# *Publishing Medical Context of Neurological Drug Indications as a Knowledge Graph*

*Jinzhou Yang*, Remzi Celebi*, Leoni Bücken, Sarah Chenine, Vincent Emonet, Michel Dumontier*

*Institute of Data Science, Maastricht University, Maastricht, The Netherlands*

**ABSTRACT**

**Motivation:** Understanding the medical context of therapeutic intervention is crucial to its successful use in people. However, this contextual information is not recorded in a machine-readable manner, thereby limiting its use in query answering, clinical decision support, and computational drug discovery. Here, we describe a semi-automated approach to capture drug indications and their medical context. Our approach involves i) a pre-screening of relevant terms using natural language processing tools, and ii) the development and use of Nanobench semantic templates to facilitate data curation with support for term auto-completion from vocabulary standards. We apply our method to create the NeuroDKG, a knowledge graph for Neuropharmaceuti-cal Drugs, which is available as a set of nanopublications.

**Availability:** The NeuroDKG is available at https://github.com/MaastrichtU-IDS/neuro_dkg

**Contact:** j.yang@maastrichtuniversity.nl
remzi.celebi@maastrichtuniversity.nl

## 1 INTRODUCTION

Accurate capture of therapeutic uses of drugs is important for predicting new drug indications and characterizing lead compounds, as well as recommending personalized therapy (Li et al. 2019). The contextual information regarding the therapeutic uses of an approved drug is generally present in the structured product label (SPL). The structured product label is available in free narrative text from databases, such as DailyMed, which can be retrieved in XML format. However, the information about the medical context of the drug indication such as the target group's age, sex, pre-existing conditions and symptoms, is not in fully structured form. Most of the time, physicians take a "one size fits all" approach when prescribing medications as it is not always clear which other drugs may have a more targeted profile for the patient (Nelson et al. 2017). To avoid this issue, the therapeutic intent of specific pharmacological agents should be readily available for physicians as well as scientists to be used for prescription and research, respectively. Developing methods of identifying therapeutic intent will likely improve existing digital health, resulting in more precision medicine.

There have been attempts to extract drug indication information from structured product labels using natural language processing (NLP) (Khare, Wei, and Lu 2014). Yet extracting these indications from unstructured text using NLP technologies still faces challenges in identifying disease mentions in the text and standardizing the disease into standard terms in the ontologies (Khare et al. 2013). Some structured resources such as DrugBank and DrugCentral provide high quality, manual curated information about the therapeutic usage of drugs. However, these resources lack the specific medical context upon which the indication is based on. In addition, because the indications in these resources are partly mined from literature in an automated fashion, their quality is often inconsistent (Moodley et al. 2021). Therefore, there is a need for an accurate representation of therapeutic intent with its medical context in a machine-interpretable manner to facilitate reliable decision making and drug repurposing (Gamberger et al. 2008).

Here, we attempt to address the issue of capturing medical context regarding drug indications by proposing a semi-automated annotation approach. We use a knowledge graph data model that utilizes graph structure to describe the relations between entities (objects, or abstract concepts) in RDF (Resource Description Framework) format. We present NeuroDKG, the Neuropharmaceutical Drugs Knowledge Graph, which contains a set of indications and their medical context related to nervous system drugs curated from DailyMed pages. We populated the curated data to the NeuroDKG using a user interface (UI) template within Nanobench (Kuhn et al. 2021). This template will increase FAIR (Findable, Accessible, Interoperable, and Reusable) sharing of such data as it supports the assignment of persistent and globally unique identifiers and is based on community standards. The template will allow for other researchers to publish the medical context of drug indications and increase the completeness and maintainability of the knowledge graph. In our workflow design, we adhere to FAIR principles (Wilkinson et al. 2016) to ensure the reproducible and

* To whom correspondence should be addressed.

reusable data would benefit the other members of the scientific community.

Our contributions are as follows: i) the development of a data model to capture the medical context of a drug indication, and ii) the development of a Nanobench semantic template to facilitate the manual curation of a drug indication and its medical context, iii) the development of a FAIR knowledge graph for neuropharmaceuticals, termed NeuroDKG.

## 2 METHOD

The approach consists of two parts. First, automated tools are used to annotate the target text with key terms. Second, the content is manually curated using a semantic template with autocomplete support to create a rich, machine-readable description of drug indications and their medical context.

### 2.1 Data Acquisition

We extracted drug indications from DailyMed, a free and complete source of all drug prescription data approved by the U.S. Food and Drug Administration (FDA). Drug indications and any pertinent medical context are specified as free text in the Indication and Uses section of each drug product label.

In the preliminary phase, we searched for neurological drugs (N0X) using the Anatomical Therapeutic Chemical Classification System (ATC) tree[1] in the DrugBank database. This search yielded 344 drug results. We selected only approved neurological drugs, and searched for their drug labels on DailyMed. Some drug labels with no contextual and relevant text were filtered out for further investigations in the protocol. We retrieved 101 drugs with a minimum of two drugs from each subclass of neurological drugs (i.e. N01-N07) for our analyses. This procedure ensures that representative samples were selected from all subclasses of neurological drugs. Furthermore, we referred to (Fuentes, Pineda & Venkata, 2018) which provided a comprehensive review of the top 200 prescribed drugs in the United States. From those 200 most prescribed drugs, we selected the neurological associated drugs that have drug labels.

### 2.2 Medical Context Curation

To identify relevant concepts and relations for medical contexts, we conducted annotation experiments with two scenarios: using a semi-automated pre-annotation approach and a fully manual annotation approach.

**Pre-Annotation using Automated Tools**: To reduce the manual annotation effort, we used automated tools to identify relevant concepts and relations for the context of the therapeutic intent. We used three main tools, MetaMap, SemRep and BioPortal Annotator, to annotate sentences from the labels with indications and therapeutic use information. For each drug in the dataset, the task of the annotators was composed of four sub tasks:

(1) Finding a relevant drug label from DailyMed by searching for each selected drug.

(2) Using the MetaMap Annotator to annotate the medical concepts in the drug labels. MetaMap Annotator provides access from biomedical text to concepts within the unified medical language system (UMLS) Metathesaurus. We selected relevant concept IDs and types from drug labels.

(3) We used the SemRep tool to detect relations between identified concepts and extract relevant UMLS IDs. SemRep is a UMLS-based programme which extracts semantic predications from biomedical text. These predicates come in triples in the form of a subject, object and a relation which connects the previously identified concepts.

(4) Lastly, we searched for other relevant terms from existing ontologies, such as Human Disease Ontology, using NCBO BioPortal Annotator service[2].

**Manual Annotations**: Two annotators manually annotated drug labels. Initially, annotators skimmed the drug labels and highlighted the contextual information. Concepts were then annotated manually and their respective concept IDs as well as types were recorded. When all annotations were finally collected, the annotations identified were recorded in an excel file. These annotations include a concept, concept ID and types. We extracted the following medical concepts based on information from DailyMed: 'therapyType', 'drug', 'disease', 'age range', 'symptom' and 'health condition'. After annotating the concepts, the annotators extended the spreadsheet to include 'Concept ID', 'MatchType', 'Quote', and 'Link'. The Concept IDs were identified using Bioportal and DrugBank for disease/symptom and drug IDs, respectively. 'MatchType' refers to whether the extracted Concept ID is an exact match to the object, or if it is a close match, meaning that the wording of the identified ID slightly differs from the object. 'Quote' refers to the quotes from DailyMed that were used for the triple identification. 'Link' shows the links to the exact DailyMed entry used to annotate the triples from.

[1] https://go.drugbank.com/atc

[2] https://bioportal.bioontology.org/annotator

## 2.3 Creating Ontology

Ontologies are a knowledge representation formalism in which concepts in a domain, their properties, relations, and restrictions are formally represented. We created an ontology using existing standards to capture the medical context of a drug indication. Figure 1 shows key entities and their relations between instances.

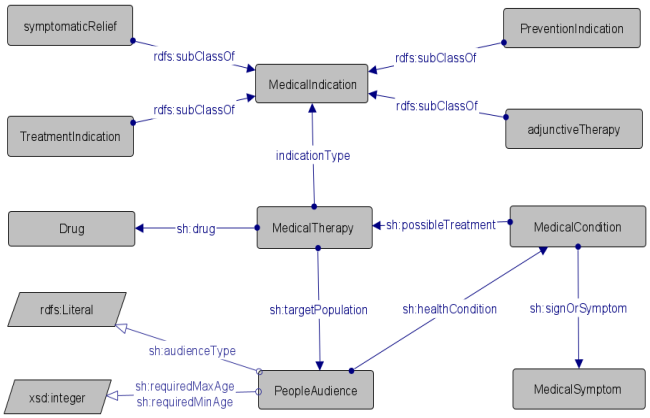


**Fig. 1.** The NeuroDKG data model.

We used *Schema.org* which is community-maintained vocabulary and uses W3C semantic web standards. We have used the following classes and properties from *Schema.org*:.

**Classes**: *Drug, MedicalTherapy, MedicalCondition, MedicalSymptom, PeopleAudience, MedicalIndication (subclasses: TreatmentIndication and PreventionIndication).*

**Properties**: *drug, targetPopulation, healthCondition, signOrSymptom, possibleTreatment, requiredMinAge and requiredMaxAge.*

The *Schema.org* vocabulary with additional types and properties were extended to capture contextual information of drug indications. The indicationType property was added to capture different types of treatment, i.e. *indication, symptomatic relief, prevention, adjunctive therapy* for *MedicalTherapy*. And two additional indication types were added to our ontology as subclasses of *MedicalIndication: SymptomaticRelief, AdjunctiveTherapy*.

We used identifiers from common biomedical ontologies and vocabularies such as DrugBank for drugs, MedDRA and Human Disease Ontology for conditions and symptoms.

## 2.4 Creating Nanopublications

Nanopublications are a way to publish scientific findings in a manner that is machine-readable, can be cited and verified. They are based on the idea of a scientific paper being a core scientific statement with associated context (Groth et al. 2010). Nanopublications can be given trusty URIs for verification of the digital artifact, which is an important requirement for complying with FAIR data principles.

Nanobench provides a web UI to easily publish contextual statements as nanopublications. We created a Nanobench template for publishing the medical context of drug indication (illustrated in Figure 2).

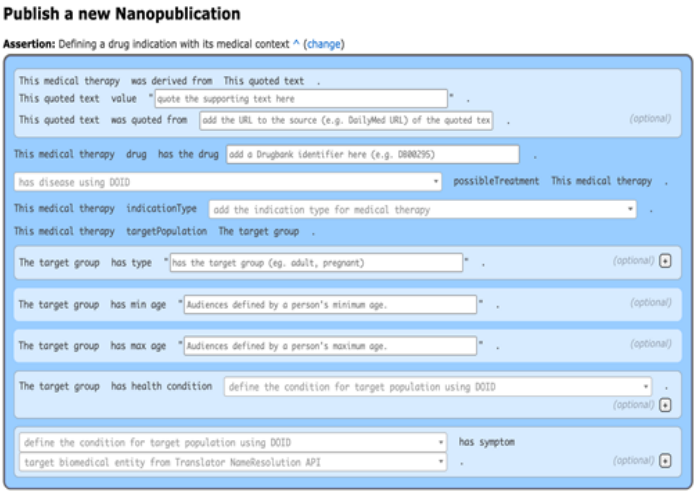


**Fig. 2.** Nanobench template for drug indication with its medical context

The Nanobench template helps ORCID-authenticated annotators to find and use standard identifiers to describe a drug/disease, and publish valid linked data following a predefined structure.

The resulting nanopublications can be accessed using the nanopublication API[3], or using the Nanobench search web UI (e.g. grlc API call to retrieve indications). The nanopublications network is a decentralized public knowledge graph with strong support for statements provenance, such as who published it, and when. Additionally, users can publish statements about existing drug indications, add supporting or refuting evidence.

# 3 RESULTS

A total of 2,397 triples, 460 entities, 13 properties, and 10 classes are defined in the NeuroDKG knowledge graph. Table 1 lists the relation that holds between types of subjects and types of objects in NeuroDKG. The descriptive statistics of the knowledge graph were computed by the d2s-cli (Emonet et al. 2018) tool using the HCLS dataset description.

[3] http://purl.org/np/RA56KBJXCfPObqVb3hNCSbbohzz02_VEW2PxnmyIQ0Hus

**Table 1.** Descriptive statistics of NeuroDKG

| # Subjects | Subject Class | Relation | Object Class | # objects |
|---|---|---|---|---|
| 192 | schema: Medical Therapy | schema: drug | schema: Drug | 101 |
| 192 | schema: Medical Therapy | schema: indication Type | schema: Medical Indication | 4 |
| 59 | schema: Medical Therapy | schema: target Population | schema: People Audience | 59 |
| 75 | schema: MedicalCondition | schema: possible Treatment | schema: Medical Therapy | 192 |
| 10 | schema: MedicalCondition | schema: signOr Symptom | schema: Medical Symptom | 28 |
| 22 | schema: PeopleAudience | schema: heath Condition | schema: Medical Condition | 13 |

## DISCUSSION

Access to high quality machine-readable descriptions of drug indications and their medical context is key to downstream computational uses. We used state-of-the-art tools to represent drug indications and their associated context. The constructed ontology by building upon Schema.org vocabulary that encapsulates relationships among diseases, drugs, and other contextual information. Creating web UI interface using Nanobench template based on this ontology where users can freely publish their own curated drug indications as nanopublications and cite others' nanopublications.

Nanobench interface incorporates services such as normalization to help users choose the right standard identifiers and create structured data in the form of a knowledge graph. Furthermore, users can define the medical context of other indications, and contribute to the scientific knowledge facilitating clinical decision making and precision medicine applications.

To the best of our knowledge, no such data model or structured template has been proposed for neuropharmaceutical drug indications in the literature. The proposed data model may not capture the more complex medical context. In this case, it requires an update in the current Nanobench template. However, this could also be an opportunity to develop a Nanobench template creation based on schema/data model.

Although we have applied some FAIR data principles such as assigning globally unique identifiers to data and metadata and providing the data openly via an API, a further assessment is needed to ensure that NeuroDKG is fully FAIR-compliant.

## REFERENCES


Li Z, Huang Q, Chen X, *et al*. Identification of Drug-Disease Associations Using Information of Molecular Structures and Clinical Symptoms via Deep Convolutional Neural Network. *Front Chem*. 2020;7:924. Published 2020 Jan 10. doi:10.3389/fchem.2019.00924

Nelson, S. J., Oprea, T. I., Ursu, O., Bologa, C. G., Zaveri, A., Holmes, J., Yang, J. J., Mathias, S. L., Mani, S., Tuttle, M. S., & Dumontier, M. (2017). Formalizing drug indications on the road to therapeutic intent. *Journal of the American Medical Informatics Association*, 24(6), 1169-1172. https://doi.org/10.1093/jamia/ocx064

Khare R, Wei CH, Lu Z. Automatic extraction of drug indications from FDA drug labels. *AMIA Annu Symp Proc.* 2014 Nov 14;2014:787-94. PMID: 25954385; PMCID: PMC4419914.

Ritu Khare, Jiao Li, and Zhiyong Lu. 2013. Toward Creating a Gold Standard of Drug Indications from FDA Drug Labels. In Proceedings of the 2013 IEEE International Conference on Healthcare Informatics (ICHI '13). *IEEE Computer Society*, USA, 30–35. DOI: https://doi.org/10.1109/ICHI.2013.11

Moodley, K., Rieswijk, L., Oprea, T.I. *et al*. InContext: curation of medical context for drug indications. *J Biomed Semant 12*, 2 (2021). https://doi.org/10.1186/s13326-021-00234-4

Gamberger, D., Prcela, M., Jovic, A., Šmuc, T., Parati, G., Valentini, M., Kawecka-Jaszcz, K., Styczkiewicz, K., Kononowicz, A., Candelieri, A., Conforti, D., & Guido, R. (2008). Medical Knowledge Representation within Heartfaid Platform. HEALTHINF.

Kuhn T, Taelman R, Emonet V, Antonatos H, Soiland-Reyes S, Dumontier M. 2021. Semantic micro-contributions with decentralized nanopublication services. *PeerJ Computer Science* 7:e387 https://doi.org/10.7717/peerj-cs.387

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data 3*, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Fuentes, A.V.; Pineda, M.D.; Venkata, K.C.N. Comprehension of Top 200 Prescribed Drugs in the US as a Resource for Pharmacy Teaching, Training and Practice. *Pharmacy 2018*, 6, 43. https://doi.org/10.3390/pharmacy6020043

Groth, Paul & Gibson, Andrew & Velterop, Johannes. (2010). The Anatomy of a Nano-publication. *Information Services and Use*. 30. 10.3233/ISU-2010-0613.

Emonet, Vincent; Malic, Alexander; Zaveri, Amrapali; Grigoriu, Andreea; Dumontier, Michel (2018): Data2Services: enabling automated conversion of data to services. *Semantic Web Applications and Tools for Healthcare and Life Sciences*. Journal contribution. https://doi.org/10.6084/m9.figshare.7345868.v1