Journal of Pathology Informatics 14 (2023) 100332

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Modelling digital health data: The ExaMode ontology for computational pathology

Laura Menotti [a], Gianmaria Silvello [a,*], Manfredo Atzori [b,c], Svetla Boytcheva [d], Francesco Ciompi [e], Giorgio Maria Di Nunzio [a], Filippo Fraggetta [f], Fabio Giachelle [a], Ornella Irrera [a], Stefano Marchesin [a], Niccolò Marini [b], Henning Müller [b], Todor Primov [d]

[a] Department of Information Engineering, University of Padua, Padova, Italy
[b] Information Systems Institute, University of Applied Sciences Western Switzerland, Delémont, Switzerland
[c] Department of Neuroscience, University of Padua, Padova, Italy
[d] Sirma AI, Sofia, Bulgaria
[e] Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[f] Pathology Unit Gravina Hospital Caltagirone ASP, Caltagirone, Italy

ARTICLE INFO

Keywords:
Computational pathology
Ontology
Semantic integration
Histopathology

ABSTRACT

Computational pathology can significantly benefit from ontologies to standardize the employed nomenclature and help with knowledge extraction processes for high-quality annotated image datasets. The end goal is to reach a shared model for digital pathology to overcome data variability and integration problems. Indeed, data annotation in such a specific domain is still an unsolved challenge and datasets cannot be steadily reused in diverse contexts due to heterogeneity issues of the adopted labels, multilingualism, and different clinical practices.

Material and methods: This paper presents the ExaMode ontology, modeling the histopathology process by considering 3 key cancer diseases (colon, cervical, and lung tumors) and celiac disease. The ExaMode ontology has been designed bottom-up in an iterative fashion with continuous feedback and validation from pathologists and clinicians. The ontology is organized into 5 semantic areas that defines an ontological template to model any disease of interest in histopathology.

Results: The ExaMode ontology is currently being used as a common semantic layer in: (i) an entity linking tool for the automatic annotation of medical records; (ii) a web-based collaborative annotation tool for histopathology text reports; and (iii) a software platform for building holistic solutions integrating multimodal histopathology data.

Discussion: The ontology ExaMode is a key means to store data in a graph database according to the RDF data model. The creation of an RDF dataset can help develop more accurate algorithms for image analysis, especially in the field of digital pathology. This approach allows for seamless data integration and a unified query access point, from which we can extract relevant clinical insights about the considered diseases using SPARQL queries.

Introduction

In the past 20 years, we have witnessed a significant increment in the volume and complexity of the data produced in the life science domain.[1] The progress in the high-throughput experimental techniques and the development of new diagnostic methods, therapies, and medications have become a precious information flow used to support medical applications.[2] Representing information as an open, machine-readable, and shared ontology can enhance data reuse and limit issues related to data heterogeneity in the text produced in clinical practice.[3] To this end, an ontology is a semantic data model defining the types of concepts and objects that exist in a given domain or subject area, as well as the properties that can be used to describe them. It offers a shared representation of a domain of interest concepts and relationships that can be read and, ideally, understood by computers.[4–6] Nowadays, ontologies are used in many applications relying on domain-specific terms, including Natural Language Processing (NLP)[1,5] and deep learning.[7]

Much effort has been made to produce standard ontologies and thesauri in the medical and biological domains.[5,8] In this context, BioPortal[1] [9] and Ontobee[2] [10] are 2 of the most comprehensive repositories of biomedical ontologies, comprising over 700 biomedical ontologies including the widely adopted Open Biomedical Ontologies (OBO). National Cancer Institute Thesaurus (NCIT)[3] [11,12] has been developed by the National Cancer

* Corresponding author.
E-mail address: gianmaria.silvello@unipd.it (G. Silvello).

[1] https://bioportal.bioontology.org/.
[2] https://ontobee.org/.
[3] http://purl.obolibrary.org/obo/ncit.owl.

http://dx.doi.org/10.1016/j.jpi.2023.100332
Received 9 May 2023; Received in revised form 14 July 2023; Accepted 16 August 2023
Available online 22 August 2023

Institute to provide reference terminology in the biomedical domain. It is a widely recognized resource for biomedical reference since it covers a wide range of vocabulary for clinical care, especially for cancer and related diseases. NCIT is not a proper ontology, but it is more of a nomenclature with ontological features. Medical Subject Heading (MeSH)[4] is a thesaurus produced by the U.S. National Library of Medicine (NLM) which comprises health-related information of subject headings that appears in MEDLINE/PubMed, the NLM catalog, and other NLM databases. Unified Medical Language System (UMLS)[5] is a metathesaurus organized by meaning, and it links synonyms for the same concept from nearly 200 different vocabularies. It aims to promote the creation of more effective and interoperable biomedical information systems and services.[13] One of the main reasons to develop an ontology is to model a part of reality that can be imported and reused in other ontologies, where we extend the existing one to represent our particular domain and task best.[14] Thus, one can use external ontologies to model some medical and clinical aspects of general purpose and define components specific to the considered domain.

Among the different disciplines in the life science field, *Computational Pathology* can significantly benefit from ontologies to standardize the employed terminology and concepts and help with (semi-)automatic knowledge extraction processes.[7] Computational pathology is an emerging domain centered on computer-assisted diagnosis tools to automatically analyze histopathology data in images and text. This field revolves around *digital pathology*, a process where specialized hardware is used to generate substantial high-resolution digital images—i.e., whole slide images (WSI)—of histological sections. WSIs can be processed by image analysis tools[15] to train deep learning (DL) models[16] which can aid the diagnostic process.[7,17] However, the few large image datasets publicly available usually have some drawbacks preventing their use to train DL algorithms since annotations can be few, sparse, unbalanced,[16] and highly variable depending on the pathologist who performed them. Across different datasets, the same concepts might be annotated with variable labels hampering machine interpretability, interoperability, and reuse of training and test data.[18] One recently explored solution for constructing annotated image datasets is to employ the diagnostic text (containing diagnoses and other expert observations) often associated with the WSI slides to extract so-called weak labels to automatically annotate the WSI.[19,20] Currently, this process is often manual and time-consuming due to the high volumes of data and the lack of a standard shared terminology and structure between institutions.[21,22] Indeed, automatic knowledge extraction algorithms need a common and machine-readable reference point to understand the concepts identified in the text. To this end, an ontology for digital pathology can overcome data heterogeneity and integration problems as it provides a shared terminology that can be used to produce high-quality annotated image datasets.

Little effort has been made to develop specific ontologies in the *Digital Pathology* area [16]. Serra et al.[4] developed the Cancer Cell Ontology (CL), which represents a variety of vertebrate cell types with special attention to hematopoietic cell types. However, this ontology focuses on diagnosing hematologic malignancies thus, it does not provide an ontology for histopathology diagnostics. Gurcan et al[23] introduce the Quantitative Histopathology Image Ontology (QHIO), an ontology for Quantitative Histopathology Image (QHI) that eases interoperability across datasets of pathological image data. This resource integrates different types and subtypes of pathological images with imaging processes and techniques. However, it focuses on images and not on diagnosis. SNOMED-CT[6] is an international clinical reference terminology[24–26] which was appointed as the best ontology for annotation labeling of WSIs in Lindman et al.[16] Nevertheless, SNOMED-CT is a general-purpose biomedical ontology that does not specifically model the diagnosis of histopathology images.

This paper presents the ExaMode ontology,[27] modeling the diagnosis process using WSIs reports in histopathology. Compared to previous efforts, the ExaMode ontology focuses on diagnosing histopathology exams, defining components related to the annotation process of WSIs. Moreover, the proposed ontology is *multilingual* since components are labeled in 3 different languages: English, Italian, and Dutch. It comprises 5 semantic areas grouping components related to the same aspect of the diagnostic process: clinical case reports (i.e., general aspects), diagnosis results, other tests performed, interventions or surgical procedures employed to retrieve the specimen, and the anatomical location of the specimen. This classification provides an ontological template that can be used to model any disease in the histopathology domain. We modeled 4 largely diffused and studied diseases: 3 cancer diseases (colon, cervical, and lung tumors) and celiac disease. However, thanks to the modularity of this approach, we can easily expand our ontology to include more use cases. The ontology design followed a bottom-up approach starting from anonymized clinical reports about the 4 considered diseases provided by Azienda Ospedaliera per l'Emergenza Cannizzaro (AOEC) in Italy and the Radboud University Medical Center (RUMC) in The Netherlands. We analyzed these textual records and worked together with the pathologists and physicians, employing a co-design methodology to accurately identify the classes and relations to be included in the ontology. The ExaMode ontology is designed to meet the OBO principles to ease interoperability with other biomedical ontologies. Hence, we maximized the reuse of concepts defined in already available and well-known biomedical ontologies and vocabularies, limiting the creation of new classes and relations to a minimum. The ontology was developed in the context of the Extreme-scale Analytics via Multimodal Ontology Discovery & Enhancement (ExaMode) project[7], co-financed by the European Commission[8]. This project aims to allow weakly supervised knowledge discovery of multimodal heterogeneous data, limiting human interaction. Important pathological concepts are extracted from medical reports, used to weakly annotate WSIs associated with the records themselves, and used to train prediction algorithms such as convolutional neural networks (CNNs) which finally translate to healthcare decision-making applications.

The rest of the paper is organized as follows: The "Methods" Section introduces the domain requirements, the methodology used to design the ontology, and the source datasets. The "ExaMode Ontology" subsection presents a detailed overview of the ExaMode ontology. The "Results" Section describes downstream applications where the ontology is currently employed as a common semantic layer. The "Discussion" Section illustrates how histopathology diagnoses are modeled in the ontology and provides some relevant queries for the clinicians supported by the ontology. Finally, The "Conclusions and future work" Section presents some final remarks and future work.

Methods

Domain requirements

ExaMode focuses on *histopathological diagnosis* of tissues to detect cancer-related diseases. Taking into account the future cancer incidence and mortality burden worldwide, which is predicted to be increasing by 62% from today until 2040,[28] it has been decided to focus on 3 cancer diseases and 1 non-cancerous disease: (i) colon cancer; (ii) cervix cancer; (iii) lung cancer; and (iv) celiac disease.

Colon cancer

The estimated number of colon cancer incidence from today to 2040 will increase by up to 75%, for both sexes and all ages.[29] The American Cancer Society (ACS) recommends regular screening for colon cancer for people over 45 years.[30] The screening can be done either with a stool-based molecular test or a visual exam. At this stage, the screening process does not include a histopathological examination. The majority of colorectal cancers derives from precursor lesions which can be identified using endoscopic procedure (colonoscopy), leading to the excision of these lesions, known as *polyps*.[31] Good endoscopic practice, together with an accurate

[4] https://www.nlm.nih.gov/mesh/meshhome.html.
[5] https://www.nlm.nih.gov/research/umls/index.html.
[6] https://www.nlm.nih.gov/healthit/snomedct/index.html.
[7] https://www.examode.eu/.
[8] https://cordis.europa.eu/project/id/825292.

histopathological diagnosis, decreases the incidence of colorectal cancer. There are different precursor lesions with different diagnostic and prognostic significance.[32]

The main task for a pathologist is to detect cancerous polyps (e.g., for population screening) and to identify the degree of dysplasia. Moreover, in the microscopic analysis of colon excisional biopsy sample, pathologists provide data about:

1. *Type* of the polyp. It is important to distinguish between the 2 main types of polyps: adenoma-serrated polyps and malignant polyps. From the diagnostic perspective, it can also be useful to know the *number* of each type of polyp.
2. The *grade* of the dysplasia if a polyp is present (low, medium, or high grade).
3. In case of malignant polyps (considered as cancer, e.g., colon adenocarcinoma or metastatic adenocarcinoma), several critical *histological features* need to be assessed, which include tumor type, histological tumor grade, lymphovascular invasion, and margin involvement.

This information is a prognostic factor leading to the decision about the patient's management. For example, polyps with a negative polypectomy margin, low-grade histology, and no lymphovascular invasion can be safely treated with endoscopic polypectomy. On the other hand, positive margin, high-grade (poorly differentiated) histology, and lymphovascular invasion are associated with an increased risk of adverse outcomes, and surgical resection is indicated.[33]

Lung cancer

According to the International Agency for Research on Cancer, lung cancer in both sexes and at all ages is estimated to increase by 72% from today to 2040. In general, lung cancer is the second most common cancer in both men and women—about 13% of all new cancers diagnoses are lung cancers. Moreover, lung cancer is the leading cause of cancer death among men and women (18% of all cancer deaths), being the leading cause of cancer death in men.[34,35] The average survival rate for metastatic lung cancer is very low, whereas early stages have higher survival rates. The treatment of low-stage lung cancer is complete surgical resection. Instead, for metastatic lung cancer, the surgical option is often impossible. An accurate diagnosis from lung biopsies targets the most correct prognostic and therapeutic management for the patient.

The identification of new therapeutic targets over the past decade resulted in an urgent need for a classification system for non-resection specimens (particularly small biopsies) and cytology samples. For this reason, an accurate and specific pathology report is important to establish the diagnosis and patient's treatment. Starting from the analysis of lung biopsies, the microscopic analysis section of the clinical report on lung cancer biopsy samples must provide the following information:

1. Histologic type.
2. Histologic grade.
3. Spread Through Air Spaces (STAS)—information about the presence of micropapillary clusters, solid nests, or single cells of tumor extending beyond the edge of the tumor into the air spaces of the surrounding lung parenchyma.
4. Visceral pleura invasion.
5. Direct invasion of adjacent structures.
6. Margins—information about the involvement of the tissue margins, indicating a negative outcome.
7. Lymphovascular invasion—provides information about vascular/lymphatic vessel invasion.
8. Pathologic stage classification—based on the classification system proposed by the WHO.
9. Extranodal extension—indicates the presence of metastasis.

Uterine cervix cancer

Cervical cancer is the fourth most common cancer in women and the eighth most commonly occurring cancer overall, with an estimated 604 k new cases in 2022.[36] Worldwide, approximately 570 k cases of cervical cancer and 311 k deaths from the disease occurred in 2018.[37] The ACS recommends regular screening starting at age 25 through age 65, involving an Human Papilloma Virus (HPV) test. Indeed, a strong association between cervical precursor lesions and HPV infection has been observed, where low-grade squamous intraepithelial lesion (LSIL) is strongly associated with low intermediate-risk HPV, and high-grade squamous intraepithelial lesion (HSIL) is associated with high-risk HPV.[38] Therefore, the first feature that has to be identified and reported is the presence and the grade of dysplasia with possible HPV association.

The cervical biopsy (colposcopy) is a procedure done when previous tests provide evidence of precancerous/abnormal or neoplastic lesions in the uterine cervix. The cervical tissue removed has to be analyzed by an expert pathologist to identify if the tumor lesions are present or not. If present, the pathology report provides diagnostic information and works as a prognostic tool for the patient's treatment. Thus, the histopathologist aims to recognize and identify these precursor lesions, known as cervical intraepithelial neoplasia (CIN), which displays the proliferation of atypical basaloid cells.[39] Based on proliferation spread, the World Health Organization (WHO) classification categorizes cervical dysplasia into 3 grades: CIN1 (mild dysplasia) corresponds to LSIL, whereas CIN2 (moderate dysplasia) and CIN3 (severe dysplasia or carcinoma in situ) correspond to HSIL.

In the presence of cervical carcinoma, main microscopic features and measurements of uterine cervix colposcopy biopsy are identified, and they should be provided in the pathology report:

1. Histologic type.
2. Histologic grade.
3. Stromal invasion—provides information about cancer invasion into stromal tissue.
4. Margins—indicate a negative outcome.
5. Lymphovascular invasion—provides information about vascular/lymphatic vessel invasion.

Also, the immunohistochemistry (p16 and Ki-67 staining) assists in the differential histological diagnosis of precursors to the reactive and metaplastic epithelium. For invasive cervical carcinoma, the stage is the strongest prognostic factor.[39]

Celiac disease

Celiac disease is now recognized as a global disease affecting about 0.7% of the world's population,[40] resulting from environmental (gluten) and genetic factors.[41] For these reasons, celiac disease was chosen as a non-cancerous disease to be included in the ExaMode priority list. Celiac disease is an immune-mediated disease with chronic outcome and genetic predisposition to an intolerance to gluten and its proteins. This intolerance leads to an abnormal immune response, followed by chronic inflammation and alteration of the small intestinal mucosa. The diagnosis of this pathology is based on the description of the histopathological alterations of the small intestine (after duodenal biopsy) by expert pathologists.[42]

Microscopic analysis of small colon biopsy sample for celiac disease provides information about:

1. Orientation of biopsy—indicates biopsy position on cellulose acetate filter and is very important for the diagnostic criteria.
2. Normal intestinal mucosa description—includes information about: villi, enterocytes, intra-epithelial lymphocytic infiltrate, and glandular crypts. The absence or alteration of these structures must be reported.
3. Pathological intestinal mucosa—including features that have to be reported and well-described, with particular attention to increased intraepithelial T lymphocytes, decreased enterocyte height, crypt hyperplasia, and villous atrophy.
4. Pathologic Stage Classification—based on the classification system proposed by Marsh-Oberhuber[43,44] and Corazza-Villanacci,[45] in presence of intestinal mucosa alterations previously described.

Design principles

The ExaMode ontology has been developed in the context of the *ExaMode* Project[9], where different European partners were involved. For this reason, we adopted a co-design approach, collaborating with pathologists and clinicians to embed their knowledge in the ontology and validate the design choices. To achieve this goal, we iteratively discussed with medical partners to validate newly defined components so that they correctly describe the corresponding real-world concepts.

In this section, we describe how the ExaMode ontology complies with the OBO principles[10].

- **Open**: The ontology is publicly available on the documentation web page[11] and Zenodo[12], in 3 different serialization formats: RDF/XML, Turtle, and JSON-LD.
- **Common Format**: The ontology is defined according to the OWL 1.2 *Common Format*.
- **URI/Identifier space**: All components defined in the ontology are identified by the namespace https://w3id.org/examode/ontology/. In the following sections, we use the prefix "`exa`" when referring to the ExaMode ontology namespace.
- **Versioning**: Different versions are described as part of the documentation of the Examode ontology. In particular, for each version of the ontology, we provided an ontology version IRI in the documentation and previous versions can be accessed in zenodo.
- **Scope**: The ExaMode ontology is meant to model the histopathology process of diagnosis using WSIs, focusing on 4 use cases: 3 cancer diseases (colon, cervix, and lung tumor), and celiac disease.
- **Textual definitions**: We associate textual definitions with the definition source with each ontology component to favor its reusage.
- **Relations**: None of the relations defined in the ExaMode ontology have the same meaning of *Relations* available in the Relations Ontology (RO). Thus, they could not have been replaced with one of the RO.
- **Documentation**: A detailed documentation of the ontology is available on its web page.
- **Documented plurality of users and Commitment to collaboration**: The ExaMode ontology has been developed in the context of the ExaMode Project, which includes multiple European partners. Thus, its collaborative nature is intrinsic to the development and usage of ontology.
- **Locus of authority**: The ExaMode ontology identifies its *Locus of authority* into its creators, who are indicated in the documentation of the ontology.
- **Naming conventions**: We define the naming conventions followed during ontology design in Section 2.4.
- **Maintenance**: The ExaMode consortium is working on the ontology's maintenance, and a European research grant supports it.

Source datasets

The ExaMode ontology has been developed in collaboration with 2 European medical centers namely AOEC and RUMC. The AOEC clinical reports include diagnostic reports written in Italian for colon, cervix, lung cancer, and celiac disease. The RUMC clinical case reports are written in Dutch and consist of diagnostic reports concerning colon and cervix cancer cases. Each diagnosis is associated with the corresponding WSIs in both cases, allowing a one-to-many correspondence between diagnoses and WSIs. Note that the case reports span several years, different pathologists, and work modalities—e.g., RUMC reports are created via speech recognition, whereas AOEC reports are usually more structured and concise. Diagnostic reports contain the outcome of pathology tests, and they follow the College of American Pathologists (CAP) international guidelines[13] for pathology reports.[46,47] Each report comprises the patient's personal and clinical-specific information, the description of the analyzed specimen at the microscope, and the final histopathology diagnosis. We provide some samples of the diagnostic reports we handled in Section 4.1; the provided reports have been anonymized and modified to prevent any information leakage about patient data. Table 1 reports the total number of diagnostic reports for each of the 4 diseases we consider and each medical center. This work presents the ExaMode ontology, which models the histopathology diagnostic process by analyzing data from AOEC and RUMC. In this study, we do not release any diagnostic reports from the medical centers that participated in the ontology design because the data is confidential.

Implementation

This section describes the shared guidelines we adopted to design the ontology components concerning external referencing, annotation properties, naming conventions, and modeling choices. When modeling medical information, it is preferable to reuse entities and properties already defined in other ontologies[48] or thesauri to provide correspondence with largely adopted medical terminology. For this reason, we maximized the reuse of concepts defined in well-known biomedical knowledge resources, limiting the creation of new classes and relations to a minimum. Under the term "knowledge resource" lies a multitude of different and heterogeneous resources, which share common characteristics. Depending on the complexity of the underlying model and the relations considered, different types of knowledge resources can be defined. In this context, we focus on 4 types: nomenclatures, thesauri, taxonomy, and ontologies. Note that we use the terms "terminology" and "vocabulary" as synonyms for "nomenclature". A nomenclature (lit. "list of names") is a naming system for a given domain, formed according to strict linguistic rules. Nomenclatures are composed of terms collected by domain specialists and approved by scientific authorities. The purpose of nomenclatures is to standardize the use of the domain language to avoid ambiguity. A thesaurus is a controlled vocabulary and terminology, which denotes concepts and relations in a specific domain or subject area. It consists of systematized lists of synonyms, antonyms, and otherwise-related terms. Thesauri use preferred terms to refer unambiguously to concepts, avoiding the need to impose additional model constraints. Thesauri can form part of ontologies. Terms can be grouped in a taxonomy that formalizes the hierarchical relations among concepts. There are many definitions of an ontology. For the purposes of this paper, referring to https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html, an ontology is a formal explicit description of concepts in a domain of discourse (classes), and properties of each concept describing various features and attributes of the concept.

In general, external referencing can be realized by using the URI of the term in the original taxonomy or by using an annotation property to state the mapping between the 2 URIs. In the ExaMode ontology, we used the ExaMode namespace to define all classes. All the concepts also defined in relevant external vocabulary are related to the ExaMode class via the "`rdfs:isDefiniedBy`" annotation property. We defined all the other components—e.g., individuals and annotation properties—by using the URI of the taxonomy of reference. The main external reference we relied on is NCIT.[12] NCIT is used as a reference thesaurus in a wide variety of ontologies in the medical domain.[49–52] We also employed other well-known resources when no information is available in NCIT or to ensure compatibility with already-developed tools within the ExaMode consortium.[14] Besides NCIT, the external resources used in the ExaMode ontology are Mammalian Phenotype Ontology (MP)[15,53] BRENDA Tissue Ontology (BTO)[16,54] MONDO

[9] https://www.examode.eu/.
[10] The description of OBO principles is available at https://obofoundry.org/principles/fp-000-summary.html.
[11] http://examode.dei.unipd.it/ontology/.
[12] https://doi.org/10.5281/zenodo.7669237.
[13] https://www.cap.org/protocols-and-guidelines.
[14] In this case, we add the reference to the NCIT term using the annotation property "rdfs:seeAlso".
[15] https://www.ebi.ac.uk/ols/ontologies/mp.
[16] https://www.ebi.ac.uk/ols/ontologies/bto.

Table 1
ExaMode data size. For each medical center, we report the number of provided diagnostic reports for each of the 4 diseases we considered. Symbol "–" indicates that the medical center did not share data about the specific use case.

	Language	Colon cancer	Cervix cancer	Lung cancer	Celiac disease
AOEC	Italian	4020	4810	2077	1965
RUMC	Dutch	14 147	5861	–	962

Disease Ontology[17,55] Human Phenotype Ontology (HPO)[18,56] Gene Ontology (GO)[19,57,58] UNIPROT[20,59] UBERON,[21 60] and Foundational Model of Anatomy (FMA).[22 61]

All components of the ExaMode ontology have metadata in the form of annotation properties. We defined a list of essential properties to add when a new resource is added to the ontology. Table 2 reports all the required annotation properties and their values for the example class "Surgical Procedure". We require these metadata to be filled for each class, but we can have classes with additional properties based on specific cases. We recall the ExaMode ontology is a multilingual resource. Hence, each component has 3 different labels corresponding to the 3 languages we model: English ("@en"), Italian ("@it"), and Dutch ("@nl"). We also add a comment for each class that briefly defines the concept and its source, if available. As we explained above, if the class we defined has a corresponding term in NCIT or in any of the resources we listed before, we also add a reference to the term URI with the annotation property "`rdfs:isDefinedBy`". Most biomedical vocabularies are defined in the UMLS Metathesaurus with a Concept Unique Identifier (CUI). Thus we map each class to the corresponding CUI using the property "`dcterms:conformsTo`", defined by the Dublin Core (DC) Metadata Terms.[23] Finally, we recall our ontology is divided into 5 semantic areas. Thus, we assert each class's semantic area with the custom annotation property "`exa:hasSemanticArea`". Regarding other components, we still require the 3 language labels, the comment, the semantic area, and the UMLS mapping if available. Even though the ontology models 4 specific diseases, its modularity enables the expansion to all the other diseases in the field. To this end, we created an annotation property called "`exa:associatedDisease`" which is used in all components created for a specific use case to differentiate between the different diseases. For instance, "Colon biopsy" is a component specifically instantiated for the colon use case; thus it is associated to "Colon cancer."

We used self-explanatory labels for objects, annotation, and data properties regarding naming conventions. In particular, object properties have labels where we include the property range, and the comment explains the relationship between the 2 classes involved. For instance, the object property "`exa:detectedHPV`" describes the object property that links an outcome from a cervix clinical case to the HPV. Similarly for data properties, from the property name "`exa:duodenitisSeverity`" we can already infer its meaning. Concerning data properties, we also provide consistency in datatypes, for instance, all literal properties have datatype "`xsd:string`".

As mentioned, we use external taxonomies and thesauri to model common terminology in the medical domain. In the ExaMode ontology, the taxonomies of terms are modeled by employing the Simple Knowledge Organization System (SKOS) data model[24]; thus following a consolidated modeling practice in the semantic web community. The list of taxonomies modeled in the ExaMode ontology via the SKOS data model is listed in Table 3. Given a specific taxonomy, say the positive outcomes for colon cancer diagnoses, we model a single term, say "Adenocarcinoma", as a named individual belonging to 2 classes: "`skos:Concept`" and "`exa:ColonPositiveType`", which is the class describing for which type of colon cancer the clinical case tested positive. This is possible because, in these cases, we are interested in modeling the abstract concept behind the medical term. The ExaMode ontology models diagnostic reports of pathology exams; hence we have no additional information about the specific disease affecting a patient or the procedure performed to obtain the specimen. We employed such a design principle for each class that refers to a set of abstract terms for which we do not have additional information describing the peculiarity of the specific term for the specific clinical case, i.e., ontology components without specific object or data properties. This approach reduces the complexity of our queries and allows us to save space since we reduce the number of needed URIs.

To better explain this design choice, Fig. 1 shows how the taxonomy related to the "Cervix Surgery Type" is modeled. We can see that we are considering 4 terms, modeled as named individuals belonging to classes "`skos:Concept`" and "`exa:CervixSurgeryType`". "Hysterectomy" is a broader concept than "Radical Hysterectomy" and "Total Abdominal Hysterectomy"; this hierarchical dependency is modeled with the "`skos:broaderTransitive`" property. "Conization" is another type of cervix surgery, but unrelated to hysterectomy. The SKOS data model allows us to express the hierarchical relationships between the taxonomy terms. On top of this, all the concepts are also related to class "`exa:CervixSurgeryType`" which defines to which taxonomy these concepts belong.

Table 2
List of the mandatory annotation properties for the ExaMode classes. We report the values for the example class "Surgical Procedure" (URI: https://w3id.org/examode/ontology/SurgicalProcedure).

Annotation property	Value
rdfs:label	"Surgical Procedure"@en
rdfs:label	"Intervento Chirurgico"@it
rdfs:label	" Chirurgische Ingreep"@nl
rdfs:comment	A diagnostic or treatment procedure performed by manual and/or instrumental means, often involving an incision and the removal or replacement of a diseased organ or tissue; of or relating to or involving or used in surgery or requiring or amenable to treatment by surgery. [Definition Source: NCI]
rdfs:isDefinedBy	http://purl.obolibrary.org/obo/NCIT_C15329
dcterms: conformsTo	http://linkedlifedata.com/resource/umls/id/C0543467
exa: hasSemanticArea	https://w3id.org/examode/ontology/procedure

Table 3
List of classes modeled using the SKOS data model. For each class, we specify the corresponding semantic resource of reference which we use to define the majority of concepts.

Class	Resource of reference
Gender	National Cancer Institute Thesaurus (NCIT)
Onset	Human Phenotype Ontology (HPO)
Disease	Mondo Disease Ontology (MONDO)
Intervention type	National Cancer Institute Thesaurus (NCIT)
Anatomical entity	UBERON, Foundational Model of Anatomy (FMA)
Positive outcome type	National Cancer Institute Thesaurus (NCIT)
Finding	National Cancer Institute Thesaurus (NCIT)

The ExaMode ontology

The ExaMode ontology has been developed using Protégé editor[25 62] and is publicly available in several serialization formats along with its documentation.[26] In the following section, we provide an overview of the ExaMode ontology, focusing on the general aspects concerning each

[17] https://www.ebi.ac.uk/ols/ontologies/mondo.
[18] https://hpo.jax.org/app/.
[19] http://geneontology.org/.
[20] https://www.uniprot.org/.
[21] https://www.ebi.ac.uk/ols/ontologies/uberon.
[22] https://bioportal.bioontology.org/ontologies/FMA.
[23] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[24] https://www.w3.org/TR/2009/REC-skos-reference-20090818/.
[25] https://protege.stanford.edu/.
[26] http://examode.dei.unipd.it/ontology/.

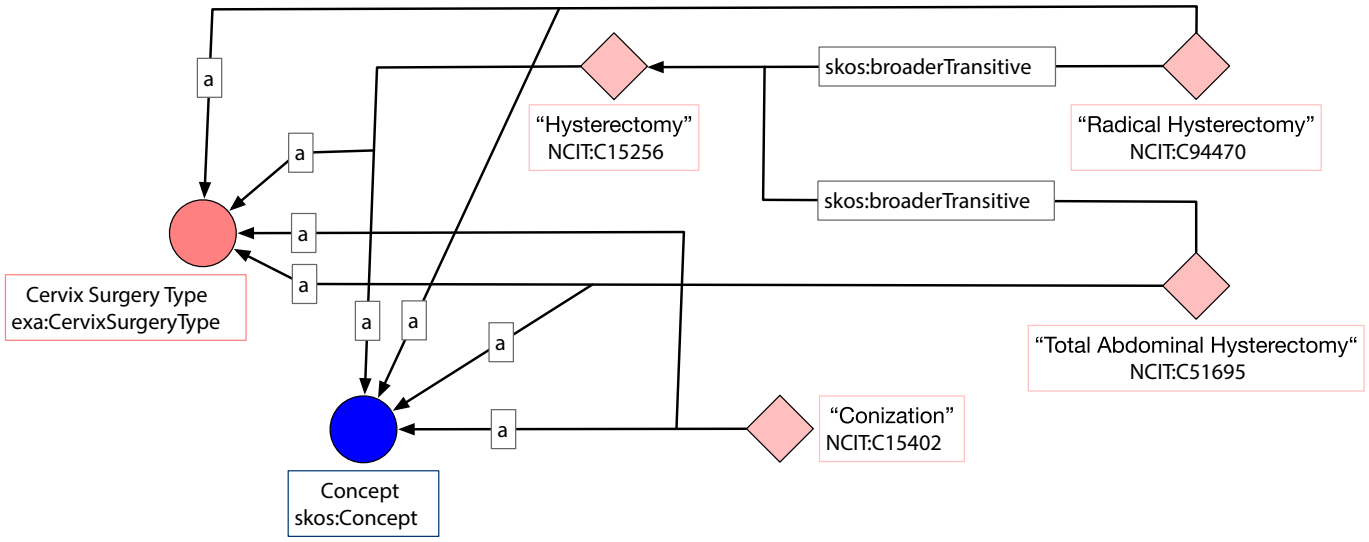

Fig. 1. Example of a taxonomy modeled following the SKOS data model. The hierarchical relationships between the terms are asserted using the object property "skos: broaderTransitive". Classes are represented with circles, while rhombuses identify named individuals.

semantic area to describe the modeling schema underlying any use case. To clarify how we deployed such a model in practice, for each semantic area we illustrate a use case in detail and provide some insights on the other use cases' peculiarities. The online documentation provides a detailed description of all the ontology components for all the modeled diseases.

The ExaMode ontology is organized into 5 semantic areas, each concerning different aspects of the histopathology process. The *General Area* comprises entities that are common to all use cases, such as patients' information and clinical case reports. We report the schema of the general area referring to colon cancer in Fig. 2, yet the following reasoning applies to all use cases. As we mentioned before, to ensure modularity in the ExaMode ontology, each use case has its own classes which are subclasses of a general-purpose class. For instance, clinical case reports concerning colon cancer are instances of the Class "`exa:ColonClinicalCaseReport`", which is a subclass of the general-purpose class "`exa:ClinicalCaseReport`". The same reasoning is applied to most classes, such as outcomes, interventions, and locations. Therefore, each clinic case report is classified based on the use case and we associated the corresponding disease through the object property "`exa:isAboutDisease`". Clinical case reports are described by means of some data properties concerning the corresponding diagnosis text, an identifier, and information about the associated images. Each report is associated with a patient that presents information about gender and the age at which they manifested the first symptoms, which complies with the domain requirements. Indeed, since the ExaMode ontology models the histopathology diagnosis process, the focus is on the clinical case report rather than on the patient. We link the WSIs corresponding to the specific clinical case report through the object property "`exa:hasSlide`", which ranges to instances of the class "Slide Device", describing the specific WSI. Finally, we associate each clinical case report to the organization that provided it, e.g., AOEC or RUMC in our case, through the property "`foaf:maker`", imported from the Friend-Of-A-Friend (FOAF) ontology.[27]

The *Diagnosis Area* describes the diagnosis associated with a specific clinical case report. In particular, this area contains all classes related to the different outcomes one can expect from a clinical case report concerning a specific use case. All use cases have the same structure, yet the components describing positive outcomes differ based on the considered disease. We report the modeling of the diagnosis area for colon cancer in Fig. 3. In general, each clinical case report is associated with its outcome using the object property "`exa:hasOutcome`" and outcomes can be negative, inconclusive, or positive. Classes concerning negative and inconclusive outcomes are the same for all use cases, while positive outcomes differ based on the disease. Positive outcomes describe the type of cancer or disease that has been diagnosed in the examined specimen and it can be associated with additional information for some types of diseases. For instance, in the colon cancer use case, if the specimen has been diagnosed as "Polyp of Colon" or any of its subclasses, we might have additional information about the degree of dysplasia the polyp presents. We refer to this additional data as *"Annotations to the Case"* and we modeled them as subclasses of the general-purpose class called "`exa:Finding`". For lung cancer, additional information regards the presence of necrosis or metastasis in a specimen of "lung carcinoma" or any of its subclasses. In celiac disease, a positive outcome can be enriched with information about the presence of immune cells such as granulocytes or lymphocytes, the presence of intestinal abnormalities such as edema or intestinal fibrosis, or data about the villi status, such as their atrophy degree or length. Finally, for cervix cancer, we could detect the presence of koilocytes, or the specimen could test positive for HPV.

The *Location Area* models the anatomical locations of the retrieved specimen used for diagnosis or where an intervention has been performed. All classes include locations specific for each use case and are modeled as subclasses of the general-purpose class "`exa:AnatomicalEntity`". The specific components, i.e., anatomical locations, of this semantic area highly depend on the different use cases. Fig. 4 reports the location semantic area for colon cancer. The class "`exa:ColonAnatomicalEntity`" comprises all anatomical locations concerning colon cancer, such as the ileum, colon, and its subparts.

The *Procedure Area* describes the intervention or surgical procedure performed to obtain the specimen used for the diagnosis. As for the other semantic areas, the general-purpose classes are "`exa:Intervention`" and "`exa:SurgicalProcedure`", which is actually a subclass of the former. All interventions or surgeries performed for a specific clinical case are instances of the corresponding classes specific for the considered use case. For instance, interventions concerning cervix clinical cases are instances of the class "`exa:CervixIntervention`". The type of intervention or surgery is modeled following the SKOS data model and we use as taxonomy of reference NCIT. The classification between interventions or surgical procedures also complies with the NCIT taxonomy. Fig. 5 reports the procedure semantic area for the cervix cancer use case. Each outcome can be associated with an intervention using the object property "`exa:hasIntervention`" and we can add information about the anatomical location of the procedure with the property "`exa:hasLocation`".

The *Test Area* comprises all classes corresponding to additional tests on the specimen to identify additional information useful for the diagnosis. In

[27] http://xmlns.com/foaf/0.1/.

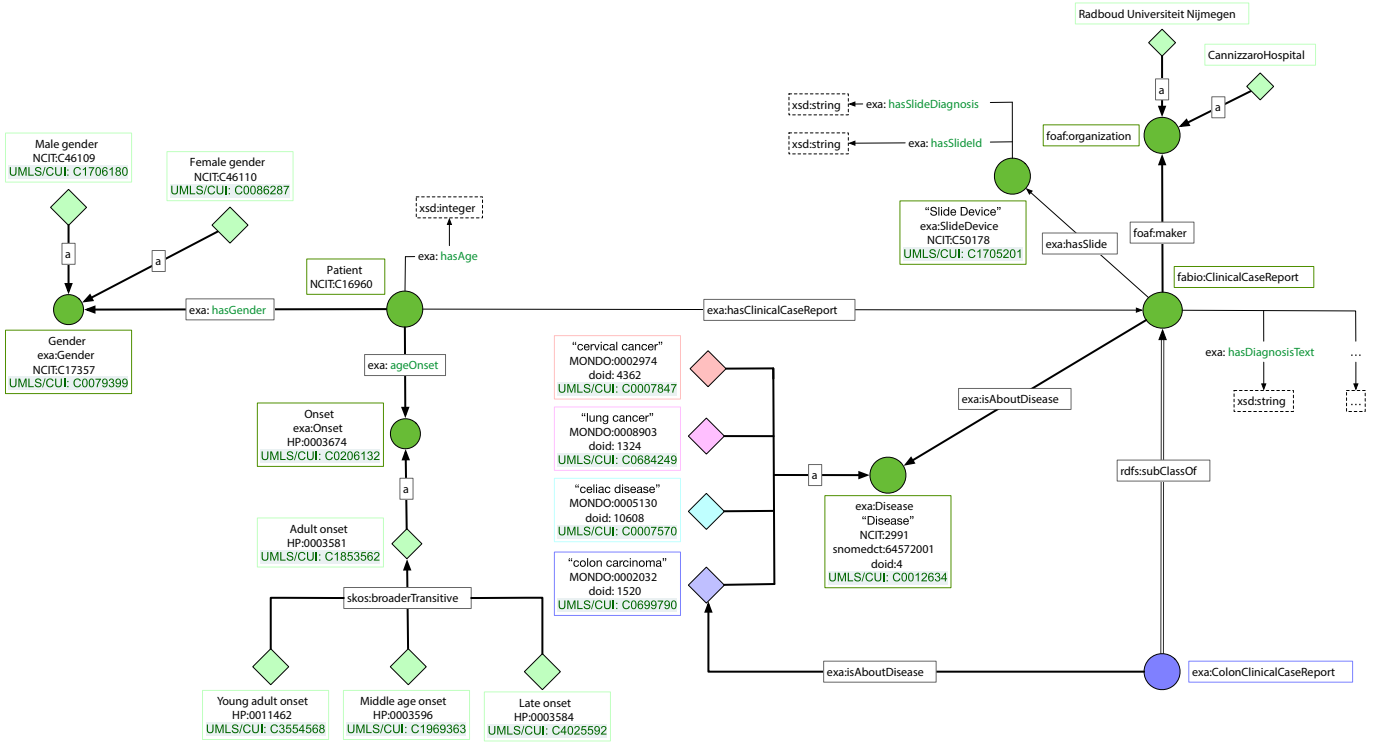

Fig. 2. General semantic area. Circles specify ontology's classes while rhombus-shaped components are modeled with named individuals. The core of this area is the "Clinical Case Report" together with data about the patient and the associated images. These classes are common for all use cases since they do not embody information specific to a disease. The only exception is the "Colon Clinical Case Report" Class which refers to the studied disease. In this figure, for instance, we report the modeling of the colon cancer use case. Components related to all use cases are identified by the green color, while purple identifies components specific to the colon cancer use case.

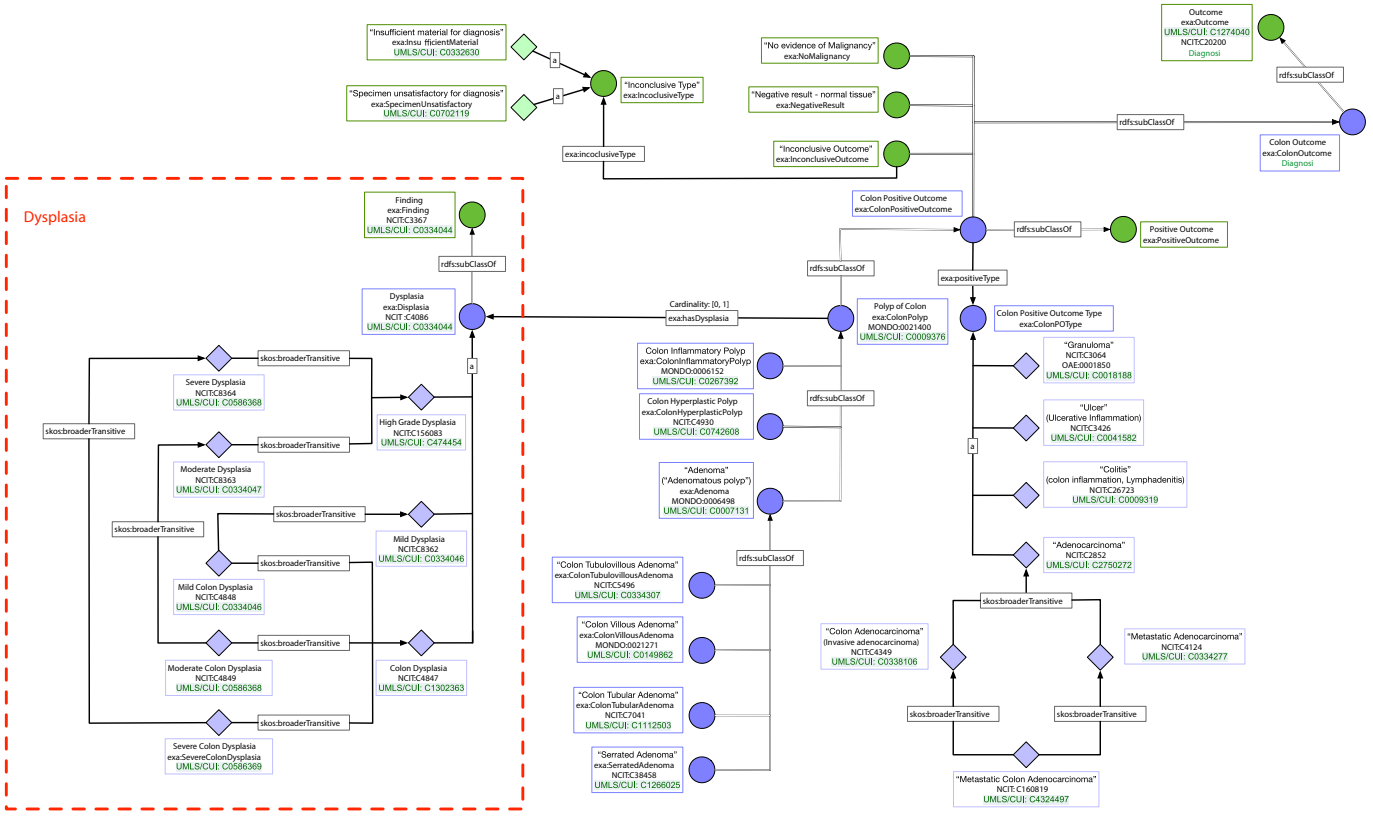

Fig. 3. Diagnosis semantic area for colon cancer. Each colon clinical case report is associated with an outcome, that can be inconclusive, negative, or positive. Classes for inconclusive and negative outcomes are the same for all use cases, while the positive outcome one, i.e., "exa:ColonPositiveOutcome", differs based on the disease. In fact, different diseases have different positive outcome types. We also model some annotations specific to the use case, i.e., the presence of dysplasia in colon polyps.

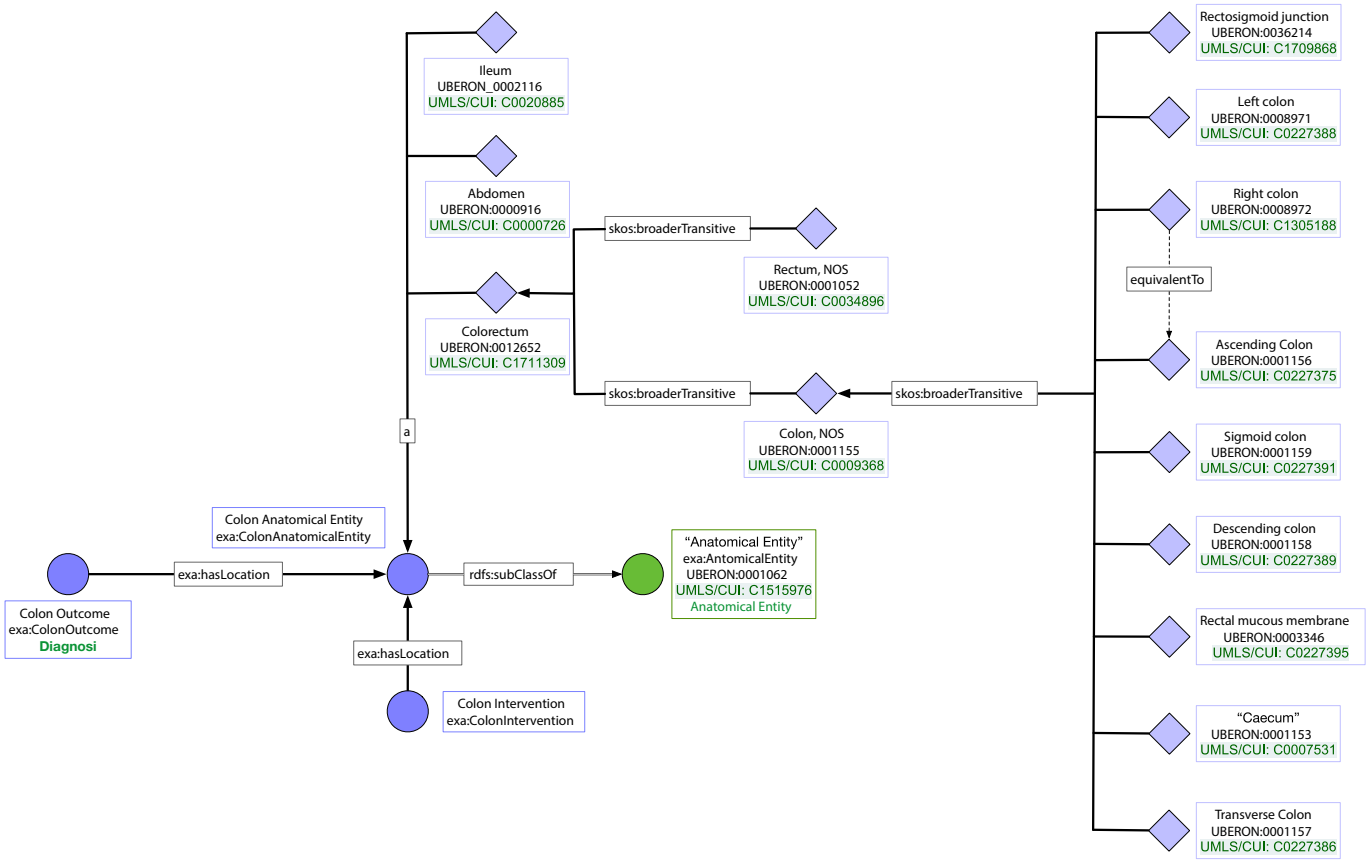

Fig. 4. Location semantic area for colon cancer. Each colon outcome or intervention can be associated with an anatomical location. The specific locations follow the SKOS data model and they depend on the considered use case.

fact, immunohistochemistry (IHC) is an important field of study that determines the tissue distribution of antigens and it is widely used for the diagnosis of diseases, especially for identifying the type of cancer in a tissue.[63] For this reason, we identify a general-purpose class called "`exa:ImmunohistochemicalTest`" and all the specific tests for each use case are modeled as subclasses of it. We link an outcome with the additional test through the object property "`exa:hasTest`" and each test can have a boolean, float, or string-valued result embodied in 3 different data properties corresponding to the different result datatype. We report the test semantic area for lung cancer in Fig. 6.

We added some restrictions to allow for consistency checking and data cleaning leveraging the ontology. In particular, following OWL syntax, we added some universal restrictions for some subclasses to avoid the ingestion of noisy data. For instance, if we try to instance a "Colon

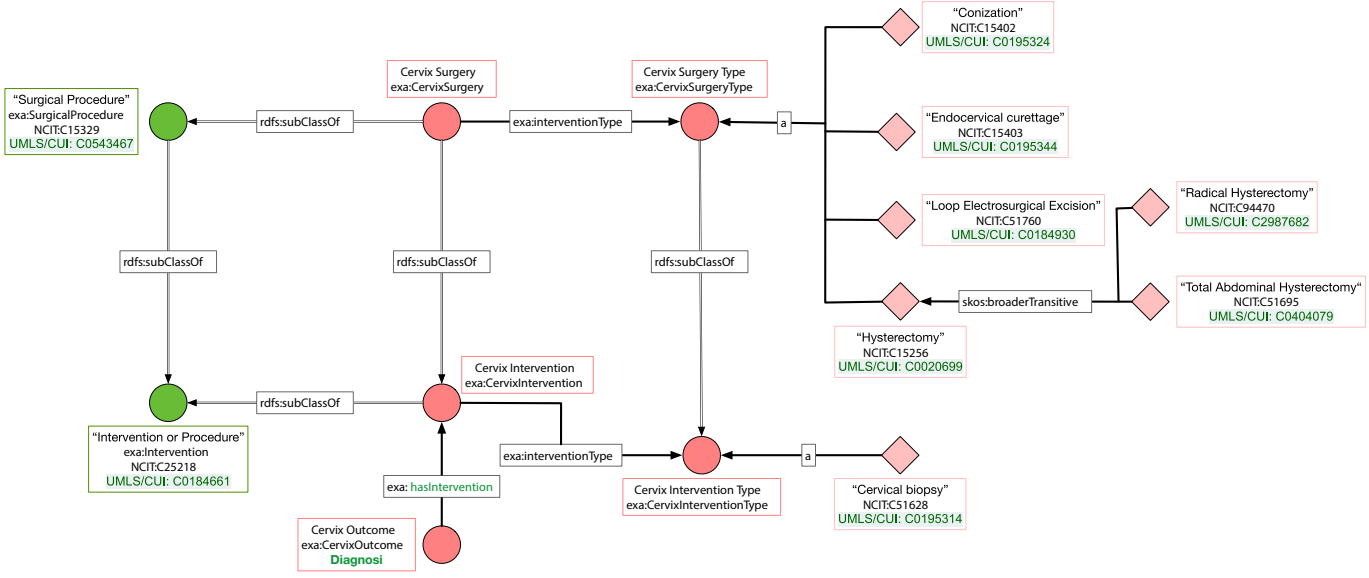

Fig. 5. Procedure semantic area for cervix cancer use case, identified by color peach. Each cervix outcome can be associated with the intervention performed for retrieving the specimen used for diagnosis. The specific type of interventions or surgeries follow the SKOS data model and they depend on the considered use case.

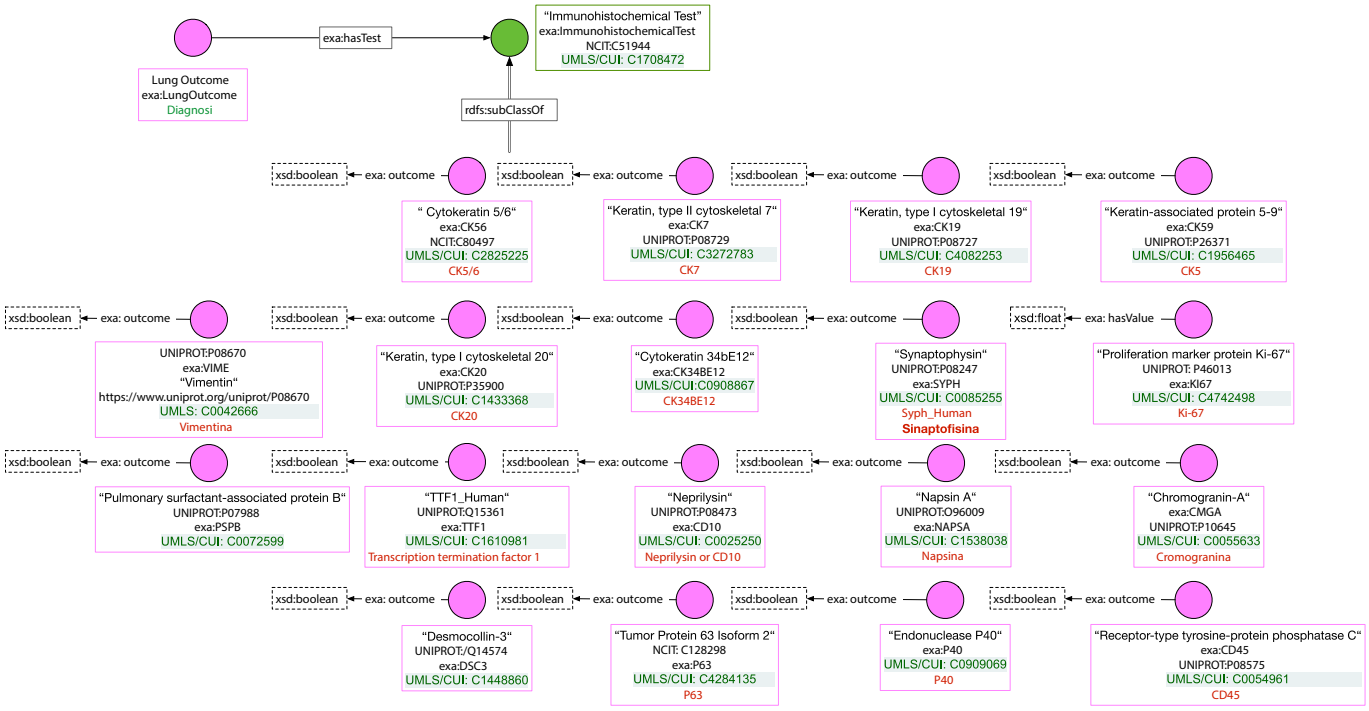

Fig. 6. Test semantic area for lung cancer use case, identified by color pink. Each lung outcome can be associated with a test for different receptors, each resulting in a boolean or float-valued result.

Clinical Case" with a diagnosis of a type of cancer belonging to the cervical cancer domain, a reasoner running on the ontology will raise an "InconsistenOntologyException".

Results

Entity linking

One of the most relevant tasks that can be performed with an ontology is entity linking (EL).[64] EL is the task of assigning unique meanings to entities mentioned within text. In a nutshell, the aim of EL is to determine if a given (extracted) entity refers to a specific concept within a reference ontology. To this end, Marchesin et al[65] developed the Semantic Knowledge Extractor Tool (SKET), an unsupervised hybrid knowledge extraction system that combines a rule-based expert system with pre-trained machine learning models to extract relevant concepts from pathology reports. The system presents a modular architecture, where different components or methods can be easily plugged-in. First, if needed, SKET performs (English) translation and curation over pathology reports. For translation, the system exploits the open-source, pre-trained Marian Neural Machine Translation (NMT) models,[66] which rely on a transformer-based encoder–decoder architecture.[67] Due to the complexity of the task, automated approaches introduce systematic translation errors that, if propagated, could affect the quality of the extraction process. Therefore, SKET performs a curation step in which recurring, manually identified translation errors are corrected through the use of handcrafted rules. Once the reports are translated, SKET combines pre-trained Named Entity Recognition (NER) models with unsupervised EL methods to extract relevant entities from translated reports and link them to the ExaMode ontology. To handle negated entities, SKET integrates NegEx[68] within the NER component. NegEx is a negation detection algorithm that evaluates whether extracted entities are negated within text or not. It uses regular expressions to identify the scope of trigger terms that are indicative of negation, such as "no" or "ruled out". Then, the entities extracted within the scope of a trigger term are marked as negated. In this way, SKET identifies—and removes—those entities that NegEx considers as negated. For instance, in the phrase "free of dysplasia", NegEx identifies the trigger term "free of" and marks "dysplasia" as negated, which is then removed. SKET has been adopted to extract concepts from pathology reports coming from the clinical workflow of AOEC and RUMC medical centers, considering reports from the use cases modeled by the ExaMode ontology. To evaluate its effectiveness, SKET has been tested on 500 colon cancer and 500 cervix cancer reports, manually annotated by experts, coming from both AOEC and RUMC medical centers, together with 250 lung cancer reports from AOEC. SKET achieved high performance on each use case, with a (micro) F1 score of 0.8861 on colon cancer, 0.8322 on cervix cancer, and 0.9375 on lung cancer. Given its high performance, we deployed SKET on all the ExaMode data. Table 4 reports the statistics of the extraction process for each use case, including celiac disease. The extracted concepts can then be used in different applications as: (i) weak annotations to train models for computer-aided diagnosis[20] or (ii) automatic annotations to support semi-automatic tagging.[69] Indeed, the use of pre-trained NER models and unsupervised EL methods makes SKET suitable for weak supervision tasks.[20,65] Therefore, SKET can be used in the clinical workflow without the expensive and time-consuming cost of human annotations. This demonstrates the importance of developing medical ontologies, which empower many different computer-aided diagnosis tools.[70,71]

Decision-support systems and tagging

The designed ExaMode ontology, alongside with other domain ontologies (ICD-10, ICD-O, UMLS, SNOMED, Human Disease Ontology[28], PathLex[29], Mondo[30], and ProstateCancer[31]) was integrated with patient data into the "HistoGrapher demonstrator"[32] for decision-support systems and triaging. Such integration with other well-known biomedical resources provides further validation for the ExaMode ontology. HistoGrapher is a software platform for building holistic solution integrating multimodal histopathology data. This solution supports pathologists in making more informed decisions based on a larger amount of data (judging from

[28] https://disease-ontology.org/.
[29] https://bioportal.bioontology.org/ontologies/PATHLEX.
[30] https://mondo.monarchinitiative.org/.
[31] https://bioportal.bioontology.org/ontologies/PCAO.
[32] http://examode.ontotext.com/.

Table 4
SKET extraction statistics. Columns represent, from left to right, the considered hospital, the considered use case, the total number of reports, the total number of concepts, the maximum number of concepts per report, and the mean number of concepts per report. The "–" symbol represents the lack of pathology reports for the considered hospital and use case.

Hospital	Use case	N. of reps.	N. of concs.	Max concs./rep.	Mean concs./rep.
AOEC	Colon	4020	17 056	15	4.24
	Cervix	4810	28 110	26	5.84
	Lung	2077	11 924	22	5.74
	Celiac	1965	9716	21	4.94
RUMC	Colon	14 147	57 411	47	4.06
	Cervix	5861	24 847	27	4.24
	Lung	–	–	–	–
	Celiac	962	4747	10	4.93

similarity to other cases in their clinical practice or the identified likelihood in scientific literature). Prioritization of cases is considered, so that the cases with higher severity and likelihood of specific diseases will be presented for confirmation first, while the cases with smaller likelihood—later when there is enough time. The data model of the HistoGrapher platform is based on the defined ExaMode ontology, which is used for information extraction and semantic data normalization of medical synopsis texts provided by the laboratory information management system (LIMS). The platform applies machine translation in order to translate the source text data, provided in Italian and Dutch, into corresponding English representations. All translated text fields from the synopsis record are processed with a NLP pipeline (Fig. 7). The pipeline is based on the ExaMode ontology and all created semantic annotations in-lined in the source text are referred to concepts from the ontology or to other biomedical terminologies, which are part of the platform semantic solution (Fig. 8). The output of the NLP pipeline, in the form of RDF triples, is imported in GraphDB[33], which is the semantic RDF triples store used to build the knowledge graph behind the Histographer platform. As the NLP pipeline output is fully harmonized with the ExaMode ontology, the extracted data can be further explored in the context of the ontological model and the terminologies included in the knowledge graph of the system. The HistoGrapher platform provides fully configurable semantic search and data exploration dashboards that are adapted to the specific data schema and ontologies used. The system provides capabilities for intuitive faceted filtering for relevant clinical case records and both semantic (concept) and free-text search. The focused data exploration dashboards provide different views over the data in the knowledge graph. An important dashboard component is the case report similarity widget used to retrieve similar case reports. The similarity of case reports is based on graph embedding configured per certain knowledge graph class and its `DataType` and `ObjectType` properties.

MedTAG

MedTAG[69] is an open-source collaborative biomedical annotation tool whose aim is to support, organize, and speed-up the annotation process. Indeed, semantic annotators and NLP methods for NER and EL require lots of training and test data, especially in the biomedical domain. However, despite the abundance of unstructured biomedical data, there is a lack of richly annotated datasets. On top of that, manual annotation of biomedical documents by experts is a costly and time-consuming operation. MedTAG is a tool easy to install and supported by every platform that eases manual annotation of unstructured information. It has been employed in the histopathology domain by physicians and experts to manually annotate thousands of clinical reports in 3 different languages (Dutch, English, and Italian) from 2 healthcare institutions—AOEC and RUMC. MedTAG annotations rely on terms defined in the ExaMode ontology, which allows interoperability among medical centers and clinical experts who use this tool to annotate unstructured biomedical data. Table 5 reports statistics about the manual annotation process, displaying the number of diagnostic reports annotated per language and use case.

Discussion

The development of the ExaMode ontology allows modeling information about clinical case reports, WSIs, and diagnostic outcomes for the 4 use cases considered within the ExaMode project. We defined 5 semantic areas populated with relevant components for each considered use case. As a result, the ExaMode ontology is easily extensible to new components of interest and can be applied to any disease studied in the histopathology domain. According to the RDF data model, the ontology ExaMode is a key means of storing data in a graph database. Creating an RDF dataset can help develop more accurate algorithms for image analysis, especially in digital pathology. In the following section, we provide examples of how clinical reports data are stored based on the ExaMode ontology. We also show some queries that can extract relevant clinical insights about the considered diseases.

Diagnosis

In Fig. 9 we can see a diagnosis for each of the 4 diseases we modeled. In particular, for each case, we report the plain text of the diagnosis and the semantic graph created via entity extraction and linking to the ExaMode ontology. We recall the ExaMode ontology models clinical reports containing information about the diagnosis, patients, and WSIs, but we focus only on the diagnosis part for this example.

Concerning the colon cancer use case, Fig. 9a, we instantiate a specific colon clinical case report (individual "`exa:report colon1`"), an outcome (individual "`exa:outcome colon1`"), and an intervention (individual "`exa:intervention colon1`"). The specimen tested positive for *tubular adenoma*, so the specific outcome is an instance of the class "`exa:ColonTubolulovillousAdenoma`". Since the adenoma has *low grade dysplasia*, following the SKOS data model, the specific outcome is linked with the instance "Mild Colon Dysplasia". The specimen was retrieved through a *colon biopsy* located in the *transverse colon*. The type of intervention is stored using the object property "`exa:interventionType`" pointing at "Colon Biopsy". Following the SKOS data model, the location information is embedded using the object property "`exa:hasLocation`" pointing at "Transverse Colon".

About the cervix cancer use case, Fig. 9b, we instantiate a specific cervix clinical case report (individual "`exa:report cervix1`"), an outcome (individual "`exa:outcome cervix1`"), and an intervention (individual "`exa:intervention cervix1`). Since the diagnosis was inconclusive, the specific outcome is an instance of the class "`exa:InconclusiveOutcome`". The result was inconclusive due to insufficient material, so the specific outcome is linked with the named individual "`exa:insufficientMaterial`" using the object property "`exa:inconclusiveType`". The specimen was retrieved through a *cervix biopsy* with no additional information about the location. The type of intervention is stored using the object property "`exa:interventionType`" pointing at "Cervical Biopsy", modeled following the SKOS data model.

Regarding the lung cancer use case, as reported in Fig. 9c, we instantiate a specific lung clinical case report (individual "`exa:report lung1`"), an outcome (individual "`exa:outcome lung1`"), an intervention (individual "`exa:intervention lung1`"), and 2 tests, (individual "`exa:test lung1`") and (individual "`exa:test lung2`"). The specimen tested positive for *adenocarcinoma*. Thus, "`exa:outcome lung1`" is a named individual of the class "`exa:LungAdenocarcinoma`". The outcome is located in the pleura, which presents *metastasis*. The specific outcome is linked with the instance "Metastasis", using the object property "`exa:presenceOf`". The location information is embedded using the object property "`exa:hasLocation`" pointing at "Pleura". The specimen was retrieved through a *bronchial biopsy* located in the *left main bronchus*. The type of intervention is stored using the object property "`exa:interventionType`" pointing at

[33] https://www.ontotext.com/products/graphdb/.

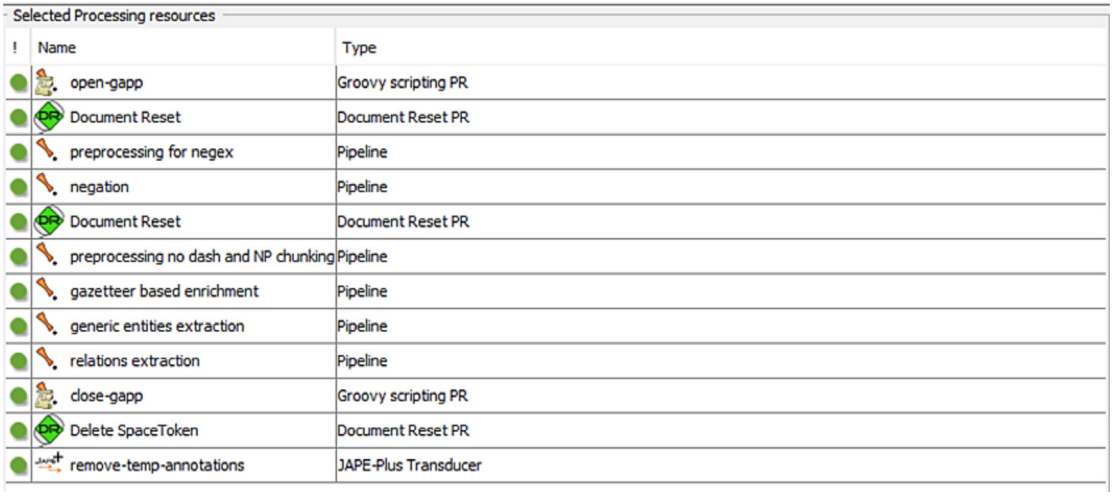

Fig. 7. NER pipeline consists of multiple text pre- and post-processing resources (PRs) and ontology-based semantic annotation PR "gazetteer-based-enrichment", which normalizes the identified phrases to ontology terms.

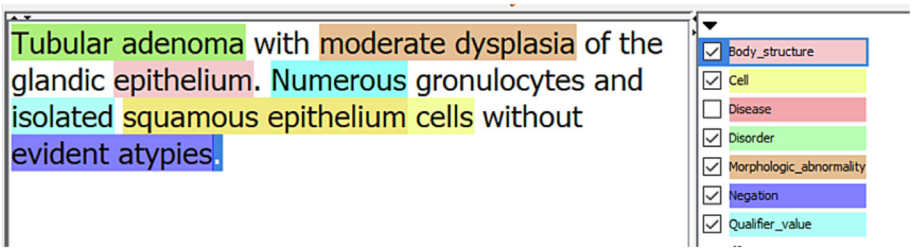

Fig. 8. Semantically annotated synopsis report with highlighter terms from various annotation classes (in different colors).

Table 5
MedTAG statistics on manual annotations. We report the number of diagnostic reports annotated per language and use case. Symbol "–" represents the lack of annotated reports for the considered language and use case.

Language	Total	Use case			
		Colon cancer	Cervix cancer	Lung cancer	Celiac disease
Dutch	2322	889	1433	–	–
English	9893	2996	4880	2017	2576
Italian	4911	1132	1828	1951	–
Total	17 126	5017	8141	3968	2576

"Bronchial biopsy", and the location information is embedded using the object property "`exa:hasLocation`" pointing at "Left Main Bronchus", modeled following the SKOS data model.

As for the celiac disease use case, Fig. 9d, we instantiate a specific celiac clinical case report (individual "`exa:report celiac1`"), an outcome (individual "`exa:outcome celiac1`"), an intervention (individual "`exa:intervention celiac1`"), and 1 laboratory finding (individual "`exa:finding celiac1`"). The specimen tested positive for *celiac disease (type 3c of marsh-oberhuber)*. Thus, the specific outcome is an instance of the class "`exa:PositiveToCeliacDisease`" and the marsh classification is stored as a data property called "`exa:marshStage`". The location information, i.e., *duodenal mucosa*, is embedded using the object property "`exa:hasLocation`" pointing at "Duodenal Mucosa". The specimen presented also *erosion*, *flattening of the villi*, and *severe villi atrophy*. The first information can be stored by linking the specific outcome to the concept "Erosion" with the object property "`exa:presenceOfIntestinalAbnormality`". Data about villi status is embedded in the specific laboratory finding instance through 2 data properties. In particular, we set the data property "`exa:hasFlatVilli`" to "True", and we set the "`exa:villiAtrophy`" property to "severe degree of atrophy". The specimen was retrieved through a *duodenal biopsy*. The type of intervention is stored using the object property "`exa:interventionType`" pointing at "Biopsy of Duodenum". The location is embedded using the object property "`exa:hasLocation`" pointing at "Duodenum", which is inferred from "*duodenal biopsy*".

Example queries

The ExaMode ontology models histopathology clinical case reports. Thus, diagnoses are stored in the form of RDF graphs, as we reported in Fig. 9. This approach allows for seamless data integration and a unified query access point, from which we can extract valuable information using SPARQL queries. For instance, we can investigate the most common anatomical location for a specific type of cancer, or check if our dataset confirms the correlation between HPV and cervical cancer. All insights we can extract from a dataset modeled using the ExaMode ontology can support decisions on cancer-prevention policies or can highlight an unexpected research path. In the following section, we provide some example queries related to the use cases modeled in the ExaMode ontology that extract some valuable information regarding histopathology diagnoses. Due to the confidential nature of the data we used in this study, we cannot release the diagnostic reports coming from the involved medical centers (i.e., AOEC and RUMC). Nevertheless, for each query, we report a table with the statistics derived from the diagnostic reports.

Query 1 investigates the distribution of colon adenocarcinoma based on the anatomical location. In fact, it could be useful to see the most common anatomical location where a specific type of cancer develops for determining prevention strategies. For instance, research has shown that serrated adenomas are more commonly found in the proximal colon, while polypoid adenomas are more common in the distal colon.[72] In the ExaMode

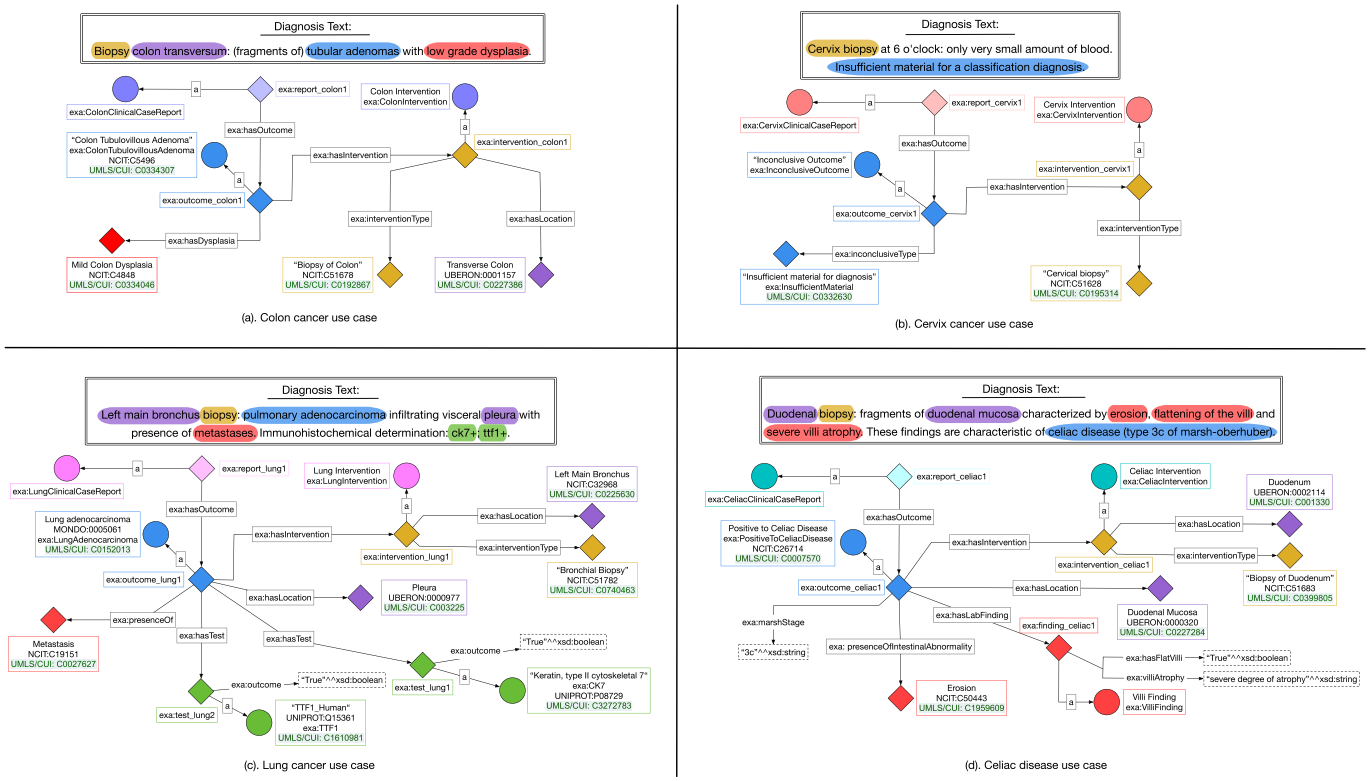

Fig. 9. Diagnoses modeled using the ExaMode ontology. Tag colors are based on the use case and the ExaMode semantic areas. In particular, yellow identifies the interventions, purple identifies the locations, green highlights the tests performed, blue identifies the outcome, and red identifies additional information about the outcome (i.e., the so-called "annotations to the case").

ontology, "colon adenocarcinoma" is modeled following the SKOS data model and it is stored with its NCIT URI, i.e., "`NCIT:C4349`". To extract this type of information, we perform a COUNT operation, grouping the cases with outcome type "colon adenocarcinoma" based on the anatomical location where the diagnosis has been assessed. Table 6 reports the query result obtained from the colon diagnostic reports.

to see if such a correlation is expressed in a dataset modeled using the ExaMode ontology. Query 2 counts the number of clinical case reports diagnosed with cervical dysplasia of grade 2 or higher and groups the result based on the detection of HPV infection. "Cervical dysplasia" is a type of positive outcome for cervix cancer, thus we extract all cases that had an outcome of the above-mentioned type or any of its narrower

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX exa: <https://w3id.org/examode/ontology/>
PREFIX ncit: <http://purl.obolibrary.org/obo/NCIT_>

SELECT ?location (COUNT(?case) AS ?adenoCases)
WHERE {
    ?case a exa:ColonClinicalCaseReport;
          exa:hasOutcome ?outcome.
    ?outcome exa:positiveType ncit:C4349;
             exa:hasLocation ?locationURI.
    ?locationURI rdfs:label ?location.
    FILTER(langMatches(lang(?location), 'en'))
}
GROUP BY ?location
ORDER BY DESC (?adenoCases)
```

Query 1: Investigate the most common locations for colon adenocarcinoma: Count the number of cases of colon adenocarcinoma (NCIT:C4349) grouped by anatomical location.

HPV has been discovered to be one of the major causes of cervical cancer and cervical dysplasia.[73] For this reason, it can be interesting

concepts and we count such cases based on the presence of an HPV infection. Note that the presence of HPV can be assessed by the existence of the object property "`exa:detectedHPV`" pointing at the individual "HPV Infection". Table 7 reports the query result obtained from the cervix diagnostic reports.

```
PREFIX exa: <https://w3id.org/examode/ontology/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX mondo: <http://purl.obolibrary.org/obo/MONDO_>

SELECT ?hpvDetected (COUNT(DISTINCT ?case) AS ?numCases)
WHERE{
    ?case a exa:CervixClinicalCaseReport;
          exa:hasOutcome ?outcome.
    ?outcome exa:positiveType ?type.
    FILTER(?type = ?neoplG23)
    {
        SELECT ?neoplG23 WHERE{
            ?neoplG23 skos:broaderTransitive* mondo:0006137.
        }
    }
    BIND(IF(
    EXISTS {?outcome exa:detectedHPV mondo:0005161.},
    "HPV", "noHPV") AS ?hpvDetected)
}
GROUP BY ?hpvDetected
```

Query 2: Investigate HPV correlation with cervix cancer: Count how many cases of moderate or severe dysplasia (mondo:0006137 or any of its narrower concepts) resulted positive to HPV (mondo:0005161) and how many did not.

Query 3 investigates the distribution of different types of lung carcinoma diagnoses. There are several types of lung cancer, such as non-small cell lung cancer (NSCLC) or small cell lung cancer (SCLC), with very different incidences and courses of treatments. In fact, SCLC is a highly malignant type of cancer and accounts for 15% of lung cancer cases, the remaining 85% of cases present a form of NSCLC or any of its pathologic subtypes, i.e., adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.[74] The 5-year survival rate is much greater in early-stage cancers; thus better screening methods could help prevent late-stage discovery. For instance, if we aim at developing software for the early detection of lung cancer cells, it is crucial to have a representative dataset for training models. For this reason, it could be useful to see what is the incidence of different types of lung carcinoma in a dataset modeled using the ExaMode ontology and check data is aligned with cancer studies statistics. To do so, we count the cases with a diagnosis of “lung carcinoma" or any of its subclasses and group the results based on the type of cancer. In this case, it was not possible to model “lung carcinoma" following the SKOS data model because we can have additional information about the presence of necrosis and metastases in the specimen. Therefore, we modeled lung carcinoma as a class with URI belonging to our namespace and we reference external resources, such as UMLS and MONDO. Table 8 reports the query result obtained from the lung diagnostic reports.

```
PREFIX exa: <https://w3id.org/examode/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?diagnosis (COUNT(DISTINCT ?case) AS ?numCases)
WHERE{
    ?case a exa:LungClinicalCaseReport;
          exa:hasOutcome ?outcome.
    ?outcome a ?lungCarcinoma.
    FILTER(?lungCarcinoma = ?descendants)
    {
        SELECT ?descendants WHERE{
            ?descendants rdfs:subClassOf* exa:LungCarcinoma.
        }
    }
    ?lungCarcinoma rdfs:label ?diagnosis.
    FILTER(langMatches(lang(?diagnosis), 'en'))
}
GROUP BY ?diagnosis
ORDER BY DESC(?numCases)
```

Query 3: Investigate the incidence of different types of lung carcinoma: Count the number of cases of lung carcinoma grouped by its types, i.e., lung carcinoma or any of its subclasses.

Celiac disease has a highly variable clinical expression, the most common pathologic lesions are a flattened small intestinal mucosa with lymphocytic infiltrate, crypt hyperplasia, and villous atrophy.[75] There is a correlation between the presence of flattened villi and celiac disease, thus we can investigate how high is such a correlation by means of data modeled using the ExaMode ontology. Query 4 counts how many cases of celiac disease present flattened villi or not. Such information is embedded in a

boolean data property called "exa:hasFlatVilli" whose domain is the class "Villi Finding". Thus, we count cases with celiac disease and check the ones that have a "Villi Finding" instance connected with "exa:hasFlatVilli" set to "True". Table 9 reports the query result obtained from the celiac diagnostic reports.

```
PREFIX exa: <https://w3id.org/examode/ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?flatVilli (COUNT(?case) AS ?numPositiveCases)
WHERE{
    ?case a exa:CeliacClinicalCaseReport;
          exa:hasOutcome ?outcome.
    ?outcome a exa:PositiveToCeliacDisease.
    BIND(IF(EXISTS {
        ?outcome exa:hasLabFinding ?finding.
        ?finding a exa:VilliFinding;
                 exa:hasFlatVilli "true"^^xsd:boolean.},
        "flattenedVilli", "notFlattenedVilli") AS ?flatVilli)
}
GROUP BY ?flatVilli
```

Query 4: Investigate the correlation between flattened villi presence and celiac disease diagnoses: Count how many cases positive to celiac disease presented flattened villi and how many did not.

Conclusions and future work

In this work, we presented the ExaMode ontology modeling the histopathology diagnostic process by considering 4 diseases: colon cancer, cervix cancer, lung cancer, and celiac disease. In particular, we modeled the components related to the annotation process of WSIs, storing information about clinical case reports, diagnoses, histopathology images, anatomical locations, and interventions. The ontology was developed in the context of the *ExaMode project*, which aims to allow weakly supervised knowledge discovery of multimodal heterogeneous data, limiting human interaction. For this reason, the ontology design followed a bottom-up approach starting from anonymized clinical reports provided by 2 European medical centers in Italy and The Netherlands. For modeling the ontology, we followed an iterative co-design process with continuous feedback and validation from pathologists and clinicians. To ease interoperability, we designed the ExaMode ontology to meet the OBO principles and we defined some basic guidelines to adopt when defining each component to guarantee semantic consistency. We also connected every defined concept to widely used medical taxonomies and thesauri as UMLS and SNOMEDCT. The ExaMode ontology is organized into 5 semantic areas, which group

Table 6
Result sample for Query 1: "Investigate the most common locations for colon adenocarcinoma: Count the number of cases of colon adenocarcinoma (NCIT:C4349) grouped by anatomical location.". Each column is named after the selected variables in the query.

Location	adenoCases
Rectal mucous membrane	1052
Colon, NOS	845
Sigmoid colon	424
Abdomen	311
Rectum, NOS	102
Rectosigmoid junction	76
Right colon	74
Ascending colon	68
Cecum	63
Ileum	52
Descending colon	25
Left colon	16
Colorectum	9
Transverse colon	2

Table 8
Result sample for Query 3: "Investigate the incidence of different types of lung carcinoma: Count the number of cases of lung carcinoma grouped by its types, i.e., lung carcinoma or any of its subclasses.". Each column is named after the selected variables in the query.

Diagnosis	numCases
Lung adenocarcinoma	991
Non-small cell squamous lung carcinoma	564
Lung carcinoma	520
Malignant lung neoplasm	401
Non-small cell lung carcinoma	371
Metastatic neoplasm	322
Small cell lung carcinoma	140
Lung large cell carcinoma	64
Clear cell adenocarcinoma	5

Table 7
Result sample for Query 2: "Investigate HPV correlation with cervix cancer: Count how many cases of moderate or severe dysplasia (mondo:0006137 or any of its narrower concepts) resulted positive to HPV (mondo:0005161) and how many did not.". Each column is named after the selected variables in the query.

hpvDetected	numCases
noHPV	4044
HPV	795

Table 9
Result sample for Query 4: "Investigate the correlation between flattened villi presence and celiac disease diagnoses: Count how many cases positive to celiac disease presented flattened villi and how many did not.". Each column is named after the selected variables in the query.

hpvDetected	numCases
flattenedVilli	19
notflattenedVilli	558

together components related to the same aspect of the diagnostic process: *General Area*, *Diagnosis*, *Intervention*, *Location,* and *Test*. We applied this methodology to model 3 types of cancer and celiac disease. The ExaMode ontology is openly available on the Web[34] and it is currently being used as a common semantic layer in several downstream applications. In particular, the ExaMode ontology is used in an entity linking tool called SKET, which is a hybrid knowledge extraction system useful for the extraction of concepts from pathology reports. Histographer is a software platform for building holistic solutions integrating multimodal data, integrated several biomedical resources, including the ExaMode ontology, to develop a decision-support system for pathologists. Finally, MedTAG is a web-based collaborative annotation tool for histopathology reports whose annotations rely on terms defined in the ExaMode ontology.

Comprising 5 semantic areas, the structure of the ExaMode ontology provides an ontological template that can be used to model any disease in the histopathology domain. As future work, we plan to integrate other high-profile diseases, such as breast and prostate cancer. In this regard, we simply need to define specific concepts related to each semantic area. To model all the possible outcomes related to breast cancer, for instance, we can define a new class called “BreastCancerOutcome" as a subclass of “`exa:Outcome`" and include specific diagnoses as subclasses or individuals. A similar procedure can be applied to prostate cancer as well. Furthermore, the modular design of the ExaMode ontology allows not only the integration of additional diseases but also the expansion of the ontology itself. In this regard, we can easily broaden the scope of the ontology by including new concepts or adding new semantic areas. For instance, we can model different staining procedures for WSIs, such as IHC staining, in addition to the Hematoxylin and Eosin (H&E) one; or we can add a whole new semantic area concerning molecular genetics, where all the corresponding information would be modeled via classes and properties.

Regarding downstream applications, we plan to integrate SKET in the “HistoGrapher demonstrator” to leverage its semantic similarity capabilities and negation detection features. By performing semantic match, as opposed to exact match, SKET can bypass the hurdles associated with typos and identify concepts that are misspelled. At the same time, integrating SKET within HistoGrapher will empower the latter with negation detection mechanisms, which can improve its performance and provide more valuable insights about synopsis records.

Funding

Gianmaria Silvello reports financial support was provided by the European Commission ExaMode project, as part of the EU H2020 program under Grant Agreement no. 825292. The European Union’s Horizon 2020 research and innovation programme did not play a role in the design or development of the ExaMode Ontology, nor the development of this report.

Contributions

L.M. contributed to the ontology’s encoding, query formulation, the extension of SKET, and manuscript preparation. G.S. coordinated the teamwork, designed the ontology, contributed to the ontology encoding, wrote the first draft of the paper, and revised it before submission. M.A. contributed to designing the project, discussed ontology design choices, and validated the final ontology. S.B. contributed to the validation of the ontology and the development of histographer. F.C. provided medical records, analyzed them, and contributed to ontology validation. G.M.D.M. contributed to the ontology design with a specific focus on the multilingual aspects. F.F. provided medical records, analyzed them, and contributed to ontology validation. F.G. contributed to the design and development of MedTag and SKET and validated the ontology encoding. O.I. contributed to the design and development of MedTag and SKET and validated the ontology encoding. S.M. developed SKET, wrote the code to extract medical records knowledge and create semantic graphs based on the ExaMode ontology, and contributed to the design of the ontology. N.M. discussed ontology design choices and validated the final ontology. H.M. contributed to designing the project, discussed ontology design choices, and validated the final ontology. T.P. contributed to the validation of the ontology and its encoding, validated the semantic graphs created from the medical records, and curated the development of histographer. All the authors contributed to the revision of the manuscript.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Gianmaria Silvello reports financial support was provided by European Commission. Filippo Fragetta is an author of the paper and a member of the editorial board of JPI.

Acknowledgments

The authors wish to thank Dennis Dosso who contributed to the first version of the ontology in the early stages of the ExaMode project.

References

1. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform 2015;16(6):1069–1080. https://doi.org/10.1093/bib/bbv011.
2. Konopka BM. Biomedical ontologies—a review. Biocybern Biomed Eng 2015;35(2):75–86. https://doi.org/10.1016/j.bbe.2014.06.002.
3. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–175. https://doi.org/10.1016/j.media.2016.06.037. 20th anniversary of the Medical Image Analysis journal (MedIA).
4. Serra LM, Duncan WD, Diehl AD. An ontology for representing hematologic malignancies: the cancer cell ontology. BMC Bioinform 2019;20-S(5):231–236. https://doi.org/10.1186/s12859-019-2722-8.
5. Freitas F, Schulz S, Moraes E. Survey of current terminologies and ontologies in biology and medicine. RECIIS-Elect J Commun Inform Innov Health 2009;3(1):7-18. https://doi.org/10.3395/RECIIS.V3I1.239EN.
6. Turner JA, Mejino JL, Brinkley JF, et al. Application of neuroanatomical ontologies for neuroimaging data annotation. Front Neuroinform 2010;4:10. https://doi.org/10.3389/fninf.2010.00010.
7. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
8. Ivanovic M, Budimac Z. An overview of ontologies and data resources in medical domains. Expert Syst Appl 2014;41(11):5158–5166. https://doi.org/10.1016/j.eswa.2014.02.045.
9. Whetzel PL, Noy NF, Shah NH, et al. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39(Web-Server-Issue):541–545. https://doi.org/10.1093/nar/gkr469.
10. Ong E, Xiang Z, Zhao B, et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res Oct 2016;45(D1):D347–D352. https://doi.org/10.1093/nar/gkw918.
11. Golbeck J, Fragoso G, Hartel FW, Hendler JA, Oberthaler J, Parsia B. The national cancer institute’s thésaurus and ontology. J Web Semant 2003;1(1):75–80. https://doi.org/10.1016/j.websem.2003.07.007.
12. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40(1):30–43. https://doi.org/10.1016/j.jbi.2006.02.013.
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database-Issue):267–270. https://doi.org/10.1093/nar/gkh061.
14. Noy NF, McGuinness DL, et al. Ontology Development 101: A Guide to Creating Your First Ontology. [Online]. Available: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf 2001.
15. Evans AJ, Salama ME, Henricks WH, Pantanowitz L. Implementation of whole slide imaging for clinical purposes: issues to consider from the perspective of early adopters. Arch Pathol Lab Med Apr 2017;141(7):944–959. https://doi.org/10.5858/arpa.2016-0074-OA.
16. Lindman K, Rose JF, Lindvall M, Lundstrom C, Treanor D. Annotations, ontologies, and whole slide images – development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue. J Pathol Inform 2019;10(1):22. https://doi.org/10.4103/jpi.jpi_81_18.
17. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology 2017;70(1):134–145. https://doi.org/10.1111/his.12993.

[34] http://examode.dei.unipd.it/ontology/.

18. Romanchikova M, Thomas SA, Dexter A, et al. The need for measurement science in digital pathology. J Pathol Inform 2022;13:100157. https://doi.org/10.1016/j.jpi.2022.100157.
19. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8): 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.
20. Marini N, Marchesin S, Otálora S, et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. npj Digit Med 2022;no. 5:102. https://doi.org/10.1038/s41746-022-00635-4.
21. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare J 2019;6(2):94. https://doi.org/10.7861/futurehosp.6-2-94.
22. Dhrangadhariya A, Otálora S, Atzori M, Müller H. Classification of noisy free-text prostate cancer pathology reports using natural language processing. Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I; 2021. p. 154–166. [Online]. Available: https://doi.org/10.1007/978-3-030-68763-2_12.
23. Gurcan MN, Tomaszewski J, Overton JA, Doyle S, Ruttenberg A, Smith B. Developing the Quantitative Histopathology Image Ontology (QHIO): a case study using the hot spot detection problem. J Biomed Inform 2017;66:129–135. https://doi.org/10.1016/j.jbi.2016.12.006.
24. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279. [Online]. Available: https://ebooks.iospress.nl/publication/9130.
25. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc Jul 2013;21(e1):e11–e19. https://doi.org/10.1136/amiajnl-2013-001636.
26. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc Jun 2021;28(9):2017–2026. https://doi.org/10.1093/jamia/ocab084.
27. Stefano M, Laura M, Silvello G. *The Examode Ontology (Full Version, v.2) (2.0)*. 2023. https://doi.org/10.5281/zenodo.7669237.
28. International Agency for Research on Cancer [Internet]. Cancer Tomorrow: A Tool that Predicts the Future Cancer Incidence and Mortality Burden Worldwide from the Current Estimates In 2020 Up Until 2040. [cited 2021 July 13] Available from: https://gco.iarc.fr/tomorrow/en.
29. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Translat Oncol 2021;14(10):101174. https://doi.org/10.1016/j.tranon.2021.101174.
30. Wolf AM, Fontham ET, Church TR, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the american cancer society. CA Cancer J Clin 2018;68(4):250–281. https://doi.org/10.3322/caac.21457.
31. Fleming M, Ravula S, Tatishchev SF, Wang HL. Colorectal carcinoma: pathologic aspects. J Gastroint Oncol 2012;3(3):153. https://doi.org/10.3978/j.issn.2078-6891.2012.030.
32. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. New Engl J Med 2012;366(8):687–696. https://doi.org/10.1056/NEJMoa1100370.
33. Butte JM, Tang P, Gonen M, et al. Rate of residual disease after complete endoscopic resection of malignant colonic polyp. Dis Colon Rectum 2012;55(2). https://doi.org/10.1097/DCR.0b013e3182336c38.
34. World Health Organization, International Agency for Research on Cancer. Latest Global Cancer Data: Cancer Burden Rises to 19.3 Million New Cases and 10.0 Million Cancer Deaths in 2020, Press Release No. 292. [cited 2021 July 13] Available from: https://www.iarc.who.int/uploads/uploads/2020/12/pr292_E.pdf 2020.
35. Travis WD, Brambilla E, Noguchi M, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. J Thoracic Oncol 2011;6(2):244–285. https://doi.org/10.1097/JTO.0b013e318206a221.
36. World Health Organization, International Agency for Research on Cancer. Cervical Cancer. [cited 2023 May 9] Available from: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer 2022.
37. Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Global Health 2020;8(2):e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6.
38. Lowy DR, Schiller JT. Reducing hpv-associated cancer globally. Cancer Prev Res 2012;5(1):18–23.
39. Lax SF. Histopathology of cervical precursor lesions and cancer. Acta dermatovenerologica Alpina, Pannonica, et Adriatica 2011;20(3):1 25–133. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22131112/.
40. Brown JRG, Singh P. Coeliac disease. Paediat Int Child Health 2019;39(1):23–31. https://doi.org/10.1080/20469047.2018.1504431.
41. Gujral N, Freeman H, Thomson A. Celiac disease: prevalence, diagnosis, pathogenesis and treatment. World J Gastroenterol WJG Nov 2012;18:6036–6059. https://doi.org/10.3748/wjg.v18.i42.6036.
42. Villanacci V, Ceppa P, Tavani E, Vindigni C, Volta U. On behalf of the "Gruppo Italiano Patologi Apparato Digerente (GIPAD)" and of the "Società Italiana di Anatomia Patologica e Citopatologia Diagnostica"/International Academy of Pathology, Italian division (SIAPEC/IAP) Coeliac disease: the histology report. Digest Liver Dis 2011;43: S385–S395. https://doi.org/10.1016/S1590-8658(11)60594-X.
43. Marsh MN. Gluten, major histocompatibility complex, and the small intestine: a molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). Gastroenterology 1992;102(1):330–354. https://doi.org/10.1016/0016-5085(92)91819-P.
44. Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur J Gastroenterol Hepatol 1999;11(10): 1185–1194. https://doi.org/10.1097/00042737-199910000-00019.
45. Corazza G, Villanacci V. Coeliac disease. J Clin Pathol 2005;58(6):573–574. https://doi.org/10.32074/1591-951X-157.
46. J. R. Srigley, T. McGowan, A. MacLean, M. Raby, J. Ross, S. Kramer, and C. Sawka, "Standardized synoptic cancer pathology reporting: a population-based approach," J Surg Oncol, vol. 99, no. 8, pp. 517–524. [Online]. Available: https://doi.org/10.1002/jso.21282
47. Ellis D, Srigley J. Does standardised structured reporting contribute to quality in diagnostic pathology? the importance of evidence-based datasets. Virchows Archiv 2016;468: 51–59. https://doi.org/10.1007/s00428-015-1834-4.
48. Simperl E. Reusing ontologies on the semantic web: a feasibility study. Data Knowl Eng 2009;68(10):905–925. https://doi.org/10.1016/j.datak.2009.02.002.
49. Zhang W, Zeng B, Yang M, et al. ncrnavar: a manually curated database for identification of noncoding rna variants associated with human diseases. J Mol Biol 2021;433(11): 166727. https://doi.org/10.1016/j.jmb.2020.166727.computation Resources for Molecular Biology.
50. Gogate N, Lyman D, Bell A, et al. COVID-19 biomarkers and their overlap with comorbidities in a disease biomarker data model. Brief Bioinform May 2021;22(6):bbab191. [Online]. Available: https://doi.org/10.1093/bib/bbab191.
51. Bairoch A. The cellosaurus, a cell-line knowledge resource. J Biomol Techniq JBT Jul 2018;29:25–38. https://doi.org/10.7171/jbt.18-2902-002.
52. Sayre RR, Wambaugh JF, Grulke CM. Database of pharmacokinetic time-series data and parameters for 144 environmental chemicals. Scientific Data Apr 2020;7. https://doi.org/10.1038/s41597-020-0455-1.
53. Smith CL, Goldsmith C-AW, Eppig JT. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 2005;6:1–9. https://doi.org/10.1186/gb-2004-6-1-r7.
54. Gremse M, Chang A, Schomburg I, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res Oct 2010;39(suppl_1):D507–D513. https://doi.org/10.1093/nar/gkq968.
55. Vasilevsky NA, Matentzoglu NA, Toro S, et al. Mondo: unifying diseases for the world, by the world. medRxiv 2022. https://doi.org/10.1101/2022.04.13.22273750.
56. Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. Nucleic Acids Res Dec 2020;49(D1):D1207–D1217. https://doi.org/10.1093/nar/gkaa1043.
57. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–29. https://doi.org/10.1038/75556.
58. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res Dec 2020;49(D1):D325–D334. https://doi.org/10.1093/nar/gkaa1113.
59. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res Nov 2018;47(D1):D506–D515. https://doi.org/10.1093/nar/gky1049.
60. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012;13(1):1-20. https://doi.org/10.1186/gb-2012-13-1-r5.
61. Rosse C, Mejino J. The foundational model of anatomy ontology. Anat Ontol Bioinform Princip Pract Jan 2008;6. https://doi.org/10.1007/978-1-84628-885-2_4.
62. Musen MA. The protégé project: a look back and a look forward. AI Matters 2015;1(4): 4-12. https://doi.org/10.1145/2757001.2757003.
63. Duraiyan J, Govindarajan R, Kaliyappan K, Palanisamy M. Applications of immunohistochemistry. J Pharmacy Bioall Sci 2012;4(suppl 2):S307. https://doi.org/10.4103/0975-7406.100281.
64. Shen W, Wang J, Han J. Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans Knowl Data Eng 2015;27(2):443–460. https://doi.org/10.1109/TKDE.2014.2327028.
65. Marchesin S, Giachelle F, Marini N, et al. Empowering digital pathology applications through explainable knowledge extraction tools. J Pathol Inform 2022;13:100139. https://doi.org/10.1016/j.jpi.2022.100139.
66. Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, et al. Marian: fast neural machine translation in C++. Proc. of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations. ACL; 2018. p. 116–121. [Online]. Available: https://aclanthology.org/P18-4020/.
67. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Proc. of the 30th Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA; 2017. p. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
68. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–310. https://doi.org/10.1006/jbin.2001.1029.
69. Giachelle F, Irrera O, Silvello G. Medtag: a portable and customizable annotation tool for biomedical documents. BMC Med Inform Decision Making 2021;21(1):352. https://doi.org/10.1186/s12911-021-01706-4.
70. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34–49. https://doi.org/10.1016/j.jbi.2017.11.011.
71. Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016;69(11):949–955. https://doi.org/10.1136/jclinpath-2016-203872.
72. Carethers JM. Risk factors for colon location of cancer. Translat Gastroenterol Hepatol 2018;3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232066/.
73. Harro CD, Pang Y-YS, Roden RBS, et al. Safety and Immunogenicity Trial in Adult Volunteers of a Human Papillomavirus 16 L1 Virus-Like Particle Vaccine. JNCI J Natl Cancer Inst Feb 2001;93(4):284–292. https://doi.org/10.1093/jnci/93.4.284.
74. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin Chest Med 2011;32(4):605–644. https://doi.org/10.1016/j.ccm.2011.09.001.
75. Dewar DH, Ciclitira PJ. Clinical features and diagnosis of celiac disease. Gastroenterology 2005;128(4, Supplement 1):S19–S24. https://doi.org/10.1053/j.gastro.2005.02.010.