantiate manually the ontology. We instantiate UMLS and MedDRA concepts with OntoTox toxicities leveraging the PymedTermino library [10]. An example of the instantiation of OntoTox is illustrated in Figure 1. OntoTox is available at https://github.com/TeamHeka/OntoTox.git .

Provenance encoding

OntoTox serves as a global schema for the integration of information from various sources. In such a context, preserving the information regarding the provenance is crucial. We used the PROV-O ontology [11], an ontology dedicated to express provenance in virtually all the realms of science. Here, we consider the chemotherapy toxicity extraction as an entity generated by our extraction algorithm, which itself is an Activity. The input of the extraction algorithm is an entity that is, in the context of our use case, either a free text, a table or a form.

Extraction and integration of heterogeneous toxicities guided by OntoTox

Recognition of entities of interest

In this initial stage of our work, we only identified two types of entities of interest: toxicities and grades. Both are recognized by the same method independently from the considered source. We created a toxicities dictionary based on our toxicities sources terminologies (see Materials), and enriched it with synonyms to 4,038 terms mapped on 835 UMLS concepts thanks to PymedTermino. We relied on QuickUMLS [12], to extract toxicity entities using this dictionary. QuickUMLS is a Python tool leveraging Simstring [13] for approximate string matching. It was applied with a length overlapping criteria, jaccard distance similarity and a 0.9 threshold parameter. In addition to the identification of the toxicity, the dictionary approach also provided us with UMLS Concept Unique Identifier, enabling some normalization of the toxicities. We used a regular expression to detect grades, and normalized them according to their numeric value.

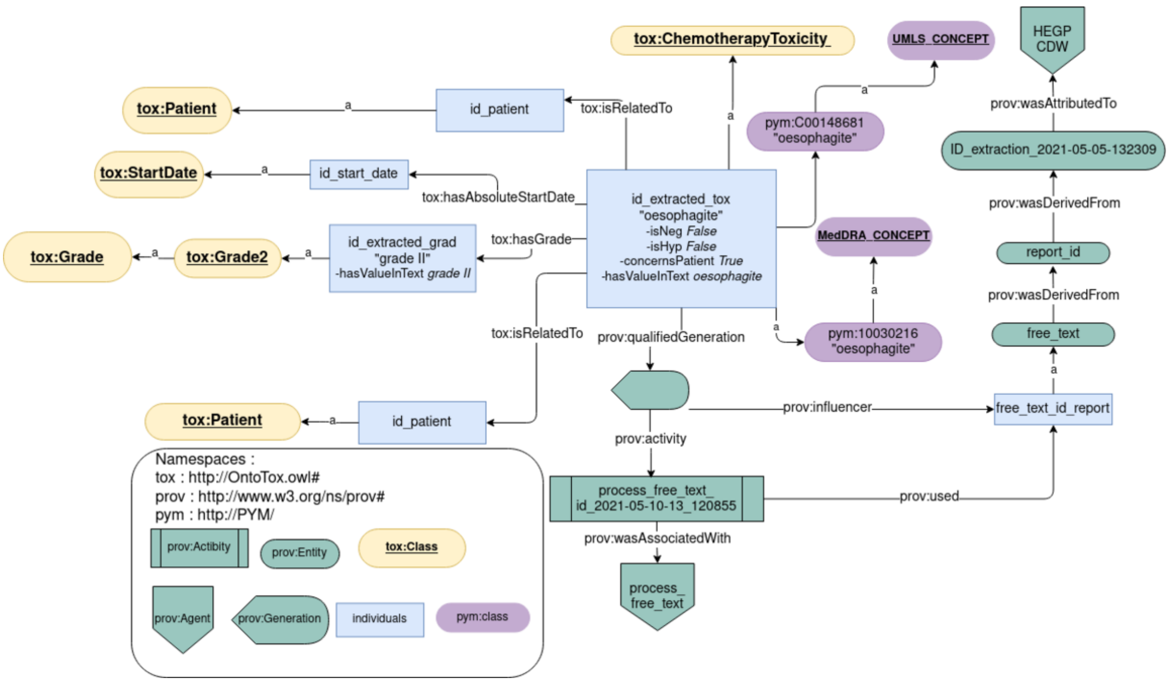

Figure 1 – Main structure of OntoTox, and an example of its instantiation with the following free text: "oesophagite de grade II"

Extracting data from toxicities EHR questionnaires to instantiate OntoTox

We identified two EHR questionnaires directly related to chemotherapy toxicities. The questionnaires are composed of toxicity items (questions), and their values (answers) correspond to grade levels. We parsed the questionnaires and instantiated OntoTox accordingly.

Extracting data from narrative reports to instantiate OntoTox

Free text extraction. We leveraged QuickUMLS and PyMedExt[1], a Python library designed to process clinical text, to process the text. PyMedExt includes annotators to detect polarity (negation, affirmation), the experiencer (patient, family) and hypothesis in French. We leveraged the Stanza dependency parser to link toxicity entities and their grades [14] Stanza is a Python natural language analysis library that uses the Universal Dependencies formalism [15]. We processed all the sentences that contained at least one toxicity entity. The dependency parser provided the syntaxical structure of the sentences.

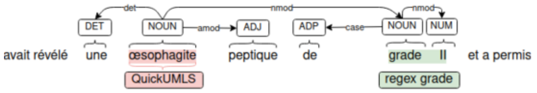

Figure 2 – Link toxicity to its grade thanks to dependency parsing.

We selected recursively all the entities that were under the head of the toxicity entity (see Figure 2). We linked the toxicity and grade if a path exists between the two entities.

Tables extraction. Our method to process free-text is not suitable to semi-structured tables. We identified and extracted all the tables from the original documents. We identified tables related to toxicities by searching for the terms *effets indésirables* (adverse event), *grade*, *liée au traitement* (induced by treatment) *date de début* (starting date) *date de fin* (ending date), in the header of the tables. These terms are found in the default template of the oncology department. We mapped each toxicity to a UMLS concept with QuickUMLS, and parsed the other pieces of information.

Unifying information from the three sources in OntoTox

We instantiated the ontology with inputs from the three sources. Each toxicity extracted from the three sources is thus an instance of the OntoTox ChemotherapyToxicity class. If this toxicity has been found to be related to a grade, a grade subclass is instantiated according to the associated number that is also extracted and normalized. Furthermore, we instantiated Patient and StartDate classes with questionnaire and document metadata. Depending on the source, data property of the ChemotherapyToxicity individual may differ according to specific attributes. For instance, individuals ChemotherapyToxicity instantiated with free text have Boolean data properties covering the information about negation, context and hypothesis.

We compared the distinct toxicities observed per month and per patient to evaluate the contribution of the three sources (see Figure 3). A toxicity is considered to be present in two distinct sources if the two sources share an instance of a specific toxicity within the same year and month.

[1] https://github.com/equipe22/pymedext_core.

Results

Extraction and integration of heterogeneous toxicities guided by OntoTox

Table 1 – Number of instantiation of the OntoTox classes, per data source

OntoTox classes	questionnaires	free text	table
ChemotherapyToxicitiy	**53,510**	**54,420**	**2,366**
Grade	53,510	6,366	400
Grade1	9,981	2,100	87
Grade2	1,832	1,996	52
Grade3	191	817	23
Grade4	19	422	0
Grade5	0	2	1
GradeNull	0	433	85
Grade0	41,487	596	152
Patient	330	330	330
StartDate	1112	2782	372

Unifying information of OntoTox in the three sources

To generate the graph below, we selected UMLS concepts per Patient and per StartDate month grouped by sources by SPARQL querying OnoTox.

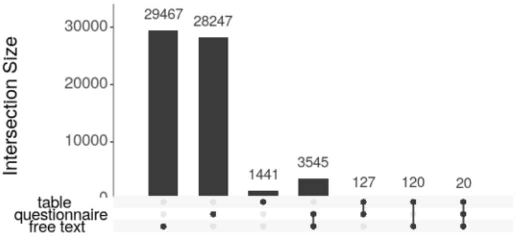

Figure 3 – UMLS concept Per-patient and per-month intersection sets between the three sources

Discussion

Enriching OntoTox

With comparison use-case, we illustrated that classes and properties of OntoTox may be used to guide the integration of chemotherapy toxicities information of various provenances. In this initial effort, we focused on the extraction and integration of the type and grade of toxicities. Of course, more qualification of the toxicity, such as its associated treatment, and its duration are required to provide a view of the potentially available information. We plan on enriching progressively OntoTox, while we develop scripts adapted to various toxicity attributes.

Evaluating the quality of the extraction

Note that this work is not aiming at evaluating or achieving the best performances in recognizing entities related to toxicity, but rather at showing their possible integration in OntoTox. However, we are conscious that our approach, based on dictionary matching and regular expressions, is simple. Advanced approaches for named entity recognition and relation extraction would enable us to improve the quality of our extraction. To assess probable gain in performance and choose most adequate approaches, we plan on manually annotating EHR and its various sources of toxicity. Such annotation could also serve as a training set for supervised or semi-supervised approaches.

Despite our lack of evaluation, our experiments let us think that the use of dependency graphs help in disambiguating the extraction of the grade. Indeed, in oncology, grades may also qualify tumor stages. We avoid such false positives thanks to dependency graphs that verify that grades are connected to toxicity, and not tumors. In particular, additional statistics showed that only 7,809 of the 53,687 recognized grades were linked to a toxicity entity, recognized by our dictionary. Moreover, we think that dependency graphs could be beneficial to detect other qualifiers of interest. For instance, in Figure 2, we note the adjective “peptique” that could help qualify our toxicity. Possibly, we could also combine dependency graphs with a temporal tagger such as HeidelTime [16] to extract StartDate with better precision. Similarly, we could detect drug entities inspiring Lerner and al [17], and instantiate a Drug class.

Instantiating and unifying OntoTox with the three sources

Table 1 shows the number of OntoTox class instantiations by source. We note that free text and questionnaires brought far more information about toxicities than tables. However, toxicities free text extraction are mostly not associated with the grade. This table highlights that questionnaires are structured data. Indeed, a grade instance is always associated with a toxicity instance.

Figure 3 summarizes the contribution of the three sources. We note that there is far more information brought by free text and questionnaires alone. This can be explained by different reasons. The StartDate is a metadata that is not precise enough. Chemotherapy treatments are constituted of alternative few days cure and few weeks inter-cure events. Side effects can occur all along the treatment. However, both questionnaires we selected were used by caregivers to collect toxicity events the day before their chemotherapy cure. Thus, the toxicity events collected in these questionnaires could have occurred in a different date. On the scale of the chemotherapy treatment, the gap between the two dates matters. Furthermore, maybe UMLS concept normalisation is not sensitive enough. For instance, “souffle court” (C0013404 “breathless”) and “difficulté à respirer” (C0013428 “difficulty breathing”) have a different identifier, whereas, in our context, they should be gathered.

Why an ontology?

We chose to create and instantiate an ontology rather than another data model to represent the field of chemotherapy toxicity. One reason for this choice is that we could easily link our ontology to other knowledge models, as PROV-O MedDRA and the UMLS. Furthermore, this enables the further use of a reasoner. For instance, SWRL rules could be implemented to deduce CTCAE criteria knowing the grade and MedDRA concept of a toxicity.

Conclusion

In this article we introduced OntoTox, a simple ontology to represent chemotherapy toxicities. We show that this ontology can guide the integration of information from various data sources. OntoTox is rather small, but aim at being enriched to enable integrating a maximum of information qualifying chemotherapy toxicities and response that can be found in EHRs. OntoTox constitutes the seed of a valuable resource for oncology research and will further serve as a brick of a clinical decision support software.

References

[1] E. Basch, B.B. Reeve, S.A. Mitchell, S.B. Clauser, L.M. Minasian, A.C. Dueck, T.R. Mendoza, J. Hay, T.M. Atkinson, A.P. Abernethy, D.W. Bruner, C.S. Cleeland, J.A. Sloan, R. Chilukuri, P. Baumgartner, A. Denicoff, D. St. Germain, A.M. O'Mara, A. Chen, J. Kelaghan, A.V. Bennett, L. Sit, L. Rogak, A. Barz, D.B. Paul, and D. Schrag, Development of the National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE), *JNCI: Journal of the National Cancer Institute*. **106** (2014).

[2] E.G. Brown, L. Wood, and S. Wood, The Medical Dictionary for Regulatory Activities (MedDRA), *Drug-Safety*. **20** (1999) 109–117.

[3] J.L. Warner, A.J. Cowan, A.C. Hall, and P.C. Yang, HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals, *JOP*. **11** (2015) e336–e350.

[4] S. Datta, E.V. Bernstam, and K. Roberts, A frame semantic overview of NLP-based information extraction for cancer-related EHR notes, *Journal of Biomedical Informatics*. **100** (2019) 103301.

[5] A. Neuraz, I. Lerner, W. Digan, N. Paris, R. Tsopra, A. Rogier, D. Baudoin, K.B. Cohen, A. Burgun, N. Garcelon, B. Rance, A.-H.C.-19 R. Collaboration, and A.-H.C.C. Initiative, Natural Language Processing for Rapid Response to Emergent Diseases: Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic, *Journal of Medical Internet Research*. **22** (2020) e20773.

[6] P. Monnin, J. Legrand, G. Husson, P. Ringot, A. Tchechmedjiev, C. Jonquet, A. Napoli, and A. Coulet, PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison, *BMC Bioinformatics*. **20** (2019) 139.

[7] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *Journal of the American Medical Informatics Association*. **17** (2010) 124–130.

[8] J.-B. Lamy, Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies, *Artificial Intelligence in Medicine*. **80** (2017) 11–28.

[9] The protégé project: a look back and a look forward: AI Matters: Vol 1, No 4, (2021).

[10] J.-B. Lamy, A. Venot, and C. Duclos, PyMedTermino: an open-source generic API for advanced terminology services, (n.d.) 5.

[11] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, PROV-O: The PROV Ontology, (2013).

[12] L. Soldaini, and N. Goharian, QuickUMLS: a fast, unsupervised approach for medical concept extraction, (2016) 4.

[13] N. Okazaki, and J. Tsujii, Simple and efficient algorithm for approximate dictionary matching, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010: pp. 851–859.

[14] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C.D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020: pp. 101–108.

[15] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajicˇ, C.D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, Universal Dependencies v1: A Multilingual Treebank Collection, (2016) 8.

[16] J. Strötgen, and M. Gertz, HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010: pp. 321–324.

[17] I. Lerner, J. Jouffroy, A. Burgun, and A. Neuraz, Learning the grammar of prescription: recurrent neural network grammars for medication information extraction in clinical texts, *ArXiv:2004.11622 [Cs]*. (2020).

Address for correspondence

Correspondences should be addressed to Alice Rogier
alice.rogier@inserm.fr