International Journal of Medical Informatics 181 (2024) 105284

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

HeNeCOn: An ontology for integrative research in Head and Neck cancer

Liss Hernández [a], Estefanía Estévez-Priego [a], Laura López-Pérez [a], María Fernanda Cabrera-Umpiérrez [a], María Teresa Arredondo [a], Giuseppe Fico [a,*], the BD2Decide Consortium [b]

[a] Universidad Politécnica de Madrid-Life Supporting Technologies Research Group, ETSIT, 28040 Madrid, Spain
[b] Life Supporting Technologies research group, Photonics Technology and Bioengineering department, School of Telecommunications Engineering, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Ontology
Head and neck cancer
Data modeling
Data standardization
Clinical information system
Ontology-based knowledge
Protégé

ABSTRACT

Background: Head and Neck Cancer (HNC) has a high incidence and prevalence in the worldwide population. The broad terminology associated with these diseases and their multimodality treatments generates large amounts of heterogeneous clinical data, which motivates the construction of a high-quality harmonization model to standardize this multi-source clinical data in terms of format and semantics. The use of ontologies and semantic techniques is a well-known approach to face this challenge.
Objective: This work aims to provide a clinically reliable data model for HNC processes during all phases of the disease: prognosis, treatment, and follow-up. Therefore, we built the first ontology specifically focused on the HNC domain, named HeNeCOn (Head and Neck Cancer Ontology).
Methods: First, an annotated dataset was established to provide a formal reference description of HNC. Then, 170 clinical variables were organized into a taxonomy, and later expanded and mapped to formalize and integrate multiple databases into the HeNeCOn ontology. The outcomes of this iterative process were reviewed and validated by clinicians and statisticians.
Results: HeNeCOn is an ontology consisting of 502 classes, a taxonomy with a hierarchical structure, semantic definitions of 283 medical terms and detailed relations between them, which can be used as a tool for information extraction and knowledge management.
Conclusion: HeNeCOn is a reusable, extendible and standardized ontology which establishes a reference data model for terminology structure and standard definitions in the Head and Neck Cancer domain. This ontology allows handling both current and newly generated knowledge in Head and Neck cancer research, by means of data linking and mapping with other public ontologies.

1. Introduction

Head and neck cancer (HNC) encompasses a diverse spectrum of tumors located on the upper aerodigestive tract. The annual incidence of HNC reaches more than 700,000 new cases and over 350,000 deaths per year, and its prevalence is expected to increase up to 30% in the following years [1,2]. Due to its complexity and diversity of locations, HNC cases usually require multimodality approaches for treatment and biomarker identification [3]. The increasing prevalence and miscellany of HNC generate large amounts of heterogeneous data from different clinical centers, which is one of the main challenges when conducting an analytics study. Hence, there is a need for a high-quality harmonization model that standardizes these clinical data from multiple sources in terms of format and semantics [4,5]. This task grants the accurate reuse of comprehensive HNC datasets in further studies to encourage interoperability, enable personalized medicine, and facilitate integrative research.

In this sense, the use of ontologies in combination with semantic techniques is a powerful solution to address data harmonization and integration [6–9], as demonstrated in the biomedical field by the Open Biological and Biomedical Ontology (OBO) Foundry [10,11]. International standards have been developed to provide multilingual clinical

* Corresponding author.
E-mail address: gfico@lst.tfo.upm.es (G. Fico).

https://doi.org/10.1016/j.ijmedinf.2023.105284
Received 14 April 2023; Received in revised form 14 July 2023; Accepted 1 November 2023
Available online 8 November 2023

healthcare terminology, such as SNOMED-CT[1] [12] for content mapping in electronic health records, ICD-10[2] [13] for disease classification, and LOINC[3] [14] for medical laboratory observations. In parallel, field-specific ontologies have been implemented to model chronic diseases (e.g., breast [15,16], prostate [17,18], thyroid [19], liver [20] and lung [21,22] cancers). Other transversal works with special significance in HNC research are the Gene Ontology (GO) [23,24], Radiation Oncology Ontology (ROO) [25], and Radiomic Ontology [26]. However, there is currently no ontology specifically focused on HNC.

Efforts to unify multi-source data are essential to foster advances in our understanding of HNC and improve prevention measures and therapy strategies while supporting data-sharing. For that purpose, building a robust ontology that represents the knowledge acquired in the HNC domain requires close collaboration with specialized clinicians, a meticulous harmonization process, and must contain collected terms and relationships that are certain, trustable, and standard in terms of semantic meaning, to become the cornerstone of clinical decision support systems, predictive modeling, disease modeling, and other bioinformatic applications [27,28] in HNC management. This work presents the Head and Neck Cancer Ontology (HeNeCOn), developed as part of the BD2Decide project [29], as a clinically verified data model for HNC processes during prognosis, treatment, and follow-up phases of the disease.

2. Methods and materials

The HeNeCOn ontology was developed as part of the European project BD2Decide, which aimed to provide patient-specific prognosis and tailored treatments for better clinical outcomes. This project integrated multicenter data from 1537 HNC patients, to be explored within a Decision Support System (DSS). The BD2Decide cohort consisted of patients from five clinical centers and a total of 396 clinical, pathological, and demographic parameters were collected. Detailed in a previous work [29], these parameters include descriptive information about the tumor; the ASA,[4] ECOG,[5] and ACE-27[6] classification scores, among others; major risk factors; familiar history of malignancies; diagnostic data from Computerized Tomography (CT), Magnetic Resonance Imaging (MRI) and Diffusion-Weighted Imaging (DWI); genomic and radiomics data; treatments, toxicity, and follow-up. Besides the BD2Decide cohort, patient data comprising 106 variables were collected from external registries of the RARECAREnet project [30]. The total of 502 variables served to define the ontology requirements in terms of categories and relations between elements.

The scope of the HeNeCOn ontology covers the complete healthcare plan of HNC patients, focusing on diagnosis and prognosis. The development process of HeNeCOn was organized into three steps: (1) taxonomy creation, (2) semantic definition and data linking proceeding, and (3) ontology validation, as shown in Fig. 1.

2.1. Taxonomy creation

The taxonomy creation started by determining the scope and selecting the *meta*-characteristics for the terms of interest [31] as described in a related work [32]. Relevant-to-scope HNC terms were identified based on individual-level data and organized hierarchically following the class structure of the Ontology for Biomedical Investigations (OBI) [33] and the NeoMark ontology on oral cavity cancer [34,35] that paved the way towards a more detailed HNC-based work.

2.2. Semantic definition and data linking

Once the taxonomy was established, three phases were conducted using the Protégé ontology editor [36]: (I) Insertion of semantic meaning for every identified term, (II) identification of similarities between terms, and (III) link to related external ontologies.

2.2.1. Insertion of semantic meaning

This process began by gathering and validating the semantic definitions and relationships for every term and was performed through annotations describing each meaning based on clinical glossaries such as the National Cancer Institute (NCI) Dictionary of cancer terms [37], the Medical Subject Headings (MeSH) thesaurus [38] and Head and Neck glossary [39].

2.2.2. Identification of similarities between data

To enrich the list and definition of the terms with a qualitative analysis, the similarity between terms was assessed based on three conditions: meaning, values, and format. When two terms have the same meaning, values, and format, they are linked *directly* to expand the characteristics of individual eligible patients (i.e., from individual-level data) with additional information (i.e., from external registries). Conversely, when data terms are related but differ either in meaning, values, or format, they are linked by validated *equivalence rules*.

2.2.3. Linking to external ontologies

The data linking process was extended by iteratively including pre-selected external ontologies from public repositories, which required an initial standardization and conversion to Terse RDF (Resource Description Framework) Triple Language (Turtle) format. The mapping is based on the likelihood between terms' definitions. Some terms had multiple mapping within the same external ontology, while others were mapped across different external ontologies. These links allow a semantic search of standard terminology and complex relations with all external ontologies referenced in the HeNeCOn ontology.

2.3. Ontology validation

The ontology creation process was supervised by physicians that participated in the BD2Decide project to guarantee an ontology compliant with the following quality attributes: robust, to arrange similar terms enabling their differentiation using data linking; comprehensive, through accurate designations; explanatory, by facilitating the finding of terms based on possible values; interoperable, by mapping with external ontologies; and extensible, allowing the inclusion of new terms in a continuously evolving field. Consequently, and following ontology evaluation practices, HeNeCOn covered evaluation domains as correctness, lexical, taxonomic, semantic, structural, and interoperability. Based on their expertise, the physicians validated all the ontology sections, as specified in Table 1. Following the validation process by clinicians in their specific domains, all experts conducted a thorough review of the entire work.

2.4. Public access

The HeNeCOn ontology was published in the BioPortal repository following open-source principles and can be retrieved locally from bioportal.bioontology.org/ontologies/HENECON.

3. Results

3.1. Taxonomy structure

The HeNeCOn taxonomy is built upon the class structures from the OBI and NeoMark ontologies. We first selected those pre-existing classes relevant to HeNeCOn: *data item, post-treatment* and *treatment*. *Virtual*

[1] Systematized Nomenclature of Medicine - Clinical Terms.
[2] International Classification of Diseases.
[3] Logical Observation Identifiers Names and Codes.
[4] American Society of Anesthesiologists.
[5] Eastern Cooperative Oncology Group.
[6] Adult Comorbidity Evaluation-27.

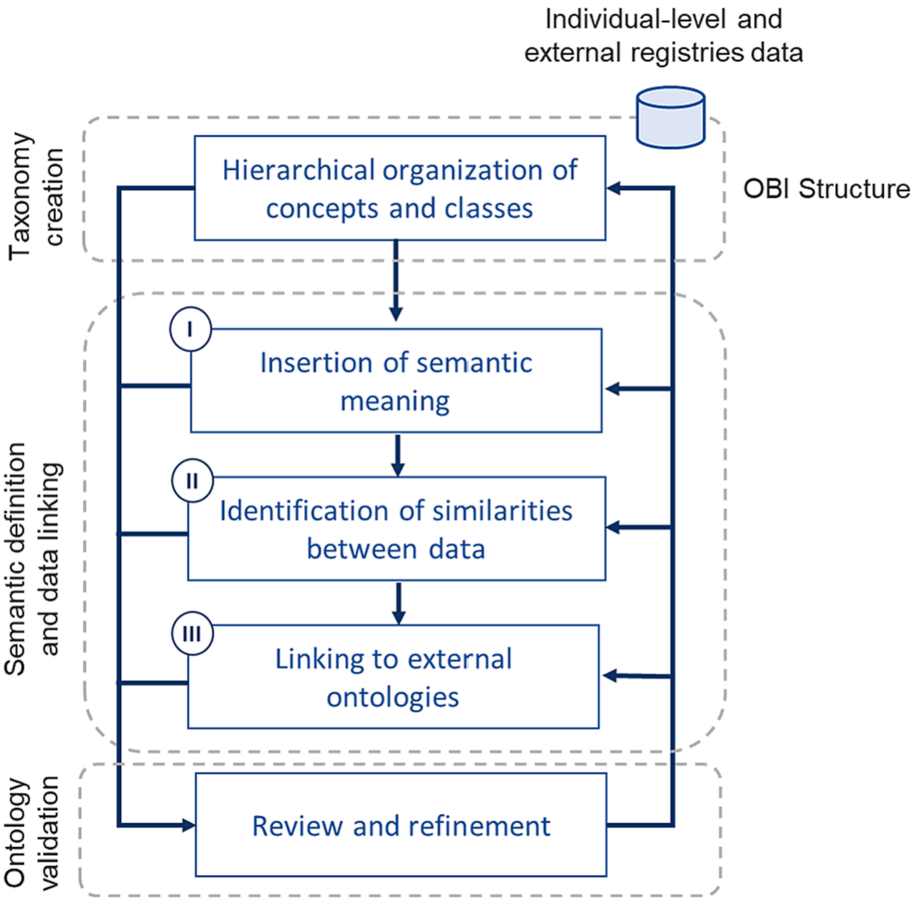

Individual-level and external registries data
Taxonomy creation
Hierarchical organization of concepts and classes
OBI Structure
Semantic definition and data linking
I
Insertion of semantic meaning
II
Identification of similarities between data
III
Linking to external ontologies
Ontology validation
Review and refinement

Fig. 1. Methodology flowchart for the HNC-based ontology. First, the taxonomy is created based on collected datasets, then the semantic definition and data linking are split into three phases: (I) Insertion of the semantic meaning of the terms, (II) identification of similarities between data, (III) identification of related ontologies for knowledge linking. All phases were iteratively reviewed and validated by clinicians.

Table 1
Clinical centers responsible for validating the different sections of the ontology depending on their field of expertise.

Clinical center (Country)	Expertise	Ontology sections
Istituto Nazionale dei Tumori di Milano (Italy)	Oncology, molecular biology, epidemiology, statistics	Chemotherapy, toxicity and genomics
Maastro Clinic (Netherlands)	Imaging, radiology, radiation oncology	Imaging and radiotherapy
Düsseldorf University Hospital (Germany)	Otorhinolaryngology	Surgery
Azienda Ospedaliero Universitaria of Parma (Italy)	Surgery	Clinical TNM, pathology, demographic, risk factors, follow-up and quality of life

Patient and *Cancer Registries Data* are two HeNeCOn subclasses added to *data item*; the subclasses *Follow-up* and *Quality of Life* were included in *post-treatment;* and 62 new variables within the surgery, chemotherapy, and radiotherapy classes were inserted into the *treatment* class [32] (Fig. 2).

In this context, *Virtual Patient* corresponds to the root subclass with patient information about HNC diagnosis and prognosis. This includes 315 variables regarding clinical, demographic, pathological, genomics, imaging, radiomics, risk factors, and toxicity data. *Cancer Registries Data* models the information from a public dataset of cancer registries. The treatment section contains data related to treatments utilized in HNC and are divided into two classes: the non-surgical treatments class for chemotherapy, radiotherapy, immunotherapy, and palliative care, and the surgical treatment class which also includes surgical procedures and reconstruction. The *post-treatment* section comprises two classes: *Follow-up* and *Quality of Life*, with 13 and 3 new variables included, respectively. The *Follow-up* class includes data related to the recurrence of cancer and the status of the patient after the treatment is completed, while the *Quality of Life* of patients includes terms such as deglutition and respiration capabilities. This structure covers all the necessary terms identified during the taxonomy definition.

3.2. Semantic annotation

Semantic annotations were applied to 283 variables referring to the HNC patient characteristics. *Chemotherapy* is described using 23 variables as the number of cycles performed, agent name, dose, and best tumor response, among others. *Radiotherapy* is described through 19 variables including overall treatment time, target volume, dose, and settings. *Surgical treatment* contains 20 variables including reason, site, complications, and actions taken. *Post-treatment* section encloses *follow-up* variables such as status of the patient, date and type of recurrence, and cause of death. This section, with 16 variables, also contains *quality of life* indicators based on deglutition, respiration capabilities, and standard questionnaires: HN30, HN35 [40], and EQ-5D [41]. The *pathology* section includes 42 variables describing the pathologic features of the tumor such as maximum diameter, thickness, pattern of invasion,

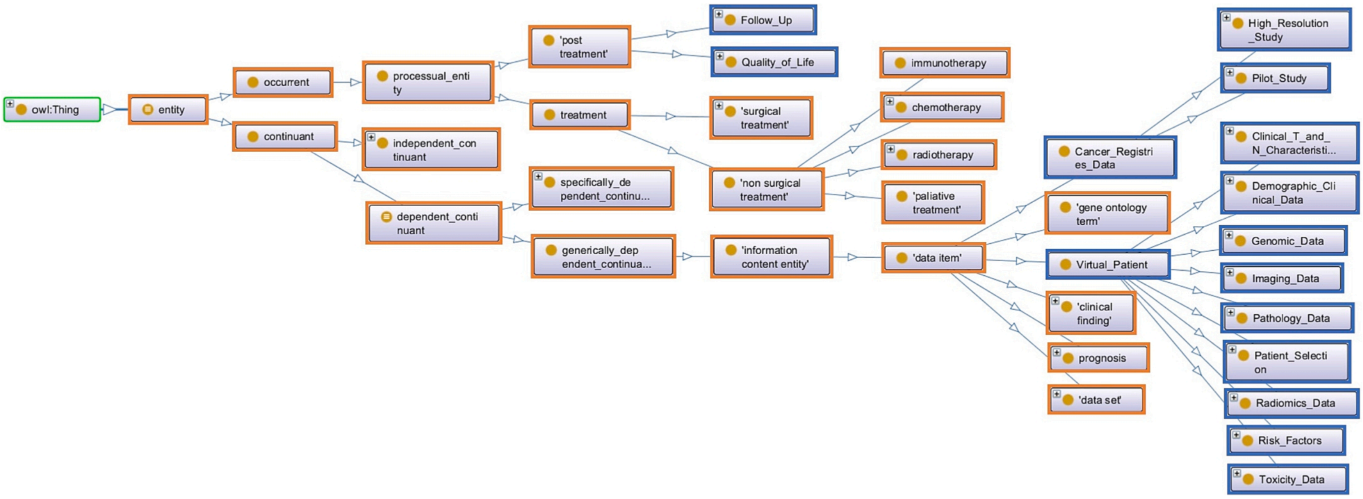


Fig. 2. Hierarchical representation of the HeNeCOn ontology. The ontology is represented as a graph where nodes are classes and the arrows are their relationships (i.e. subclass). Starting from the root class *owl:Thing*, the new classes created for the HeNeCOn ontology (blue) are defined as subclasses of the pre-existing terminology from OBI and Neomark ontologies (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HPV[7] infection, and pathologic TNM staging. The *imaging* data definitions with 54 variables relate information from radiological TNM, CT, MRI, DWI, and corresponding configuration settings and analysis. *Demographic* and *clinical* data comprises 27 variables such as year of birth, age at diagnosis, gender, ethnicity, and clinical performance assessment. The *toxicity* section contains 41 variables of medullary and hepatic toxicity, infections, xerostomia, and dysphagia, among others. *Risk factors* with 16 variables cover information about smoking, alcohol consumption, oral hygiene habits, and family history of malignancies. Finally, the clinical data dictionary contains 21 variables for tumor region, anatomical location, depth of invasion, and TNM staging for the 7th and 8th editions, 4 genomic variables representing genes relevant for HNC prognosis, Binary Alignment and Map (BAM), and FASTQ format files for raw and processed data.

The semantic definitions were included in the ontology as annotations using Protégé (Fig. 3). Every variable can be selected (Fig. 3A) to include semantic annotation (Fig. 3B) and revise the relationships between other ontology components (Fig. 3C). All this information is written in the Web Ontology Language (OWL) file (Fig. 3D). For some pre-selected variables, semantic terms definitions were skipped due to the impossibility of finding a standard definition associated with the selected concept. Those cases were included using a not standard but common definition under the supervision of the five clinical centers.

3.3. Similarity between data

Similarities were identified for clinical and demographic data through a comparative analysis between datasets. The mapping between these variables does not necessarily mean they share the same subordinates or same hierarchical relevance, but rather the same semantic meaning or equivalent values.

The matching between datasets (Table 2) was possible in a total of 9 variables with direct equivalence, 5 variables with required equivalence rules, and 3 variables describing the hospital center were discarded due to the lack of similarities.

Direct matching corresponds to a straightforward linkage between terms. As shown in Table 2, although the terms 'tumour region' and 'site' have different labels, they correspond to the same data item (i.e., the ICD-10 classification), therefore requiring a direct relationship. While direct matching does not imply further actions, equivalence rules are carefully assessed. Table 3 shows how equivalent data and possible values are identified for the ontology class *Stage at Diagnosis*, and the subsequent equivalence rules are defined to indirectly match external registries. External cancer registry data correspond to the RARECARE.net pilot studies: Study 1 (High-resolution study) and Study 2 (Pilot study) comprising 106 variables. In such a way, the ontology class *Stage* uses the TNM staging system, where each code (from 0 to IVC) describes the size of the tumor, the depth of invasion, and spreading to lymph nodes or other body parts. However, the external registries did not apply the TNM staging system but rather a descriptive approach and custom numeric associations. The definition of equivalence rules and similarities allows information systems (i.e., DSS) to add relevant characteristics from external cancer registries and complement individual patient datasets.

3.4. Link to other ontologies

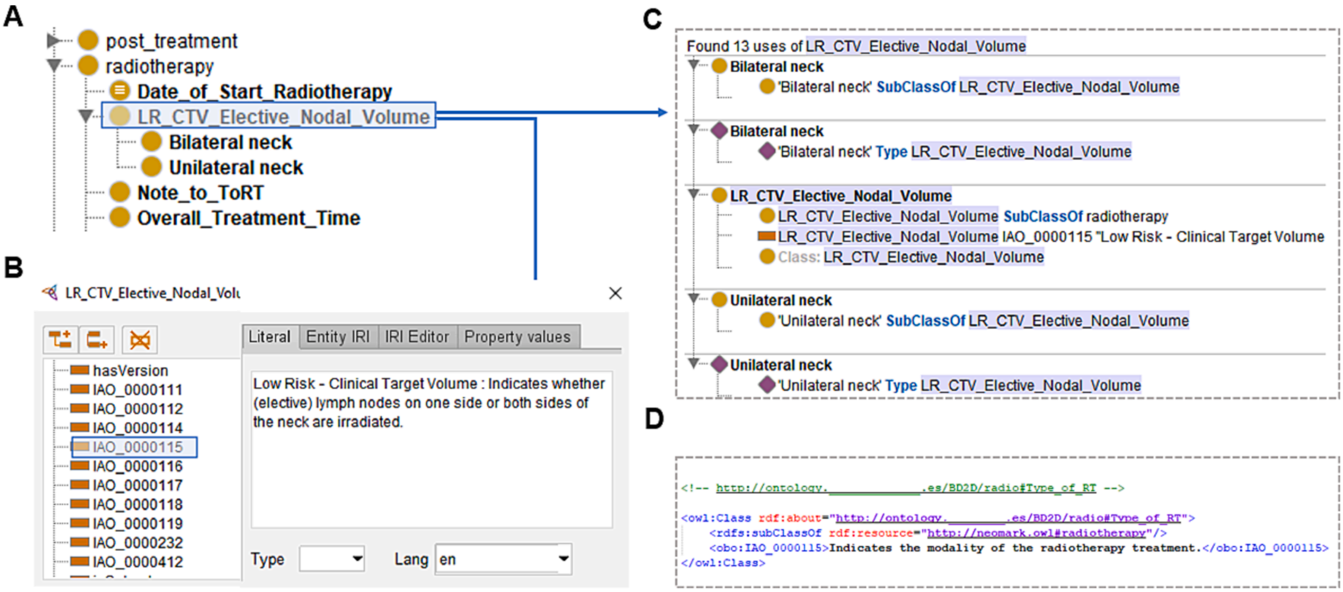
The linking with external ontologies relates terms with similar semantic meanings independently of their values. Similar terms from external ontologies have been linked with the classes included in the HeNeCOn ontology, as shown with two examples in Table 4. This relation allows the extraction of knowledge from different standardized sources and provides additional information and semantic meaning to the ontology terms. Six ontologies were selected for being HNC content-related: NCI Thesaurus, ICD10CM,[8] OBI, ROO, Exposure Ontology (EXO), and SNOMED-CT. The list of mapped terms in OWL format the corresponding code number and the name of the external ontology they belong to, e.g., 'Anatomical Tumor Location' corresponds to 'Thesaurus: C13717′ from the NCI Thesaurus ontology. This distributed approach ensures scalability by allowing updates to be made solely to the associated mapping file when new terms are added or modified in an external ontology.

3.5. Ontology creation roadmap

From the resulting HeNeCOn ontology, we propose a roadmap for ontology creation (Fig. 4). This roadmap can be used as a methodology

[7] Human papillomavirus.

[8] International Classification of Diseases, Tenth Revision, Clinical Modification.



Low Risk - Clinical Target Volume : Indicates whether (elective) lymph nodes on one side or both sides of the neck are irradiated.

```
Found 13 uses of LR_CTV_Elective_Nodal_Volume
Bilateral neck
    'Bilateral neck' SubClassOf LR_CTV_Elective_Nodal_Volume
Bilateral neck
    'Bilateral neck' Type LR_CTV_Elective_Nodal_Volume
LR_CTV_Elective_Nodal_Volume
    LR_CTV_Elective_Nodal_Volume SubClassOf radiotherapy
    LR_CTV_Elective_Nodal_Volume IAO_0000115 "Low Risk - Clinical Target Volume
    Class: LR_CTV_Elective_Nodal_Volume
Unilateral neck
    'Unilateral neck' SubClassOf LR_CTV_Elective_Nodal_Volume
Unilateral neck
    'Unilateral neck' Type LR_CTV_Elective_Nodal_Volume
```

```
<!-- http://ontology.___________.es/BD2D/radio#Type_of_RT -->

<owl:Class rdf:about="http://ontology.________.es/BD2D/radio#Type_of_RT">
    <rdfs:subClassOf rdf:resource="http://neomark.owl#radiotherapy"/>
    <obo:IAO_0000115>Indicates the modality of the radiotherapy treatment.</obo:IAO_0000115>
</owl:Class>
```

Fig. 3. Semantic description included in the HeNeCOn terms. (A) Hierarchical organization of variables: LR_CTV_Elective_Nodal_Volume depends on the radiotherapy class and contains two instances, bilateral neck and unilateral neck. (B) The variable annotations follow the IAO:0000115 (OBI digital entity definition). (C) The correspondence with other classes, variables and instances is fully described in the Usage section. (D) The information is read from the OWL file in xml format.

Table 2
Matching of individual-level data with external registries based on the type of mapping.

Type of mapping	Individual-level data	External registries
Direct match	Year of birth	Date of birth
	Date of diagnosis	Date of diagnosis
	Hospital of diagnosis	Hospital of diagnosis
	Gender	Gender
	Tumour region	Site
	Anatomical tumour location	Tumour location
	CT and CN of TNM	CT and CN of TNM
	Last day of follow-up	Last day of follow-up
	Vital status of patient	Vital status of patient
Equivalences	Stage at diagnosis	Stage
	Grade at diagnosis	Grade
	Surgical margins	Margin status
	Chemotherapy treatment / radiotherapy treatment / surgery	Type of treatment ctscandone / mriscandone
	Image type	
No match	N/A	Geographic area
	N/A	Level of specialization
	N/A	University hospital (Yes / No)

Table 3
Example of equivalence rules established between external registries and patient level datasets.

Source	Patient level data	Study 2	Study 1	Equivalence rules
Class label	Stage at diagnosis	Stage	Stcond4	I ≡ 1 ≡ early II or III ≡ 2 or 3 ≡ advanced
Possible values	0, I, II, III, IVA, IVB, IVC	1 = localized, 2 = locally advanced, 3 = N+, 4 = M+	Early, advanced, metastatic	IV ≡ 4 ≡ metastatic

for developing and clinically validating disease-specific ontologies, being consistent with pre-existing resources in biomedical and cancer-related fields while expanding the range of action and impact on future research projects.

Table 4
Examples of ontology mapping with external ontologies.

HeNeCOn term	Ontologies mapping
Total number of fractions	**SNOMED-CT:** ● ID: https://purl.bioontology.org/ontology/SNOMEDCT/228862004 ● Preferred Name: Number of fractions. ● SNOMEDID: R-42A40. **ROO:** ● ID: https://www.cancerdata.org/roo/100354 ● Preferred Name: Number of Radiotherapy Fractions. ● Definition: The count of radiotherapy fractions given in a certain time interval (e.g., per day) or in a certain coherent context (e.g., per treatment).
Tissue invasion – ExtraNodal Extension (ENE)	**NCI Thesaurus:** ● ID: https://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C103442 ● Preferred Name: Target ExtraNodal Tumor Identification ● NCI Thesaurus Code: C103442 ● Definition: The identification of a target tumor located outside of or independent of the lymph node. **SNOMED-CT:** ● ID: https://purl.bioontology.org/ontology/SNOMEDCT/396644006 ● Preferred Name: Extra-capsular extension of nodal tumor present. ● SNOMEDID: F-004F1 ● ID: https://purl.bioontology.org/ontology/SNOMEDCT/396643000 ● Preferred Name: Extra-capsular extension of nodal tumor absent. ● SNOMEDID: F-004EF

3.6. HeNeCOn application

The validated ontology was applied to leverage a web-based DSS customized for HNC patient management, where the outcome of

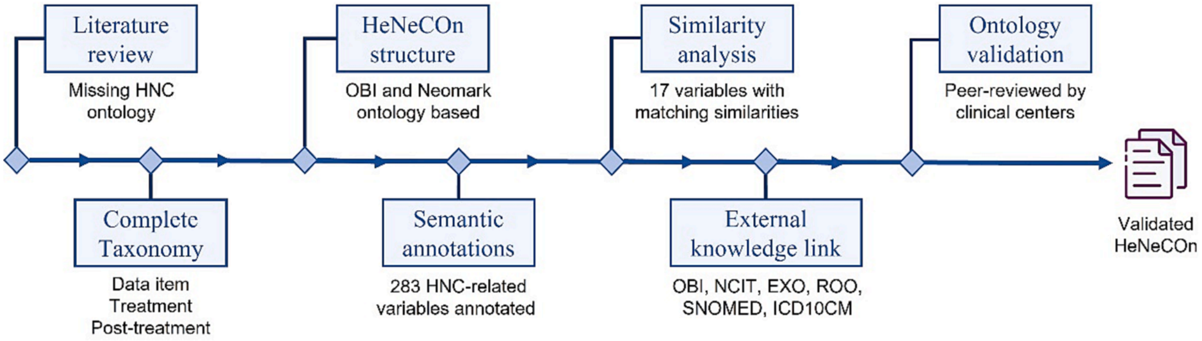


Fig. 4. Roadmap and outcomes of the HeNeCOn ontology creation. Starting from a thorough literature revision, through the taxonomy and hierarchy construction, until the final validated product for supporting integrative research in Head and Neck cancer.

prognostic models is visualized by HNC clinical experts. Prognostic models relied on the ontology structure to normalize the data gathered from multiple clinical centers, promoting harmonization and interoperability. The researchers benefit from the semantic dictionary and relationships map provided by the HeNeCOn ontology, through semantic queries implemented in the system using SPARQL, to explore HNC-specific data and retrieve relevant knowledge.

4. Discussion and conclusions

The management of HNC cases requires the use of an extensive and complex terminology map and typically large amounts of non-standardized data. The main goal of constructing the HeNeCOn ontology is to provide a trustable and standardized resource for clinicians and bioinformaticians when working with HNC. We identified 502 terms from both patient-level data and external cancer registries, enriched with semantic meaning, mapped across 6 related pre-existing ontologies and established similarities among them. This work was validated by physicians and statisticians from 4 international medical centers to ensure trustworthiness and clinical applicability.

Broadly, the definition and implementation of HeNeCOn adds up to the increasing efforts of the data science community to facilitate interoperability between medical centers [42,43], and promote integrative research, by including new fields of knowledge into the publicly available realm of ontologies. The resulting taxonomy, definitions, and data linking process grew into a reusable and extendible multi-source data model aligned with previous health-related works through similarity assessments and equivalent rules definitions.

Disease-specific ontologies set explicit statements and non-misleading information in medical domains where unambiguity is extremely important. This is particularly significant for heterogeneous and multi-factor diseases as HNC: different treatments and curative modalities [3], high probabilities of recurrence [44], convoluted follow-up procedures [3], the pursuit of patient's QoL [45] and the development of personalized and precision medicine [46,47].

4.1. Advantages and innovations

HeNeCOn is a domain-specific ontology with hierarchical organization, semantic data annotation and detailed relations between terms that are essential for information extraction and knowledge handling, combining blocks of sorted knowledge among external sources and previous related works. Previous ontologies focused on general medical science [12,48,49] research processes [23,33] and other major initiatives [24] do not include disease-specific terms. Some domain-based ontologies are dedicated to concrete cancer types [9,20–22], but only a few are linked with other ontologies or data resources [33,48], which might lay a gap for further interoperability activities.

While some studies relied on physicians' domain knowledge and expertise [9,24,33,48], like HeNeCOn, others did not request any clinical validation process [20–23].

Many studies provide details about the implementation performed [12,20–22,24], but few describe a clear methodology for ontology development guidance [48]. As there is not a strictly defined methodology for building ontologies [50], the hierarchical structuring and mapping procedure followed for HeNeCOn can inspire other medical-related research works. In this context, HeNeCOn serves as a reference resource for HNC clinicians' consultation.

The submitted ontology is the outcome of an iterative process of inspection, validation and refinement that aims to support foreseen applications. Following ontology evaluation practices [51], HeNeCOn covered evaluation domains as correctness (formal language), lexical, taxonomic, semantic, structural and interoperability (link with other ontologies) ensuring the quality, scalability and applicability of the ontology. Dedicated assessments in various HNC data-driven projects have proven the usefulness of a HNC-specific data model to be applied for a DSS [29] and a HNC multisource data harmonization work with adherent quality rules [52].

4.2. Limitations and future perspectives

The HeNeCOn ontology can be enriched through formal statements to enable automatic reasoning, thus improving the quality of data interpretation and promoting additional application-driven solutions [53]. An enhanced data modeling strategy shall incorporate an automatic continuous updating procedure able to add additional levels and supplementary instances if new related content is found [54]. Incidentally, due to the constant new data generation, updates in healthcare guidelines, advances in biomedical technology, and the emergence of new ontologies, HeNeCOn will need to be incorporated into novel data models in the future. However, as of today, it established a formalized knowledge basis to assist HNC medical professionals and researchers in their everyday practice.

4.3. Conclusions

The urgent need for big data standardization in terms of semantics and conceptualization set in motion a cooperative attempt to develop ontologies and semantic techniques to allow the management of increasing amounts of multisource data. Within this framework, the HeNeCOn ontology exhibits a reference model that can expand the research possibilities in the HNC field, to assist and improve current healthcare procedures. The proposed methodology for ontology development provides a landmark for future integration works in the HNC domain.

5. Summary table

What was already known on the topic.

1. The Head and Neck Cancer management (i.e., diagnosis, prognosis, treatment and follow-up) generates large amounts of heterogeneous data.
2. Ontologies are widely used tools to model information, especially in the biomedical and healthcare field.
3. The harmonization of data from multiple sources is a challenge that requires a multidisciplinary team.

What this study added to our knowledge.

1. HeNeCOn is the first ontology dedicated to Head and Neck Cancer disease and was created to be easily extensible to other related fields.
2. HeNeCOn serves as a reference data model for Head and Neck Cancer research and clinical disease management.
3. The presented methodology is applicable to developing other clinically-validated disease-specific ontologies.
4. HeNeCOn facilitates the management of both existing and emerging knowledge by integrating and mapping other publicly available ontologies.

CRediT authorship contribution statement

Liss Hernández: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Estefanía Estévez-Priego:** Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Laura López-Pérez:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **María Fernanda Cabrera-Umpiérrez:** Conceptualization, Funding acquisition, Supervision. **María Teresa Arredondo:** Funding acquisition, Supervision, Project administration. **Giuseppe Fico:** Conceptualization, Methodology, Supervision, Writing – review & editing.

The BD2Decide Consortium

Tito Poli[1], Silvia Rossi[1], Elena Martinelli[1], Lisa Licitra[2,8], Stefano Cavalieri[2], Loris De Cecco[3], Silvana Canevari[3], Kathrin Scheckenbach[4], Ruud H. Brakenhoff[5], Irene Nauta[5], Frank J.P. Hoebers[6], Frederik W. R. Wesseling[6], Annalisa Trama[7], Gemma Gatta[7].

[1]Unit of Maxillofacial Surgery, Department of Medicine and Surgery, University of Parma – University Hospital of Parma, Parma, Italy.

[2]Head and Neck Medical Oncology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.

[3]Integrated Biology Platform, Department of Applied Research and Technology Development, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.

[4]Department of Otolaryngology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

[5]Vrije Universiteit Amsterdam, Otolaryngology/Head and Neck Surgery, Amsterdam UMC, Cancer Center Amsterdam, Amsterdam, The Netherlands.

[6]Department of Radiation Oncology (MAASTRO), Research Institute GROW, Maastricht University, Maastricht, The Netherlands.

[7]Department of Preventive and Predictive Medicine, Evaluative Epidemiology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.

[8]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy.

Funding

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements BD2Decide No. 689715 and BD4QoL No. 875192.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World Health Organization, "Global Cancer Observatory. International Agency for Research on Cancer." 2022.

[2] H. Sung, et al., Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA Cancer J. Clin. 71 (3) (May 2021) 209–249, https://doi.org/10.3322/caac.21660.

[3] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and neck squamous cell carcinoma," *Nature Reviews Disease Primers*, vol. 6, no. 1. Nature Research, Dec. 01, 2020. 10.1038/s41572-020-00224-3.

[4] E. Roelofs, A. Dekker, E. Meldolesi, R.G.P.M. Van Stiphout, V. Valentini, P. Lambin, International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining, Radiother. Oncol. 110 (2) (2014) 370–374, https://doi.org/10.1016/j.radonc.2013.11.001.

[5] B. Rolland, et al., Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach, Am. J. Epidemiol. 182 (12) (Jul. 2015) 1033–1038, https://doi.org/10.1093/aje/kwv133.

[6] J. L. McCarthy, D. Warzel, E. Kendall, B. Bargmeyer, H. Solbrig, K. Keck, and F. Gey, "Data modeling and harmonization with OWL: Opportunities and lessons learned.," in *CEUR Workshop Proceedings, 524*, 2009, pp. 86–97.

[7] H. Zhang, et al., An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival, BMC Med. Inf. Decis. Making 18 (Jul. 2018), https://doi.org/10.1186/s12911-018-0636-4.

[8] C. Tao, G. Jiang, W. Wei, H. R. Solbrig, and C. G. Chute, "Towards Semantic-Web Based Representation and Harmonization of Standard Meta-data Models for Clinical Studies.".

[9] Y. Chen, et al., PCLiON: An Ontology for Data Standardization and Sharing of Prostate Cancer Associated Lifestyles, Int. J. Med. Inf. 145 (Jan. 2021), https://doi.org/10.1016/j.ijmedinf.2020.104332.

[10] O. Bodenreider, Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support, Yearb. Med. Inform. 67–79 (2008).

[11] B. Smith, et al., The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (11) (Nov. 2007) 1251–1255, https://doi.org/10.1038/nbt1346.

[12] S. El-Sappagh, F. Franda, F. Ali, K.S. Kwak, SNOMED CT standard ontology based on the ontology for general medical science, BMC Med. Inf. Decis. Making 18 (1) (Aug. 2018), https://doi.org/10.1186/s12911-018-0651-5.

[13] D. Sonntag, M. Möller, P. Ernst, Modeling the International Classification of Diseases (ICD-10) in OWL GeAR-Gelingensbedingungen beim Experimentieren mit Augmented Reality View project Kognit View project Modeling the International Classification of Diseases (ICD-10) in OWL, Article in Communications in Computer and Information Science (2013), https://doi.org/10.1007/978-3-642-29764-9-16.

[14] A. Srinivasan, N. Kunapareddy, P. Mirhaji, and S. W. Casscells, "Semantic Web Representation of LOINC: an Ontological Perspective." [Online]. Available: http://www.cdc.gov/epo/dphsi/nndsshis.htm.

[15] J. Homepage and U. Teknologi Malaysia Johor Bahru, "International Journal of Innovative Computing Development of Breast Cancer Ontology Based on Hybrid Approach Fatimatufaridah Jusoh Roliana Ibrahim Mohd Shahizan Othman Norshafarina Omar.".

[16] A. Bulzan, "Breast Cancer Grading Ontology https://bioportal.bioontology.org/ontologies/BCGO/," 2010.

[17] C. Yu, Q. Wei, and B. Shen, "Prostate Cancer Ontology, https://bioportal.bioontology.org/ontologies/PCAO," 2019.

[18] H. Min, F.J. Manion, E. Goralczyk, Y.N. Wong, E. Ross, J.R. Beck, Integration of prostate cancer clinical data using an ontology, J. Biomed. Inform. 42 (6) (Dec. 2009) 1035–1045, https://doi.org/10.1016/j.jbi.2009.05.007.

[19] X. Liu, "Thyroid Cancer Ontology, https://bioportal.bioontology.org/ontologies/TCO/," 2020.

[20] R. Messaoudi, et al., Ontology-Based Approach for Liver Cancer Diagnosis and Treatment, J. Digit. Imaging 32 (1) (Feb. 2019) 116–130, https://doi.org/10.1007/s10278-018-0115-6.

[21] M. B. Sesen, R. Banares-Alcantara, J. Fox, T. Kadir, and J. M. Brady, "Lung Cancer Assistant: An Ontology-Driven, Online Decision Support Prototype for Lung Cancer Treatment Selection.".

[22] J. Sirisha and Dr. M. B. Reddy, "An Ontology Based Expert System for Lung Cancer : OBESLC," *Int J Eng Adv Technol*, vol. 9, no. 2, pp. 4622–4626, Dec. 2019, 10.35940/ijeat.B5116.129219.

[23] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology The Gene Ontology Consortium*," 2000. [Online]. Available: http://www.flybase.bio.indiana.edu.

[24] S. Carbon *et al.*, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Res*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, 10.1093/nar/gky1055.

[25] A. Traverso, J. van Soest, L. Wee, and A. Dekker, "The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. 45(10)," *Med Phys*, pp. e854–e862, 2018.

[26] A. Traverso, "Radiomics Ontology, http://bioportal.bioontology.org/ontologies/RO," 2017.

[27] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, Yearb. Med. Inform. (2008) 67–79.

[28] A. Galopin, J. Bouaud, S. Pereira, B. Seroussi, An Ontology-Based Clinical Decision Support System for the Management of Patients with Multiple Chronic Disorders, Stud. Health Technol. Inform. 216 (2015) 275–279.

[29] S. Cavalieri, et al., Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: The Big Data to Decide EU Project, Head Neck (2020), https://doi.org/10.1002/hed.26515.

[30] B.A.C. Van Dijk, et al., Rare cancers of the head and neck area in Europe, Eur. J. Cancer 48 (6) (Apr. 2012) 783–796, https://doi.org/10.1016/j.ejca.2011.08.021.

[31] R.C. Nickerson, U. Varshney, J. Muntermann, A method for taxonomy development and its application in information systems, Eur. J. Inf. Syst. 22 (2013) 336–359.

[32] L. Hernández, L. Lopez-Perez, A.M. Ugena, M.T. Arredondo, G. Fico, Designing an ontology for Head and Neck Cancer research, in: IEEE EMBS International Conference on Biomedical & Health Informatics, 2019, pp. 1–5.

[33] A. Bandrowski *et al.*, "The Ontology for Biomedical Investigations," *PLoS One*, vol. 11, no. 4, Apr. 2016, 10.1371/journal.pone.0154556.

[34] D. Salvi, et al., Merging person-specific bio-markers for predicting oral cancer recurrence through an ontology, I.E.E.E. Trans. Biomed. Eng. 60 (1) (2013) 216–220.

[35] T. Poli, et al., "Biomarkers in NeoMark European Project for Oral Cancers", Biomark, Cancer (2014) 1–19, https://doi.org/10.1007/978-94-007-7744-6_12-1.

[36] N.F. Noy, M. Sintek, S. Decker, M. Crubezy, R. Fergerson, Creating Semantic Web Contents with Protégé-2000, IEEE Intell. Syst. 16 (2) (2001) 60–71, https://doi.org/10.1109/5254.920601.

[37] "NCI Dictionary of Cancer Terms. Accessed 2021. https://www.cancer.gov/publications/dictionaries/cancer-terms/".

[38] "Home - MeSH - NCBI. Accessed 2022. https://www.ncbi.nlm.nih.gov/mesh/".

[39] "Cancer Glossary Resources THANC Guide. Accessed 2022 https://thancguide.org/resources/glossary/".

[40] K. Bjordal *et al.*, "Quality of Life in Head and Neck Cancer Patients: Validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35," 1999.

[41] R. Rabin, F. De Charro, EQ-5D: A measure of health status from the EuroQol Group, Ann. Med. 33 (5) (2001) 337–343, https://doi.org/10.3109/07853890109002087.

[42] R.K. Saripalle, Fast health interoperability resources (FHIR): Current status in the healthcare system, Int. J. E-Health Med. Commun. 10 (1) (Jan. 2019) 76–93, https://doi.org/10.4018/IJEHMC.2019010105.

[43] S. Mathur and J. Sutton, "Personalized medicine could transform healthcare (Review)," *Biomedical Reports*, vol. 7, no. 1. Spandidos Publications, pp. 3–5, 2017. 10.3892/br.2017.922.

[44] L. Nissi, S. Suilamo, E. Kytö, S. Vaittinen, H. Irjala, H. Minn, Recurrence of head and neck squamous cell carcinoma in relation to high-risk treatment volume, Clin Transl Radiat Oncol 27 (Mar. 2021) 139–146, https://doi.org/10.1016/j.ctro.2021.01.013.

[45] J. Ringash, "Quality of Life in Head and Neck Cancer: Where We Are, and Where We Are Going," *International Journal of Radiation Oncology Biology Physics*, vol. 97, no. 4. Elsevier Inc., pp. 662–666, Mar. 15, 2017. 10.1016/j.ijrobp.2016.12.033.

[46] D. J. Patil and R. Nagaraju, "Personalised Precision Medicine-a Novel Approach for Oral Cancer Management." [Online]. Available: www.intechopen.com.

[47] E. Ong *et al.*, "Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project," *Nature Reviews Nephrology*, vol. 16, no. 11. Nature Research, pp. 686–696, Nov. 01, 2020. 10.1038/s41581-020-00335-w.

[48] M. Thandi, S. Brown, S.T. Wong, Mapping frailty concepts to SNOMED CT, Int. J. Med. Inf. 149 (May 2021), https://doi.org/10.1016/j.ijmedinf.2021.104409.

[49] A. Benis, et al., Medical informatics and digital health multilingual ontology (MIMO): A tool to improve international collaborations, Int. J. Med. Inf. 167 (Nov. 2022), https://doi.org/10.1016/j.ijmedinf.2022.104860.

[50] H. Zhang, et al., A scoping review of semantic integration of health data and information, Int. J. Med. Inf. 165 (Sep. 2022), 104834, https://doi.org/10.1016/j.ijmedinf.2022.104834.

[51] M. Amith, Z. He, J. Bian, J.A. Lossio-Ventura, C. Tao, Assessing the practice of biomedical ontology evaluation: Gaps and opportunities, J. Biomed. Inform. 80 (2018) 1–13, https://doi.org/10.1016/j.jbi.2018.02.010.

[52] ERA-LEARN, "Project: Supporting Personalized Treatment Decisions in Head and Neck Cancer through Big Data" SuPerTreat (Reference Number: ERAPERMED2019-281).

[53] F.Z. Smaili, X. Gao, R. Hoehndorf, Formal axioms in biomedical ontologies improve analysis and interpretation of associated data, Bioinformatics 36 (7) (Apr. 2020) 2229–2236, https://doi.org/10.1093/bioinformatics/btz920.

[54] S. Althubaiti, Ş. Kafkas, M. Abdelhakim, and R. Hoehndorf, "Combining lexical and context features for automatic ontology extension," J. Biomed. Semant. vol. 11, no. 1, Jan. 2020, 10.1186/s13326-019-0218-0.